P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia
Science & Technology Asia
Vol.28 No.1 January - March 2023
Page: [120-135]
Original research article

Manga Face Detection on Various Drawing Styles Using Region Proposals-Based CNN

Kittinun Aukkapinyo[1], Seiji Hotta[1], Worapan Kusakunniran[2,*]

[1]*Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Tokyo 184-0012, Japan*
[2]*Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom 73170, Thailand*

Received 2 June 2021; Received in revised form 11 November 2022
Accepted 14 November 2022; Available online 20 March 2023

ABSTRACT

Faces of characters in comic books can be used as meta-features for manga analytics. Manga character faces are not easy for a machine to detect when compared to human faces due to the high variation of drawing styles from various distinct authors. There exist several convolutional neural network-based (CNN-based) frameworks that can achieve high accuracy in an object detection task. However, their drawback is time and resource consuming to perform data modeling due to the nature of deep learning. Thus, this paper is to propose a method to develop a model using Mask R-CNN, which is one of the CNN-based frameworks, with the transfer learning technique in order to reduce training time and resources while maintaining high performance in the manga character face detection task. The proposed method could achieve the average precision of 87% in the manga character face detection tasks on both seen and unseen drawing styles. It significantly outperforms the existing conventional methods. Moreover, pre-trained weights from MS COCO dataset are transferable to manga character face detection tasks. Therefore, a well-performed manga character face detector could be developed using a limited amount of training data and time.

Keywords: Comic analysis; Manga face detection; Mask R-CNN; ROI detection

1. Introduction

In this era, many comics are available in the digital format via either digitizing of physical books or originally illustrating in digital format. There are many images of comic book pages that are in a bunch of image files such as JPEG or PNG. There exists an enormous number of manga books in the world that have a high variance in strokes and drawing styles for various creators. It might be very hard for humans to

***Corresponding author:** worapan.kun@mahidol.edu
doi: 10.14456/scitechasia.2023.11

analyze information or compare the drawing styles of each manga. Hence, the implementation of a system that can extract character faces from volumes of comics for further analysis can be helpful to fulfill this task. There are several analyses of comics using the faces of characters that can be done. For example, comic genre analysis, appearance for each character in a comic, and character's facial expression that might help to analyze the emotion of this comic. Another possible application is to automate the annotation of unlabeled manga character faces. It can accurately generate bounding boxes and masks as ground-truths for other tasks such as training a machine to illustrate manga using Generative Adversarial Network (GAN) [1–3].

The manga character faces mostly are a human face. However, the existing human character face detection techniques cannot be used to detect and localize the manga character's faces. This is because there are many key differences between real-world and manga human faces such as unreal facial expressions, unnatural facial styles, and illustrated organs. There exist several manga characters face detectors that can generate bounding boxes around the faces of manga characters. The existing work technique can be classified into local binary patterns (LBP) classifiers and CNN-based detectors described in the next section.

1.1 Related Works

The first existing related project is Image::AnimeFace. [4] It is the project that uses a local binary pattern (LBP) based classifier algorithm from the OpenCV library [5] to create a manga face detector. A detector is trained on hundreds of thousands of positive and negative samples. Positive samples are manga character face images, and negative samples are non-face images. It does not take color information in the training process since it is trained using images in grayscale mode. This approach does not take much time to train and has a robustness against geometric transformation such as occlusion and rotation. It generates bounding boxes that cover manga character faces in each frame. This detector has a fast speed of detection; however, it has less precision and recall since it is not generalized enough towards variations of drawing styles.

Xiaoran Qin et al. [6] proposed the second technique which is the Faster R-CNN-based comic character faces detector which is one of the CNN-based approaches for an object detection task. It consists of two neural networks that share some convolutional layers. They are region proposal network (RPN) and convolutional neural network (CNN), which is one kind of deep neural network used for image classification. It is robust to scaling and rotation. It also provides very high accuracy if the amount of training data is sufficient. The RPN will generate all candidate regions of interest (ROIs) in an image that is likely to contain faces. Then, all proposed ROIs will be parsed into the CNN to classify each proposal into face and non-face. Finally, bounding boxes with their corresponding confidence score will be generated for each detected face.

The Faster R-CNN detector was trained from scratche with big datasets of scanned comics pages. They randomly created the dataset called JC2463, which contains 2463 manga pages from Manga109 [7, 8]. They also randomly created AEC912, which is the dataset that contains 912 American comics. So, it should take a long time and consume resources for computation and modeling data. Their model can achieve the

highest precision of 91% in their testing set of JC2493 and AEC912 which are the seen drawing styles. They do run their detector through eBDtheque [9], which is the dataset that does not contain the drawing style in the training dataset to measure generalization of the detector and achieve a significantly dropped precision of 75%. Since the model is trained from a large number of training examples, it leads to the need for high computational resources and times.

Several related works on human face detection which is similar to our task, have been introduced. Most proposed methods are region-based convolutional neural networks and one-shot detectors [10–14]. Face Detection Data Set and Benchmark (FDDB) [15] is the common dataset used in their evaluations. It is a dataset of human faces in various unconstrained settings and environments. The one-shot detector is more suitable with real-time face detection since it has only one convolutional neural network to localize the human face in images with a small amount of computational time required. There are several related applications and benefits, such as person identification, face recognition, and facial expression analysis [16–18].

The method proposed in this paper[14] aims to detect human faces in crowds using Mask-RCNN. The implementation of Mask-RCNN is also applied in the method proposed in this paper. This work adjusted hyperparameters of the network to improve detection on obscured faces. For example, the set of anchors for the RPN is set to be 8, 16, 32, 64, and 128. The data modeling and evaluation is done on several face detection benchmark datasets such as Labeled Faces in the Wild (LFW) and Face Detection Data Set and Benchmark (FDDB).[15, 19] The used evaluation metric is average precision (AP) at Intersection-over-Union (IoU) of 0.5. The author claimed that it could outperform the current state-of-the-art using Max-Margin Object Detection (MMOD)[20] and cascade classifiers with much better average precision value.

1.2 Contribution

This paper is the further development of Image::AnimeFace[4] and Faster R-CNN-based comic character faces detector[6]. In Image::AnimeFace[4], the authors apply LBP to character face detection from comics. It is the traditional approach in computer vision that is used to create an object detection system. It used less training and detecting time; however, it cannot handle high variance of unseen drawing styles due to its bottleneck. In other words, adding more data cannot improve its performance of the face detection measured by precision and recall. Hence, its detection performance is poor when compared to the work that applies a deep learning framework to develop a manga character face detector which is Faster R-CNN-based comic character faces detector[6].

The deep learning framework improves the data modeling process. More generalization and accuracy of the model can be achieved by adding more data. At the same time, their Faster R-CNN framework requires considerable time in its training process since they execute a data modeling process to train the model from scratch. In other words, they don't use any pre-trained model to initialize weights in each neuron. In contrast, the Faster R-CNN consumes a little time when performing a detection task from a trained model and yields much better accuracy as it is more robust to unseen drawing styles when compared to the traditional LBP classifier in Image::AnimeFace [4]. Their evaluation met-

rics are precision and recall. To compare our performance with these existing methods, 2 additional metrics which are precision and recall are made for the comparison based on the best setting of our experiment. Our proposed detector can achieve a precision of 93% and recall of 89% towards unseen drawing styles which is higher than Image::AnimeFace [4] and Faster R-CNN-based comic character faces detector [6] by focusing on frontal face detection using Mask R-CNN and transfer learning.

Mask R-CNN [21] is the deep learning framework that can be used for object detection and instance segmentation tasks. Its network architecture supports a pixel-level segmentation task which could generate masks and bounding boxes as the output. Our work is applying the Mask R-CNN framework and transfer learning technique to improve the performance of manga character face detection and reduce the training time in the training process in a deep learning framework. The differences in the architecture between our related work, Faster R-CNN, and our Mask R-CNN are shown in Table 1.

The first difference is the backbone architecture. Our Mask R-CNN uses Residual Network (ResNet) and Feature Pyramid Network (FPN) to serve a feature extractor while they use only VGG-16 as a backbone. Both convolutional neural network architectures can extract features ranging from low-level to high-level features from parsed images. However, VGG-16 has an enormous number of parameters and requires large weight space which could result in memory space-consuming and slow speed in training. It also has a gradient vanishing problem when more layers are added to the network while the residual network can overcome this issue. The residual network has a much smaller weight space since

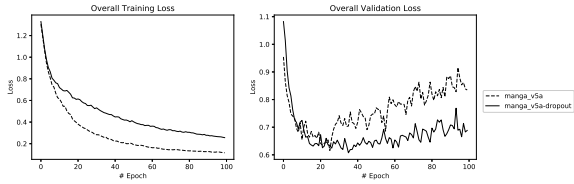

Fig. 1. Learning graph visualization for manga_v5a dataset (x-axis: # epoch, y-axis: loss).

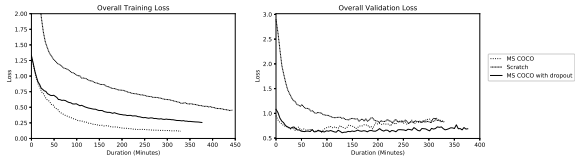

Fig. 2. Learning graph visualization for manga_v5a dataset (x-axis: # epoch, y-axis: loss).

it has skip connections to reduce complexity while having more depth, when compared to VGG-16. Its residual connection is the key capability that helps to endure the vanishing gradient problem when the depth of the network increases. In addition, the Feature Pyramid Network enables our backbone to learn features of an object at different scales which could help to make our Mask R-CNN robust against scaling.

The second difference is the dropout regularization. Some dropout layers are added inside the backbone architecture to temporarily drop some neurons out from the network. This could help to regularize the trained model not to be too sensitive to unseen drawing style. The visualization of the data modeling process on the dataset that has the highest variation in drawing styles is shown in Fig. 1. Although dropout layers might cause a model to take more time to converge, it has better generalization during the validation process in each epoch.

The next difference is the outputs from the detector. Both detectors can out-

put bounding boxes and confidence scores for each detected face. However, our Mask R-CNN can output pixel-level masks of manga character faces along with their corresponding bounding boxes and scores. It helps in extraction of manga character faces as pixel-level masks.

Another difference is in the pooling method. Our Mask R-CNN uses RoIAlign or Region-of-Interest(RoI)Align, which is the pooling method proposed in Mask R-CNN [21] that can scale down all-region proposals into one pre-defined size with less loss in quantization than RoIPool or Region-of-Interest(RoI) Pooling since RoIPool performs the max pooling on parsed feature maps. It can deal with misalignment of boundaries in the result caused by a round-off error.

The last difference is the application of the transfer learning technique. The pre-trained weights from the MS COCO dataset [27], which is provided by matterport along with their Mask R-CNN Python library [26], are used as initial weights for our backbone and classifier. The main benefit is that the accurate and generalized models can be trained much faster with fewer resources from the well-trained model. The learning graph shown in Fig. 2 is the learning graph of manga_v5a, which is one of our datasets. From the overall training loss graph, it takes much longer to reduce the training loss when a detector is trained without applying the transfer learning technique. The training time is significantly improved once a detector is trained with the pre-trained weights of MS COCO. Although overall validation loss from different scenarios regarding applying pre-train weights is almost identical as the elapsed time, a dropout regularization can help in generalization along with pre-train weights.

The rest of this paper is organized as follows. The proposed method is explained in Section 2. Experimental results and comparisons are shown in Section 3. Discussions and future works are concluded in Section 4. Then, conclusions are drawn in Section 5.

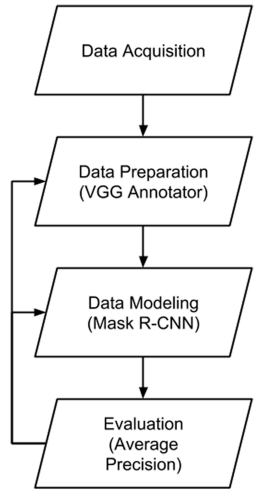

Fig. 3. The overview of proposed method.

2. Proposed Method

The proposed framework for training a model of the Manga face detector is shown in Fig. 3. It is an iterative process that consists of 4 subprocesses. The first subprocess is to acquire the dataset for training the detector. This subprocess is important since the quality of the dataset affects the performance and precision of our detector. The second subprocess is data preparation. Using all data from the acquired dataset might result in time and resources consumption and it is not necessary that training with all data will result in having a good model. So, many sub-datasets will be created from the acquired dataset to find the best combination of manga volumes that will make a model that is generalized enough toward

Table 1. Structural difference of our related work's Faster R-CNN and our Mask R-CNN framework.

Difference	Faster R-CNN	Our Mask R-CNN	Advantages
Backbone	VGG-16[22]	-ResNet50[23] -Feature Pyramid Network[24]	-Smaller weight space -Solve gradient vanishing problem when more layers are added.
Dropout Regularization[25]	No	Yes	Less complex network architecture improves the generalization of the trained model
Outputs	-Bounding Boxes -Confidence level	-Bounding Boxes -Confidence level -Mask	Can output mask along with bounding boxes and confidence level
Pooling Method	RoIPool	RoIAlign	No quantization in scaling down region proposals
Pre-trained weights	No	MS COCO pre-trained weights from matterport[26]	Much faster training time and less resource is used

unseen drawing styles.

The next subprocess is the data modeling. Our Mask R-CNN detector will be trained using the data from the two previous subprocesses. The last subprocess is an evaluation. The average precision will be used to measure the performance of models. Since our proposed method is an iterative process, more sub-datasets can be created if the learning and evaluation results are not satisfied. The details of each step are explained in the following subsections.

The proposed method differs from [4, 6] in the data modeling and evaluation process. In the data modeling, the Mask R-CNN framework is used to implement our manga character faces detector instead of using LBP classifier and Faster R-CNN which provides benefits and better performance, for example, less training time and a more generalized data model. Our method can generate bounding boxes along with their confidence level and pixel-level mask. In the evaluation section, the proposed Mask R-CNN detector is evaluated using mAP (mean Average Precision)

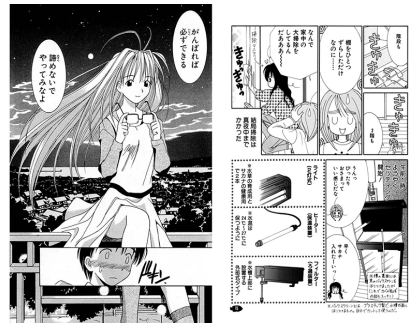

Fig. 4. An example of training samples from Manga109 dataset. ©Akamatsu Ken and Kuniki Yuka

in PASCAL VOC style at the intersection over union (IoU) threshold of 0.5 instead of using precision and recall [28]. It is the standard metric in the object detection task in Computer Vision. It can evaluate a detector from its average of the maximum precisions at different recall levels.

2.1 Data Acquisition

The dataset for this work should be the volumes of manga in digitized comics that are readable by a machine. There

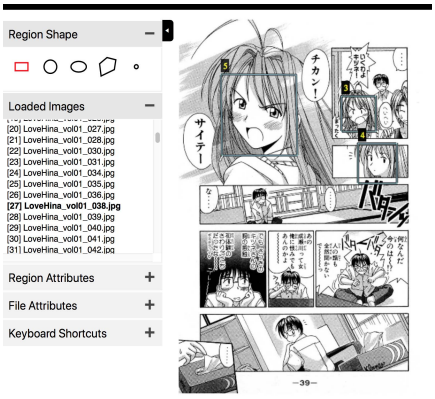

Fig. 5. Annotation of character's face using VGG image annotator. ©Akamatsu Ken (A manga face from "Love Hina" by Akamatsu Ken)

should be adequate variance drawing styles of manga character faces in the dataset to enable them to represent most drawing styles of manga in this world. It will significantly contribute to the performance and generalization of the model that will be used as a core of the Mask R-CNN detector.

2.2 Data Preparation

Several sub-datasets are created from various combinations of manga volumes from Manga109 [6, 7]. Examples of training samples are shown in Fig. 4. It is to find the best dataset for robust manga face detection on various drawing styles, based on empirical evidence. First, one manga volume is randomly picked as a base volume. After that, other manga volumes in Manga109 are added to create different composite datasets. Then, our detector is trained with composite datasets and evaluated on both seen and unseen drawing styles. The generalization of the model can be measured when the evaluation is done on unseen drawing styles. The process is repeated until the satisfying result is achieved. All frontal faces on each page of digitized manga are annotated as ground-truths for data modeling. In each dataset, digitized comics pages are randomly split into the training set, validation set, and testing set with a proportion of 60:20:20 respectively. The testing set is unseen to the Mask R-CNN during the training phase but it contains the same drawing styles as in the data modeling process.

The VGG Annotator Tool (VIA)[29] is used as an annotation tool in this paper. It allows us to define ground-truth regions and their corresponding class in an image file. Annotations are saved as JavaScript Object Notation (JSON) and comma-separated values (CSV) files to parse annotated ground-truths into the network. The example of a manga character's face annotation via the VGG annotator tool is shown in Fig. 5.

2.3 Data Modeling

Mask R-CNN is one of the state-of-the-art for object detection tasks using the deep learning approach. Therefore, it is used for the proposed data modeling procedure. It is a framework that is an extension of Faster R-CNN that can perform object detection and pixel-level instance segmentation. Our library provides two modes of Mask R-CNN which are the training mode and inference mode.

Mask R-CNN is created in the training mode for the data modeling process. The proposed Mask R-CNN has an architecture as shown in Fig. 6. It consists of several components that play different roles that contribute to the detection. One important component is the backbone that serves as feature extractors. Our Mask R-CNN uses the Residual Network-50 (ResNet50) with Feature Pyramid Network (FPN) as a backbone architecture. ResNet50 employs a residual connection to prevent gradient vanishing as shown in Fig. 7. Its

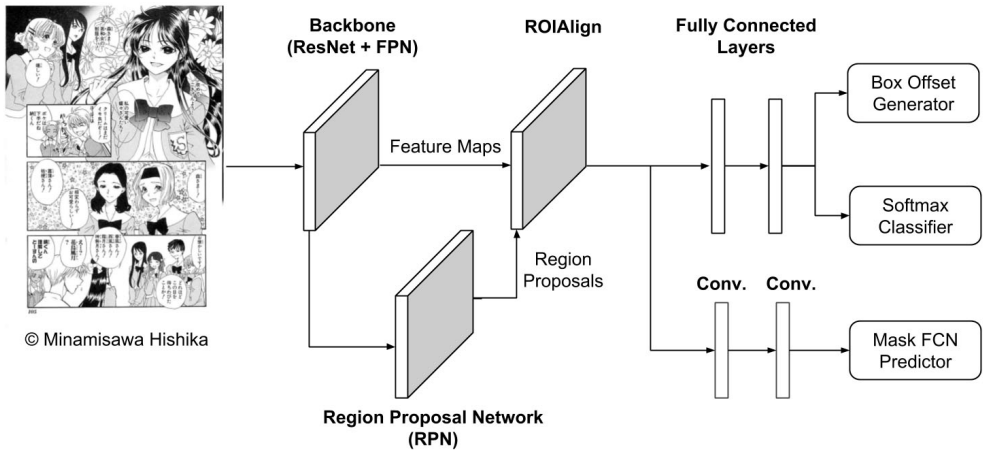

Fig. 6. The architecture of the proposed Mask R-CNN based solution.

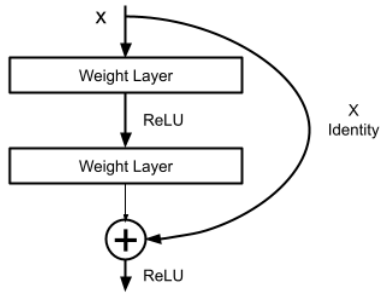

Fig. 7. A residual connection in ResNet50.

skip connection improves feature extraction of a backbone. They generate feature maps for Region Proposal Network (RPN) to learn to generate candidate region proposals that are likely to contain manga character faces. Then, feature maps are parsed into RoIAlign which is a component that does a re-scale of all-region proposals into one pre-defined size. This scheme helps Mask R-CNN to use only one size of feature maps for all candidate regions proposals.

After that, region CNN features are parsed into fully connected layers to locate and classify each face of the manga character in an input image. Bounding boxes, masks, and confidence levels for each face in an input image are generated as outputs. The loss function used for bounding boxes is smooth L1-loss. Cross entropy loss is used as a loss function for classification of class ID from softmax classifier at the end of the network. Also, it is used for calculating a loss of each pixel in a mask.

The transfer learning[30] technique is applied during the training process of Mask R-CNN as well. It allows us to transfer knowledge from one task into one similar task by transferring neural networks weights between similar tasks. It helps in the reduction of time and resources used in data modeling with the deep neural network. The pre-trained network weights are used as initial weights in the retraining of some layers in the network for a new task.

2.4 Evaluation

The Mask R-CNN is running through each manga volume which have different drawing styles in the inference mode with trained models and weights. In other words, testing datasets drawn in unseen drawing styles are necessary for measuring the gen-

eralization of the model. This repeats what the previous sentence says. It can be deleted. The evaluation metric used to evaluate the performance is the Average Precision with IoU (Intersection over Union) from Pascal VOC Challenge [28]. The true positive, which is a frontal manga character face, is counted when a predicted proposal overlaps with ground-truth region reach defined IoU threshold. This approach is used commonly for comparing performance between two object detectors. It used 11-points interpolated an average precision to calculate AP from average of precision values at 11 recall levels as shown in Eq. 2.1

$$AP = \frac{1}{11} \sum_{r \in \{0,0.1,\ldots,1\}} p_{interp}(r), \quad (2.1)$$

where AP represents an average precision from each image; the r denotes a set of recall levels ranges from 0 to 1, $p_{interp}(r)$ represents the maximum precision interpolated at recall r. It is one approach to calculate AUC (Area Under Curve) of the Precision-Recall curve. As the level of recall increase, the precision will decrease. The AP for each image is calculated, then the average of APs is used as AP of a model on each test set.

Table 2. Information for each sub-dataset used in the experiment.

Dataset	Number of Training Faces	Training Data
manga_v1	36	LoveHina_vol01
manga_v2	79	LoveHina_vol01 + Nekodama
manga_v3	128	LoveHina_vol01 + Nekodama + MagicStarGakuin
manga_v4	159	LoveHina_vol01 + Nekodama + MagicStarGakuin + MagicianLoad
manga_v5	405	LoveHina_vol01 + SaladDays_vol01
manga_v5a	1231	LoveHina_vol01 + SaladDays_vol01 + EverydayOsakanaChan + BakuretsuKungFuGirl
manga_v6	654	LoveHina_vol01 + SaladDays_vol01 + YumeiroCooking

Note: All images in a dataset are manually annotated only for frontal faces.

3. Experiments

The dataset used in this project is Manga109 [6,7] from the Aizawa Yamasaki Laboratory at the University of Tokyo. It contained digitized 109 manga volumes from different 96 manga authors during the 1970s to 2010s. The top three genres are humor, science fiction, and romance. The average number of pages in each volume is 195 pages. It contains 118,715 faces including both frontal and rear faces. Due to the high variance in drawing styles and aspects on faces, many small datasets were created from the Manga109 dataset to find the best performing model. Moreover, the greater the amount of training data, the more time and resources for training are required. The details about the training data used in each dataset are shown in Table 2. Comic volume names used in each training dataset are shown in the training data column. For example, manga_v2 consists of pages from Love Hina and Nekodama.

This paper focuses on frontal face detection since it can help to reduce the variation of faces, where the frontal face is more beneficial for further use. In each experiment, some hyperparameters such as the learning rate and weight decay are fixed. The learning rate is 0.001 and weight decay is set to 0.0001. The learning rate is set to a small value since an optimal solution can be skipped and weight decay is just a basic regularization in the deep neural network. Since each dataset has a different volume of manga and number of images, the number of steps per epoch is varied by the number of manga pages in each training dataset.

3.1 Results with pre-trained weights

The first experiment is to model the data using pre-trained weights of the residual network (ResNet) of MS COCO dataset from matterport [26]. All datasets were pre-

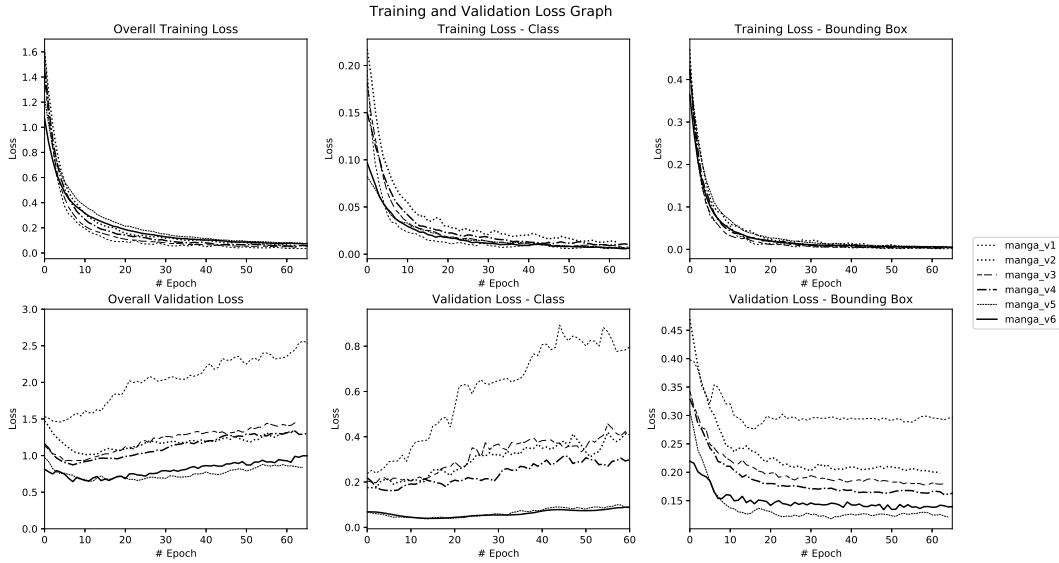

Fig. 8. Learning graph visualization for manga dataset version 1 to 6 (x-axis: # epoch, y-axis: loss).

pared and annotated for frontal faces manually. Each dataset is composed of training and validation dataset. The head part of Mask R-CNN is re-trained with prepared training datasets and pre-trained weights of MS COCO dataset. The training loss and validation loss are calculated at the end of each epoch to see how our trained weights fit the data. The result from training is shown in Fig. 8. All loss graphs during the training phase with manga dataset version 1 to 6 are almost identical. Training losses are decreased as the number of epochs increases. In contrast, the validation loss of all datasets in the validation phase starts to increase after 30 epochs are passed.

After the data modeling is terminated due to the stability of losses, all models are evaluated using Average Precision (AP) at IoU of 0.5. They are evaluated on their own testing dataset which is the seen drawing style. After that, the evaluation on the unseen drawing styles is done on 3 volumes of the manga. The evaluation result is shown in Table 3. All Average Precision of all models is dropped when they are evaluated on unseen drawing styles.

3.2 Results with pre-trained weight after applying overfitting reduction technique

In this experiment scenario, it attempts to improve generalization of our model by applying the overfitting technique in our data modeling process. One of the common techniques is to add dropout layers to reduce the complexity of the network. It is one of the regularization techniques used for preventing neural networks from overfitting. Its concept is to randomly ignore some neurons with probability 1 - p during training. In other words, some nodes will be dropped out from the network. This technique can help to reduce codependency amongst each neuron during training which can lead to overfitting.

In this scenario, The manga_v5 dataset is used as an experimental dataset. It is the dataset that consists of one volume of comics called LoveHina and SaladDays. Dropout layers are added to the network with different position and dropout rate as

Table 3. The evaluation table of all dataset in experimental scenario 3.1.

Dataset	Seen Drawing Style	Unseen Drawing Styles		
	manga_vN testing set	Donburakokko	YouchienBoueigumi	Count3DeKimeteAgeru
	AP	AP	AP	AP
manga_v1	0.820	0.515	0.354	0.357
manga_v2	0.867	0.684	0.508	0.547
manga_v3	0.720	0.583	0.418	0.470
manga_v4	0.598	0.425	0.360	0.342
manga_v5	0.818	0.662	0.320	0.591
manga_v6	0.788	0.674	0.456	0.642

Note: Intersection over Union (IoU) used in an evaluation is 0.5.

an experiment.

The learning result of the model after applying dropout regularization is shown in Fig. 9. The training loss of models with dropout layers is slowly reduced when compared with the normal training. The validation loss of models with some dropout configurations starts to increase after several epochs passed by. The best setting is the insertion of dropout layers before the activation layer in the residual network (ResNet), softmax classifier, and mask predictor with probabilities of 0.2, 0.2, 0.2, respectively. It has the best validation loss when compared to the rest of the configurations. The line that represents the learning of the network in this configuration is manga_v5_dropout5. The probability for dropout layers should not be high since the more neurons that are dropped out, the fewer features that the network can learn in each epoch. The high number of dropped neurons can lower the performance of a backbone.

Dropout layers help to decrease a validation loss or overfitting in our trained weights in case that they were placed in the right locations with the right dropout rate. They can significantly reduce the validation loss although it might be still high due to high variation in drawing styles of manga character faces in the dataset. Some samples of results from this and previous experiment scenario are shown in Fig. 10. Even though not all faces of characters in this page are detected, there is an improvement in the detection after the dropout regularization is applied.

Our Mask R-CNN with trained models work is evaluated using the Average Precision metric as mentioned in the previous section. There are three test datasets which are unseen to our Mask R-CNN during training. In other words, they have unseen drawing styles of faces to measure generalization of our detectors. In addition, our proposed detector is compared against 2 existing methods in the related works [4, 6]. Since Image::AnimeFace published their source code, their performance can be evaluated using the same metric as us. For the Faster R-CNN detector, it is implemented with the same architecture as our related work as shown in Table 2. It is trained on the manga_v5 and manga_v5a datasets with the similar hyperparameters of our related work [6]. From the evaluation, the proposed Mask R-CNN with dropout regularization outperforms both traditional LBP Cascade and Faster R-CNN. The compari-

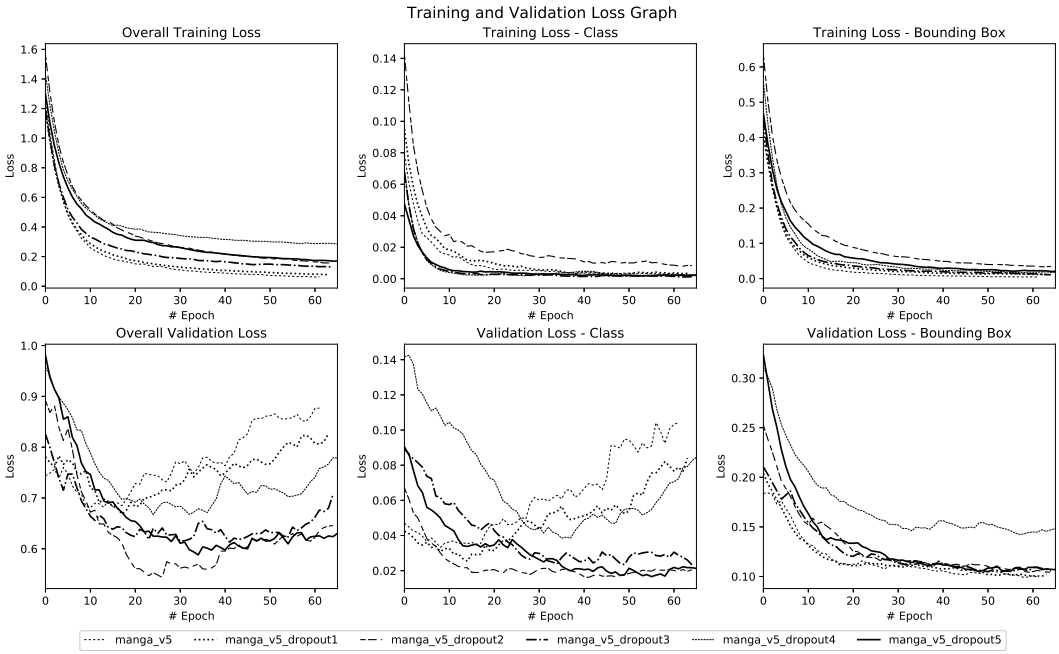

Fig. 9. Learning graph visualization of manga model version 5 (manga_v5) (x-axis: # epoch, y-axis: loss).

Fig. 10. Samples of result images before and after applying dropout layers of Mask R-CNN detector trained using manga_v5 sub-dataset. ©Tenya and Nakanuki Eri

son of our proposed detector and LBP cascade classifier is shown in Table 3.

From the table, all Mask R-CNN that are trained using pre-trained weights with the manga_v5 dataset, which consists of only two frontal face drawing styles and few training examples, can outperform LBP Cascade one. Although there is no improvement for Count3DeKimeteAgeru dataset, the AP of our Mask R-CNN detector for Donburakokko and Youchien-Boueigumi testing dataset are getting better if dropout layers are added into the network. Also, there is an attempt to increase variation of drawing styles and training examples with manga_v5a, which is the extension of our manga_v5 datasets that consists of two more distinct drawing styles and has training examples for frontal faces of 1231 in total. The result from training a backbone using the manga_v5a dataset is improved AP on test datasets which have unseen drawing

styles.

4. Discussion

The manga face detector can be trained by retraining the network with the pre-trained weight of COCO provided by matterport. However, the trained model suffered from the overfitting issue. The high validation loss can indicate the overfitting of our model. As a result from overfitting, a model cannot detect some unseen manga character faces in some input images as shown in Fig. 11. The more difference in drawing styles, the harder for a detector to detect faces since manga character faces might not share something in common like human faces. Their strokes and structure of faces are varied by authors. This fact might be due to high variance in the drawing style of manga character faces of testing datasets. From Table 3, all models have high performance in detecting their seen drawing styles. They can detect frontal faces with seen drawing styles and their highest Average Precision is 87%. However, their performance is significantly reduced when an evaluation is done on unseen drawing styles. The highest Average Precision is reduced to 69%. Hence, the model is not generalized enough toward unseen drawing styles.

Our next attempt is to reduce overfitting issues by using dropout regularization technique. It helps our model to cope with an overfitting issue. The ignoring of some neurons during training can improve the generalization of the model as the validation loss of the model decreases in each epoch. The complexity of the network is partially reduced and contributes to an improved generalization of the model. However, it still has an issue of overfitting, but it is better than just training using pre-trained weights. The dropout regularization can help us to reduce overfitting in training.

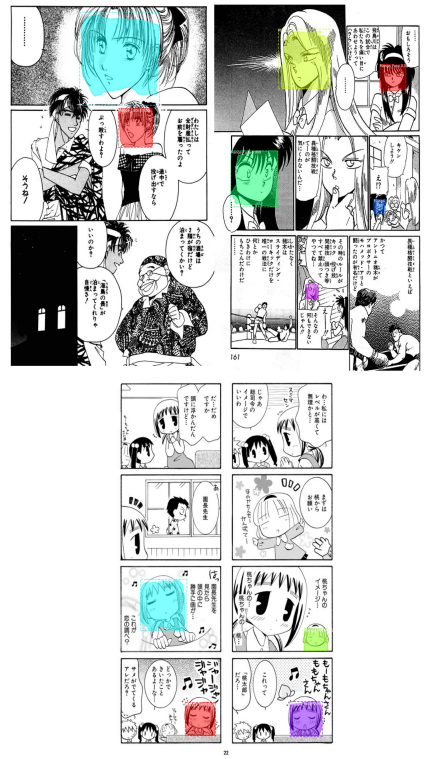

Fig. 11. Samples of result images of unseen drawing styles with loss detection ©Nakanuki Eri, Omiya Naoi, and Tenya.

In the end, trained weights from manga_v5a, which is a dataset that consists of four distinct drawing styles of faces and has 1231 annotated frontal faces, is selected as weights in our Mask R-CNN backbone to identify which locations in input images contain faces of comic characters. The whole network of Mask R-CNN is retrained using manga_v5a with dropout layers inside the backbone and classifier part of the network. Its performance is the best among every detector in Table 4. Thus, it is the most generalized detector that can detect unseen faces drawing styles.

Table 4. The evaluation table of detectors.

Testing Dataset	Manga Face Detector	AP @ IoU = 0.5
Donburakokko	LBP Cascade (lbpcascade_animeface)	0.24
	Faster R-CNN (manga_v5)	0.32
	Faster R-CNN (manga_v5a)	0.58
	Mask R-CNN (manga_v5)	0.66
	Mask R-CNN (manga_v5 with 5th dropout config.)	0.77
	Mask R-CNN (manga_v5a with 5th dropout config.)	0.87
YouchienBoueigumi	LBP Cascade (lbpcascade_animeface)	0.41
	Mask R-CNN (manga_v5)	0.32
	Faster R-CNN (manga_v5)	0.35
	Faster R-CNN (manga_v5a)	0.53
	Mask R-CNN (manga_v5 with 5th dropout config.)	0.60
	Mask R-CNN (manga_v5a with 5th dropout config.)	0.82
Count3DeKimeteAgeru	Faster R-CNN (manga_v5)	0.22
	LBP Cascade (lbpcascade_animeface)	0.32
	Faster R-CNN (manga_v5a)	0.57
	Mask R-CNN (manga_v5)	0.59
	Mask R-CNN (manga_v5 with 5th dropout config.)	0.77
	Mask R-CNN (manga_v5a with 5th dropout config.)	0.81

5. Conclusions

This paper proposed a manga character face detection method by using Mask R-CNN and transfer learning techniques. The proposed methods are an iterative process that consists of data acquisition, data preparation, data modeling, and evaluation. It begins with data acquisition, which Manga109 dataset is used as our dataset. Sub-datasets and images were annotated in the data preparation procedure. Mask R-CNN is used to model prepared sub-datasets, and COCO pre-trained weights are used. Then, Mask R-CNN is created in an inferencing mode with trained models to measure its performance using Average Precision at IoU of 0.5. The testing dataset is volumes of the manga which have unseen drawing styles than the training data. Subprocesses in the proposed method are repeated until a model with the best Average Precision of 0.87 from sub-dataset manga_v5a is found. The pre-trained weights of MS COCO dataset can be used as initial weights for this project task which is the manga character face detection.

In the future, the more various combinations of manga from Manga109 dataset might be created to find the one that can produce more generalized weights. When more data and combination of drawing styles are used for training, overfitting issues can be overcome. Moreover, higher Average Precision (AP) towards unseen drawings styles can be achieved. Another future work might be a creation of hair, face and neck detectors for each manga character separately. The results from each detector can be combined for detecting a manga character face region.

References

[1] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434. 2015.

[2] Chen Y, Lai YK, Liu YJ. Cartoongan: Generative adversarial networks for photo cartoonization. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 9465–74.

[3] Liu Y, Qin Z, Luo Z, Wang H. Autopainter: Cartoon image generation from sketch by using conditional generative adversarial networks. arXiv preprint arXiv:170501908. 2017.

[4] nagadomi. AnimeFace 2009. GitHub; 2009. https://github.com/nagadomi/animeface-2009.

[5] Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools. 2000.

[6] Qin X, Zhou Y, He Z, Wang Y, Tang Z. A Faster R-CNN Based Method for Comic Characters Face Detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01; 2017. p. 1074–80.

[7] Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, et al. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications. 2017 Oct;76(20):21811–38. Available from: https://doi.org/10.1007/s11042-016-4020-z.

[8] Ogawa T, Otsubo A, Narita R, Matsui Y, Yamasaki T, Aizawa K. Object Detection for Comics using Manga109 Annotations. Computing Research Repository. 2018;abs/1803.08670. Available from: http://arxiv.org/abs/1803.08670.

[9] Guérin C, Rigaud C, Mercier A, Ammar-Boudjelal F, Bertet K, Bouju A, et al. eBDtheque: A Representative Database of Comics. In: 2013 12th International Conference on Document Analysis and Recognition; 2013. p. 1145–9.

[10] Jiang H, Learned-Miller E. Face detection with the faster R-CNN. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE; 2017. p. 650–7.

[11] Sun X, Wu P, Hoi SC. Face detection using deep learning: An improved faster RCNN approach. Neurocomputing. 2018;299:42–50.

[12] Wang Y, Ji X, Zhou Z, Wang H, Li Z. Detecting faces using region-based fully convolutional networks. arXiv preprint arXiv:170905256. 2017.

[13] Yang W, Jiachun Z. Real-time face detection based on YOLO. In: 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII). IEEE; 2018. p. 221–4.

[14] Kneis B. Face Detection for Crowd Analysis Using Deep Convolutional Neural Networks. In: International Conference on Engineering Applications of Neural Networks. Springer; 2018. p. 71–80.

[15] Jain V, Learned-Miller E. FDDB: A Benchmark for Face Detection in Unconstrained Settings. University of Massachusetts, Amherst; 2010. UM-CS-2010-009.

[16] Tong SG, Huang YY, Tong ZM. A Robust Face Recognition Method Combining LBP with Multi-mirror Symmetry for Images with Various Face Interferences. International Journal of Automation and Computing. 2019;16:671.

[17] Wu H, Chen ZW, Tian GH, Ma Q, Jiao ML. Item Ownership Relationship Semantic Learning Strategy for Personalized Service Robot. International Journal of Automation and Computing. 2020;17:390.

[18] Lian Z, Li Y, Tao JH, Huang J, Niu MY. Expression Analysis Based on Face Regions in Real-world Conditions. International Journal of Automation and Computing. 2020;17:96.

[19] Huang GB, Mattar M, Berg T, Learned-Miller E. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. Marseille, France: Erik Learned-Miller and Andras Ferencz and Frédéric Jurie; 2008. Available from: https://hal.inria.fr/inria-00321923.

[20] King DE. Max-margin object detection. arXiv preprint arXiv:150200046. 2015.

[21] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017. p. 2980–8.

[22] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computing Research Repository. 2014;abs/1409.1556. Available from: http://arxiv.org/abs/1409.1556.

[23] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8.

[24] Lin T, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 936–44.

[25] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014 Jan;15(1):1929–58. Available from: http://dl.acm.org/citation.cfm?id=2627435.2670313.

[26] Abdulla W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. Github; 2017. https://github.com/matterport/Mask_RCNN.

[27] Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55.

[28] Everingham M, Gool LV, Williams C, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 2010 Jun;88(2):303–38. Available from: https://doi.org/10.1007/s11263-009-0275-4.

[29] Dutta A, Gupta A, Zissermann A. VGG Image Annotator (VIA); 2016. Version: 1.0.6, Accessed: 12-06-2018. http://www.robots.ox.ac.uk/vgg/software/via/.

[30] Yosinski J, Clune J, Bengio Y, Lipson H. How Transferable Are Features in Deep Neural Networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14. Cambridge, MA, USA: MIT Press; 2014. p. 3320–8. Available from: http://dl.acm.org/citation.cfm?id=2969033.2969197.