Politecnico di Torino
Master's Degree in Data Science and Engineering

Synthetic-to-Real Domain Transfer with Joint Image Translation and Discriminative Learning for Pedestrian Re-Identification

Master Thesis Summary

Supervisors:
Prof. **Barbara Caputo**
Dott. **Mirko Zaffaroni**

Candidate:
Antonio Dimitris Defonte

1 Overview and Background

Autonomous methods for pedestrian re-identification provide additional intelligence for surveillance and security applications in high-risk areas. This task has seen a growing interest in the computer vision community, with recent developments in cross-domain techniques. In this work, we adopted a generative approach for synth-to-real pedestrian re-identification with the objective of generalizing from our synthetic dataset, *GTASynthReid*, to real-world data. This thesis completes the internship in which GTASynthReid was created by exploiting only the graphic engine of *Grand Theft Auto V*. Both the internship and this thesis were completed at the *Links Foundation* in the *Data Science For Industrial & Societal Applications* area. The Links Foundation operates at a national and international level on topics such as industry 4.0, smart mobility, agritech, space economy, security and intelligence. The following subsections provide a broad overview of the field to contextualize our contributions. In the subsequent sections instead, we first introduce the building blocks of our architecture and then show the evaluation protocol for both quantitative and qualitative results. Finally, the last section is devoted to remarks and future directions.

1.1 Generative adversarial neural networks

Generative adversarial neural networks (GANs) include a discriminator and generator. The former is a standard neural network for binary classification (other options are possible), while the latter samples from a random vector and, using transpose convolutions, outputs an image. They are trained in an adversarial fashion, meaning that the generator will try to fool the discriminator into believing that the fake generated images are real, while the discriminator learns to distinguish whether an image is real or fake. Ideally, this process will converge to an equilibrium where the generator outputs images indistinguishable from the training data and the discriminator always guesses with a probability of 0.5 that the images it receives are real or fake.

1.2 Pedestrian re-identification

Person re-identification is a challenging task where one has to match the same individual across several recorded videos. It is essential to account for pose, viewpoint and illumination variations. For this task, datasets contain already extracted pedestrian bounding boxes. The identities are then splitted between the training and test set. It is crucial to notice that these two partitions do not share any identity. The test partition is further divided into query and gallery sets. Most person re-identification pipelines extract pedestrian feature descriptors via supervised learning on the training partition. Then, using the learned model, they first extract image features from all the bounding boxes in the test set and finally perform retrieval. For each probe in the query, we have to compute the distance with all the samples in the gallery and then rank by decreasing similarity. To evaluate such ranking, common metrics are the *Mean average Precision* (MaP) and the *Cumulative Matching Characteristics* (CMC).

1.3 Cross-domain transfer

When a model is trained and tested on the same dataset, it will usually perform poorly when we evaluate it on different data. This stems from several reasons. For instance, since real-world datasets are recorded in a given time window, it could happen that they only capture weather and illumination conditions that do not overlap. Other issues include viewpoint position, occlusions and person-to-person interactions. Besides this, annotating data in different environments is much more expensive than just asking for unlabeled images. For these motivations, we have recently seen a surge in models that adapt from a source to a target environment. This scenario is known as *unsupervised domain adaptation* (UDA). During training, we can exploit the whole source data and the unlabeled training partition of the target for cross-domain techniques. For the evaluation, the retrieval is performed on the test target identities (which were never seen by the model). One will notice that this procedure differs from UDA in other tasks, where we usually validate on a third dataset.

1.4 Synthetic data

Collecting and annotating real-world data is error-prone and time-consuming. Moreover, when recording a dataset for this task, we cannot expect to capture all the desired variations for pose, illumination and weather conditions. Synthetic data aims to solve these issues since one can model computer-generated worlds at will, automating the labeling process and combining the characteristics of more datasets. Another issue concerns data ethics. Pedestrians are often recorded without informed consent, unaware of the data collection purposes (*e.g.*, sensitive applications for surveillance and intelligence). Software-generated data also frees us from this burden since we can exploit synthetic identities. However, we have now introduced an additional layer of difficulty. In a practical scenario we would still want to adapt the synthetic data to real-world environments.

1.5 Contributions

Our contributions are as follows:

- we extended a Pytorch framework[1] [4] for person re-identification to allow learning via translation from synthetic data;

[1] The original work can be found at https://github.com/KaiyangZhou/deep-person-reid.

- to the best of our knowledge, we are the first to employ the *contrastive unpaired translation* (CUT) framework [3] for pedestrian re-identification. CUT avoids the cycle consistency loss [5] and double architecture structure common to many other methods;
- we designed a similarity loss inspired from [2] that matches features at different discriminator layers when the discriminator itself is fed with a pair of source and translated images or a pair of target and translated images (when switching the perspective from the discriminator's point of view to the generator's point of view);
- we were able to achieve a better performance compared to a Resnet50 [1] baseline trained for direct transfer;
- although we could not surpass the current state of the art (given also the limited amount of pedestrians of our dataset), we achieved better results than some earlier real-to-real adaptation methods with a lighter, one-stage synth-to-real model;
- we also measured the distance by means of FID between the real and our GTASynthReid datasets before and after the translation, observing a lower value after the translation.

2 Architecture

There are many ways in which one can approach this task, ranging from *iterative pseudo labeling techniques* to *generative adversarial networks* (GANs). We embraced a generative approach that translates the images of our synthetic source to match the style of a target real dataset. During the translation, we jointly learn discriminative features from embeddings of the source identities injected with target-domain information. For the image translation, we adopted the CUT framework of work [3].

2.1 Domain Mapping

CUT employs a single GAN for one-way source-to-target translation. The generator has an encoder-decoder structure since it first encodes each source image into a low-level representation exploited by the decoder to generate the output. The generator tries to fool the discriminator by learning to output images with the same style as the target, while the discriminator tries to predict their domain. However, instead of classifying an entire image, the discriminator applies this process to patches of the input images. Since the discriminator has a less deep architecture, we argue that it should capture general concepts of its input images. Thus, inspired by earlier works [2], we designed a feature matching similarity loss applied to each discriminator's layer. When optimizing for the generator, we ask the extracted features of a translated and a target image to be similar. When optimizing for the discriminator instead, we expect the extracted features of a source and translated image to be similar.

2.2 Relationship preservation

Unpaired image translation can be seen as a disentanglement problem, where one wants to separate the content from the style of an image. While we seek to preserve the content of each source image, we wish to match the style of the target dataset without having pairs of source-target samples. In adversarial works, *cycle consistency* is a common mechanism that guides the generative process by mapping from one domain to the other and vice versa (with a pair of GANs). CUT gives up on the restrictive bijection relationship between the two domains and employs a clever one-way translation method to constrain the image generation. The intuition is that, after the translation, body parts such as hands, legs and torso should preserve their semantics and remain in the same region. CUT addresses this by comparing corresponding input and output image patches, asking them to be similar. Identifying as *query* an output patch, as *positive* the corresponding input patch and as *negatives* all the remaining input patches, we want the dot product between the query and the positive to be higher than all the dot products between the query and the negatives. CUT frames this as a classification problem where the logits are the dot products. Regarding the image patches, we can directly extract them from the generator's encoder. When considering a specific layer, each spatial location of the output feature map looks at a different region in the input image (receptive field arithmetic). In this sense, each spatial location is a feature vector (along the channel dimension) of the corresponding input patch. As in CUT, we repeat this process at different depths of the encoder. For each selected layer, the feature embeddings are further encoded by a smaller network before computing the dot products. Additionally, one can also consider the identity mapping loss. This term will constrain the generator to output an image as close as possible to the input one when fed with samples from the target dataset. This procedure should help preserve better pedestrian characteristics such as the clothing colors, shape and texture.

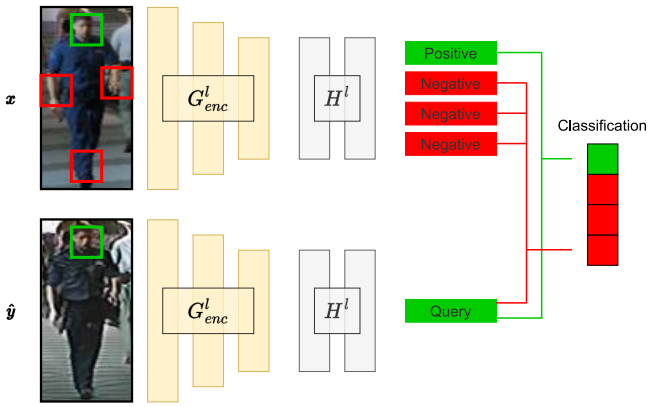

Figure 1: For each source and translated image, we extracted learned feature vector for each input image patch from the generator's encoder. Those features are further processed by feeding them to smaller networks. The resulting embedding is then employed for the patch classification (output query, input positive and negatives).

2.3 Discriminative learning

While translating the images of GTASynthReid to a target real-world dataset, we exploit the generator's features to influence the discriminative module. The features that come from the last layer of the encoder hold some low-level information about the source images without being specialized in recognizing pedestrian identities. At the same time, these are the features responsible for the translation, holding target-domain information. Following this direction, we pass the resulting feature maps to a network trained to classify the source identities, employing a Resnet50 from the second convolutional block. We show that exploiting these feature representations of the original source images returns better results than a direct synth-to-real transfer on a Resnet50 baseline.

3 Results

We adapted our synthetic dataset to three real-world targets: Market1501 (Market), DukeMTMC-Reid (Duke) and CUHK03. In this section, we explain the evaluation pipeline of our models and we also provide some insights about the quality of the image generation.

3.1 Quantitative results

To compute the ranking, we extracted feature embeddings of all query and gallery images by feeding them to the generator's encoder and then passing the outputs to the re-identification network

trained on the source identities. Compared to a standard Resnet50 baseline trained for direct transfer, we achieved better results. We carried out an ablation analysis about the contribution of the identity mapping and feature matching losses. Over several experiments, without considering the image generation quality, including the feature matching losses in place of the identity mapping one returned the best-performing models. Most methods in this field only evaluate the learned source-to-target model on the target dataset for the cross-domain analysis. Besides following this procedure, we also tested a more general synth-to-real adaptation. For each GTASynthReid-to-target learning configuration, besides evaluating the model on the target test set, we performed the retrieval also on the remaining two real datasets. In this way, one can see whether or not our model can generalize on other real-world data with different characteristics from the specified target set. Table 1 shows the performance of the baseline and the three GTASynthReid-to-target cross-domain models when evaluated on the Market dataset. Although GTASynthReid-to-Market returns the best performance, GTASynthReid-to-Duke and GTASynthReid-to-CUHK03 achieve similar results. We then compared for each target dataset our best-performing model with other real-to-real generative methods and general synth-to-real works. Recent generative techniques adopt more complex models than ours, while synthetic approaches use far larger datasets. For these reasons, they achieve better performance. However, when compared to earlier synth-to-real and real-to-real works, we achieve similar and even better results, using less data and a simpler model.

Model	**Target**	Market	
		MaP	R1
Baseline	-	23.3	39.4
Ours	Market	**42.6**	**61.2**
Ours	Duke	40.7	59.5
Ours	CUHK03	41.6	60.0

Table 1: We compare to a Resnet50 baseline for direct transfer each GTASynthReid-to-target translation evaluated on the Market dataset. GTASynthReid-to-Market achieved the best results although the other translations have comparable performance.

3.2 Qualitative results

Our model adapts from synthetic to real-world data by learning mappings that translate each source image to match the style of the target ones. We then exploit the features learned in the translation for the re-identification task. Nonetheless, one could also be interested in evaluating the quality of translated images. Although the best performing configuration of our models returned good images, we found out that, for the sole purpose of the image generation quality, adopting the identity mapping loss in place of the feature matching one delivered results with fewer artifacts. In figure 2 there are some examples. There are many ways one can test whether a translated image feels more realistic or not. For instance, we could ask a group of people to label our outputs with online services such as *Amazon Mechanical Turk*. Another option relies instead on using similarity metrics. We used the *Fréchet inception distance* (FID) to measure the similarity between our dataset before and after the translation. We first measure how distant GTASynthReid is with each target. Then, we translated our images to each target and computed the FID again, noticing smaller distances.

4 Conclusions

The very heart of our implementation is the adoption of CUT [3] for cross-domain person re-identification. This framework translates the source data to match the target style by asking for similar corresponding input-output patches. It represents a one-way translation that does not require any cycle consistency loss or double architecture. We exploited it in a synth-to-real scenario, adapting our computer-generated GTASynthReid to three other datasets. We combined CUT with discriminative learning to allow for a one-stage training procedure and designed a feature matching loss that improved our results. Recent generative works disentangle appearance from pose across domains via cycle consistency. They achieve better performance, although their models are harder to train. Compared with earlier generative methods that only work with appearance information, we provide a simpler model with similar or better results. Moreover, for hardware limitations, existing works sample more images than we did, in particular when compared to recent synth-to-real models that use far more pedestrians. Instead of just evaluating each translation on the specific target, we also tested our generalization capabilities by evaluating on all the three chosen real datasets. On a qualitative level, we noticed that the generated images have a closer style to the target data. In this sense, we tried to output more realistic images, transcending pedestrian re-identification.

For future improvements, one could try to increase the number of identities by designing new characters for the video game Grand Theft Auto V. Recent works show that the availability of more pedestrians can render the re-identification task more challenging, improving generalization capabilities. Another extension would be to design a method that disentangles, for instance, appearance and pose information without cycle consistency. However, unless we redefine the concept of patch similarity or design other losses, CUT can be inconsistent with this kind of objective. On a more general level, we feel that working with software-generated environments will inspire new interesting works and help to address existing challenges.

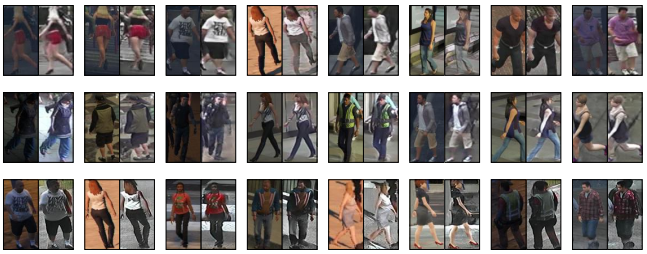

Figure 2: For each translated pedestrian we show the original (left) and translated (right) images from our GTASynthReid to Market, Duke and CUHK03 (first, second and third rows).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90

[2] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[3] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020.

[4] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3069237

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