Evaluating Person Re-identification Performance on GAN-enhanced Datasets

Daniel Hofer and Wolfgang Ertel

Institute for Artificial Intelligence, University of Applied Sciences Ravensburg-Weingarten, Doggenriedstrasse, 88250 Weingarten, Germany

Keywords: Person Re-identification, GAN (Generative Adversarial Network), Data Enhancement.

Abstract: Person re-identification remains a hard task for AI systems because high intra-class variance across different cameras, angles and lighting conditions make it difficult to create a reliable re-identification system. Since only small datasets for person re-id tasks are available, in recent years Generative Adversarial Networks (GANs) have become popular to improve intra-class variance to train more robust re-identification frameworks. In this work we evaluate an Inception-ResNet-v2 using triplet loss, introduced by (Weinberger and Saul, 2009), which works very well for face re-identification and use it for full-body person re-identification. The network is trained without GAN generated images to get a baseline accuracy of the network. In further experiments, the network is trained by adding constantly rising amounts of synthetic images produced by two image generators using different generating approaches.

1 INTRODUCTION

The task of person re-identification proposes an interesting challenge. It involves very high intra-class variance due to different lighting conditions, poses, cameras and even clothes. Furthermore, the training and test sets do not necessarily share any persons in common. This means, that classifier learning is not a viable option. A network needs to be able to learn how to generate a representation that can be compared without further processing. There is already a network architecture available that was built for exactly such an use-case, the so called Siamese neural network (Bromley et al., 1994). This architecture tends to need a lot of training data to perform well(Schroff et al., 2015). In the area of full body[1] person re-identification there is no dataset available, which can compete with the size of face re-identification datasets (e.g. (Cao et al., 2018)) but there are many different GANs available that are able to generate images based on existing person pictures to enhance datasets. This work evaluates two different image generators by training a improved FaceNet(Schroff et al., 2015) for person re-identification with datasets enhanced by various amounts of synthetic images.

2 RELATED WORK

(Schroff et al., 2015) proposes ”a system, called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity.” (Schroff et al., 2015). Their system achieved a new record accuracy of 99.63% in the widely used Labeled Faces in the Wild dataset(Huang et al., 2007), but is only able to re-identify faces and not images of a complete person with the face not visible.

(Zheng et al., 2017) introduced a re-identification system which uses a GAN to generate unlabeled images of persons. They propose to assign an uniform label distribution to those images by Label Smoothing Regularization for Outliers. Their experiments show, that the GAN generated images improve the discriminative ability of the learned embedding. (Ma et al., 2017) proposes a novel ”Pose Guided Person Generation Network (PG^2)” (Ma et al., 2017) that is able to generate images of persons in arbitrary poses. The output image is based on an input image of a person and a target pose.

(Bak et al., 2018) addresses the issue of the lack of diversity in lighting conditions in currently available datasets. They introduce a new synthetic dataset created with the Unreal Engine 4 game engine. 100 virtual humans are placed in modeled environments

[1] Full-body in this context means an image of the whole or nearly whole body.

Hofer, D. and Ertel, W.
Evaluating Person Re-identification Performance on GAN-enhanced Datasets.
DOI: 10.5220/0010060200770081
In *Proceedings of the International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS 2020)*, pages 77-81
ISBN: 978-989-758-479-4

with realistic outdoor and indoor lighting conditions and are captured from different viewpoints. (Ge et al., 2018) uses a pose map and an input image from a person to create a new image of that person in the given pose. This way the authors want to force a network to learn identity related and pose unrelated features. (Liu et al., 2018) approaches the problem of insufficient pose coverage of existing re-id training datasets. Therefore the authors propose ”a pose-transferrable person ReID framework which utilizes pose-transferred sample augmentations (i.e., with ID supervision) to enhance ReID model training.”(Liu et al., 2018). In addition to the conventional GAN-discriminator, they introduce a novel guider-sub-network which guides the generated data towards better satisfying the re-id loss. (Qian et al., 2018) also addresses the problem of huge pose variations in the person re-id task. They propose a novel image generation model to generate realistic person images conditioned with a pose. This way they can learn re-id features without the influence of a pose.

(Zheng et al., 2019) merges two images of persons by taking the appearance related features like clothing, shoes etc. and maps them to a person on a different image taking hairstyle, posture and face from the second image.

3 METHOD

The goal of this work is to verify whether real world datasets complemented with synthetically generated images can improve the accuracy of a re-identification system. The mentioned system is working on images showing the complete person and not only small parts, e.g. the face. Figure 1 shows some sample images from the Market-1501 dataset. Persons are visible in different poses and angles meaning a re-id system needs to be robust against those changes.

Figure 1: Example images from Market-1501 dataset(Zheng et al., 2015).

To evaluate, if GAN generated images can be helpful to increase the robustness of a re-id pipeline, an Inception-ResNet-v2(Szegedy et al., 2017) is trained. The complete training process is based on FaceNet(Schroff et al., 2015) which is a record breaking face re-id system. They propose to use the Triplet-Loss function first introduced in (Weinberger and Saul, 2009). This way, the squared L2-distances directly correspond to face similarity. The complete system is trained in an end-to-end manner. Regarding the FaceNet implementation[2], the used deep neural network was changed to a more recent one, the Inception ResNet-v2. (Szegedy et al., 2017) states, that this network architecture needs lesser training data because it converges faster and therefore benefits this work. Also the stem was exchanged for the Inception-ResNet-v1 stem to reduce dimensionality.

To get a baseline to compare later results to, the network is trained without generated images first.

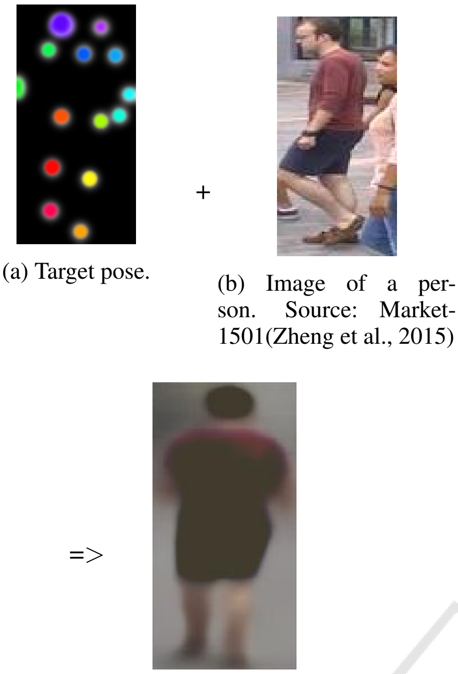
To verify, if artificial images can improve the accuracy, two different GANs are used to enhance existing real world datasets with newly generated images. The first GAN integrated in our pipeline is FD-Gan(Ge et al., 2018). This system uses a target pose map and an image of a person as input. The target pose map consists of key points of the human skeleton. The image generator then proceeds to generate an image of the person in the target pose. To generate new images, an image was chosen at random and combined with a pose extracted from a random image of a different person. By using this system, new images of the same person are generated. The person will have a new pose, but is still wearing the same outfit. Figure 2 shows an example of the image generation process. On the left side there are a pose map and an image of a person. The resulting image on the right side shows the person in a new pose.

To compare the training results with a completely different image generation approach, DG-Net(Zheng et al., 2019) was chosen. There, people put on new cloths, so to speak. As input, the generator needs two images of different people. The outfit of one person will be transferred to a different person. This way, the network should be directed into learning, that the identity of a person is not related to one’s outfit but other features like the body shape, hair color and so on.

Figure 3 shows the inputs and output of the DG-Net image generator. It is clearly visible, that the outfit of the person in the second image was mapped to the person in the first input image.

To evaluate the proposed framework, Mar-

[2] Our implementation is based on this work:https://github.com/davidsandberg/facenet

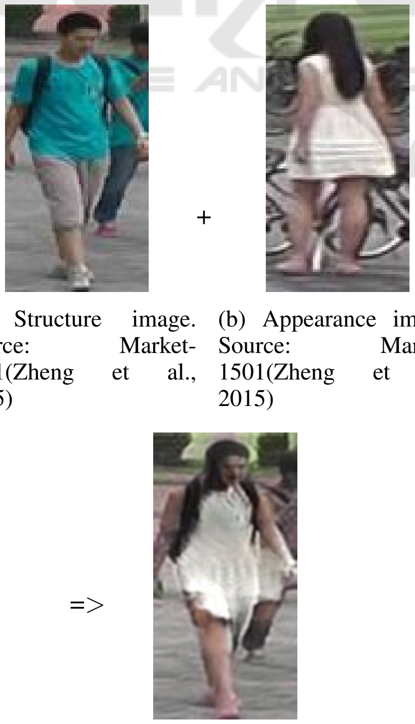


(a) Target pose.

(b) Image of a person. Source: Market-1501(Zheng et al., 2015)

(c) Newly generated image

Figure 2: Input and output of FD-GAN for image generation. Those image are generated by the author using the weights provided by the paper authors.



(a) Structure image. Source: Market-1501(Zheng et al., 2015)

(b) Appearance image. Source: Market-1501(Zheng et al., 2015)

(c) Newly generated image

Figure 3: Input and output of DG-Net for image generation.

Table 1: Ratios of real to generated images in the used training datasets.

Dataset	Ratio
	Real:Generated
1	1:0
2	1:0.5
3	1:1
4	1:2
5	1:10
6	0:1

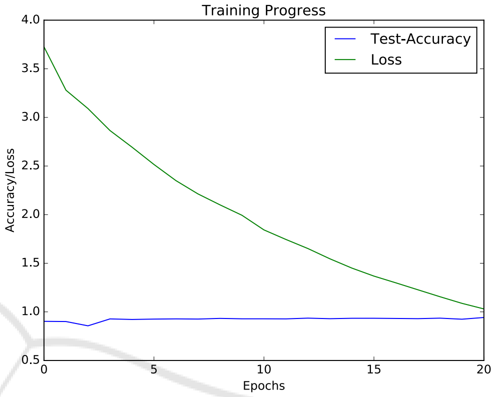


Figure 4: Accuracy and loss during training on Market-1501 with zero synthetic images.

ket1501 (Zheng et al., 2015) and DukeMTMC-reID (Ristani et al., 2016) datasets were chosen because they are widely used. Therefore it will be easy to compare the results of our work to others. Market-1501 contains 12936 training images of 751 persons and 19732 images for testing. DukeMTMC-reID training set contains 16522 images of 702 different identities. The test-set contains 17661 images in 1110 different classes.

To test if the GAN generated images can improve the accuracy of the trained network, the existing datasets were enhanced by different portions of generated images. Table 1 gives an overview of the various ratios. Dataset 6 was created containing 1000 generated images for each identity. If synthetic images improve the network's accuracy, the results on this dataset should be very good, if not, it should be clear that the network is underperforming. Before the actual evaluation, the threshold, which yields the best accuracy is chosen. The threshold is the upper limit for the distance between two images showing the same person.

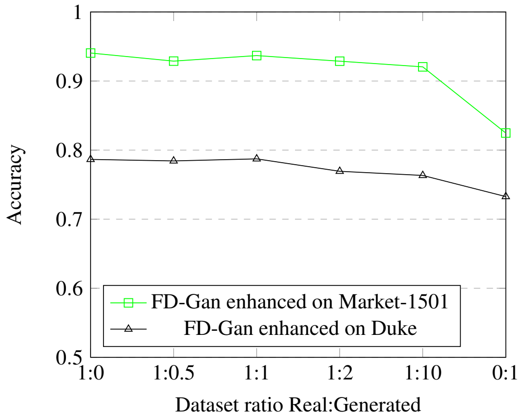


Figure 5: Accuracy over different real to generated image ratios.

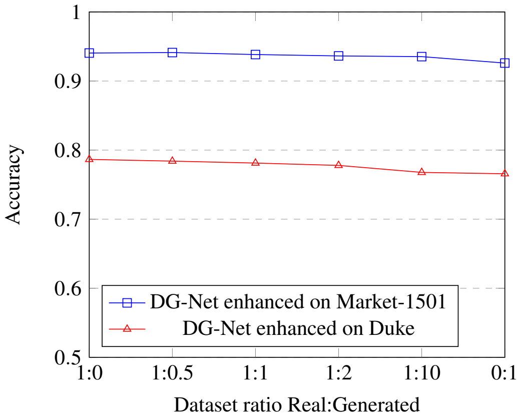


Figure 6: Accuracy over different real to generated image ratios.

3.1 Preprocessing and Training

Before training the images are resized to 160 times 160 pixels to match the input size of the network. Afterwards the network is trained for 20 epochs using the ADAGRAD optimizer. The learning rate started at 0.05 with a decay factor of 0.98 every 4 epochs. The small number of epochs is justified in the fact, that, as shown in figure 4 afterwards the training process was stalled out and did not yield any progress. Stopping it there is preventing further over fitting to training data. To get a baseline, the network was trained on the Market-1501 without generated images.

4 RESULTS

The baseline, retrieved by testing a network only trained on real data, achieves a accuracy of 94.05% on the Market-1501 and 78.65% on the DukeMTMC-reID dataset.

To evaluate, if GAN enhanced datasets can improve the re-id accuracy, enhanced datasets were used for training the network. Figure 6 shows the results of the network over different ratios of real to generated images using the image generator from DG-Net. On the other hand, figure 5 shows the results of the network over different ratios of real to generated images using the image generator from FD-Gan.

For the evaluation of the first GAN, FD-GAN, the Market-1501 dataset was enhanced by artificially generated images according to the ratios in Table 1. No improvement is visible, the accuracy is worse by 0.06% compared to the baseline results. The drop in accuracy when using 1000 syntactical images per identity indicates that the generated images are not as good as the real ones. Figure 2c visualizes, that the generated images are quite blurred. The trained re-id pipeline does not seem to handle that very well.

If enhancing the Market-1501 with the generator from DG-Net, the best network was trained on a ratio of 1:0.5 and achieved an accuracy of 94.12% in Market-1501 and 78.4% in Duke. On Market this is a small improvement of 0.07% compared to the baseline. For comparison, DG-Net reaches 87.4% accuracy on Duke-MTMC-reID and 98.5% on Market-1501 and FD-GAN achieves 86.28% accuracy on DukeMTMC-reID datasets and 98.41% on the Market-1501 dataset in peak performance. Figure 3 shows, that the image generation is not always perfect. The trained re-id system is not as affected by those errors as by blurred images because the drop in accuracy is not nearly as high(shown in Figure 6 compared to Figure 5) as with blurred generated images when using 1000 synthetic images per identity for training.

For this evaluation, both network weights were taken from the authors of the original works.

5 CONCLUSION

The generated images produced with two different image generators do not improve the accuracy of the trained re-identification system. Neither new images of existing persons in new poses nor other clothes yielded significant improvements in accuracy. The amount of images per class does not seem to be the problem. We assume, that there are too few identities, hence classes to train on. So the better approach would be to add more classes to improve the performance of the network.

Additional future work would be the usage of an image set containing images from several different datasets for training to improve the cross-dataset performance. Also a changed training process could be executed. The network would be trained on imageNet as a classifier. Before inference, the softmax, dropout and avg. pooling layers would be stripped. This way the amount of available training data is not the problem and the network could still work as feature extractor.

REFERENCES

Bak, S., Carr, P., and Lalonde, J.-F. (2018). Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 189–205.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al. (2018). Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1222–1233.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., and Hu, J. (2018). Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Van Gool, L. (2017). Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416.

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.-G., and Xue, X. (2018). Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

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

Zheng, Z., Zheng, L., and Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762.