A Novel Method for Handling Partial Occlusion on Person Re-identification using Partial Siamese Network

Muhammad Pajar Kharisma Putra[1], Wahyono[2]*
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta Indonesia[1, 2]
Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Lampung, Indonesia[1]

***Abstract*—Person-reidentification (Re-ID) is one of the tasks in CCTV-based surveillance system for verifying whether two detected objects are the same person or not. Re-ID visually matching one person or group in various situations obtained from different cameras or on the same camera but at different times. This method replaces the task of surveillance through surveillance cameras that was previously carried out conventionally by humans because it is prone to errors. The challenge of Re-ID is the pose of varied objects, occlusions, and the appearance of people who tend to be similar. Occlusion issues receive special attention since the performance of Re-ID can decrease due to partial occlusion. This can occur because the re-identification process relies on features of the person such as the color and pattern of clothing. The occlusion resulted in the feature not being caught by the camera resulting in a re-identification error. This paper proposed to overcome this problem by dividing the image into several parts (partial) and then processed in different neural network (NN) but with the same architecture. The research conducted is applying the CNN algorithm with the Siamese network architecture and applying the contrastive loss function to calculate the similarity distance between a pair of images. The test results show that the partial process obtained an accuracy of 86%, 77%, 68%, and 56% for occlusion data of 20%, 40%, 60%, and 80%. This accuracy is three to five percent higher than images without partial.**

Keywords—CCTV; CNN; video-surveillance; NN; contrastive-loss

I. Introduction

Today’s surveillance cameras are common in public places such as shopping centers, airports, schools or offices. This technology has been used widely in applications related to vision such as video surveillance. With a surveillance camera, every event captured by the camera can be monitored or stored by the operator; therefore it can be analyzed if needed.

One application of video surveillance is for criminal problems. Often a criminal investigator needs to find out where certain people appear based on pictures taken by the camera [1]. However this is not without problems. Relying on human intervention performance for surveillance of more than one cam- era requires expensive and inaccurate costs. The operator assigned to monitor more than one camera and carry out a manual matching process, is prone to errors. In addition, human performance is determined subjectively based on the experience of each operator therefore it can lead to differences in performance between operators [2]. To handle this problem, monitoring activities are carried out by computers using the person re-identification (Re-ID) method.

Re-ID is the process of matching certain people through different cameras [3]. Unlike identification which aims to get the identity of the object identified, the purpose of Re-ID is to match the same person on different cameras or at different times but on the same camera [4]. The first stage of Re-ID is to detect people on the camera, which is then followed by identification. The identification stage is the most important stage of Re-ID, since at this stage it is concluded whether the person is the same person or not.

The challenge of Re-ID is the pose of varied objects, occlusions, and the appearance of people who tend to be similar [3]. Occlusion issues receive special attention since the performance of Re-ID can decrease due to partial occlusion. This problem is difficult to overcome since some parts of the body of people are covered by other people or objects in the environment [5]. On the other hand, performing a process of matching between detected images with a large number of images requires a long computational time [6].

Therefore, in this paper, the problem of partial occlusion at the re-identification stage can be dealt by dividing the image partially so that people can still be recognized even though some attributes are blocked from the camera. The image is divided into three parts, namely the head, body, and legs and then each part will be trained on a different neural network. This partial process can improve the accuracy of the data by occlusion [7]. Then the reidentification process uses the Siamese Network algorithm. The Siamese Network algorithm as a feature extraction is better in terms of accuracy compared to other methods, since this model has identical network layers and fewer datasets. The last layer of the Siamese Network as a layer to find the value of similarity between two objects with a distance function [8]. Siamese Network uses contrastive loss as a function of distance that is able to distinguish between objects in pairs [6].

Overall, this paper provides following major contributions: (1) Proposed a new strategy to apply partial regions division in Siamese Network based on human body proportion. (2) Utilize constructive loss for matching two difference objects. (3) Provide more detail investigation regarding to the effect of partial occlusion on person re-identification problem.

*Corresponding Author

II. Related Work

Some research on Re-ID has been done in recent years with various methods to overcome various problem. Ku *et al.* in 2018 doing research on Re-ID based on features extracted with CNN and manually selected [4]. In 2017, Gu *et al.* combined DPM and SVM for person reidentification [1]. J Liu *et al.* in 2017 proposed method a novel multi-part compact bilinear convolutional neural network (CNN) model, which consists of a bilinear CNN and two part-networks aiming to learn the global features and the finer local features simultaneously [9]. However, these method weak against occlusion since occlusion can block images from camera so that features from images can't be extracted. Another research about Re-ID for overcome occlusion problem was carried out by Song *et al* in 2017 by dividing the image into three equal parts [7] and Liu *et al* in 2017 by dividing the image into four parts with overlapping [10] then processed in a separate neural network.

Dividing the image into several parts of the same size allows an occlusion that blocks one part will have a significant impact on other parts because no part is prioritized. Based on previous research, our proposed method will modify images partial process by dividing images into three parts based on human body proportion consisting of head, body, and leg. This research also proposes the process of determining which parts are considered priority or have more features based on accuracy during the training process. Thus, occlusion that occurs in parts that are not a priority will not have a significant impact on the results of re-identification.

III. The Method

The system aims to re-identify human objects on the surveillance camera. This system implements the siamese network algorithm which is partially processed in each sub-region to overcome the occlusion problem in Re-ID. The research flow consists of data collection, pre-processing, feature extraction with CNN, similarity measurement with contrastive loss, and testing. While the design of the built algorithm is divided into two algorithms (i.e training and testing). The details of these stages will be explained in the following subsections.

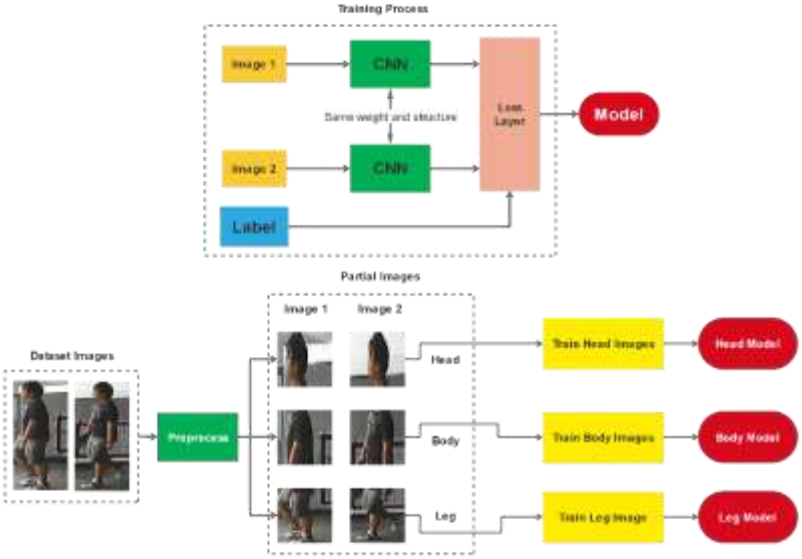


Fig. 1. Flowchart of the Proposed Method for Training Stage.

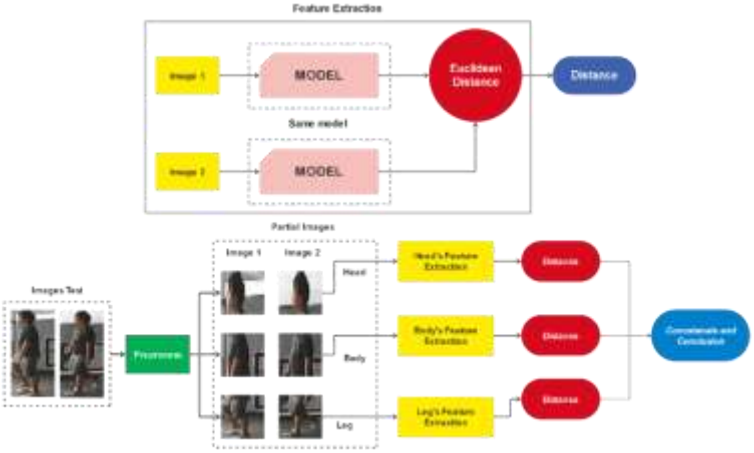


Fig. 2. Flowchart of the Proposed Method for Testing Stage.

Fig. 1 and 2 show the design of the training and testing algorithm. In the training algorithm, the two images are extracted using an identical CNN architecture [11]. The features obtained will be calculated using a loss function. This process will be carried out continuously until certain conditions or iterations. Each subregion is processed separately. The model produced at the training stage will be used at the testing stage. In the testing phase, the test image pair feature will be extracted using the trained model. Then the feature distance of this image pair will be calculated using the Euclidean Distance.

A. Data Collection

The dataset used in this research was obtained from CUHK03 dataset [12]. In the CUHK03 dataset there are 1360 identities and 13164 images, consisting of manual crop images and automatic detection. Fig. 3 shows the example of CUHK03 dataset.

B. Preprocessing

The input image will be divided into several sub-regions then resize so that both images have the same resolution as shown in Fig. 4. Later each of these sub-regions will be processed separately on different neural networks. The purpose of this stage is to avoid occlusion. The image will be divided into eight sub-regions vertically with the same size. The division of these eight sub-regions is based on the proportion of human body height that is equal to the height of the human head itself [13].

C. Feature Extraction using CNN

Feature extraction is the process of extracting features such as colors and textures from person images. CNN is the algorithm used in the process of extracting this feature. The CNN architecture design used is shown in Fig. 5.

Fig. 5 is an initial design that will be used in this study adopted from Liu dan Huang (2017) [10]. Overall the architecture consists of two inception modules, one convolution layer, and one fully connected layer. The first Inception module consists of 16 convolution kernels of 5×5 and 16 kernels of 3×3. The resulting map feature will be combined with input followed by 2×2 max pooling. The second inception module similar as the first inception module, except the convolutional kernel sizes are 3×3 dan 1×1. Different kernel sizes in the two inception modules aim to extract features from different resolutions. The last convolution layer consists of 16 kernels of 1×1 aims to reduce the depth of feature maps.

D. Similarity Measurement using Contrastive Loss

Similarity measurement is a function that provides the real value of the calculation results of the similarity between two objects. Similarity measurement is done to calculate the distance of similarity between a pair of person's images. The contrastive loss function used in this research will calculate the distance between images. The architecture of constrastive loss is shown in Fig. 6.

Contrastive loss can measure the distance between two images [6]. This function will encourage similar images to have a close distance and different images have at least an ecludian distance of a predetermined margin [14]. Contrastive loss is defined as follows:

$$l = \frac{1}{2} l D^2 + \frac{1}{2}(1-\mathrm{l})\{max(0, m-D)\}^2 \quad (1)$$

where l is image label, m is margin, and D is Euclidean Distance of pair images. The output of this process is float value. The smaller number produced shows the closer similarity distance between two images and vice versa.

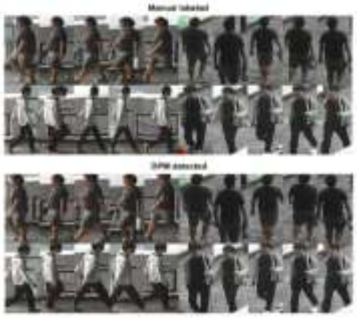


Fig. 3. Selected Samples of CUHK03 Dataset [12].

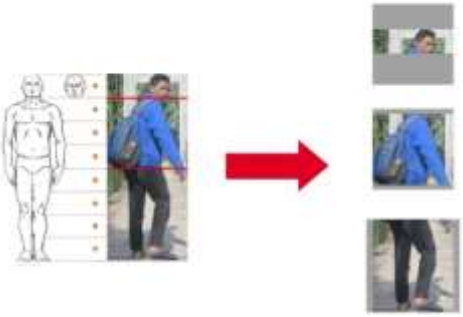


Fig. 4. Preprocessing Illustration.

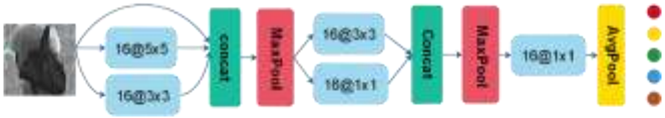


Fig. 5. The Proposed CNN Architecture.

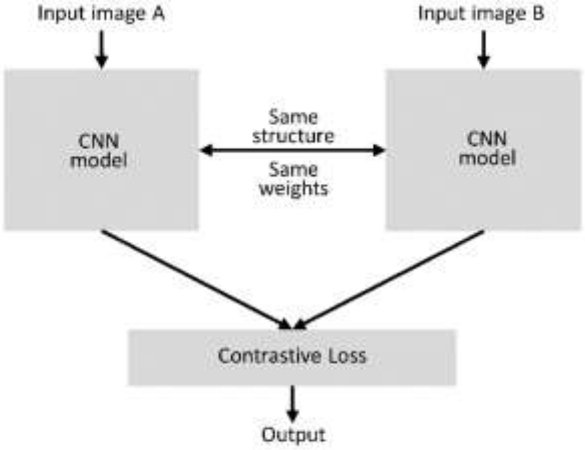


Fig. 6. Contrastive Loss Architecture.

In addition to contrastive loss, there is one other loss function that can be used, namely triplet loss. Training with triplet loss is done by comparing the baseline vector (anchor image) with a positive vector (truth image) and a negative vector (false image), so that three images are needed as input [15]. By applying a margin between pairs of faces with the same identity and faces with different identities, triplet loss keeps faces of the same identity closer than faces of different identities in the embedding space as in Fig. 7.

Triplet loss is defined as follows:

$$Loss = \max(d(A,P) - d(A,N) + margin, 0) \quad (2)$$

Where *d(A,P)* is distance between anchor and positive images, *d(A,N)* is distance between anchor and negative images.

E. Testing

The testing phase is the stage of evaluating the performance of the proposed algorithm. This stage will compare the performance of the proposed method with a model without a partial process. The measured performance is the ability of the model in reidentifying data with occlusion. This data is an augmentation image with occlusion of 20%, 40%, 60%, and 80%. Occlusions are randomly placed at the top or bottom of the image. An example of an occlusion image is shown in Fig. 8.

Test image using CUHK02 and RAiD dataset. CUHK02 was collected at the Chinese University of Hong Kong using two cameras. Overall, this dataset has 7,264 images consisting of 1816 identities [16]. RAiD dataset was collected at the Winstun Chung Hall of UC Riverside. It is a four camera dataset with two indoor and two outdoor cameras. The cameras are numbered as 1, 2, 3 and 4 where cameras 1 and 2 are indoor while cameras 3 and 4 are outdoor. 43 people walked in these camera views resulting in 6920 images. Among the 43 persons 41 people appeared in all the 4 cameras whereas person 8 is not present in camera 3 and person 34 is not present in camera 4 [17]. However only 250 images were used at the test stage.

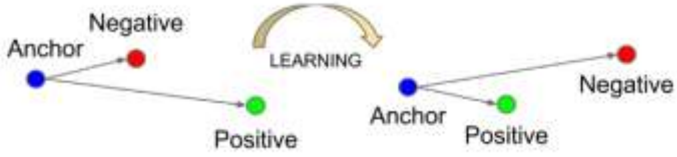


Fig. 7. Triplet Loss Illustration.

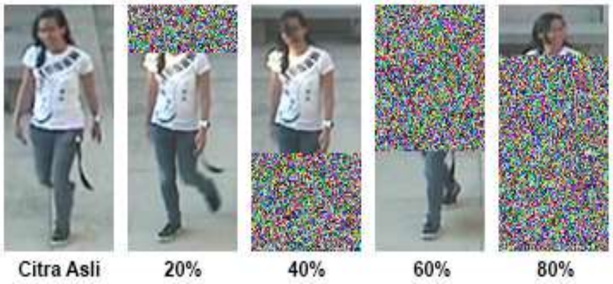


Fig. 8. Example Data with Occlusion.

F. Concatenate

Concatenate is a process that is done in the testing phase. Concatenate will combine each distance from each sub-region (i.e., head, body, and leg). Each sub-region has a different probability based on the accuracy value obtained after the model has been trained; the higher accuracy of the sub-region, the greater probability of the sub-region. Each distance and probability value from each sub-region will be multiplied then added to obtain the final score. The next step is to determine the threshold obtained during the training process. Two images that have a final score below the threshold will be summed up as the same person. The process of determining probabilities and final scores is shown in the following equation:

$$p_i = \frac{acc_i}{\sum_{i=1}^{n} acc_i} \quad (3)$$

where p_i is a probability of each sub-region dan acc_i is an accuracy of each sub-region. While the process of determining the final score is shown in the following equation:

$$D = \sum_{i=1}^{n}(d_i * p_i) \quad (4)$$

where D is final distance between pair images, d_i is distances of each sub-region, dan p_i is probability of each sub-region. Finally, the process of determining the image based on the threshold is shown in the following equation:

$$f(D) = \begin{cases} similar;\ D < threshold \\ disimilar;\ D \geq threshold \end{cases} \quad (5)$$

IV. Experiment and Results

At the training phase the model has been determined several hyperparameters that are considered optimal, namely batch size 32 and Adam optimizer. Large batch size (more than 512) decreases the quality of the model [18]. This is measured by the ability of the model to generalize data. While Adam's optimizer provides better performance in reducing error values compared to other optimizers such as SGD, AdaGrad, and RMS [19].

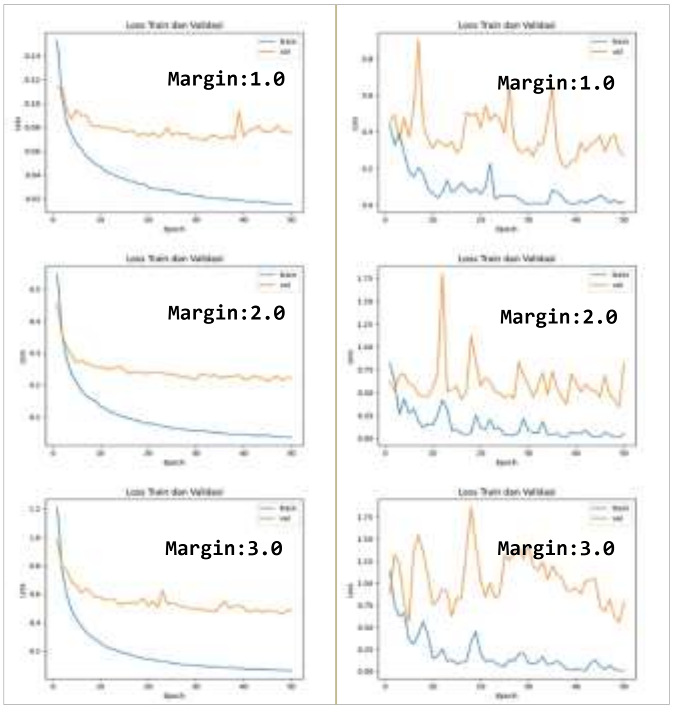
A. Loss Function Testing

As mentioned before, there are two common loss functions in Siamese network, namely, contrastive loss and triplet loss. This test aims to compare the performance of two loss functions and analyze the effect of margin parameters on training results. Margin defines the radius around the embedding space of the sample so that different sample pairs only contribute to the loss function if the distance between the sample pairs is within the margin [14]. The results are shown in Table I.

Table I shows that there is a significant difference between the two loss functions. Contrastive loss obtains a smaller loss value and higher accuracy in the train and validation data for every margin value except the margin value 3 for triplet loss, but requires a longer training time than triplet loss. This is because in contrastive, each data consists of two images and a label, while in triplet each data consists of three images, namely anchor, positive, and negative. Every two data on contrastive is equal to one data on triplet so contrastive loss requires a longer training time than triplet loss.

TABLE I. RESULTS LOSS FUNCTION TESTING

Loss Function	Margin	Loss Value		Acc	Time (minutes)
		Train	Validation		
Contrastive Loss	1.0	0.016	0.075	92.3	30:40
	2.0	0.036	0.222	94.5	29:12
	3.0	0.062	0.487	93.9	29:18
Triplet Loss	1.0	0.019	0.271	88.6	17:59
	2.0	0.045	0.821	89.2	17:59
	3.0	0.071	0.787	89.6	17:58



(a) Contrastive Loss

(b) Triplet loss.

Fig. 9. Training Results Graph.

Fig. 9 shows that there is overfitting in both loss functions, but the contrastive loss has decreased consistently in each epoch for the data train. Meanwhile, in the validation data, the value of margin 2 shows a decreasing graph that is more consistent than margins 1 and 3. In the triplet loss test, in addition to obtaining a loss value that is greater than the contrastive loss, the loss value does not consistently decrease. As shown in equations (1) and (2), if the resulting loss value is negative, then the value will be normalized to zero. This results in the loss of some of the information in the data. For example, data with a loss value of -2 has a more significant similarity between anchor and positive images or a significant difference between anchor and negative images than data with a loss value of -1. However, with the normalization of the negative value to zero, it causes the two data to be considered the same. This has resulted in overfitting in the triplet loss test. Even though contrastive loss also applies the same normalization, as previously mentioned, if two data in a contrastive are equal to one data in a triplet, so that when normalization occurs, the contrastive only loses information on one data, namely the distance between two images is the same, while in a triplet it will lose information. Information on the distance between the two data, namely, the distance of the anchor image - positive and the image anchor - negative.

Furthermore, in the margin parameter, the margin value of 2 gives the best results. Reidentification process can work well if applied to objects with the same class. As previously explained, the purpose of the margin is to keep the image pair actually consisting of the same object (person and person object) so that the image pair that has a distance greater than the margin will not contribute to the loss function or will be ignored. A margin value that is too small to make a small difference between the two objects will be ignored and will not contribute to the loss function, while a margin value that is too large will cause a pair of objects that are far enough apart to have an impact on the loss function because they are still within the margin radius even though the object is a objects of different classes (person and non-person objects). The impact on this loss function will affect the backpropagation process and parameter weight updates so that an incorrect margin value will reduce accuracy even though the loss value obtained is smaller. It can be concluded that contrastive loss with a margin value of 2 gives the best results.

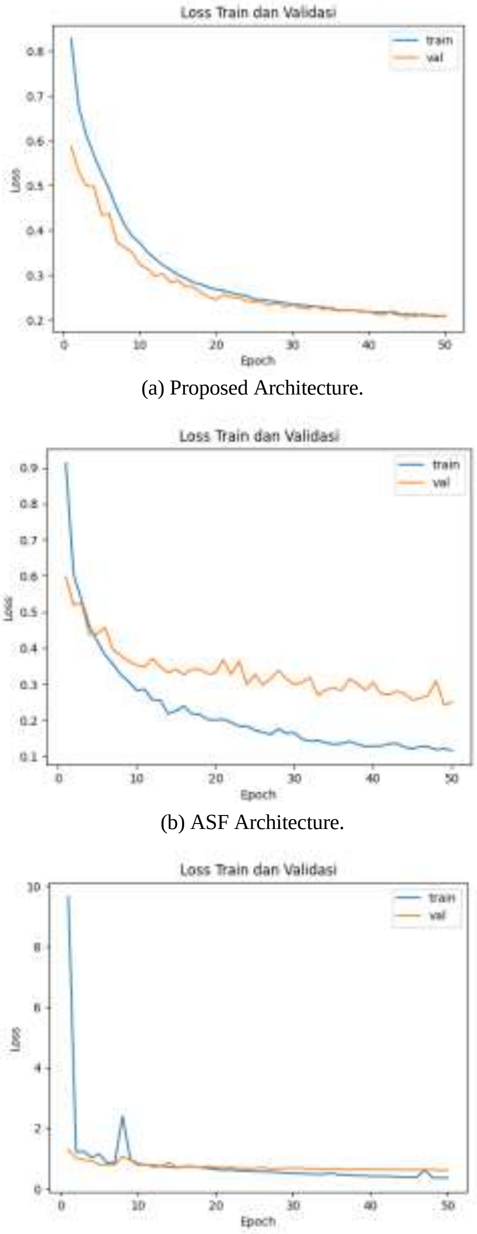
B. Model Training

Each sub-region that has been divided will be trained to obtain a model that will be used at the testing stage. The training process uses 50 epochs. This stage also compares the performance of the architecture used with the Adaptive Spatial Feature (ASF) architecture [7] and Multi-part Compact Bilinear CNN (MCBCNN) Architecture [9] using the same dataset and parameters. Loss value and training time obtained as shown in Table II.

The test results in Table II show that the ASF architecture requires the longest training time while the MCBCNN architecture requires the shortest training time. The length of time training is influenced by the number of parameters being trained. The ASF architecture has 289,470,848 parameters, while the proposed architecture has 89,160,288 parameters, and the MCBCNN architecture has 31,931,664 parameters. The number of parameters that must be trained makes the ASF architecture have the longest training time followed by the proposed architecture and the MCBCNN architecture. Meanwhile the training chart for each epoch is shown in Fig. 10.

TABLE II. CNN ARCHITECTURE TEST RESULTS

Architecture	Loss		Training Time
	Train	Validation	
Proposed	0.2076	0.2095	07:07:58
ASF	0.1167	0.2503	07:23:43
MCBCNN	0.3550	0.6209	04:27:07



(a) Proposed Architecture.

(b) ASF Architecture.

(c) MCBCNN Architecture.

Fig. 10. Loss Train and Validation.

Fig. 10 shows a gradual decrease in the loss value as the epoch increases both in the data train and validation for each architecture. However, in Fig. 10(a), it can be seen that there is no big difference between the train data and validation. Thus, the inception module applied to the convolution layer with the aim of extracting features from different resolutions proved to be able to overcome overfitting.

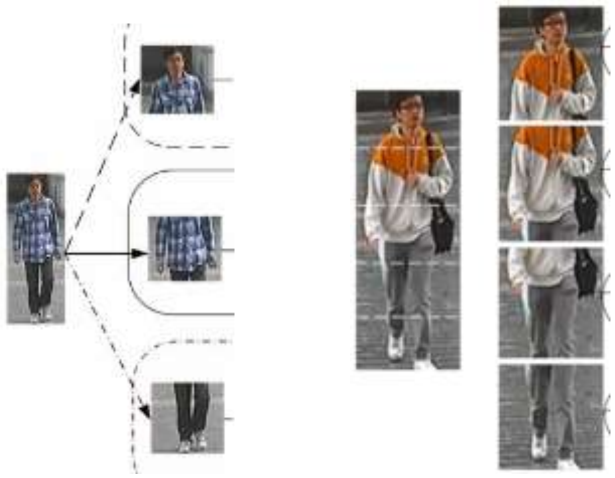
Fig. 10(a) shows the ASF architecture obtains the lowest loss value and the highest accuracy for the data train. In addition, there is also a decrease in the loss value as the epoch increases, but there is a large difference between the train and validation data so that it can be concluded that the ASF architecture is overfitting. One of the causes of overfitting is because each convolutional layer only uses one type of kernel with a small number of feature maps so that the features that can be extracted are limited. If you look at other CNN architectures that are sequential like VGG16 [20], there are 64 feature maps in the first and second convolution layers, as well as 512 feature maps in the last three covolution layers so you can still extract many features even though you only use one kernel type at each convolution layer.

C. Testing on Occlusion Data

Testing on occlusion data uses a model that has been previously trained and implemented on occlusion data as shown in Fig. 11. This test will also compare the proposed partial process with the partial process in Adaptive Spatial Feature (ASF) [7] and Body Structure Based Triplet CNN (BSTCNN) [10] using same dataset and parameter. ASF and BTCNN partial process show in Fig. 11.

Using a trained model, each sub-region obtains the accuracy, as shown in Table III.

The results of the accuracy of each sub-region in Table III show the body parts represented by sub-region two in the proposed partial process and ASF obtained higher accuracy than other parts since this section has more features such as motifs and clothing colors. The foot sub-region has the lowest accuracy of the three partial processes, since the colors and motifs that tend to be similar between one identity and another make not many features that can be extracted in this sub-region. Based on the accuracy in Table III, using equation (3) the probability of each sub-region is obtained as shown in Table IV.



(a) (b)

Fig. 11. (a) ASF; (b) BSTCNN Partial Process.

TABLE III. ACCURACY OF EACH SUB-REGION

Sub-Region	Accuracy (%)		
	Proposed	ASF	BSTCNN
1	82	85	88
2	92	89	82
3	76	87	85
4	-	-	71

TABLE IV. PROBABILITY OF EACH SUB-REGION

Sub-Region	**Probability**		
	Proposed	**ASF**	**BSTCNN**
1	0.328	0.351	0.270
2	0.368	0.367	0.251
3	0.304	0.281	0.261
4	-	-	0.220

The probability value in Table IV is a reference in determining the distance between a pair of images. The final distance of a pair of images is calculated based on the distance and probability value of each sub-region using equation (4). The accuracy of each partial process on occlusion data is shown in Table V.

Table V shows that the proposed partial process has higher accuracy than other partial processes or images without partials on all occlusion data. Image without partials still obtained higher accuracy than partial ASF and BSTCNN processes at 20% occlusion, but the subsequent occlusion obtained the lowest accuracy compared to others. This is because the greater of occlusion in the image, the fewer features that can be extracted because some features are lost due to the occlusion. The occlusion of one part of the image will greatly affect the calculation of the distance in the image without partial because all images are extracted using the same model. Unlike the image without partials, in the partial process each image will be divided into several sub-regions and processed separately using each model, so that the occlusion of one sub-region with a small probability value will not have a big impact on the calculation of the distance of a pair of images.

TABLE V. ACCURACY OF OCCLUSION DATA

Occlusion (%)	**Partial Process**	**Accuracy (%)**
20	Without Partial	82
	Proposed	86
	ASF	81
	BSTCNN	81
40	Without Partial	73
	Proposed	77
	ASF	74
	BSTCNN	73
60	Without Partial	65
	Proposed	68
	ASF	68
	BSTCNN	66
80	Without Partial	51
	Proposed	56
	ASF	55
	BSTCNN	53

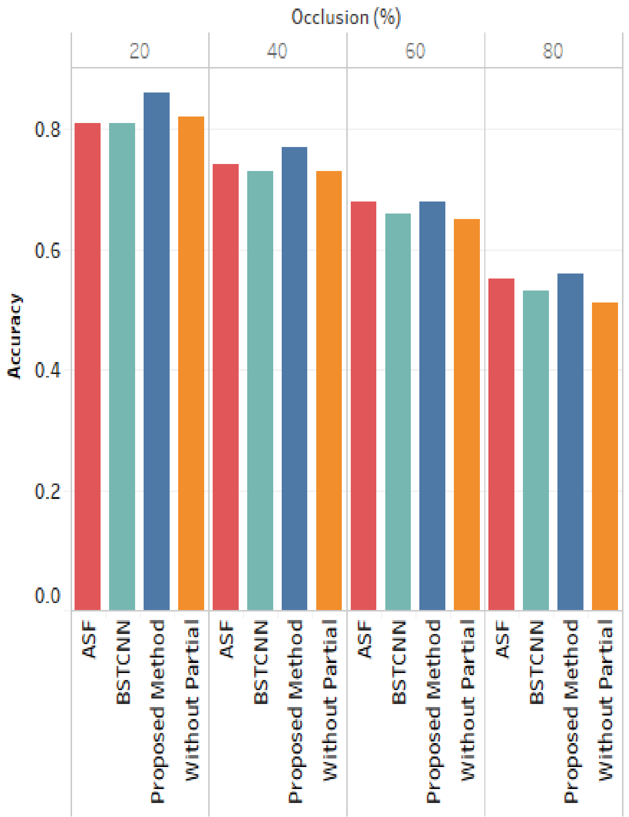


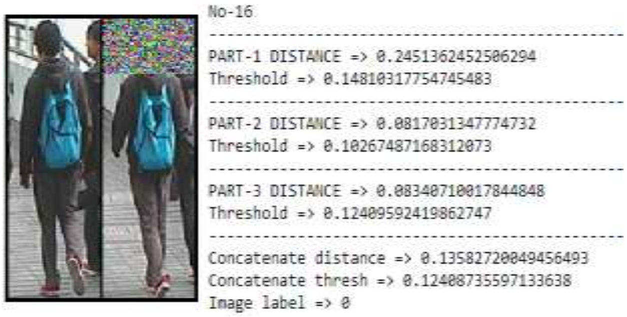
Fig. 12. Accuracy of Occlusion Data.

As shown in Fig. 12, the proposed partial process gets higher accuracy since dividing the image according to the proportions of the human body. The sub-region of the body that has the highest accuracy has a higher probability value as well because this section contains important features such as color and pattern of clothing. Two other partial processes divide the image by equal size. The small difference in probability values between sub-regions makes this partial process not much different from images without partial processing. Overall, partial processing has been shown to improve accuracy in occlusion data compared to images without partial processing.

D. Analysis and Discussion

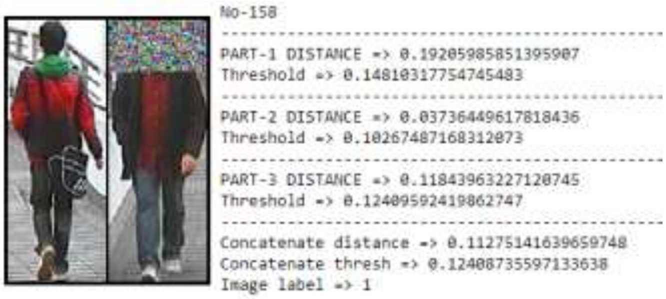
The results in Table V show that the proposed method obtained better accuracy than other methods and succeeded in increasing the accuracy compared to the model without partial processing. However, there are still some weaknesses in proposed method.

Fig. 13 shows the error that occurs in the re-identification of the same pair. Part-2 and part-3 which are parts of the body and legs get a distance below the threshold value, while part-1 which is part of the head gets a distance above the threshold since there is occlusion in that part. In some cases, the proposed method cannot re-identify if there is occlusion in all parts of one of the sub-regions or sections. This occlusion makes the distance from the sub-region far above the threshold so that it affects the concatenate process.



```
No-16
------------------------------------------------
PART-1 DISTANCE => 0.2451362452506294
Threshold => 0.14810317754745483
------------------------------------------------
PART-2 DISTANCE => 0.0817031347774732
Threshold => 0.10267487168312073
------------------------------------------------
PART-3 DISTANCE => 0.08340710017844848
Threshold => 0.12409592419862747
------------------------------------------------
Concatenate distance => 0.13582720049456493
Concatenate thresh => 0.12408735597133638
Image label => 0
```

Fig. 13. Example Error Due to Occlusion.

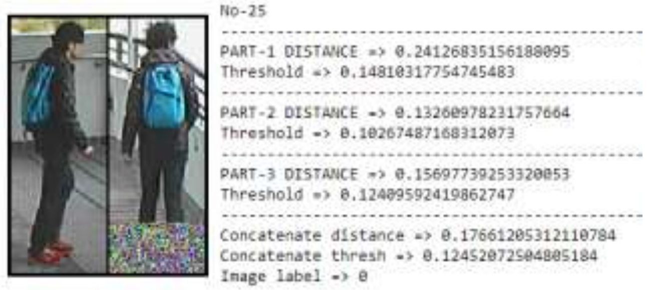


```
No-158
------------------------------------------------
PART-1 DISTANCE => 0.19205985851395907
Threshold => 0.14810317754745483
------------------------------------------------
PART-2 DISTANCE => 0.03736449617818436
Threshold => 0.10267487168312073
------------------------------------------------
PART-3 DISTANCE => 0.11843963227120745
Threshold => 0.12409592419862747
------------------------------------------------
Concatenate distance => 0.11275141639659748
Concatenate thresh => 0.12408735597133638
Image label => 1
```

Fig. 14. Example Error due to Similar Feature.

Similar features between two images with different identities can also cause errors. Fig. 14 shows an error due to a similar feature, namely color. Part-2 and part-3 get a smaller distance than the threshold, which means that the two images are the same identity even though the image is a different identity. This error occurs since both objects are wearing red and black clothes. In addition, the legs-part of the two objects use pants which tend to be similar and the background in this part is gray.

Another factor that can cause re-identification errors is the difference in camera angles in recording images. Fig. 15 shows a pair of images of people with the same identity but taken by different cameras. Part-1, part-2, and part-3 get a distance greater than the threshold. The error in calculating this distance is caused by different feature conditions due to different camera angles. For example in part-2, there is a difference in the area of the blue bag object between the left and right images. Another example in part-1, a significant difference in the background of the object can affect the distance calculation because this part is also processed during feature extraction.



```
No-25
------------------------------------------------
PART-1 DISTANCE => 0.24126835156188095
Threshold => 0.14810317754745483
------------------------------------------------
PART-2 DISTANCE => 0.13260978231757664
Threshold => 0.10267487168312073
------------------------------------------------
PART-3 DISTANCE => 0.15697739253320053
Threshold => 0.12409592419862747
------------------------------------------------
Concatenate distance => 0.17661205312110784
Concatenate thresh => 0.12452072504805184
Image label => 0
```

Fig. 15. Example Error due to taken by Different Cameras.

V. Conclusion

In this paper introduced the partial Siamese network method using CNN as a feature extraction and contrastive loss as a function of distance in overcoming occlusion in person re-identification. The CNN architecture used consists of two inception modules, one convolutional layer, and a fully connected layer. This architecture is proven to be able to get better performance than the other two architectures using the same dataset and parameters. In the testing phase, the distance of each sub-region will be multiplied by the probability value then added to obtain the final distance. An image pair that has a distance less than the threshold will be considered as a person with the same identity and vice versa. Overall, the proposed method is proven to be able to improve the accuracy of occlusion data compared to images without partiality, and two partial processes. In the future, several image enhancement methods [21] will be applied before the dataset is processed in a neural network so that it can overcome reidentification errors due to similar features.

Acknowledgments

This research was partially supported by 2020 Research Grant on Graduate Program of Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia.

References

[1] X. Gu, X. Zou, J. Liu, and L. Zhang, “Person re-identification by using a method combining DPM and SVM,” in 2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2017, 2017, vol. 2018-Febru, pp. 124–127, doi: 10.1109/ICCWAMTIP.2017.8301463.

[2] R. Layne, T. Hospedales, and S. Gong, “Person re-identification by attributes,” in BMVC 2012 - Electronic Proceedings of the British Machine Vision Conference 2012, 2012, no. June 2014, doi: 10.5244/C.26.24.

[3] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, “Adversarially Occluded Samples for Person Re-identification,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 5098–5107, 2018, doi: 10.1109/CVPR.2018.00535.

[4] H. Ku, P. Zhou, X. Cai, H. Yang, and Y. Chen, “Person re-identification method based on CNN and manually-selected feature fusion,” ICNC-FSKD 2017 - 13th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., pp. 93–96, 2018, doi: 10.1109/FSKD.2017.8393401.

[5] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “VRSTC : Occlusion-Free Video Person Re-Identification,” in CVPR, 2019, pp. 7183–7192.

[6] I. Melekhov, J. Kannala, and E. Rahtu, “Siamese network features for image matching,” in Proceedings - International Conference on Pattern Recognition, 2016, vol. 0, pp. 378–383, doi: 10.1109/ICPR.2016.7899663.

[7] Z. Song, X. Cai, Y. Chen, Y. Zeng, L. Lv, and H. Shu, “Deep convolutional neural networks with adaptive spatial feature for person re-identification,” in 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Mar. 2017, pp. 2020–2023, doi: 10.1109/IAEAC.2017.8054370.

[8] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese Neural Networks for One-shot Image Recognition Gregory,” in International Conference on Machine Learning, 2015, vol. 37, no. 5108, p. 1355, doi: 10.1136/bmj.2.5108.1355-c.

[9] J. Liu, Z. Yang, T. Zhang, and H. Xiong, “Multi-part compact bilinear CNN for person re-identification,” in 2017 IEEE International Conference on Image Processing (ICIP), Sep. 2017, vol. 1, pp. 2309–2313, doi: 10.1109/ICIP.2017.8296694.

[10] H. Liu and W. Huang, “Body structure based triplet Convolutional Neural Network for person re-identification,” in 2017 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017, pp. 1772–1776, doi: 10.1109/ICASSP.2017.7952461.

[11] M.P.K. Putra, “Person Re-identifikasi Menggunakan Partial Siamese Network untuk Mengatasi Masalah Oklusi Parsial Pada Object (Person Re-Identification Using Partial Siamese Network for Handling Partial Occlusion on Object)”, Master Thesis, Universitas Gadjah Mada, Indonesia, 2020 (In Bahasa Indonesia).

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159, doi: 10.1109/CVPR.2014.27.

[13] C. Hart, Figure It Out! Human Proportions. New York: Sixth & Spring Books, 2014.

[14] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, pp. 1735–1742, 2006, doi: 10.1109/CVPR.2006.100.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07-12-June, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[16] W. Li and X. Wang, “Locally aligned feature transforms across views,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3594–3601, 2013, doi: 10.1109/CVPR.2013.461.

[17] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, “Consistent re-identification in a camera network,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8690 LNCS, no. PART 2, pp. 330–345, 2014, doi: 10.1007/978-3-319-10605-2_22.

[18] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, “On large-batch training for deep learning: Generalization gap and sharp minima,” 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc., pp. 1–16, 2019.

[19] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14, 2015.

[21] S.S. Khan, M. Khan, and Y. Alharbi, “Multi Focus Image Fusion using Image Enhancement Techniques with Wavelet Transformation,” International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No 5, pp. 414-420, 2020. doi:10.14569/IJACSA.2020.0110555.