Multi-Modal Person Re-Identification using Lightweight Convolutional Neural Network

1st Reza Fuad Rachmadi
Department of Computer Engineering
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
fuad@its.ac.id

2nd I Ketut Eddy Purnama
Department of Computer Engineering
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
ketut@ee.its.ac.id

3rd Charles Chang
Department of Computer Engineering
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia 60111
changcenliang@gmail.com

***Abstract*—As a complement to security systems, CCTVs are increasingly used to monitor and analyze criminal acts done at a given location. However, the manual search for criminals is still prone to human error. One of the solutions to make the process more effective and efficient is with the use of re-identification. Re-identification is a computer vision and deep learning technique in which an anonymized identity of an image is matched with its owner. In this paper, we will study the method of re-identifying people with multi-modal images where the query is in the form of a body sketch drawn by several different artists. The highest Rank-1 precision achieved in this paper with the Lightweight Convolutional Neural Network is 21%.**

***Index Terms*—CCTV, Re-Identification, Multi-modal, Criminal**

I. Introduction

OVER the years, technology has improved and revolutionized our daily lives. More and more technology created to alleviate human life, such as CCTV cameras, are increasingly used to monitor public spaces. CCTV cameras have been a long-standing security measure used in public and commercial settings. The recordings from CCTV cameras provide vital visual information and can act as a witness to a crime scene. These recordings have a prominent role to play in providing evidence in criminal investigations and disputes.

According to data taken from the Central Bureau of Statistics, there is a rising crime wave throughout Indonesia. Polda Metro Jaya alone recorded the highest number of violations, namely 31,934 incidents [1]–[4]. These facts encourage research on reducing the crime rate in various ways. Automation that can reduce the costs and workloads of the police force is needed.

Person re-identification is a computer vision and deep learning technique in which an anonymized identity of a person is matched with its owner. Person re-identification can simplify many activities that were previously done manually, by establishing a person re-identification system, the inspection of recordings conducted by the police force can be carried out faster and can reduce the costs of labor.

But a query photo of the target individual is not always readily available. Previous research of multi-modal person re-identification done by Lu Pang et al. [5] defined the problem of sketch re-identification, which uses a sketch instead of an image as the query of the model. While similar to facial sketch recognition, this problem is tackled using full-body sketches which add another dimension of complexity to the model. Furthermore, this problem is a challenging task due to the domain gap between sketch and photo. Sketches lack color information that is used as a differentiator between one individual and another. This study achieved Rank-1 precision of 34% using a state-of-the-art model and cross-domain adversarial learning. However, no further study using Lightweight Convolutional Neural Network has been done.

In this paper, we investigated several lightweight Convolutional Neural Networks as a solution for sketch re-identification. We constructed the model based on the lightweight residual network used to solve the CIFAR dataset. We removed the fully connected layer of the original dataset and added two new fully connected layers, where in the first Fully Connected Layer the model will learn the discriminant features and in the last Fully Connected Layer the model will classify the identity of the person in the training data.

II. Design and Implementation

In this section, we will describe the design and implementation of lightweight Convolutional Neural Network for sketch re-identification. We will describe the experiment setting, the dataset used in the experiments, the training, and the testing processes.

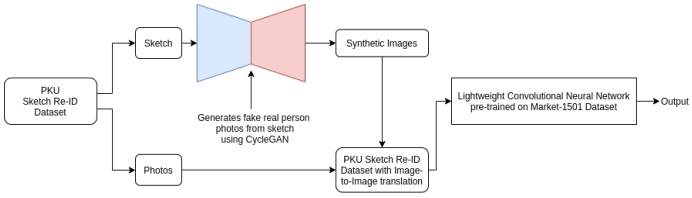

Fig. 1: Block Diagram of Working System

A. Lightweight Convolutional Neural Network

Figure 1 shows the diagram for our proposed lightweight Convolutional Neural Network classifier based on the architecture used to solve the CIFAR dataset. However, we removed the final fully connected layer and added two new ones at the end of the classifier to ensure the model learns good discriminatory features. Furthermore, unlike the original ResNet model, ours use an input resolution of 32x64x3 instead of 32x32x3.

In our experiments, we used two different lightweight residual networks, which are ResNet56 and ResNet110. Although the number of layers on these models is very deep, the number of parameters of the deepest classifier is still 1.7 million parameters.

Name	Parameters
ResNet56	0.85M
ResNet110	1.7M
GoogleNet	7M
DenseNet121	8.6M
ResNet50	23M

TABLE I: Number of Parameters for Popular Models.

B. PKU Sketch Re-ID Dataset

Fig. 2: Examples of PKU Sketch ReID Data

To evaluate the performance of our lightweight Convolutional Neural Network, we opted to use the PKU Sketch Re-ID dataset created by Lu Pang et al.. The dataset consists of 200 unique identities captured using two different cameras and one sketch corresponding to each individual, totaling 600 images. Figure 2 shows some examples of the dataset.

The dataset was then divided into 150 identities for the training set and 50 identities for the testing set, as shown in figure 3. The images are manually cropped to ensure every photo contains one specific individual. As for the sketches, there are a total of five different artists to draw each identities sketches. Each artist has his/her own art style.

Fig. 3: Training-Testing Distribution

C. Cross Domain Image-to-Image Translation

To help compensate for the lack of features in the query images, we decided to use image-to-image translation, specifically CycleGAN. CycleGAN allowed to mutually learn distributions of images, given two domains of the problems. GAN mimics the distribution of the generator given to the model which is photos and translates the style to generate synthetic photos from sketches. Essentially filling the gaps between the sketch and photos. Since the CycleGAN model is designed for unpaired translation, we created a paired testing set to ensure the style transfer is done between each unique individual.

For the CycleGAN, we used the model created by Jun-Yan Zhu et al. with the facades_label2photo pre-trained weights. The images used to train the models are the 600 images from the PKU Sketch Re-ID. The model is trained for 200 epochs with the learning rate initialized at 0.0002 and decaying to 0.0001 after 150 epochs.

D. Training and Testing Process

To ensure the consistency of our model's performance, we perform the training and testing process ten times and take the average as the final evaluation metrics. All methods are evaluated using the Rank-1 accuracy of the model, following the baseline set by Lu Pang et al..

In our experiments, we handled the lack of training data by pre-training on existing person re-id datasets. We used two different datasets which are the DukeMTMC and the Market-1501 dataset. The training process is done for 100 epochs with the learning rate initialized at 0.1. The learning rate is set to decay by a factor of 0.1 every 40 epochs. Furthermore, we used random erasing and random crop to add more training data. Figure 4 shows some examples of data augmentation using those two methods.

E. Ablation Study

To ensure the performance of the ResNet-CIFAR is at it's best, we performed an ablation study by changing the size of it's first fully connected layer. The fully connected

Fig. 4: Examples of Random Erasing

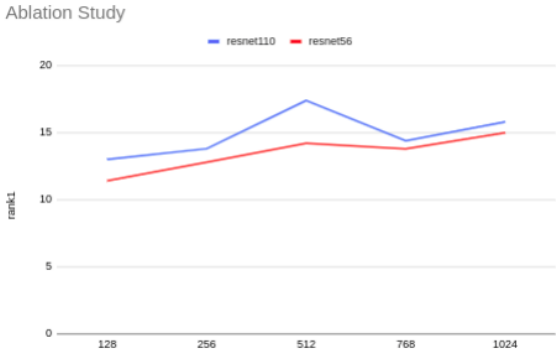

Fig. 5: Result from ablation study using size of FC

layer used are as follows; FC-128, FC-256, FC-512, FC-768, and FC-1024. Furthermore we increased the Random Erasing probability from 0% until 50% in increments of 10% on all models tested. In these and the next experiments, we only use ResNet56 and ResNet110 which in the last experiments achieved first and second highest rank1 precision.

Name	Rank-1	Rank-5	Rank-10	mAP
ResNet56 FC 128	11.4%	36.4%	50%	15.4205%
ResNet56 FC 256	12.8%	31.8%	48%	15.71973%
ResNet56 FC 512	14.2%	35%	47.8%	16.95365%
ResNet56 FC 768	13.8%	35.8%	46.8%	16.50338%
ResNet56 FC 1024	15%	35.6%	51%	17.58063%
ResNet110 FC 128	13%	35.8%	49%	15.81262%
ResNet110 FC 256	13.8%	38.4%	52.8%	16.87966%
ResNet110 FC 512	17.4%	38.8%	52.2%	18.87587%
ResNet110 FC 768	14.4%	36.8%	48.8%	16.49%
ResNet110 FC 1024	15.8%	39%	53.4%	18.0541%

TABLE II: Ablation experiments of ResNet56 and ResNet110 (averaging from ten different runs).

As shown of Figure 5, the best model performance is achieved by using an FC-size of 512, we concluded that using a FC-size of over 512 resulted in overfitting problems, and using layers lower than 512 are less optimal.

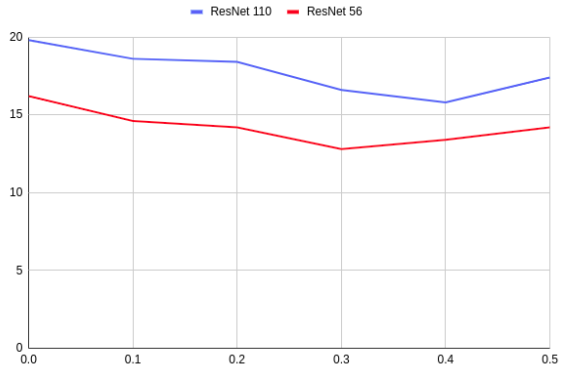

Fig. 6: Result from ablation study using probability of Random Erasing

In the second ablation study, we found out that random erasing decreased the performance of both model significantly. From figure 6, it is shown that the Rank1 accuracy of both models are decreasing when the Random Erasing is increased from 0% until 30%, and the models performance tend to get less consistent after we increased the random erasing probability to 40%. Table III shows random erasing probability experiments done on both models.

Name	Rank-1	Rank-5	Rank-10	mAP
ResNet56 RE 0%	16.2%	36.4%	50.6%	19.14881%
ResNet56 RE 10%	14.6%	38.8%	52.4%	18.6686%
ResNet56 RE 20%	14.2%	37.2%	51.2%	17.82549%
ResNet56 RE 30%	12.8%	32.2%	46.4%	16.52132%
ResNet56 RE 40%	13.4%	36%	49.4%	16.97744%
ResNet56 RE 50%	14.2%	35%	47.8%	16.9536%
ResNet110 RE 0%	19.8%	37.4%	47.8%	21.06424%
ResNet110 RE 10%	18.6%	37%	49.8%	20.12005%
ResNet110 RE 20%	18.4%	39.4%	53.8%	19.93169%
ResNet110 RE 30%	16.6%	38%	49%	18.6043%
ResNet110 RE 40%	15.8%	39.2%	52.6%	19.65465%
ResNet110 RE 50%	17.4%	38.8%	52.2%	18.87587%

TABLE III: Ablation experiments of ResNet56 and ResNet110 (averaging from ten different runs).

III. Results

Based on the results of our experiments we decided to use a droprate of 0.5, a random erasing portion of 0, and only choose to continue with the ResNet56 and ResNet110 Convolutional Neural Network for further experimentation since it yields the best results. However, because a complete ablation study has not been done yet, the hyperparameters could still be tuned to increase the performance.

To increase the performance of the classifier, we conducted an experiment making an ensemble from the tested ResNet56 and ResNet110 classifier. From the experiments conducted, the

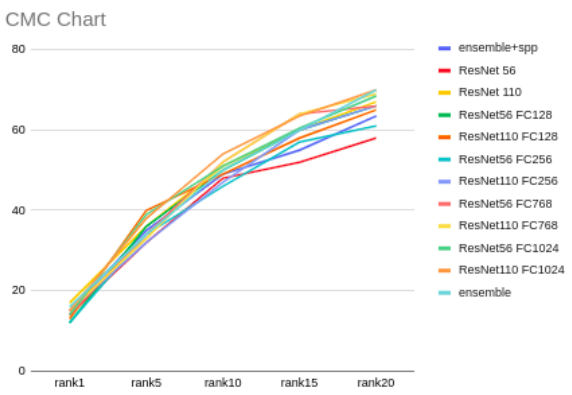

Fig. 7: CMC curve of all the model studied

ensemble created resulted in about the same performance as ResNet110 with 512 Fully Connected layers.

Name	Rank-1	Rank-5	Rank-10	mAP
Ensemble	21 %	42.4%	52.4%	22.05491%
Ensemble SPP	20.2%	42%	53%	20.94988%

TABLE IV: Experiments using Ensemble of ResNet56 and ResNet100 classifier (averaging from ten different runs).

IV. Comparison

Name	Params	Rank-1	Rank-5	Rank-10
Dense-HOG+LBP+SVM	8.6M	5.1%	16.8%	28.3%
Triplet SN	n/a	9%	26.8%	43.2%
GN Siamese	14M	28.9%	54%	62.4%
Cross-Domain Adversarial	n/a	34%	56.3%	72.5%
Ensemble	3M	21 %	42.4%	52.4%
Ensemble SPP	3M	20.2%	42%	53%
ResNet110	1.7M	19.8%	37.4%	47.8%

TABLE V: Comparison to other state of the art models

Table V shows the performance comparison of our model to several state-of-the-art models. Although the ensemble models does not have the best performance in Rank-5 and Rank-10 precision, we decided to evaluate all models using it's Rank-1 precision.

V. Conclusion

In this paper, we introduce the usage of lightweight Convolutional Neural Network to tackle the problem of Sketch re-identification. To address the difference of modality in sketch and real images, we use image-to-image translation or more specifically CycleGAN. In the training process, we managed to have better precision compared to DenseNet, a classical model with more than quadruple the number of parameters that our model has. Other than that, or model prevailed against Triplet SN, a model composed of three identical Sketch-a-Nets and is optimized by triplet ranking loss. Although our model has not achieved state-of-the-art performance, the information gained by the classifier is very high, which proves the classifier is more efficient than other methods.

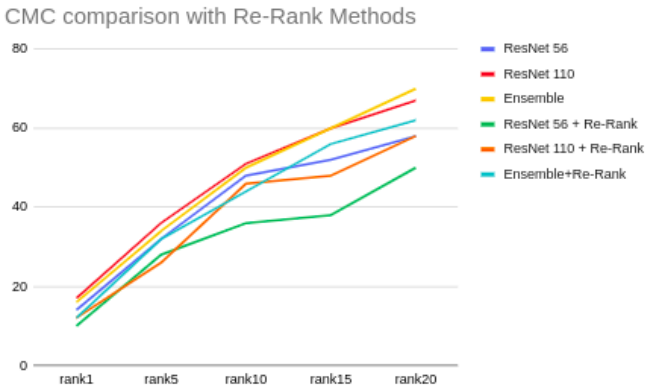

Fig. 8: CMC comparison with re-ranked methods

Other than that, re-ranking, a method of calculating the final distance between images only resulted in reducing the performance of the classifier. From previous research done using re-ranking in person re-identification, the method is supposed to increase the model's rank1 precision and mAP, whilst reducing the rank5 and rank10 precision, figure 7 shows the CMC curves comparing our lightweight models with it's reranking counterparts.

Name	Rank-1	Rank-5	Rank-10	mAP
ResNet56 FC128 + re-rank	11%	25%	37.4%	13.60%
ResNet56 FC256 + re-rank	8.2%	24%	36.2%	12.17%
ResNet56 FC512 + re-rank	8%	25.4%	36%	11.49%
ResNet56 FC768 + re-rank	8.4%	23.4%	36.6%	12.24%
ResNet56 FC1024 + re-rank	10.8%	30%	40.8%	14.49%
ResNet110 FC128 + re-rank	9.6%	26.2%	36.6%	12.72%
ResNet110 FC256 + re-rank	10.4%	25.8%	38.2%	13.43%
ResNet110 FC512 + re-rank	10.6%	26.6%	37.2%	13.64%
ResNet110 FC768 + re-rank	10.2%	27.4%	39%	14.12%
ResNet110 FC1024 + re-rank	12.8%	31%	42%	16.27%

TABLE VI: Ablation experiments of ResNet56 and ResNet110 classifier (averaging from ten different runs).

In conclusion, using lightweight methods to re-identify sketch images could not be as good as using state-of-the art models. However with much less parameters and computation time, the efficiency of this model is above most of the compared state-of-the-art models.

References

[1] "Kasus kriminal meningkat 7,04 persen dalam sepekan, salah satunya perampokan," 2020, https://nasional.kompas.com/read/2020/05/18/16253371/kasus-kriminal-meningkat-704-persen-dalam-sepekan-salah-satunya-perampokan.

[2] "Dua pekan terakhir, polri catat peningkatan kejahatan 11,80 persen," 2020, https://nasional.kompas.com/read/2020/04/20/20542321/dua-pekan-terakhir-polri-catat-peningkatan-kejahatan-1180-persen.

[3] "Ini alasan angka kriminalitas meningkat pekan lalu menurut polri," 2020, https://nasional.kompas.com/read/2020/05/18/16253371/kasus-kriminal-meningkat-704-persen-dalam-sepekan-salah-satunya-perampokan.

[4] "Dalam sepekan, polri catat peningkatan kejahatan jalanan di indonesia," 2020, https://nasional.kompas.com/read/2020/05/12/17363331/dalam-sepekan-polri-catat-peningkatan-kejahatan-jalanan-di-indonesia.

[5] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, Yonghong Tian, "Cross-domain adversarial feature learning for sketch re-identification," 2018, https://www.pkuml.org/resources/pkusketchreid-dataset.html, Last accessed on 2020-11-30.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep residual learning for image recognition," 2015, https://arxiv.org/abs/1512.03385.

[7] R. F. Rachmadi, S. M. S. Nugroho, and I. K. E. Purnama, "Lightweight residual network for person re-identification," in *International Conference on Information Technology and Digital Applications (ICITDA)*, 2020.

[8] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," 2017.

[10] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[11] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2018.

[12] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transaction on Multimedia (TMM)*, 2020.

[13] Y. Idelbayev, "Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch," https://github.com/akamaster/pytorch_resnet_cifar10, accessed: 2021-06-04.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[16] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[18] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 2016.

[19] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] N. Martinel, G. Luca Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.