Visual Watermark Identification from The Transparent Window of Currency by Using Deep Learning

Duo Tong and Wei Qi Yan
Auckland University of Technology, Auckland 1010 New Zea

ABSTRACT

Banknote identification plays an increasingly important role in financial fields due to the diffusion of automatic bank systems in terms of vending machines. Nowadays, YOLOv5 has become the state-of-the-art detector of visual objects because of its relatively outperformed accuracy with a high speed of computing. In this book chapter, the squeeze-excitation (SE) attention module is mingled with the terminal of the backbone in YOLOv5 to further improve visual watermark recognition of paper banknotes. The main contribution of this chapter is that the output precision reaches 99.99% by utilizing the novel model YOLOv5+SE.

Keywords: Banknote Identification, YOLOv5, Squeeze-Excitation (SE) Attention Module, Efficient Channel Attention Module, YOLOv5+SE, Backbone, Precision, Accuracy, Automated Banking System

INTRODUCTION

Despite the escalating commence of electronic currency, nowadays, banknotes remain galore owing to the indispensability in circulation, which means currency issuers have still confronted the menace of forging. With the prevalence of automated systems such as vending machines, currency recognition has become increasingly significant in a number of financial sectors such as currency exchange centers, shopping malls, banking systems and ticket counters (Mittal, 2018). Meanwhile, fraud techniques have been increasingly, resulting in the light of recognizing fake currency (Zhang & Yan, 2018; Yan, 2021). Besides, numerous nations suffer from the forged currency on a large scale due to its ease of printing (Trinh, *et al*. 2020). Hence the identification of counterfeit currency has become one of the most redhot topics.

As a genre of image classifications in the computer vision, currency recognition is defined as the process of identifying the denomination and the authenticity of currency (Singh, *et al*. 2010). In order to effectively determine its credibility, it is necessary for banknotes to be inspected for several specialized security features involving serial number, puzzle number, the color-changing bird, raised ink and transparent window. There are a vast variety of methods to detect currency that majorly consists of digital image processing, machine learning, and deep learning algorithms.

In recent years, deep learning has boomed in image classification and detection areas. As a kind of machine learning methods, they take use of a neural network framework consisting of multiple layers that are mainly constructed to perform classification tasks directly from sounds, images and textures. Deep learning approaches exceed conventional machine learning algorithms in precision and accuracy, though they require much data and training time. Another contributing factor of deep learning for being a popular technology in computation is that the complexity is increasingly declined with the enhancement of data and the layers of a neural network. There are various deep learning architectures in terms of VGG (Simonyan & Zisserman, 2015; Russakovsky, et al. 2015), YOLOv5 (Jocher, G. 2020), Faster R-CNN (Ren, *et al.*

2015), AlexNet (Krizhevsky, Sutskever, & Hinton, 2017) and GoogleNet (Szegedy, *et al.* 2015), which are utilized to find patterns from training data.

A state-of-the-art algorithm is selected to perform this task in this chapter. YOLOv5 is an appropriate model because of its excellent performance on object detection, acceptable precision, the first implementation on currency recognition with an attention mechanism.

Therefore, the focus of this research project is on visual watermark recognition of paper currencies through implementing deep learning algorithms involving YOLOv5 and its variants (YOLO-SE), which comprises of the SE attention block. Remarkably, the experiments have the huge size of the dataset to improve the precision and generalization ability. Hence, data augmentation consisting of cropping, flipping, rotation, colour modification, and noise addition was implemented.

Figure 1. The security features of currency, the highlighted window is the target for object detection.

This book chapter aims to achieve currency identification based on the transparent window whose phases are separated into data collection, data augmentation, denomination recognition and the analysis of the outcomes. The contributions in this research are majorly summarized into three-folds: The construction of comprehensive samples, the proposal of YOLOv5-SE, and the result analysis.

First of all, regarding the requirement of dataset in deep learning, the samples in this experiment involve the front and back sides, the changes of location, size, and others. As a result, we create a relatively full-scale dataset, which is beneficial for the experiments.

Also, YOLOv5-SE, the amalgamation of YOLOv5 and Squeeze-and-Excitation (SE) attention mechanism, is proposed to identify currency watermarks. Moreover, we compare YOLOv5 and YOLOv5-SE, analyse the likely reasons according to the experimental outcomes and the analysis from existing research work, which effectively evaluate the advantages and disadvantages. Last but not least, we conduct complementary experiments to attest the relationship between the depth of networks and noises and performance in this case.

In this chapter, following related work, methodology will be iterated. The result analysis will be explicated which leads to the final conclusion of this book chapter.

RALATED WORK

The process of currency identification includes banknote recognition and verification (Frosini, Gori, & Priami, 1996). Throughout the pertinent historical work, there are various successful methods to identify currency. In 2014, MATLAB platform was applied to recognize the credibility of currencies by utilizing red, green and blue components for segment as well as the standard deviation for evaluation (Alekhya, Prabha, & Rao, 2014).

The recognition of currency also took use of MATLAB based on PCA (principal component analysis) and LBP (local binary patterns) for the objectives of training and matching, respectively (Gautam, 2020). The images taken by the digital camera under ultraviolet light were converted into grayscale ones. Based on image processing achieved by MATLAB, the division of multiple parts by cropping, the intensity of each feature was calculated to confirm the incredibility of currency. The system acquires a high accuracy of 100% when testing the images from the dataset. Nevertheless, it fails to identify the hidden features involving latent images and watermarks. The given six features are not precisely extracted because of the variance of the size for each currency note, either.

Another method was presented to detect currency based on the features extracted from frequency domain, which applied the spatial characteristics in banknote images to accomplish the task (Shah, Vora & Mehta, 2015). The classifying process involves four phases in terms of the preprocessing for the optimal, the implementation of a two-dimensional discrete wavelet transforms, the extraction of coefficient statistical moments from the approximate efficient matrix and the utilization of serial number extraction through the deployment of OCR to detect fake currency.

Both image processing and machine learning are considered effective solutions for currency detection (Upadhyaya, Shokeen, & Srivastava, 2018). *k*-means algorithm was employed to cluster similar characteristics and an SVM classifier was taken into account to train the classifier for currency recognition (Kamal, *et al*. 2015). In 2019, HMM was employed as a robust currency detection algorithm (Kamble, *et al*. 2019). The proposed method is utilized to differentiate paper currency from various nations via modelling the texture features as a random process. To assess the performance of the algorithm, beyond 100 denominations from distinct countries were involved in the experiment, whose outcomes indicated 98% precision for currency detection.

Deep learning-based models have been also applied to currency identification. The features of currency are extracted by utilizing convolutional neural network (CNN) under the framework of single shot multi-box detector (SSD) (Zhang & Yan, 2018; Yan, 2021). Currency recognition with transfer learning having an extensive CNN pre-trained on enormous natural images was implemented to classify images from new classes (Mittal, 2018). CNN was applied to identify folded currency which involves angles, folding, damages and standard images (Jiao, He, & Zhang, 2018). Deep CNN was implemented as a feature extractor in currency identification without digital image processing which affirms the existence of security notes (Bharati & Pramanik, 2020).

Resulting from the necessity of being trained separately in every single component, object detection approaches such as R-CNN are complicated, slow, and difficult to be optimized (Redmon, *et al*. 2016). These restrictions motivate the emergence of YOLO algorithms, which was influenced by GoogLeNet model for excellent performance. Different from other models, such as two-stage algorithms that segment an image into the parts or segments, YOLO, as the name described, scans an image once to predict objects (Onyango, 2018). It treats object detection as a regression issue ranging from image pixels to the coordinates of bounding boxes as well as the probabilities of classes (Redmon, *et al*. 2016). YOLO aims to detect objects by precisely predicting the bounding box, including the instance and localizing it according to the bounding boxes. Currently, YOLO family has the updated versions from one to five.

YOLO structure has become one of the most popular models in visual object detection owing to its superiority to other traditional algorithms. First and foremost, since only neural networks are required to be run based on input images to predict objects at test time in lieu of a complicated pipeline, which performs extremely fast, the mean average precision (mAP) is over two times higher than other models (Redmon & Farhadi, 2017). Secondly, while predicting objects, YOLO infers globally from the images, which means, the whole image is taken into consideration in the period of training and testing. The final advantage of YOLO is the high generalization, which is applied to a new field or unexpected inputs.

Attention mechanism, which mimics human cognition, not only highlights the position that we need to focus on but also presents interests as well (Woo, *et al*. 2018). In accordance with attention, the modules are generally classified into vanilla attention and self-attention. Since all useful information from the input sequence must be compressed into a fixed-length vector, a standard encoder-decoder suffers from long sentence processing. Correspondingly, this shortcoming motivates the generation of vanilla attention that syndicates the standard encoder-decoder and the capability of learning joint alignment as well as translation (Bahdanau, Cho, & Bengio, 2014).

Self-attention refers to the particular attention mechanism that is related to various positions in a single sequence for the representation, which comprises Source2Token and Token2Token (Vaswani, *et al*. 2017). Source2Token self-attention was applied to show the significance of each token to a gamut of sequences in the representation (Lin, *et al*. 2018). The language translation models fulfil the-state-of-art performance through implementing token2token self-attention (Vaswani, *et al*. 2017).

Inspired by the achievement of self-attention in the NLP area, it has been one of the most predominant methods in computer vision. This kind of attention-based algorithms are mainly grouped into the altered transformers, the integration of convolutional neural networks, and a pure attention network.

Firstly, there are numerous research projects associated with the transformer applied to image classification. An image transformer, which incorporates self-attention into an autoregressive model, was proposed for image generation (Parmar, *et al*. 2018). By diminishing a number of hand-designed elements, a detection transformer was introduced for end-to-end object detection (Carion, *et al*. 2020). The vision transformer, which treats each image as a sequence of patches, implements the basic encoder with a supplementary learnable classification vector for image recognition (Dosovitskiy, *et al*. 2020). It, nevertheless, has an impediment to pixel-level dense detection, which contributes to the production of the pyramid vision transformer (PVT) (Wang, *et al*. 2021). The dense prediction transformer was proposed to compensate for the drawback of omitting feature resolution as well as granularity in deeper layers for convolutional networks (Ranftl, Bochkovskiy, & Koltun, 2021).

Secondly, it is prevalent for attention modules that incorporate into convolutional networks. Squeeze-and-excitation (SE) demonstrates a vast of potentials in advancing performance by reducing dimensionality (Hu, Shen, & Sun, 2018). SE, however, considerably hoists the computational complication, which arises from the capture of dependencies across all channels. Efficient channel attention (ECA) is a complementary approach to solve the problem (Wang, *et al*. 2020). ECA effectively shows the decline of channel dimensionality while obtaining cross-channel interaction via an extraordinary lightweight way. Both SE and ECA are channel-wise attention block unmatched with the demand of computer vision as the images are considered as the inputs with spatial architecture. Along with bottleneck attention module (Park, *et al*. 2018), convolutional block attention module (Woo, *et al*. 2018) refines convolutional features through adopting channel and spatial attention.

Although the combination of self-attention and convolutional networks has been widely employed, it is not inextricably bonded with convolutional neural networks in the success of computer vision. Pairwise

attention network (PSA), which is a variant of self-attention, is an independent block for image identification (Zhao, Jia, & Koltun, 2020).

Compared to recurrent neural networks (RNN) and convolutional neural networks (CNN), self-attention is much flexible because it has been modeled either long-range or local dependencies (Shen, *et al*. 2018). Another obvious superiority of self-attention is the ease of being facilitated due to its highly parallelizable computation (Vaswani, *et al*. 2017). Self-attention also has a number of limitations because of the complexity of memory and quadratic computation. In computer vision, the input with considerable spatial dimensions further engenders the tremendous cost of global self-attention implementation.

MATHEDOLOGY

We choose to manually produce a video through using the camera of the iPhone7s with the resolution of 1080 pixels at 30fps and then split it into digital images by using each frame regarding the requirement of data volume in deep learning. The instances involve $10NZD, $50NZD and $100NZD. Each monetary denomination has front (F) and back (B) sides, hence we have the string labels of six classes (including "10F", "10B", "50F", "50B", "100F" and "100B") in this dataset. To enhance the experimental precision, the opted images must be consistent with the following criteria:

- The images should be in high resolution.
- The currencies must be flat, the transparent window must be completely displayed in each image.
- The object should be displayed in the center of the image and have sufficient space to ensure a complete appearance when cropping and resizing.
- The currencies should be rotated in different angles and distances between the camera and the object when shooting a video.

Data labeling refers to the action of marking visual object on the given images with a bounding box whose four coordinates are stored in the corresponding "XML" file (Lee, Im, & Shim, 2019). The marked images demonstrate the recognizable patterns and tell machines which object will be detected. It plays an imperative role in deep learning as computers have no target to recognize without image annotation. In this chapter, the annotation tool called LabelImg labels the objects within a rectangle. An example is shown in Figure 2.

Figure 2. The example of data labelling

The construction of efficient and robust deep learning networks covets massive high-quality data, particularly in the situation of sharing features amongst the involved classes (Rey-Area, *et al*. 2020). Due to the heavy reliance on big data, deep learning algorithms are able to effectively eschew overfitting, which is defined as the phenomenon if the model perfectly utilizes the training data while it fails to fit supplementary data (Shorten & Khoshgoftaar, 2019).

Data augmentation is a feasible approach to remedy the scarcity of data and class imbalance. Data augmentation aims to bolster the variability of the original data so that the deep learning models acquire higher robustness to the input images collected from various environments (Bochkovskiy, Wang, & Liao, 2020). Data augmentation enriched the volume and quality of datasets by encompassing a set of image transformation techniques including geometric and photometric augmentation (Iwana & Uchida, 2021). The former refers to Affine transformations such as rotation, flipping, cropping, scaling and zooming, while the latter means color modification such as color jittering and manipulation, edge improvement, and PCA. These two types are effective ways to avoid overfitting in deep learning. The samples of image augmentation of this project is shown in Figure 3.

Figure 3. The samples of image augmentation

In this chapter, we espouse the conventional augmentations to complete the task of data preparation. The augmentation scheme we applied including flipping, rotating, cropping, color tweak and noise injection. Differentiated from the previous algorithms such as Faster R-CNN, YOLOv5 is a single-stage detector. The frame of YOLOv5 composes of three major components: Backbone, neck and output, which is described in Figure 4.

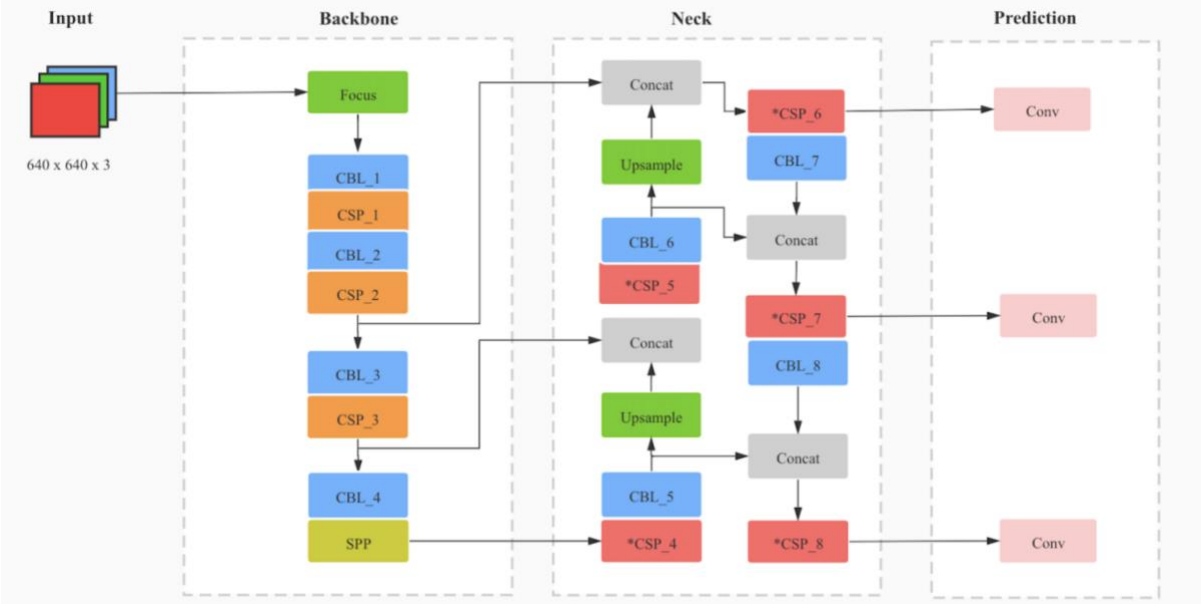


Figure 4. The architecture of YOLOv5

The input image with the resolution of 640×640×3 passes through the focus module in the backbone. It firstly takes use of slicing operations to become the feature map with resolution 320×320×12. Under 32 convolution kernels, it varies to a 320×320×32 one. Next, incorporating CSP networks into DarkNet functions as the main part of the backbone, which is utilized to extract informative features from inputs. It effectively and significantly reduces the duplication of gradient information in the optimization of convolutional neural networks, especially for large-scale backbones. Besides, it combines gradient variance and feature map to decrease the parameters and floating-point operations per second, guaranteeing inference speed and precision, whereas shrinking the model size.

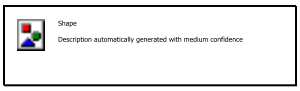
As for the neck, PANet, which is constructed with a bottom-up path based on FPN structure, is responsible for acquiring feature pyramids. From top to bottom, the FPN layer transmits solid semantic features. In the converse direction, the feature pyramid transfers positional features. This design enhances the transmission of low-level features and promotes the accuracy of locations for objects.

The head of YOLOv5, which is as same as the fourth generation, engenders feature maps with three different sizes (18×18, 36×36, 72×72) to predict targets in multiscale including small, medium, and oversized objects.

We modify YOLOv5 model by adding a SE block after the *CSP_4 module. As a computational unit, the process of SE is mainly grouped into squeeze and excitation through the operation of global average pooling and fully connected layers, respectively. Next, it takes use of the self-gating mechanism (sigmoid) to limit the output of FC to interval [0, 1], and finally multiplies this value as the scale to the channels so as to be the input data of the next stage. The principle of this structure is to enhance the important features and cripple the unimportant ones by controlling the size of the scale so as to significantly highlight the extracted features.

Global spatial statistics are squeezed into a channel descriptor by utilizing global average pooling to produce channel-wise information. The information $z \in R^c$ is produced by spatial dimensions $H \times W$. U

indicates the collection of statistics expressive for the full image (H×W×C). The *c*-th element of *z* is computed by eq.(1).



(1)

Two fully connected layers achieve the squeezing process that aims to aggregate the valuable information, followed by the excitation to catch all channel-wise dependencies. The first layer, followed by ReLU, compresses *C* channels into *C/r*, *r* means the percentage of compression) channels to reduce the spate of computation. According to the research conducted (Hu, Shen, & Sun, 2018), the SE module achieves a superior tradeoff among precision and complication if *r* equals 16. Hence, we utilize this value for the experiment. The second full connection is used to restore to *C* channels, which follows the sigmoid function.

In order to measure loss, YOLO takes use of the sum-squared error between predictions with the highest IoU and ground truth. Its loss function consists of the localization, the confidence (also known as the objectness of boxes) and the classification loss. YOLOv5 adopts GIoU to be the localization loss rather than Intersection over Union (IoU), where IoU is defined as eq.(2).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where *A* and *B* are two arbitrary convex shapes, *IoU* refers to the similarity among *A* and *B*, which allows the coordinates to be related to each other and has scale invariance, which overcomes the weakness of smooth L1 loss (Rezatofighi, *et al*. 2019). However, IoU suffers from the optimization problem if *A* and *B* have no intersection.

Alternatively, GIoU is an effective key to address the above issues, which inherits the advantages of IoU in terms of the invariance of scale and all properties of loss metrics (Rezatofighi, et al., 2019). In contrast to *IoU*, it focuses on overlapping areas and non-overlapping regions, which better reflects the intersection among *A* and *B*. The definition of *GIoU* is shown as eq.(3).

$$GIoU = IoU - \frac{|C(A \cup B)|}{|C|} \tag{3}$$

where *C* indicates the minimum encompassing convex object. Thus, *GIoU* is an appropriate replacement for *IoU* in computer vision tasks, even though it has some explicit restrictions.

In this project, the model evaluation metrics involve three indicators: Precision (*P*), recall (*R*) and average precision (*mAP*). In visual object detection, both precision and recall are the two fundamental assessment criteria. Precision refers to the percentage of accurately recognized objects amongst all detected samples, whilst recall is defined as the proportion of precisely identified objects among all positive instances detected, eq.(4) and eq.(5) are the equations for these two indices, *mAP* is a comprehensive indicator that takes precision and recall rate into consideration, which is calculated by using the average precision (*AP*) over the number of classes (*M*), *mAP* indicates the performance throughout all classes while *AP* demonstrates the performance on a given class *i*. The calculation of *mAP* is shown as eq.(6).

$$P = \frac{TP}{FP+TP} \tag{4}$$

$$R = \frac{TP}{FN+TP} \tag{5}$$

$$mAP = \frac{1}{M}\sum_{i=1}^{M} AP_i \quad (6)$$

where *AP* is defined as the region under the precision and recall curve, which is shown as eq. (7).

$$AP = \int_0^1 P(R)dR \quad (7)$$

RESULT ANALYSIS

Corresponding to the research objectives of this book chapter, three experiments were conducted. To complete the task of currency detection, the secondary work we need to cogitate is the stages of research from data generation, training data and resultant analysis. The experimental implementation can considerably benefit from the basic idea about the specific process of currency identification.

In Figure 5, the first stage is the preparation of data, which is separated into four parts, including shooting a video, the split of images by frames, label marking, and augmentation. The first two actions are applied to acquire original images for the experiment. The next process is for the computers to recognize the inputs, while the final one is to address the scantiness of data and buttress the generalization capability. After that, the outcomes of data augmentation are utilized as the input of the designed model to complete the transparent window recognition of currency, which includes training, feature extraction, dense detection, and the acquisition of the results. Specifically, the features of input images are compressed down through the backbone to complete feature extraction and forwarded to detection neck and head to accomplish feature aggregation and detection, respectively. In particular, we merge localization and classification into one step since YOLOv5 is a one-stage detector in which these two operations for every bounding box are implemented simultaneously.

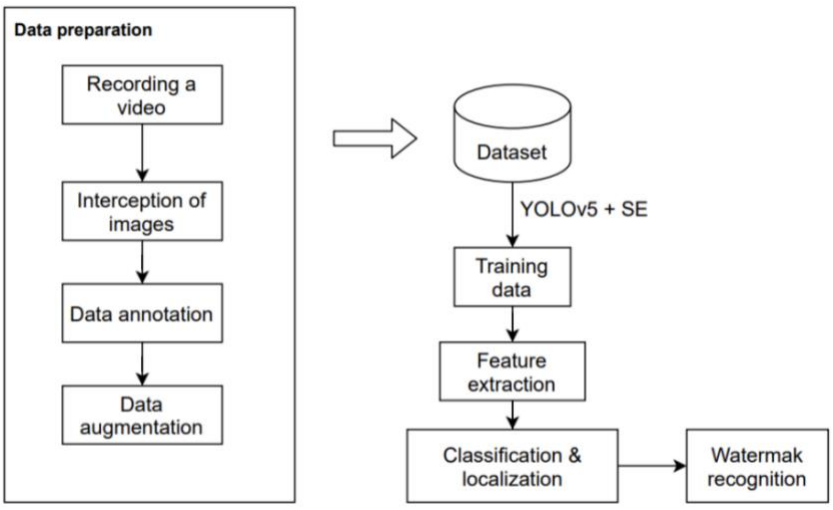


Figure 5. The phases of the proposed experiments

The outcomes of watermark identification are shown in Figure 6. The six classes are represented by bounding boxes with different colors. Specifically, the blue, light green, purple, pink, yellow and green ones show the classes with textual labels “10F”, “10B”, “50F”, “50B”, “100F” and “100B”. For instance, the string “10F” means the front side of transparent windows for $10NZD. Instead, “10B” depicts the backside of the corresponding currencies. The confidence values are ranging from 0.88 to 0.92.

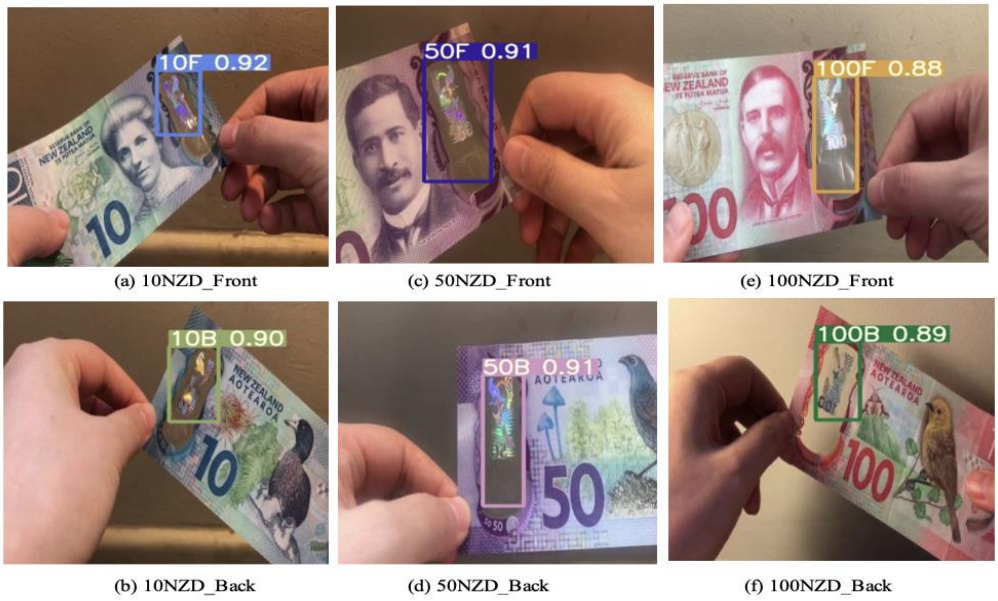


Figure 6. The results of currency watermark detection

As we all know, the only variance among different versions of YOLOv5 is the depth of networks. Therefore, we implemented the experiments, the comparison is shown in Table 1, pertaining to YOLOv5s, YOLOv5m, YOLOv5l and the updated versions that amalgamate the SE module. Table 1 indicates that YOLOv5l-SE obtains the outperforming outcomes on GIoU loss, accuracy, and mAP with the toll of training time. In other words, with the addition of SE block, each model costs longer time on training while other evaluating results such as GIoU and accuracy become superior.

For the state-of-the-art architectures, the SE module generates crucial performance promotion with the minimal expenditure on computation, which is consistent with our experimental results that more satisfying accuracy and loss value are acquired with the sacrifice of the executing speed (Hu, Shen, & Sun, 2018). Throughout utilizing global information, SE attention mechanism explicitly models dynamic and nonlinear dependencies among channels (Hu, Shen, & Sun, 2018). Thereby, the new model eases the process of learning and dramatically hikes the representative ability of the network. These are the main reasons for the superiority of the proposed algorithm. However, due to the high accuracy rate of the basic model, the attention-based algorithm has little space to improve the performance, hence the betterment of each evaluation criteria is not apparent.

As implied in the outcomes of all the modified models, by enhancing depth of networks, the model performs better except for training time. YOLOv5l-SE takes the longest time to train data compared to other models, attributing to that the deeper networks can extract much features so that it can promote the overall performance but will lead to much more computational costs and overhead.

Table 1. The outcomes of multiple models

Models	Training Time (s)	GIoU Loss	Precisions	mAP@0.5
YOLOv5s	439.56	0.02282	0.9972	0.9959
YOLOv5s-SE	461.16	0.02227	0.9980	0.9960
YOLOv5m	666.72	0.02066	0.9989	0.9959
YOLOv5m-SE	673.56	0.02047	0.9990	0.9960
YOLOv5l	1004.4	0.01995	0.9988	0.9958
YOLOv5l-SE	1100.72	0.01950	0.9990	0.9961

CONCLUSION

The main objective of this research project is to detect the watermark of paper currency based on deep learning because we integrate the SE attention module in YOLOv5 as the detector. Fortunately, the proposed model presents satisfactory outcomes.

Throughout the corresponding operations, we conclude that the attention-based model outperforms the unmodified on precision, mAP and GIoU with the exchange of time consumptions. The other finding is that the overall performance will be promoted by enhancing network layers meanwhile training time will be longer.

REFERENCES

Alekhya, D., Prabha, G., & Rao, G. (2014) Fake currency detection using image processing and other standard methods. *International Journal of Research in Computer and Communication Technology*, 3, 128 - 131.

Bahdanau, D., Cho, K., & Bengio, Y. (2014) Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations*.

Bharati, P., & Pramanik, A. (2020) Deep learning techniques - R-CNN to Mask R-CNN: A survey. In *Proceedings of Computational Intelligence in Pattern Recognition* (pp. 657 - 668), Springer.

Bochkovskiy, A., Wang, C., & Liao, H. (2020) YOLOv4: Optimal speed and accuracy of object detection, CoRR abs/2004.10934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020) End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision* (pp. 213 - 229), Springer.

Chambers, J., Yan, W., Garhwal, A., Kankanhalli, M., (2014) Currency security and forensics: A survey. *Multimedia Tools and Applications,* 74(11), 4013-4043.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020) An image is worth 16×16 words: Transformers for image recognition at scale, In *Proceedings of ICLR*.

Frosini, A., Gori, M., & Priami, P. (1996) A neural network-based model for paper currency recognition and verification. In *Proceedings of IEEE Transactions on Neural Networks* (pp. 1482 - 1490), IEEE Press.

Gautam, K. (2020) Indian currency detection using image recognition technique. In *Proceedings of International Conference on Computer Science, Engineering and Applications* (pp. 1 - 5).

Hu, J., Shen, L., & Sun, G. (2018) Squeeze-and-excitation networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132 - 7141)

Iwana, B., & Uchida, S. (2021) An empirical survey of data augmentation for time series classification with neural networks, *PLoS One*, 16(7): e0254841

Jiao, M., He, J., & Zhang, B. (2018) Folding paper currency recognition and research based on convolution neural network. In *Proceedings of International Conference on Advances in Computing, Communications and Informatics* (pp. 18 - 23).

Jocher, G. (2020) YOLOv5, Code repository, 2020, *https://github.com/ultralytics/yolov5*.

Kamal, S., Chawla, S.S., Goel, N., & Raman, B. (2015) Feature extraction and identification of Indian currency notes. In *Proceedings of National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics* (pp. 1 - 4), IEEE Press.

Krizhevsky, A., Sutskever, I., Hinton, G. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60 (6): 84 - 90.

Lee, Y., Im, D., & Shim, J. (2019) Data labeling research for deep learning based fire detection system. In *Proceedings of International Conference on Systems of Collaboration Big Data, Internet of Things & Security* (pp. 1 - 4)

Lin, Z., Feng, M., Santos, C., Yu, M., Xiang, B., Zhou, B., Bengio, Y. (2017) A structured self-attentive sentence embedding, In *Proceedings of ICLR*.

Ma, X., Yan, W. (2021) Banknote serial number recognition using deep learning. Multimedia Tools and Applications.

Mittal, S., & Mittal, S. (2018) Indian banknote recognition using convolutional neural network. In *Proceedings of International Conference on Internet of Things: Smart Innovation and Usages* (pp. 1 - 6).

Onyango, L. (2018) *Convolutional neural network to enhance stock taking*, University of Nairobi, Kenya.

Park, J., Woo, S., Lee, J., & Kweon, I. (2018) BAM: Bottleneck attention module. In *Proceedings of BMVC*.

Parmar, N., Vaswani, A., Uszkoreit, J., Ukasz, K., Shazeer, N., Ku, A. (2018) Image transformer. In *Proceedings of International Conference on Machine Learning (ICML)*.

Ranftl, R., Bochkovskiy, A., Koltun, V. (2021) Vision transformers for dense prediction. In *Proceedings of ICCV*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016) You Only Look Once: Unified, real-time object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 779 - 788).

Redmon, J., & Farhadi, A. (2017) YOLO9000: Better, faster, stronger. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7263 - 7271).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Ren, Y., Nguyen, M., Yan, W. (2018) Real-time recognition of series seven New Zealand banknotes International. *Journal of Digital Crime and Forensics (IJDCF)*, 10 (3), 50-66

Rey-Area, M., Guirado, E., Tabik, S., & Ruiz-Hidalgo, J. (2020) FuCiTNet: Improving the generalization of deep learning networks by the fusion of learned class-inherent transformations. *Information Fusion*, 63, 188-195.

Russakovsky, O., Deng, J., Su, H., et al. (2015) ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115 (3), 211 - 252.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015) *Going deeper with convolutions*. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1 - 9).

Shen, T., Zhou, T., Long, G., Jiang, J., Zhang, C. (2018) Bi-directional block self-attention for fast and memory-efficient sequence modeling, In *Proceedings of ICLR*.

Shorten, C., & Khoshgoftaar, T. (2019) A survey on image data augmentation for deep learning. *Big Data*, 6, 60.

Simonyan, K., Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conference on Learning Representations.*

Singh, S., Tiwari, A., Shukla, S., & Pateriya, S. (2010) Currency recognition system using image processing, *International Journal of Engineering Applied Sciences and Technology*.

Trinh, H., Vo, H., Pham, V., Nath, B., & Hoang, V. (2020) Currency recognition based on deep feature selection and classification. In *Proceedings of Asian Conference on Intelligent Information and Database Systems* (pp. 273 - 281), Springer.

Upadhyaya, A., Shokeen, V., & Srivastava, G. (2018) Analysis of counterfeit currency detection techniques for classification model. In *Proceedings of International Conference on Computing Communication and Automation* (pp. 1 - 6), IEEE Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017) Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*.

Wang, G., Wu, W., Yan, W. (2017) The state-of-the-art technology of currency identification: A comparative study. *International Journal of Digital Crime and Forensics*, 9 (3), 58-72

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020) ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11531 - 11539).

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021) Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of ICCV* (pp. 568-578).

Woo, S., Park, J., Lee, J.Y., & Kweon, I. (2018) CBAM: Convolutional block attention module. In *Proceedings of ECCV* (pp. 3 - 19).

Yan, W., Chambers, J. (2013) An empirical approach for digital currency forensics. IEEE International Symposium on Circuits and Systems (ISCAS), 2988-2991.

Yan, W., Chambers, J., Garhwal, A. (2014) An empirical approach for currency identification. *Multimedia Tools and Applications*, 74 (7)

Yan, W. (2021) *Computational Methods for Deep Learning - Theoretic, Practice and Applications*. Springer (ISBN 978-3-030-61080-7).

Yan, W. (2019) *Introduction to Intelligent Surveillance Surveillance Data Capture, Transmission, and Analytics*. Springer.

Zhang, Q., & Yan, W. (2018) Currency detection and recognition based on deep learning. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance* (pp. 1 - 6).
Zhang, Q., Yan, W., Kankanhalli, K. (2019) Overview of currency recognition using deep learning. *Journal of Banking and Financial Technology*, 3 (1), 59–69

Zhao, H., Jia, J., & Koltun, V. (2020) Exploring self-attention for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10073 - 10082).