Digital Archive Stamp Detection and Extraction

Xin Jin[1], Qiuyang Mu[1], Xiaoyu Chen[2], Qingyu Liu[1], and Chaoen Xiao[1](✉)

[1] Beijing Electronic Science and Technology Institute, No. 7, Fufeng Street, Fengtai District, Beijing 100070, China
xcecd@qq.com

[2] Information Center of China North Industries, Group Corporation Limited, Beijing 100089, China

Abstract. Archives contain valuable historical information and must be properly preserved. However, traditional archival materials are vulnerable to damage from water, fire, and mold, making long-term storage difficult. To address this issue, digital archives have been established for management. As a result, effective storage, detection, extraction, and utilization of archive information has become a focus of attention. This paper focuses on the feature extraction of archival stamp images, proposing a network structure of stamp extraction based on generative adversarial network for texture feature extraction of stamp images. This method aims to extract more refined texture features, improving the accuracy of stamp text recognition. An improved stamp text recognition method is proposed using PP-OCR, which can recognize text for multiple shape seals. This method effectively solves the problem of deep learning models being unable to recognize text due to the bending and tilting of the ring-shaped text in the stamp. Overall, this research aims to enhance the preservation and utilization of archival materials by improving feature extraction and text recognition methods.

Keywords: target detection · stamp extraction · image segmentation · stamp text recognition

1 Introduction

Despite in-depth research on image detection and extraction by both domestic and foreign scholars, there has been little focus on the extraction of key information, particularly seal information, from real old archives. To address this gap, we propose a digital archival seal image detection and extraction technique for seals on old archives, which has significant practical value.

This paper includes three main chapters: “Dataset construction”, “Image extraction model”, and “Seal image text recognition scheme”, in addition to “Introduction” and “Conclusions”. In “Data set construction”, we describe how we constructed the dataset. In “Image extraction model”, we explain how we trained the stamp image extraction model on the dataset. Finally, in “Seal image text recognition”, we present a scheme for seal image text recognition.

H. Lu and J. Cai (Eds.): ISAIR 2023, CCIS 1998, pp. 165–174, 2024.
https://doi.org/10.1007/978-981-99-9109-9_16

Overall, this research provides a novel approach for the detection and extraction of seal information from old archives, with the potential to significantly enhance preservation and utilization efforts.

Our contributions to the field can be summarized as follows:

- Construction of a new dataset of archival seal images
- Development of an archival seal image extraction model
- Proposal of a lightweight text recognition scheme for archival seal images

2 Dataset Construction

To address the issue of small sample sizes and uneven distribution in seal images, this paper utilized the seal making software Sedwen combined with OpenCV to manually create 2,000 seal images. Each shape of seal was represented by 500 images. The seals were added to the archives with rotation angles ranging from [−45°, 45°], accounting for 90% of the images; [−30°, 30°], accounting for 70% of the images; and [−10°, 10°], accounting for 50% of the images. This generated a large number of archival samples with seals adjacent to each other and overlapping, allowing for more diverse and realistic training data.

Additionally, the paper collected fingerprint images and added them to the archives as a background component. The resulting dataset consisted of a mixture of real archive images and generated archive images, totaling 11,068 mixed archive images. Figure 1 shows examples of the generated archive images.

Overall, this approach allowed for the creation of a larger and more diverse dataset of digitized archive seal images, providing a more comprehensive foundation for training machine learning models.

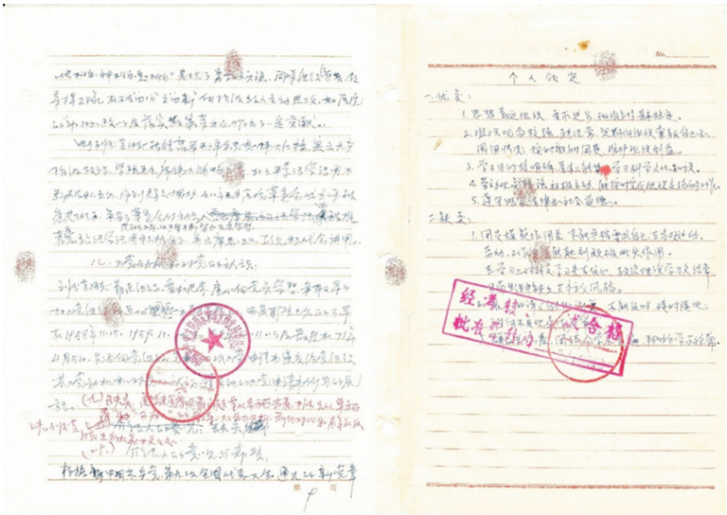


Fig. 1. Generated archive images

2.1 Archival Stamp Image Detection Model

Archival Stamp Image Detection Model Information. To facilitate subsequent deep learning training, this paper used the LabelImg open source tool to manually annotate stamp images on archives. The fully labeled digitized archival stamp image dataset was divided into a training set (7/10) and a validation set (3/10). The training images and labels were input together into the YOLOv5 [1, 7] detection model for training, resulting in a stamp detection model file with the suffix pt after 300 epochs of training. This file saved various parameter information during the model's training.

To detect stamps in new archive images, the image was input into the AI model for inference, and the location of the stamp image was marked and saved. The general framework diagram of stamp detection is shown in Fig. 2. The type of stamp detected and the probability of correctness of this type were displayed at the top of the target detection box. Based on the detection box and coordinate information obtained through inference, the target was cropped, and the stamp images were stored in separate folders.

Using the location information saved by YOLOv5 inference, the stamp area can be easily obtained in subsequent label production, enabling the creation of a large dataset of digitized archival stamp images with simple manual verification.

Overall, this approach provides an efficient and effective method for detecting and extracting stamp images from archives, enabling the creation of a more comprehensive digitized archive dataset for preservation and utilization purposes.

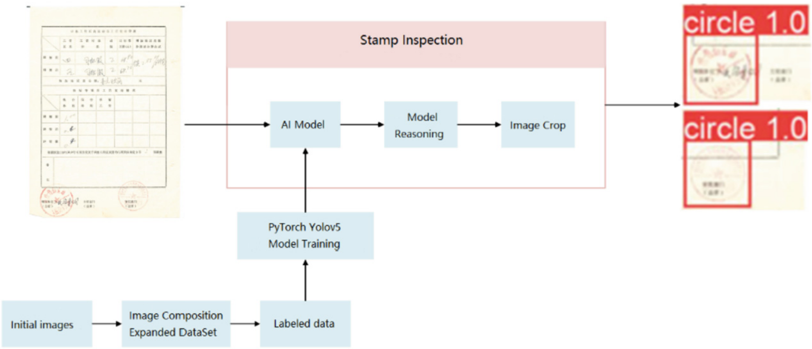


Fig. 2. General framework diagram of stamp detection

Experimental Results and Analysis. Figure 3 shows that the box_loss, obj_loss, and cls_loss for both the training and validation sets are infinitely close to 0, indicating that the stamp detection model can accurately respond to both box and target detection and classification situations. Table 1 shows that the four metrics of AP, AR, mAP_0.5, and mAP_0.5:0.95 all converge to 1 indefinitely, indicating a high level of accuracy.

Overall, the YOLOv5 network model demonstrates a high comprehensive detection rate and accuracy, making it well-suited for the field of detection and classification of stamp images.

Table 1. Combined archive image stamp detection index table

Classification	Training	Validation	Accuracy	Recall	mAP@.5	mAP@.5:.95
Circle	4151	1873	0.997	0.995	0.995	0.87
Rectangle	4062	1353	0.99	0.989	0.994	0.731
Oval	3844	1459	0.997	0.998	0.995	0.933
Rhombus	3302	1561	0.999	0.995	0.995	0.898
Total	15359	6246	0.996	0.994	0.995	0.858

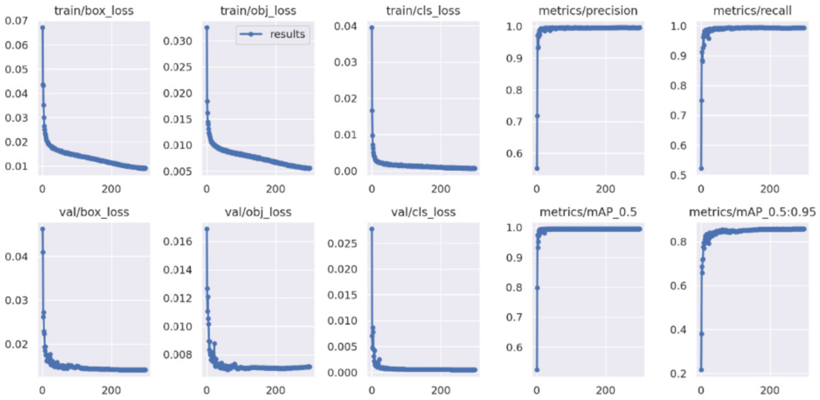


Fig. 3. Consolidation of the indicators of the archival test stamp

2.2 Construction of Stamp Extraction Dataset

To accurately extract the foreground of stamp images, this paper utilized a model based on generative adversarial network to segment image foreground and background. To improve the accuracy and reliability of the model, Ground Truth was produced to help the GAN model learn image features more effectively and generate more accurate results.

To produce a label map of the seal data, several pieces of information were recorded, including the size of the archive image (w_1, h_1, c_1), the rotation angle α of the seal, the size of the seal image (w_2, h_2, c_2), and the coordinates of the upper left corner of the stamped seal (a, b). A blank canvas with the same size as the archive image was generated, and the same seal image was stamped on the canvas at the same position with the same rotation angle, according to the three values of α, (w_2, h_2, c_2), and (a, b).

Using a series of operations such as color space conversion, binarization processing, and mask map inversion, the mask map of the canvas was obtained. The archive image and the canvas mask image were then cropped according to the coordinates of the stamped seal's position, retaining only the smallest rectangular part of the area where the stamp was located. Figure 4 provides a diagram of the label making process.

Finally, the real stamp dataset and the mask label map of the stamp necessary for generating the adversarial network, i.e. Ground Truth, were obtained. This approach allowed for the creation of more accurate and reliable data for training the generative adversarial network.

The stamp extraction dataset was constructed with 4,000 sets of stamp images, consisting of a set of images including stamp images and stamp mask label images. Of these, 2,800 sets were included in the training set, and 1,200 sets were included in the validation set. The stamp extraction dataset contained four types of images, including circle, rectangle, oval, and diamond shapes.

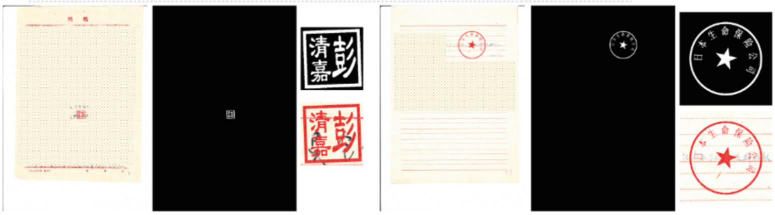


Fig. 4. Diagram of the process of making a stamp label

3 Digital Archival Stamp Image Extraction Model

3.1 Digital Archival Stamp Image Extraction Model Information

Building upon the success of the generative adversarial network, this paper designed the SealNet network structure to implement the extraction function of seal images and ensure accurate segmentation of both foreground and background parts of the seal.

The SealNet network structure diagram, shown in Fig. 5, includes a modified generator network based on U-Net [2, 8], with an attention mechanism-based CBAM module added to focus on image segmentation after the convolution operation of the U-type network downsampling process. The generator's role is to extract the foreground part of the seal image and remove the part of the seal image with the archival background. In addition, a pre-trained external VGG-19 [3] network is used for texture feature extraction of the stamp image.

The stamp image is output as a foreground mask map after extraction by the generator network. The sample mask map output by the generator is then fed into the seal texture extraction module together with the real label mask map of the seal image, allowing the network to extract texture features more finely and guide the generator to generate a mask map more similar to the label map.

The discriminator network determines whether the input data is true or false by comparing the real stamp labels with the generated virtual stamp sample features and classifies them according to the true or false category. If the discriminator judges the input data as false, it helps the generator learn how to generate more realistic samples for the discriminator to judge.

Through alternating training, learning, and parameter optimization, SealNet achieves the best performance when converging and improves the accuracy of seal image segmentation and extraction. Overall, this network structure represents a significant advance in the field of seal image extraction and has the potential to enhance preservation and utilization efforts for archives.

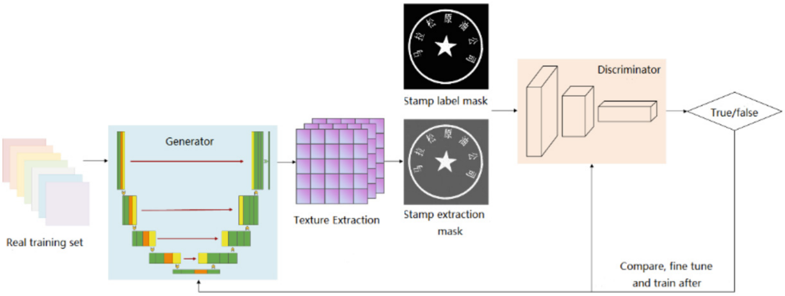


Fig. 5. SealNet network structure diagram

3.2 Experimental Results and Analysis

Figure 6 shows the extraction effect of four shapes of seals using the SealNet model. The first column of the image is the original image of the seal obtained by cropping, the second column of the image is the label mask image obtained by the SealNet model, and the third column of the image is the extracted image of the foreground of the seal obtained by overlaying the original image and the mask image. From Fig. 6, it is clear that the seal extraction model has a better segmentation effect, as the seal foreground image can be separated more completely from the red striped archival background.

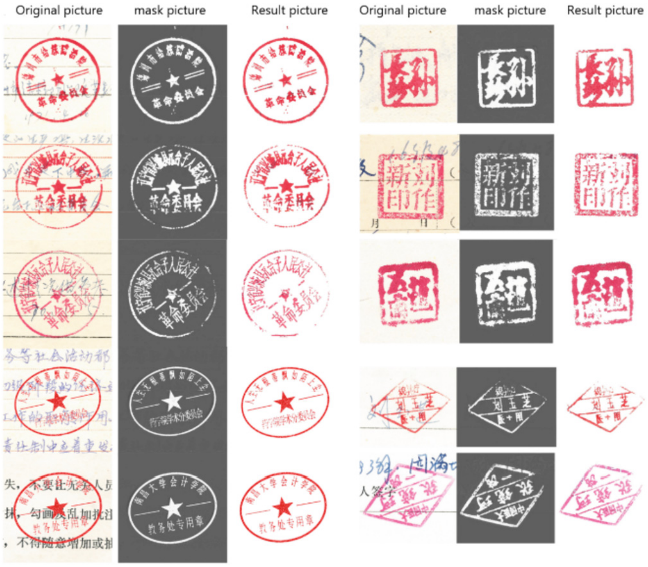


Fig. 6. Stamp extraction effect example diagram

4 Digital Archival Stamp Image Text Recognition

4.1 Digital Archival Stamp Image Text Recognition Information

After stamp detection, cropping, and extraction, the irrelevant background information is removed, and a valid region containing stamp graphics and text is obtained. First, the center point of the segmented instance of the stamp is found based on contour information. Using this center point as the rotation center, the stamp image is rotated in a specific direction and angle each time until it completes one full rotation, resulting in a series of rotated new images.

The new images are input into the PP-OCR model for text recognition in the horizontal direction for each frame in turn. The text content of individual characters and the location information of the smallest outer wrapping rectangle are recorded. The position of each character in the original stamp image is obtained by semantic continuity association calculation, and the association between each character is determined by discriminating the most similar path. Based on the trajectory information and support vector machine algorithm, the two clusters of circular text and linear text are segmented and used as the paragraph semantic output results. The improved stamp text recognition model based on PP-OCR [4–6] is shown in Fig. 7.

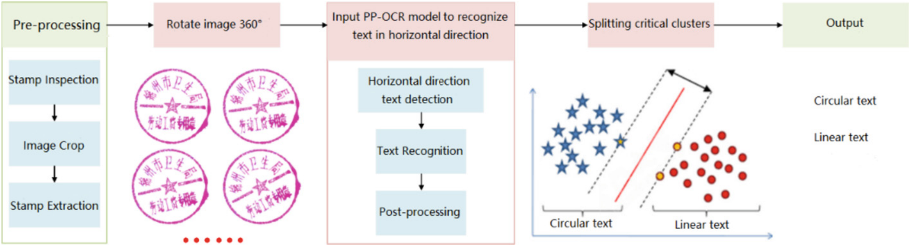
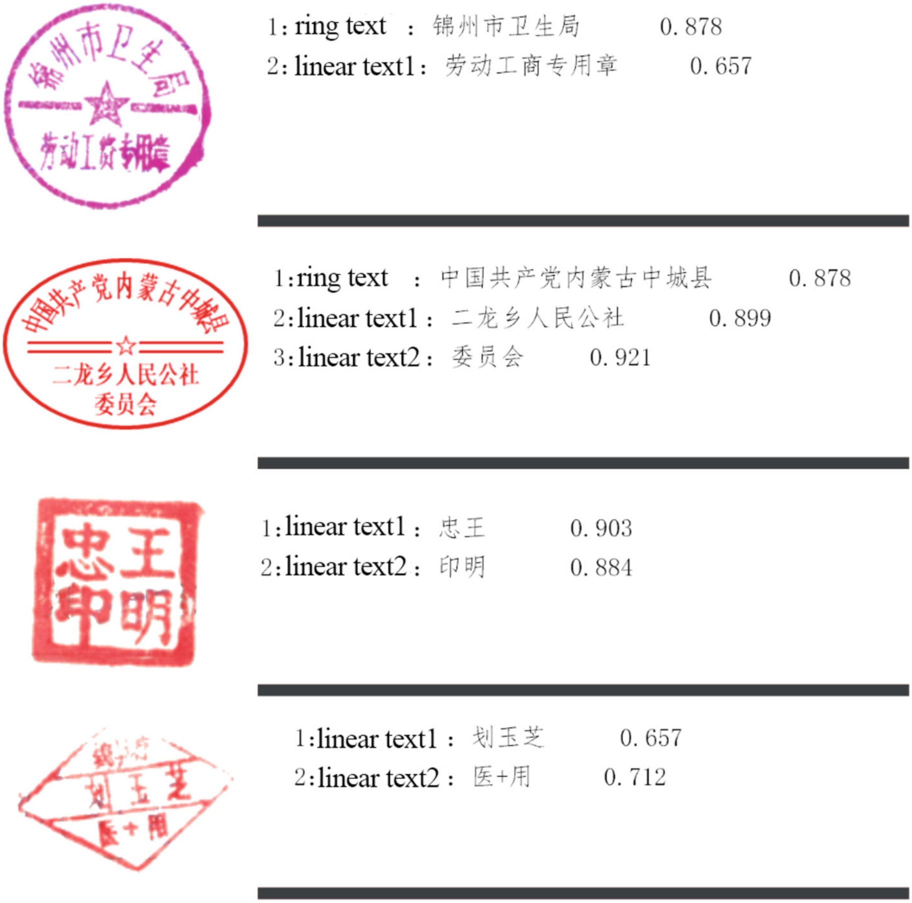


Fig. 7. Structure diagram of the improved stamp text recognition model based on PP-OCR

4.2 Experimental Results and Analysis

Figure 8 shows the stamp text recognition results, displaying the recognition results of ring text and linear text separately, along with the probability of correct text recognition results. While this algorithm can accurately recognize clear text, it struggles with fuzzy text, resulting in a higher false detection rate. For example, the recognition result of linear text 1 in the circular stamp is wrong, as it is mistakenly detected as “labor injury special stamp” instead of “labor wage special stamp”. Another example is the recognition result of linear character 1 in the diamond-shaped stamp, which is mistakenly detected as “Paddle Yuzhi” instead of “Liu Yuzhi”.

Circular moment transformation can be used to better recognize text in circular seals. This is because the circular seal image in the right-angle coordinate system can be easily mapped to the polar coordinate system. However, this method is only applicable



```
1: ring text  ： 锦州市卫生局        0. 878
2: linear text1： 劳动工商专用章       0. 657

1:ring text   ： 中国共产党内蒙古中城县     0. 878
2:linear text1： 二龙乡人民公社       0. 899
3:linear text2： 委员会       0. 921

1:linear text1： 忠王        0. 903
2:linear text2： 印明        0. 884

1:linear text1 ： 划玉芝        0. 657
2:linear text2 ： 医+用        0. 712
```

Fig. 8. Stamp text recognition result chart

to circular seals and cannot be applied to oval seals. The oval seal image after circular moment transformation is shown in Fig. 9.

The algorithm proposed in this paper requires less accuracy for the center point of the stamp in comparison to the polar coordinate expansion method. As long as the hyperplane solved by the support vector machine is guaranteed to completely separate the two rotated coordinate information points, the algorithm can be adapted to any shape of the stamp image, circumventing the problem of text image bending caused by polar coordinates unfolding the stamp image, which can lead to misrecognition and missed recognition by the detector.

In terms of execution efficiency, polar coordinates need to perform a rotational projection transformation for each point, and then use an interpolation algorithm to find the value of the middle point after the rotation. In contrast, this algorithm only needs to perform a parallelized projection transformation without calculating interpolation, which can theoretically be executed faster and substantially improve accuracy and robustness.

The advantages of this algorithm are verified by comparing the recognition results after polar coordinate expansion. Overall, this approach represents a significant advance in the field of stamp text recognition, providing a more accurate and efficient method for recognizing text in stamped images of various shapes.

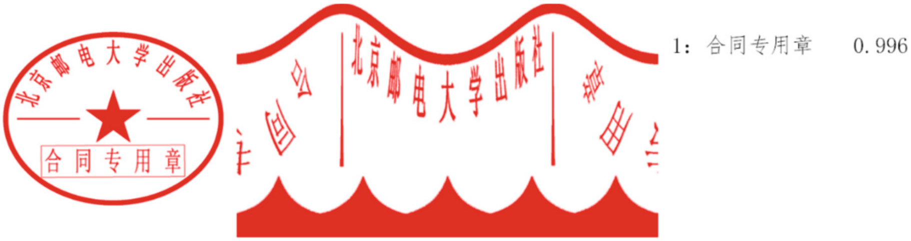


Fig. 9. Image of elliptical stamp after circular moment transformation

5 Conclusions

This paper focused on studying the key information of stamps in cadres' personnel files and developed a method for stamp image detection, extraction, and text recognition. The main research results of this paper included:

(1) Construction of a digital archive stamp image dataset.
(2) Design of a stamp image extraction network model, SealNet.
(3) Proposal of an improved stamp text recognition method based on PP-OCR.

Through this research, the key information of seals in archives was extracted and studied, and stamp target detection, stamp foreground extraction, and stamp text recognition were realized.

Moving forward, there is room for further optimization of the stamp extraction model's performance. While the current model is effective in extracting the foreground of stamp images with dark colors, it struggles to extract images with light colors and similar background colors and stamp colors. For example, it is difficult to extract the foreground of black seals in photocopies. In future research, a larger dataset could be used to allow the SealNet network to learn more detailed features and improve the accuracy and reliability of stamp extraction.

Overall, this research represents a significant advance in the field of stamp image extraction and recognition and has the potential to enhance the preservation and utilization of archives for research purposes.

Acknowledgements. We thank the ACs and reviewers. This work is partially supported by the Fundamental Research Funds for the Central Universities (3282023014), and the Project of Information Center of China North Industries Group Corporation Limited (20220100H0113).

References

1. Jiang, P., Ergu, D., Liu, F., et al.: A review of Yolo algorithm developments. Procedia Comput. Sci. **199**, 1066–1073 (2022)
2. Ronneberger, O., Fischer, P, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 Oct 2015, Proceedings, Part III 18. Springer International Publishing, pp. 234-241 (2015)
3. Mateen, M., Wen, J., Song, S., et al.: Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry **11**(1), 1 (2018)
4. Du, Y., Li, C., Guo, R., et al.: Pp-ocr: A practical ultra lightweight ocr system. arXiv preprint arXiv:2009.09941 (2020)
5. Du, Y., Li, C., Guo, R., et al.: Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. arXiv preprint arXiv:2109.03144 (2021)
6. Lan, R., Sun, L., Liu, Z., et al.: MADNet: A fast and lightweight network for single-image super resolution. IEEE Trans. Cybern. **51**(3), 1443–1453 (2020)
7. Zheng, Q., Zhu, J., Tang, H., et al.: Generalized label enhancement with sample correlations. IEEE Trans. Knowl. Data Eng. **35**(1), 482–495 (2021)
8. Lu, H., Yang, R., Deng, Z., et al.: Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM. ACM Trans. Multimedia Comput. Commun. Appl. **17**(1s), 1–18 (2021). https://doi.org/10.1145/3422668