Clothing detection using YOLOv3 in ModaNet

and

DeepFashion2 datasets.

Vijay Raya Pati, Rohan, s. Bhargava

**Abstract:** Object detection is one of the important technologies in the field of computer vision. In

the area of fashion apparel, object detection technology has various applications, such as apparel

recognition, apparel detection, fashion recommendation, and online search. The recognition task is difficult for a computer because fashion apparel images have different characteristics of clothing

appearance and material. Currently, fast and accurate object detection is the most important goal

in this field. In this study, we proposed a two-phase fashion apparel detection method named

yolov3-TPD (yolov3 Two-Phase Detection), based on the yolov3 algorithm, to address this

challenge. The target categories for model detection were divided into the jacket, top, pants, skirt,

and bag. According to the definition of inductive transfer learning, the purpose was to transfer the

knowledge from the source domain to the target domain that could improve the effect of tasks in

the target domain. Therefore, we used the two-phase training method to implement the transfer

learning. Finally, the experimental results showed that the map of our model was better than the

original yolov3 model through the two-phase transfer learning. The proposed model has multiple

potential applications, such as an automatic labeling system, style retrieval, and similarity detection.

**Keywords:** object detection; yolov3; fashion apparel; deep learning; transfer learning

1. **Introduction**

An image is worth a thousand words. The fashion apparel industry is one of the fields

that have a large image usage. These clothing images appear in newspapers, magazines,

e-commerce platforms, social media, and even advertising boards. There are various

object detection applications that extract information from images and apply it to apparel

recognition, apparel detection, fashion recommendation, and online search for the fashion

apparel industry, which can be used to flourish human life.

In recent years, deep learning technology in computer vision has many applications

and researches, for example, image classification , object detection , semantic segmentation , and instance segmentation . Among them, object detection technology

is a rapidly developing research topic. Object detection methods and other computer

vision algorithms are important to the fashion apparel industry. The clothing information

marking system based on the object detection method can automatically mark images in

social media or online stores and create information tags . The online store can use the

created tags for clothing quick search or similar clothing recommendations in the

future. Additionally, occupation recognition , fashion style recognition , and fashion

style recommendation are all extended applications in the field of computer vision.

Clothing images have a massive distinction with different characteristics of clothing

appearance, style, and posture . Different types of clothing might be similar in types

of material and color. Thus, the computer finds it difficult to recognize different types of

clothing. Due to this diffificulty, the task of detecting clothing by using computer vision

technology becomes a difficult challenge. With the improvement of GPU ’s computational

capabilities, the field of machine learning and deep learning has had a huge breakthrough

Therefore, deep learning scientists can construct deeper neural networks to solve the

unstructured data problem that machine learning algorithms cannot handle.

The mainstream object detection algorithms are based on convolution neural networks

(CNN) that are divided into one-stage detection and two-stage detection, by using different

feature extraction methods. Object detection algorithms that use a two-stage detection

method include R-CNN , Fast R-CNN , and Faster R-CNN, which divide the detection task into (1) region proposal and (2) classification. The two-stage detection method has a high detection accuracy, but the problem of high time complexity leads to a longer detection time.

Therefore, the two-stage detection method is hard to use

in real-time detecting applications. The one-stage detection method integrates region

proposal and classification into one step, which reduces the detection time. The mainstream

methods of one-stage detection are SSD , YOLO , YOLOv2 , YOLOv3 , and

yolov3 . These methods improve detection speed without losing too much accuracy

and are increasingly becoming popular object detection algorithms.

Transfer learning is the process of applying knowledge and skills learned in previous

tasks to target tasks. According to research by Panetal. , transfer learning is divided into three categories by different domains and tasks (1) Inductive Transfer Learning,

(2) Transductive Transfer Learning, and (3) Unsupervised Transfer Learning.

In this research, we proposed a fashion apparel detection model based on the You Only

Look Once (yolov3) algorithm named yolov3-TPD (yolov3 Two-Phase Detection

model for fashion apparel) by using the characteristics of inductive transfer learning.

The contributions of this paper are as follows. First, we proposed an object detection

model by using two-phase transfer learning for detecting fashion apparel images with

complex background. Second, the experimental results proved that implementing two

phase transfer learning and the CLAHE image enhancement method could improve the

accuracy of model detection. Third, the detection accuracy of our model was better than

the states of the art methods in the field of fashion apparel.

The rest of this paper is organized as follows. Section 2 reviews the related works on

apparel image recognition. Section 3 introduces the overview and steps of the yolov3-

TPD model. Section 4 describes the dataset, experimental, and results. Finally, Section 5

includes the study conclusion.

LITERATURE REVIEW

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| YEAR | AUTHOR | TITLE &  ARCHITECTURE | DOMAIN & PROBLEM STATEMENT | CONCLUSION & RESULT |
| 2016 | Joseph  Redmon,  Santosh  Divvala,  Ross  Girshick, Ali Farhadi | You Only Look Once: Unified, Real-Time Object Detection | A fast and simple approach to detecting real time images was introduced in this paper as You Only Look Once. The model was built to detect images accurately, fast and to differentiate between art and real images. | In comparison with Object detection techniques that came before YOLO, like R-CNN, YOLO introduced a single unified architecture for regression go image into bounding boxes and finding class probabilities for each box. This meant that YOLO performed much faster and also provided more accuracy. It could also predict artwork correctly. |
| 2018 | Chengji Liu, Yufan Tao,  Jiawei Liang,  Kai Li1, Yihang Chen | Object Detection Based on YOLO Network | A generalized object detection network was developed by applying complex degradation processes on training sets like noise, blurring, rotating and cropping of images. The model was trained with the degraded training sets which resulted in better generalizing ability and higher robustness. | The experiment showed that the model trained with the standard sets does not have good generalization ability for the degraded images and has poor robustness. Then the model was trained using degraded images which resulted in improved average precision. It was proved that the average precision for degraded images was better in general degenerative model compared to the standard model. |
| 2018 | Wenbo Lan,  Jianwu Dang, Yangping Wang, Song Wang | Pedestrian Detection Based on YOLO Network  Model | The network structure of YOLO algorithm is improved and a new network structure YOLO-R was proposed to increase the ability of the network to extract the information of the shallow pedestrian features by adding passthrough layers to the original YOLO network. | The YOLO v2 and YOLO-R network models were tested on the test set of the INRIA data set. The experimental results show that the YOLO-R network model is superior to the original YOLO v2 network model. The number of detection frames reached 25 frames/s, basically meeting the requirement of realtime performance. |
| 2018 | Rumin Zhang,  Yifeng Yang | An Algorithm for Obstacle  Detection based on YOLO and Light Filed  Camera | An obstacle detection algorithm in the indoor environment is proposed which combines the YOLO object detection algorithm and the light field camera and will classify objects into categories and mark them in the image. | The images of the common obstacles were labeled and used for training YOLO. The object filter is applied to remove the unconcern obstacle. Different types of scene, including pedestrian, chairs, books and so on, are demonstrated to prove the effectiveness of this obstacle detection algorithm. |
| 2019 | Zhimin Mo1,  Liding  Chen1,  Wen-jing  You | Identification and  Detection of Automotive Door Panel Solder Joints based on  YOLO | A method for identifying the solder joints of automotive door panels based on YOLO algorithm that provides the type and location of solder joints in real time. For detecting the small solder joints more precisely, this paper adopts YOLO algorithm which adopts multi-level predictions, predicting on different size feature maps and combining the prediction results to obtain the final result. | The YOLO algorithm, proposed identifies the position of the solder joints accurately in real time. This is helpful to increase the efficiency of the production line and it has a great significance for the flexibility and real-time of the welding of automobile door panels. |

1. Proposed Methodology

In related work [25], it was found that the YOLOv3 has a better detection efficiency

than R-CNN, Fast R-CNN, and Faster R-CNN in fashion clothing detection. Another

study [26] showed the YOLOv2 has a faster detection speed than Faster R-CNN, in the field

of fashion apparel detection. Synthesizing the above arguments and previous researches,

the YOLO algorithm was found to be more suitable than the two-stage detection algorithms

for fashion apparel detection. Additionally, the detection model might improve the effect of

apparel detection tasks through the characteristics of inductive transfer learning. Therefore,

we proposed a two-phase fashion apparel detection model named yolov3-TPD, based on

the yolov3 algorithm and transfer learning, to detect the apparel in complex background.

The proposed detection model needs to detect five fashion apparel categories (jacket, top,

pants, skirt, and bag) and determine the location of the target in the image, showed in

Figure 1. The detection task was to detect apparel features using its contour and appearance.

Therefore, the detection effect of the model was not affected by the color of the apparel.

Figure 2 shows the architecture of the two-phase detection model. The training process of

the proposed model was divided into the data preparation phase and the model training

phase. In the data preparation phase, data labeling and data preprocessing were performed

to prepare the training data. In the training phase, the prepared training data were used to

train the model. First, the three categories classifier was trained. Second, the five categories

classifier was trained by using transfer learning

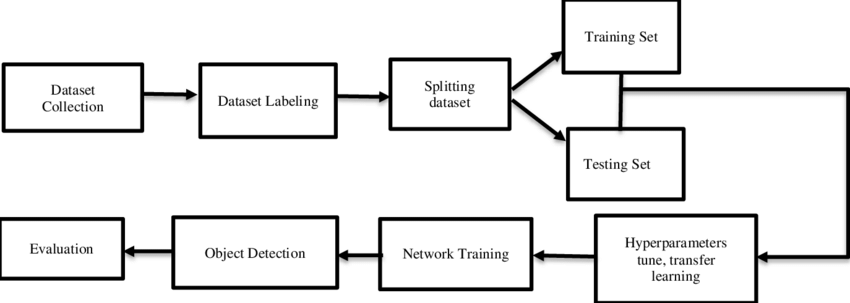


Fig 1.1 Fashion apparel detection with YOLOv3-TPD.

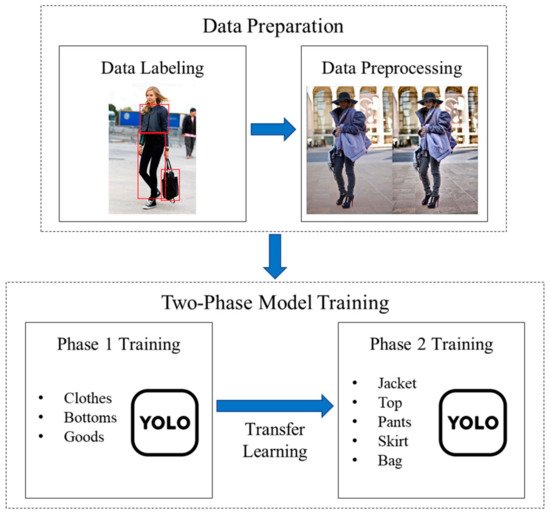


Fig 1.2 Training process of the YOLOv3-TP

*A.Data Preparation*

In this phase, the data preparation was divided into two subsections. In the first sub

section, labeling the images was needed for the training set and the testing set. The object

detection algorithm was constructed by the CNN network, which belongs to supervised

learning. The correct answer was provided to the model during the training phase. The

YOLO network required that the label information of the image must be stored in a text

format file. If there were multiple targets in one image, all information could be stored in

the same file, and each image corresponded to only one label file. The label file contained

the object number and object coordinates on this image. During the training process, the

network could determine the target in the image by using the object information stored in the label file. The two-phase detection model proposed in this research was trained by

using inductive transfer learning. Therefore, it was necessary to label the data for both

phases to provide the training data needed. Three categories of labeled data were used in

phase 1 model training, and five categories of labeled data were used in phase 2.

In the second subsection, the image was implemented through preprocessing methods

to enhance the detail of the image. The apparel images used in this research include various

kinds of apparel and accessories, such as hats, glasses, parasols, and other items that might

intercept the light. The image might have some excessive shadow or overexposure due

to the environment of the shoot. In an area with a lot of glass or metal materials, the

image might have overexposure due to reflection of light. To avoid the above potential

problems from affecting the training results of the model, this research enhanced the

image quality through the image preprocessing method of Contrast-Limited Adaptive

Histogram Equalization. Contrast-Limited Adaptive Histogram Equalization (CLAHE) [29]

is an image preprocessing technique that is used to improve contrast in images. The

CLAHE method computes several histograms that are distinct sections of the image and

redistributes the luminance values of the image. The CLAHE is suitable for improving the

local contrast and enhancing the definitions of edges.

B.*Two-Phase Model Training*

This research proposes a fashion apparel detection model based on the yolov3

algorithm. Referring to the definition of fashion apparel in [26], the classification of the

apparel in the research was similar to the actual situation in reality. Therefore, this research

aimed to detect five different categories of fashion apparel as follows—jacket, top, pants,

skirt, and bag. According to the definition of inductive transfer learning, a two-phase

training method was proposed to improve the accuracy of the detection model.

The categories of the fashion apparel were divided into two phases, as shown in

Figure 3. In the first phase, referring to the previous research [1] of the hierarchical clas

sification on fashion apparel, we simplified the five categories of the final goal into three

coarse categories of clothes, bottoms, and goods. Each first phase coarse category could be

extended to the second-phase fine categories, where “clothes” could be subdivided into

“jacket” and “top”, “bottoms” could be subdivided into “pants” and “skirt”, and “bag”

belonged to the coarse categories of “goods”.

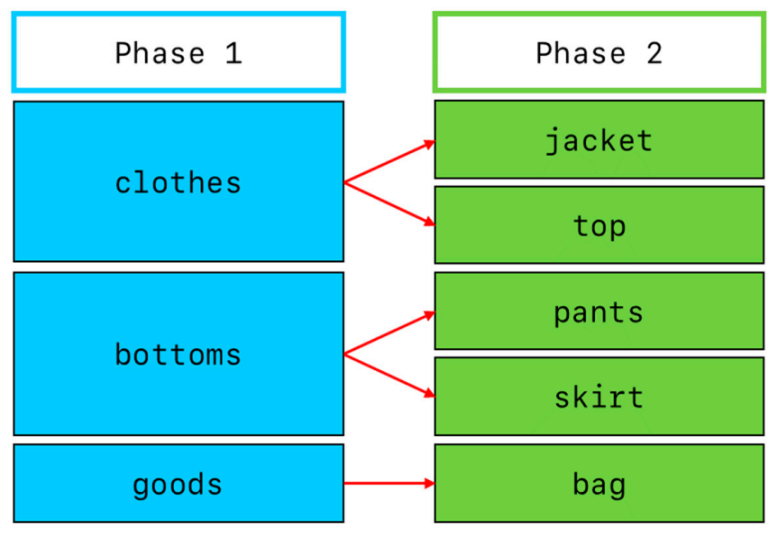


Fig 1.3 Category structure of two-phase fashion apparel model.

In the training process of the phase 1 model, the labeled data of three categories and

pre-trained model of yolov3.conv.137 were used for training. The yolov3.conv.137

was the pre-trained model provided by yolov3 authors training through the Microsoft

COCO dataset. After model training of phase 1 was over, the model weight was saved

and used as the new pre-trained model for the training of the phase 2 model. During

the phase 2 model, labeled data of fifive categories and the trained model of phase 1 were

used for transfer learning. The method proposed in this research used the characteristics

of inductive transfer learning to improve the detection effect by relearning, based on the

weight of the phase 1 model.

**Experimental Results**

In this section, the corresponding experiment materials are presented. Section 4.1

illustrates the experimental environment. Section 4.2 explains the dataset used in the

research. Section 4.3 is about the hyper parameter setting in yolov3. Section 4.4 gives the

evaluation of our proposed method. Section 4.5 concludes the experiment results.

*4.1. Experimental Environment*

The experimental environment of this research was implemented on a personal com

puter with a Windows 10 operating system. The system equipment was NVIDIA RTX 2070

super GPU using cuda 10.2 and cudnn 7.6.0, Intel i5-9600K CPU, and 32G DDR4 memory.

*4.2. Dataset*

The Clothing Co-Parsing (CCP) dataset was an open-source dataset that was con

structed by Liang et al. [30]. The CCP dataset contained 2098 high-resolution street snaps

of fashion apparel with a complex background. Each image was a full-color image and

had a uniform size (550 *×* 830). The images in this dataset included various kinds of

apparels with a complex background, which could not only suit our proposed model but

also conformed to the real situation of the street.

In the data preparation, the LabelImg software was used to label the data. The target

information could be exported to an XML format fifile including image name, path, size,

target quantity, type, and coordinates. Next, using the Python language, the XML format

was transformed to the text format, which was accepted by YOLO.

In the data preprocessing, the CLAHE method was used to enhance the image quality

and increase the contrast. Figure 4 shows the effect of the CLAHE method. The contrast of

the image on the right side was higher than the original image. The edges and contours of

the apparel were also obvious.

*4.3. Hyperparameter Setting*

This subsection lists the hyperparameters during the two-stage model training. Table 2

shows the different settings in the two-phased model. The other settings were as follows.

(1) The image size was set to 416 *×* 416 recommended by YOLO; (2) the batch size was set

to 2 due to equipment limitations; (3) the initial learning rate was set to 0.001, and (4) the

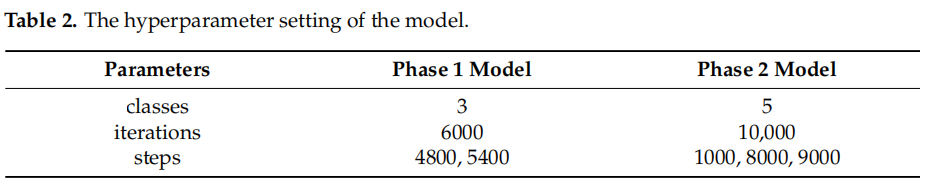
momentum and decay were referred to the original setting by the yolov3 model. In phase

1, the classes parameter was set to 3, to conform to our goal. The iterations parameter

was set to 6000 and the steps parameter was set to 4800, 5400. The learning rate would be

decreased to 0.0001 after 4800 steps and to 0.00001 after 5400 steps. Explanations for the

phase 2 model settings are the same as the phase 1 model.



*Evaluation Criterion*

Precision and recall are the most common evaluation indicators for evaluating ob

ject detection models. According to [31], the defifinition of precision and recall are in

Equations (1) and (2).

*Precision* = *TP /TP* + *FP* (1)

*Recall* = *TP /TP* + *FN* (2)

The confusion matrix divides the model detection results into the following four

categories—true positive (TP), true negative (TN), false positive (FP), and false negative

(FN). The precision measured the percentage of correct positive predictions among all

predictions made. The recall measured the percentage of correct positive predictions among

all positive cases in reality.

Intersection over Union (IoU) was the ratio between the intersection and the union of

the predicted boxes and the ground truth boxes. Referring to previous research [32], the

IoU formula is described by Equation (3).

*IoU* = *areaBP ∩ Bgt*/*areaBP ∪ Bgt* (3)

Equation (3) shows the calculation of IoU, where *Bgt* is a ground truth bounding box

and *BP* is a predicted bounding box. By calculating the IoU, we could tell that the detection

result was valid (TP) or not (FP). The most commonly used threshold was 0.5. If the IoU

was >0.5, it was considered a TP, else it was considered an FP.

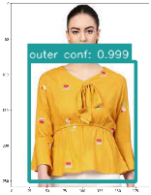
The mAP (mean average precision) was an indicator for evaluating object detection

models. The AP (average precision) was defined as the mean of the precision values and

recall values. The map hence was the mean of all average precision values across all classes

*Results :*





1. conclusion

In this research, we proposed a two-phase fashion apparel detection model based on

the YOLOv3 algorithm. The experimental results showed that the mAP of YOLOv3-TPD

was 3.03% higher than the original YOLOv4 model. The values of other validated indicators

such as recall, precision, and IoU also increased. Compared to other existing research of

clothing detection, our model had the advantage of high detection accuracy in fashion

clothing detection with complex backgrounds.

In today’s society, the types of fashion apparel are gradually increasing. It is not

enough that the apparel detection model could only detect five categories. Therefore, in

future work, we aim to increase the category of the apparel and consider detecting the

inner clothing, such as a shirt under a jacket.

1. References

1.

Seo, Y.; Shin, K. Hierarchical Convolutional Neural Networks for Fashion Image Classification. *Expert Syst. Appl.* **2019**, *116*, 328–339.

[CrossRef]

2.

Benjdira, B.; Khursheed, T.; Koubaa, A.; Ammar, A.; Ouni, K. Car Detection Using Unmanned Aerial Vehicles: Comparison

between Faster R-CNN and YOLOv3. In Proceedings of the 2019 1st International Conference on Unmanned Vehicle Systems

Oman, UVS, Muscat, Oman, 5–7 February 2019; pp. 1–6.

3.

Lyu, H.; Fu, H.; Hu, X.; Liu, L. Esnet: Edge-Based Segmentation Network for Real-Time Semantic Segmentation in Traffific Scenes.

In Proceedings of the 2019 IEEE International Conference on Image Processing, ICIP, Taipei, Taiwan, 22–25 September 2019;

pp. 1855–1859.

4.

Yi, J.; Wu, P.; Hoeppner, D.J.; Metaxas, D. Pixel-Wise Neural Cell Instance Segmentation. In Proceedings of the 2018 IEEE 15th

International Symposium on Biomedical Imaging, Washington, DC, USA, 4–7 April 2018; pp. 373–377.

5.

Wang, X.; Zhang, T. Clothes Search in Consumer Photos via Color Matching and Attribute Learning. In Proceedings of the

MM’11—2011 ACM Multimedia Conference and Co-Located Workshops, Scottsdale, AZ, USA, 28 November–1 December 2011;

ACM Press: New York, NY, USA, 2011; pp. 1353–1356.

6.

Liu, S.; Song, Z.; Liu, G.; Xu, C.; Lu, H.; Yan, S. Street-to-Shop: Cross-Scenario Clothing Retrieval via Parts Alignment and

Auxiliary Set. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,

Providence, RI, USA, 16–21 June 2012; pp. 3330–3337.

7.

Song, Z.; Wang, M.; Hua, X.S.; Yan, S. Predicting Occupation via Human Clothing and Contexts. In Proceedings of the 2011 IEEE

International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1084–1091.

8.

Yamamoto, T.; Nakazawa, A. Fashion Style Recognition Using Component-Dependent Convolutional Neural Networks. In

Proceedings of the 2019 International Conference on Image Processing, ICIP, Taipei, Taiwan, 22–25 September 2019; pp. 3397–3401.

9.

Shin, Y.G.; Yeo, Y.J.; Sagong, M.C.; Ji, S.W.; Ko, S.J. Deep Fashion Recommendation System with Style Feature Decomposition. In

Proceedings of the 2019 IEEE 9th International Conference on Consumer Electronics—Berlin, ICCE-Berlin, Berlin, Germany, 8–11

September 2019; pp. 301–305.