Vignette detection and reconstruction of composed ornaments with a strengthened autoencoder

Mohammad Khan, Rémi Emonet, Thierry Fournel

Vignette detection and reconstruction of composed ornaments with a strengthened autoencoder

Mohammad Sadil Khan, Rémi Emonet, and Thierry Fournel
Laboratoire Hubert Curien UMR 5516
Univ. Lyon, Université Jean Monnet Saint-Etienne,
CNRS, Institut d'Optique Graduate School
F-42023, Saint-Etienne, France

Abstract—**A strengthened autoencoder formed by placing an object detector upstream of a decoder is here developed in the context of the model-helped human analysis of composed ornaments from a dictionary of vignettes. The detection part is in charge to detect regions of interest containing some vignette features, and the decoding part to ensure vignette reconstruction with a relative quality depending on feature match. Images of ornaments without typographical composition are generated in order to properly assess the performance of each of the two parts.**

I. Introduction

This paper aims to develop an efficient algorithm for the decomposition and reconstruction of composed ornaments. Such ornaments which result from assembling a subset of typographical vignettes are typical of the 18th century. A better understanding of knowledge dissemination during the censorship period can be obtained by focusing on a bookseller named Marc-Michel Rey, publisher of Thinkers [1]. In addition, a dictionary of vignettes can be circumscribed for a statistical learning of the targeted task. In terms of digital humanities, decomposition with reconstruction can be helpful in the problem of book assignment by a human expert as ornament was identified as a relevant indicator [2]: missing or badly reproduced vignettes in the reconstructed image will indicate out-of-dictionary or abnormal vignettes.

In a machine learning perspective, decomposition and reconstruction can be viewed through encoding-decoding as anomaly pre-segmentation. As trained on normal images, autoencoders are expected to reconstruct abnormal inputs with higher errors in the sense of a dedicated loss function [3]. However this assumption may not be satisfied when anomalies are ”close” to some normal distortions of data. In such a case, even if usual autoencoders are well designed to learn what is a normal pattern given only normal examples, they can also well reconstruct anomalies leading to misdetection [4]. To strengthen the feature description part, we suggest to achieve a detection thanks to an object detector upstream of the decoding part in order to automatically select regions of interest containing vignettes from the dictionary, and to well reconstruct only in this case[1].

The architecture of the so-called Detector-encoder AutoEncoder (DAE) is detailed in section 2. The different parts of the model including the detector based on SSD [6], are differentiable ensuring a well-conditioned behavior. The model is learnt from images generated from a dictionary formed with vignettes used by Marc-Michel Rey. Its ability to clean so indirectly highlight out-of-dictionary vignettes is tested in an experiment described in section 3. Conclusion and perspectives are reported at the end, in section 4.

II. The proposed model

The proposed model, a Detector-encoder AutoEncoder, has two parts (Fig. 1) - (1) the detection part in the front, $\Phi(F_d(x))$, where F_d corresponds to an encoding function performing RoI detection (a Detector-Encoder dedicated to representation learning in the scope of object detection, see Fig. 2) and where Φ corresponds to a RoI Pooling transformation (Fig. 3) (2) the reconstruction part, $F_r(\phi)$, where $\phi = \Phi(F_d(x))$), a Decoder responsible for image reconstruction.

A. Detection part

In AutoEncoders, the encoder part encodes any input image from which the decoder part reconstructs the image. Our final goal is to help to detect the out-of-dictionary vignettes corresponding to anomalies in an ornament. DAE is designed to reproduce the input ornament as an output image where each normal vignette (i.e. belonging to the dictionary) is well reconstructed and each abnormal one does not appear at all or appears blurry, in place. In this way, DAE can tell which of the vignettes are normal so assigned in the dictionary, and which are not. Conventional AutoEncoders designed to perform well copy-and-paste of inputs, can generalize too much in some abnormal cases. To address this issue, we suggest to replace the encoder with a Single Shot Multibox Detector (SSD) [10] using the following feature maps for the purpose of detection with backbone classifier VGG-16 (Fig. 2):

- **Conv4_3:** Size $38 \times 38 \times 512$
- **Conv7:** Size $19 \times 19 \times 1024$
- **Conv8_2:** Size $10 \times 10 \times 512$
- **Conv9_2:** Size $5 \times 5 \times 256$
- **Conv10_2:** Size $3 \times 3 \times 256$
- **Conv11_2:** Size $1 \times 1 \times 256$

[1] Detector-encoder AutoEncoder for decomposition into M.-M. Rey's vignettes at https://ro2i.hypotheses.org

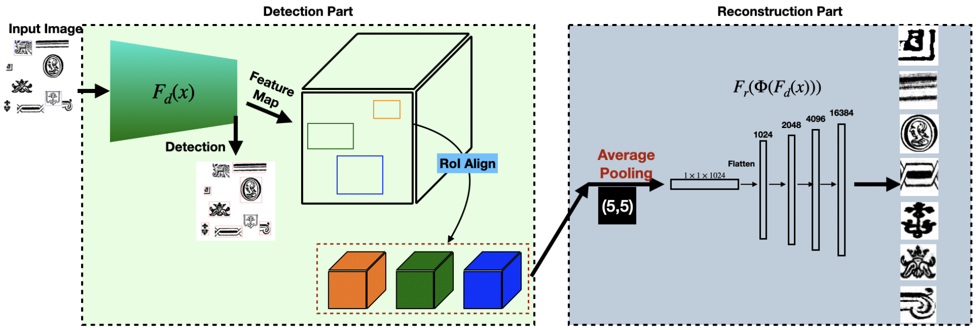

Fig. 1: Block diagram of the DAE model architecture.

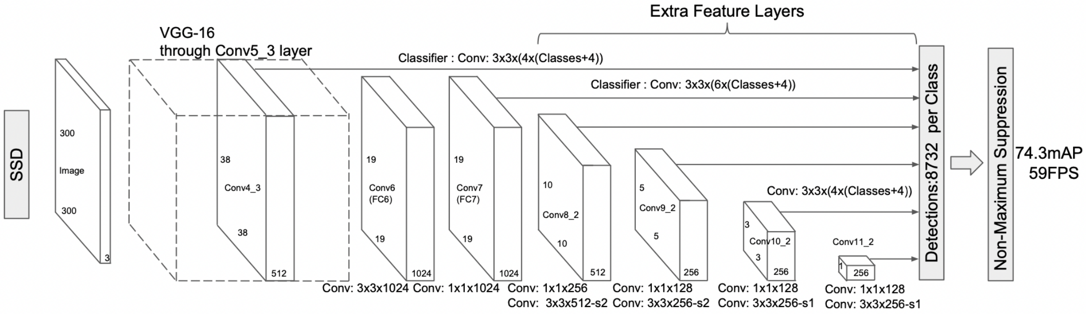

Fig. 2: SSD Architecture [10].

To optimize detector SSD, it is suggested to use an Intersection over Union (IoU) loss in replacement of l_1 loss, namely a Generalized IoU [11], Distance IoU [6] and Efficient IoU [7]. Architectures xIoU-SSDs are here trained with the dictionary (normal vignettes) in order to detect with a better localization accuracy.

The feature maps encode the characteristics of the vignettes of the dictionary. For each of the feature maps, SSD considers 4 or 6 default boxes per cell (8732[2] boxes in total). The final output of SSD for each of the default boxes are: (1) its offset $(\delta x, \delta y, \delta w, \delta h)$ (2) a probability vector $(p_0, p_1, \cdots, p_n)$ where p_i is the probability that the box contains an object of the ith class. Class i for all $i \in \{1, \cdots, n\}$, corresponds to the ith vignette in the dictionary whereas class 0 refers to the image background. A Region of Interest (RoI) is determined from the final predicted box by maximizing the probability values, which provides a confidence value with respect to the fact that it contains a certain vignette from the dictionary. The corresponding features are accessible by checking the corresponding index (Fig. 3) and this gives us the information of which feature maps the RoI is from. RoIs are as the vignettes, of different sizes and aspect ratios. By adding some layers devoted to RoI resizing after SSD, RoIs are ready to be properly processed by the reconstruction part.

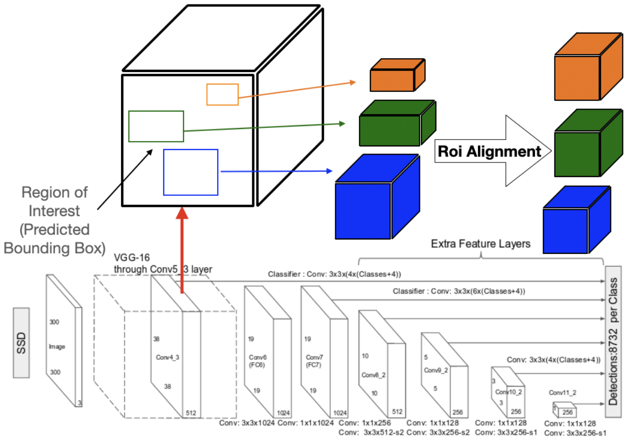

Fig. 3: RoI Alignment step in DAE architecture.

1) Bounding box regression: Bounding box regression is one of the most important components in object detection tasks. In conventional SSD, a l_n norm is used during training to evaluate the performance of the detector with the IoU (Intersection over Union) metric: $IoU = \frac{|B_G \cap B_P|}{|B_G \cup B_P|}$ where B_g and B_d are the ground and predicted bounding boxes, respectively. However there is no correlation between minimizing l_n norm and improving the loss associated to the IoU metric, $L_{IoU} = 1 - IoU(B_g, B_d)$ [5].

In Figure 4, the predicted bounding box (black rectangle)

[2] $(38 \times 38 \times 4) + (19 \times 19 \times 6) + (10 \times 10 \times 6) + (5 \times 5 \times 6) + (3 \times 3 \times 4) + (1 \times 1 \times 4) = 8732$

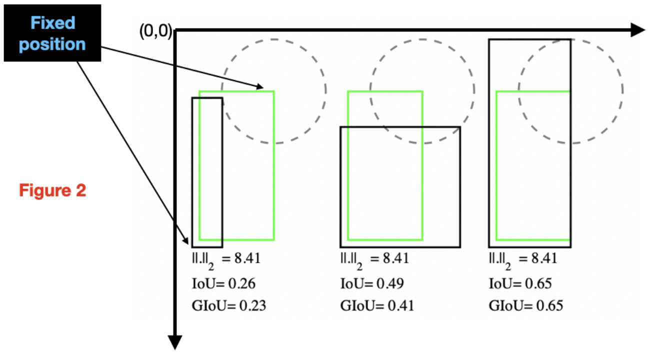

Fig. 4: Three cases where the l_2-norm distance between the representations of two rectangular bounding boxes, each given by the concatenation of the coordinates of two opposite corners, has the same value but IoU and GIoU metrics have very different values [5].

and ground truth box (green rectangle) are each represented by their top-left and bottom-right corners (pointed by arrows), and whose the Cartesian coordinates are denoted as (x_1, y_1, x_2, y_2) and (x'_1, y'_1, x'_2, y'_2), respectively. For simplicity, let us assume that the distance, e.g. $l_2 - norm$, between one of the corners of two boxes is fixed. Now, if the second corner lies on a circle with fixed radius centered on the ground truth box, then the l_2 loss between the ground truth box and the predicted bounding box is the same although their IoU values can be different depending upon the positions of top-right and bottom-left corners. So, using IOU-loss should be the best option since a bad detector will impact negatively in the reconstruction part. However IoU has two major issues as a metric, following from its definition. If two boxes do not overlap, then their IoU is zero, which does not give any indication whether they are close or far. In addition, in case of non-overlapping boxes, since their IOU is zero, the gradient is also zero, and loss L_{IoU} cannot be optimized.

A variant of this loss was suggested to address the weaknesses of the IoU metric: the Generalized IoU loss [5], $L_{GIoU} = 1 - GIoU$ given by the metric defined by $GIoU = IoU - \frac{|C \backslash (B_g \cup B_P)|}{|B_p|}$ where C is the convex hull of the union of bounding boxes B_g, B_P. Computing efficient, approximate versions were later proposed in [6]:

- the Distance IOU loss defined as $L_{DIoU} = 1 - DIoU$ where $DIoU = IoU - \frac{\rho^2(b_g, b_p)}{c_C^2}$, ρ is the euclidean distance, and c_C is the length of the diagonal of convex hull C,
- the Efficient IOU loss [7] defined as $L_{EIoU} = 1 - EIoU$ where $EIoU = IoU - \frac{\rho^2(b_g, b_p)}{c^2} - \frac{\rho^2(w_g, w_p)}{c_w^2} - \frac{\rho^2(h_g, h_p)}{c_h^2}$, ρ is the euclidean distance, and $(\mathbf{b}, w, h$ defines a box centered in point $\mathbf{b}$ having width w and height h, its diagonal length being denoted as c).

2) RoI alignment: We used RoIAlign, first proposed in [12], which allows the extraction of a $k \times k$ RoI where k is a predefined integer value, from feature maps. For any $N \times N$ RoI, RoIAlign divides the feature maps into k^2 $\frac{N}{k} \times \frac{N}{k}$ regions, named RoI bins, in each of which is computed a single value: the maximum or the average of the values at four points determined at the end by a linear interpolation.

B. Reconstruction part

We constructed the Decoder with three fully connected linear layers. The feature maps obtained from xIoU-SSD are first transformed into $1 \times 1 \times 1024$ using Average Pooling with kernel $(5, 5)$. Each of the transformed feature maps are flattened and fed into the linear layers which are as described below (See Fig. fig:mesh1):

- Layer 1: Input 1024 → Output 2048 (Activation Function: Relu)
- Layer 2: Input 2048 → Output 4096 (Activation Function: Relu)
- Layer 3: Input 4096 → Output 16384 (Activation Function: Sigmoid)

The output, any reconstructed vignette, is then reshaped into a 128×128 image.

III. Results

A. Data

Depending on their function in book composition (so their type), ornaments can be composed from a single (as fleurons at bottom of page) to about fifteen vignettes assembled together to form without overlapping a predefined geometric figure (in ornamented pages). To experiment DAE, 1,000 images were generated for training (from scratch) and validation, respectively. Each 640×640 image is composed with a number of vignettes following a Poisson distribution. The vignettes were selected in a predefined dictionary formed with 200 assigned vignettes (previously extracted from a catalog of M.-M. Rey's provider, and binarized). The vignettes were selected independently of their type and the image composition was achieved randomly in order to not bias detection. It means that the vignettes are placed randomly in the image such that they do not overlap, after having sorted them in the decreasing order of their area (their original size is kept here) to accelerate image synthesis. A generated image is composed with 11 vignettes on average. The same process were achieved for the generation of test images excepted that, on average, 2 of the composing vignettes are selected out of the dictionary. For each image in the training, validation and test datasets, the label and bounding box of the selected vignettes are stored.

B. Learning stage

The Detector-Encoder and the Decoder were trained separately. For the first one, we modified the pytorch implementation[3] of SSD in order to add the ROI alignment layers and evaluate the performance of the different IoU losses with smooth l1 loss as baseline. For optimization, we used the Adam optimizer with a learning rate initialized at 0.0001 then

[3] Kaggle SSD 300

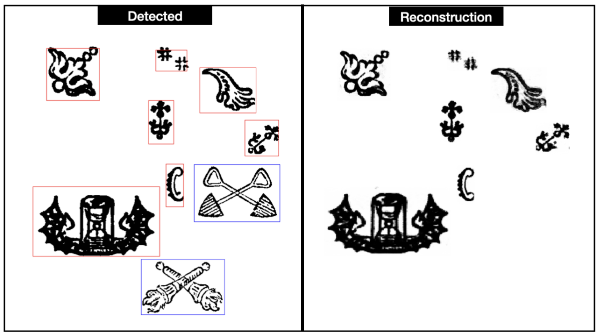

Fig. 5: DAE reconstructions with $P = 0.2$ (right) from input image (left).

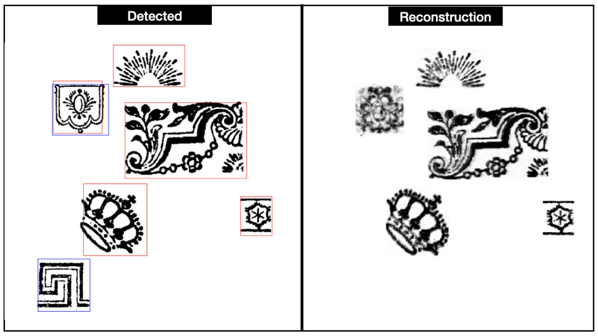

Fig. 6: DAE reconstructions with $P = 0.01$ (right) from input image (left).

decreased by factor $\frac{1}{10}$ when the loss did not decrease more than 2%. We used Google Colab with 12 GB of NVIDIA Tesla K80 GPU as hardware set-up. For the Decoder, mean squared error loss and Adam optimizer with a fixed learning rate of 0.0001 were used. We trained it with 30 epochs.

C. Tests

The Detector-Encoder part then the full DAE model with the two parts were applied on the test dataset after the learning stage in order to define the best IoU loss and to assess the ability of the Decoder to clear up ambiguities, respectively.

1) Test of the Detector-Encoder: The performance of the Decoder-Encoder with respect to a detection of normal vignettes only was measured with the mean average precision, $AP_{\%}$, for a given percentage (expressed in $[0, 100]$) used as threshold value on IoU metric. The results obtained for the different losses are reported in Table 1.

TABLE I: mAp for different SSD models(Validation Set)

Models	AP_{50}	AP_{75}	AP_{80}	AP_{85}	AP_{90}	AP_{95}
SSD_{l_1}	80.5	74.2	62.8	37.6	9.63	0.38
SSD_{GIoU}	84.0	81.8	77.1	60.1	23.8	1.37
SSD_{DIoU}	**84.06**	**82.04**	**77.88**	**63.3**	**27.8**	**1.72**
SSD_{EIoU}	83.8	82.0	77.6	61.8	26.3	1.50

Since for non overlapping boxes, GIoU first expands the predicted boxes and cover the ground-truth boxes (i.e $|B_g \cup B_p| = B_p \implies |C \setminus (B_g \cup B_P)| = \Phi$), GIoU is an IoU loss which at the end converges slowly. In reverse DIoU converges faster [6] (Table 1). Therefore, in the reconstruction part, $DIoU - SSD$ has been used.

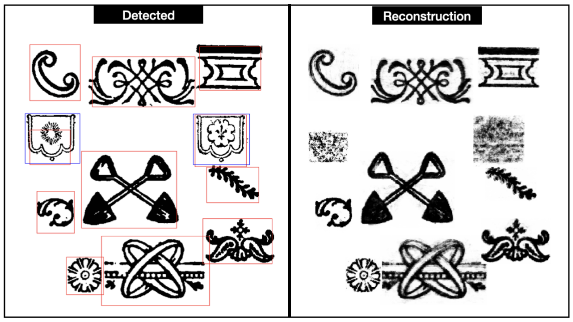

Fig. 7: DAE reconstructions with $P = 0.001$ (right) from input image (left).

2) Test of the Detector-Encoder-Decoder: Fig. 5, 6 and 7 show on left input images having two out-of-dictionary vignettes. The output images with the reconstructed vignettes delivered by the DAE model for different threshold values of the confidence value, P, are shown on right. Since SSD outputs a confidence value associated to a vignette from the dictionary per ROI, any RoI having a confidence value greater than threshold value P will be eligible for reconstruction.

In Fig. 5, the two anomalies are not detected by the Detector-Encoder so not reconstructed at all. All the other vignettes from the dictionary are detected (in red) and reconstructed with a good quality visually. In Fig. 6, threshold value P is lower, and one of the two anomalies (in blue) is detected (in red) so reconstructed but with a poor quality (statistically quantified below). In Fig. 7, the two anomalies are detected but poorly reconstructed. The lower the threshold P, the more the number of detected vignettes with a higher risk of detecting anomalies. However in the reconstruction phase, the detected anomalies are poorly reconstructed.

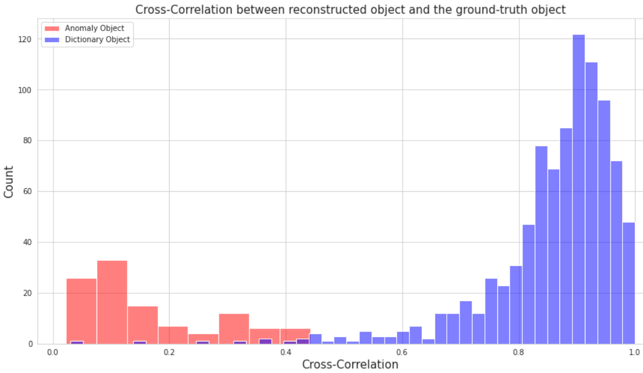

Fig. 8: Histogram of the cross-correlation between reconstructed object and the ground truth object clearly shows two different distributions with low bayes error. We took P=0.001 for the maximum number of detections by SSD.

TABLE II: Mean and Standard for the distribution of anomaly objects and dictionary objects

Object Type	Mean	Standard Deviation
Anomaly	0.16	0.12
Dictionary	0.86	0.11

As the reconstruction is in place, any correlation index can be efficiently used for quantitatively assessing the similarity between an input image and its reconstruction. The values obtained from an estimate of the mean and standard deviation of the Pearson correlation coefficient are reported for each population (normals i.e. the dictionary on one hand, and abnormals i.e. anomalies on the other hand), in Table 2. These distributions whose the absolute difference between the mean values normalized by the greater standard deviation is approximately equal to 5.8[4] are well separated (Fig. 8).

IV. Conclusion and Perspectives

In this paper, we have presented a novel architecture named Detector-encoder AutoEncoder (DAE) achieving a pattern matching associated with a more or less faithful reconstruction accessible to human expertise. The DAE model learns to reconstruct normal objects detected by the detection part, and once trained it reconstructs the detected objects fully even when the predicted bounding boxes are not so precisely positioned. The model poorly reconstructs the abnormal objects or objects abnormally distorted when detected, making it a novel approach in anomaly segmentation tasks. Although the detection part and reconstruction part have been separately trained, their combination constitutes of a fully trainable model which can be embedded in a multi-task architecture. First, the layers of a spatial transformer [13] could be added to process misaligned inputs. DAE allows decomposition of composed ornaments into vignettes from a (large) dictionary, and can help in the extension of such a dictionary as well as figuring out non-authentic vignettes (so non-authentic ornaments). Conceptually, DAE aims to strengthen autoencoders for anomaly / novelty segmentation. It has still to be compared with state-of-the-art autoencoders used for this purpose. In a future work, unsupervision and self-supervision will be explored.

Acknowledgment

The present work was achieved in the framework of the Rey's Ornement Image investigation (ROIi) project (ANR-020-CE38-0005-02) supported by the French National Research Agency.

References

[1] Bahier-Porte C., Vial-Bonacci F, *Le commerce de la librairie à la lumière de la correspondance – Marc Michel Rey, Pierre Rousseau, Charles Weissenbruch*, Journal encyclopédique aux humanités numériques. Trois siècles d'histoire du livre et de la pensée à travers le Fonds Weissenbruch, Bruxelles, Archives générales du Royaume, p. 205-222, 2019

[2] Vercruysse J., *Typologie de Marc-Michel Rey*, Wolfenbütteler Schriften zur Geschichte des Buchwesens, IV, p. 167-185, 1981

[3] Baur, C., Wiestler, B., Albarqouni, S., Navab, N., *Deep autoencoding models for unsupervised anomaly segmentation in brain MR images,* in International MICCAI Brainlesion Workshop, p. 161-169, Springer, Cham., 2018

[4] Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., Chen, H., *Deep autoencoding gaussian mixture model for unsupervised anomaly detection,* in International conference on learning representations, 2018

[5] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., *Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression,* in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 658-666, 2019

[6] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., *Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression,* in Proceedings of the AAAI Conference on Artificial Intelligence, p. 12993-13000, 2020

[7] Zhang, Y. F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T., *Focal and Efficient IOU Loss for Accurate Bounding Box Regression,* 2021

[8] Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., Yang, R., *IoU Loss for 2D/3D Object Detection,* in 2019 International Conference on 3D Vision (3DV), IEEE, p. 85-94, 2019

[9] Kosub S., *A note on the triangle inequality for the Jaccard distance*

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C., *SSD:Single-Shot Multibox detector,* in European conference on computer vision. Springer, Cham, p. 21-37, 2016

[11] Yao, Y., Yang, Y., Su, X., Zhao, Y., Feng, A., Huang, Y., Pu, H., *Optimization of the Bounding Box Regression Process of SSD Model,* in 3rd International Conference on Computer Engineering, Information Science and Application Technology (ICCIA 2019). Atlantis Press, p. 328-336, 2019

[12] He, K., Gkioxari, G., Dollár, P., Girshick, R., *Mask R-CNN,* in Proceedings of the IEEE international conference on computer vision. p. 2961-2969, 2017

[13] Jaderberg, M., Simonyan, K., Zisserman, A. *Spatial transformer networks,* in Advances in neural information processing systems, 28, 2017-2025, 2015.

[4] $\frac{|0.86-0.16|}{max(0.12,0.11)}$