Person Re-Identification Using Self-Supervised Learning

A. Nezih Kasim
Department of Computer Science
Bogazici University
Bebek, 34342 Beşiktaş/İstanbul
ahmet.kasim@boun.edu.tr

Omer Faruk Cavas
Department of Computer Science
Bogazici University
Bebek, 34342 Beşiktaş/İstanbul
faruk.cavus@boun.edu.tr

Abstract

In this work, we study on the person re-identification problem using self-supervised learning (SSL) approach. The person re-identification problem is a challenging problem because the labeled datasets are limited due to high annotation costs. It also stems from problem specific challenges such as different camera views, varying image quality, changing illumination etc. In this context, SSL methods can be used as a pretraining stage by learning feature representations from large unlabeled datasets. In this study, we utilize a state-of-the-art SSL person re-id model and try to improve it by enhancing the data augmentation in both pretraining and finetuning stages. Due to insufficient resources, the pretraining data scale and some parameters like batch size,learning rate etc. is far below the actual study, which directly affect the model performance in a negative manner. The results of our studies shows that adding specific color distortions in pretraining improves model performance while using augmentations in fine tuning stage does not yield better results. Finally, the inability to perform many experiments due to scarce resources clearly hinders achieving better results.

1 Introduction

Data annotation costs for person re-id problem is much higher than common computer vision tasks, which highlights the importance of pretraining. However, pretraining on a regular vision datasets like ImageNet is not a good choice due to high domain gap with person datasets. Accordingly, we may resort to unlabeled datasets in the context of self-supervised learning as it gains much popularity and offers high performance in current studies. LUPerson (large scale unlabeled person re-identification) dataset appears to be the best option for pretraining as it contains significant amount of person images taken from diverse range of environments. Pretraining on LUPerson dataset with a commonly used SSL framework MoCo v2, this study [6] displays great performance and obtains the second rank in the Market 1501 benchmark [13]. Inspired from this success, we attempt to increase the performance of this study by focusing on the data augmentation stage in both pre-training and fine tuning parts. Due to high gap between the resources we have and what they use, we have to work with much smaller dataset i.e Market 1501 for pretraining and minimize training parameters like batch size and learning rate. This does not allow even reproducing the current study but we try to obtain a baseline performance and then improve it with our proposed ideas.

For pretraining and fine tuning stages, we utilized the same models as our baseline i.e. MoCoV2 for the SSL framework and MGN for supervised fine-tuning. MoCoV2 is one of the most popular SSL frameworks used in vision tasks and shows high performances in benchmark datasets. MGN, on the other hands, is one of the state-of-the-art person re-id models that use both global features and local features. By keeping model architectures the same as the corresponding studies, we mostly focused

on data augmentation parts. In contrastive learning based SSL frameworks, augmented versions of the training images are used as positive and negative examples. For the supervised finetuning stage, triplet loss function tries to maximize similarities of the features generated for images coming from the same identity while minimizing that for images from different identities. Therefore, we can use data augmentations to increase the number of positive and negative examples both in SSL pretraining and supervised fine tuning stages.

Initially, we want to observe the difference between using a large scale dataset and a relatively small dataset in pretraining stage. Experiments reveals considerable performance degradation in mAP and CMC-1 scores when using Market 1501 rather than LUPerson dataset. mAP decrease from 64.6% to 31.6% in 10 percent fine tuning setting, and from 88.8 to 73.1 in 70 percent fine tuning setting. Upon these results, we studied two groups of augmentations: color distortion and affine transformation. Color distortion includes changing the brightness and the contrast values of images. Affine transformations consists of rotation, translation, scale and shear components. Simulating different camera angles and various illumination conditions is at the root of the motivation behind selecting these augmentations. We studied these augmentations in two ways. First, we applied pretraining on Market 1501 with default augmentations, color distortion augmentations and affine transformations. Then, we used these pretrained models on supervised finetuning with small scale Market 1501 dataset (i.e 10 % of the training set). Second, we used already pretrained models with LUPerson dataset and applied augmentations in supervised finetuning stage to see the effect of these augmentations on finetuning. Our experiments show that adding problem specific color distortions such as changing brightness and contrast in the pretraining stages improves cmc-1 score by 1%. On the other hand, affine transformations in pretraining does not change model performance and using augmentations in fine tuning stage does not give better results. Actually, we couldn't conduct enough experiments to find out optimal parameters for augmentations due to high training times, which was the most annoying issue in our studies.

2 Related work

2.1 Person re-identification problem

If we define the person re-identification problem, given a query person image or video, the problem is to find the matching images/videos from a gallery which is created by a multi-camera system. The matching images/videos correspond to the same identity with the query image/video. Image-based person re-identification has dominated the literature. We are also working on image-based person re-identification problems.

2.2 Person re-id approaches

When we look at the literature, we have seen that this problem is handled in three ways: Supervised, unsupervised and self-supervised learning.

2.2.1 Supervised learning

The first one is supervised Person Re-ID. Most studies fall into this category, where the person identities are available in the training stage. In this category, there are different approaches to apply the supervision. One of the most common ways is performing supervision through a classification loss in which each identity corresponds to a class and the problem becomes a multi-class classification problem. Another way is using a triplet loss function to ensure similarity between embeddings of the same person and dissimilarity between embeddings of different people. Feature learning is obtained globally or in a part-based way. In global feature learning, the features of the whole image are extracted while in part-based approaches, the features of different partitions of the image are extracted. There are also some studies like MGN [1] which apply both global feature learning and local feature learning.

2.2.2 Unsupervised learning

The Person Re-ID problem is also tackled in an unsupervised manner. The first option in this category is Unsupervised Domain Adaptation (UDA). In UDA, firstly supervised learning is applied

on a labeled source dataset, then the information obtained from this task is utilized in the unlabelled target dataset in an unsupervised manner.

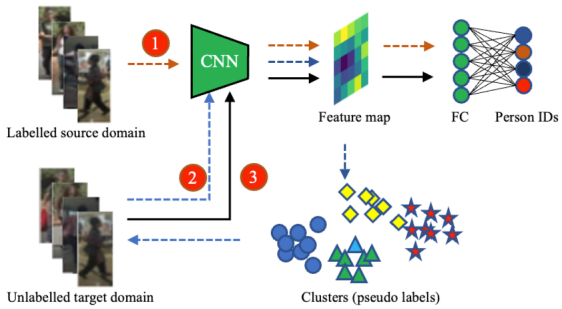

Figure 1: Overview of a UDA model for person re-id

Figure 1 shows the general structure of a UDA model. The dashed red lines show the supervised learning in the source dataset. The supervision in this figure is obtained by a multiclass classification. Yet, it could be different like using triplet loss as we mentioned before. After the pretraining on labeled data, features of the unlabelled target dataset are extracted through this model to identify the clusters, as shown in blue dashed lines. Finally, the clusters are used as pseudo labels for the supervised learning on the target dataset. The quality of the pseudo labels directly affects the supervised learning, which is a common challenge in unsupervised Person Re-ID.

The second option for the unsupervised Person Re-ID, is directly using unlabelled target dataset, without exploiting external Person Re-ID datasets. Generating robust person re-id representations in a fully unsupervised manner is expected to be more challenging than other approaches. Studies generally focus on identifying good clusters from unlabelled dataset to later use them as pseudo labels for supervision. They generally define the pseudo labels progressively to better apply the supervision later on.

2.2.3 Self supervised learning

Self-supervised learning (SSL) is an approach based on the idea of exploiting features from large unlabeled data that can be used in the actual problem. Due to high availability of unlabeled data, SSL could be a good way of learning representations and for transfer learning. This pretraining idea could be realized in different ways. Although various methods are proposed in the literature, we can divide them into three categories. The first category is pretext task-based models in which an auxiliary task is solved to learn representations from unlabeled data. These tasks can be created in many ways and there are different suggestions. One of which is the rotation task in which degree of the rotated image is predicted as a multiclass classification problem. Relative positioning, another proxy task, is to predict the position of a crop relative to another crop in the same image. The examples of pretext task-based models are abundant in the literature.

In the second category, generative based models, the input data is tried to be reproduced by the model so that it learns the representations to be used in another problem. Autoencoders can be seen under this category. The latent space representations can be used for the downstream task. Another one is context encoder in which a patch of the image is cut out and we train the model to generate this missing part.

The last category, contrastive learning, includes training a model to distinguish between similar and different input data. Similar images are called positive examples while dissimilar ones are called negative examples. Generally, positive pairs contain two augmented versions of the same image while negative pairs consist of two augmented versions of different images. The augmentation can be random crops of the original image, flipping the image or color distortion etc. In general, a contrastive loss function is used so that the similarity between embeddings of the positive samples is maximized while the similarity between embeddings of the negative ones minimized. In this way, the model learns the representations of the image by comparing it with negative and positive samples. In

Datasets	#images	#scene	#person
Market1501	32,668	6	1501
MSMT17	126,441	15	4101
LUPerson	4,180,243	46,260	>200k

Table 1: Person re-id datasets

literature, there are various contrastive learning frameworks such as SimCLR [2], MoCo [3], BYOL [4], DINO [5] etc. In general, they are based on the contrastive loss using augmentations of images as positive and negative pairs. The backbone model may change as in the case of DINO which uses Vision Transformers (ViT). Yet, some frameworks like BYOL modify the general SSL structure so that the model does not need negative examples. It adds a multi-layer perceptron (MLP), after the projection head of the first network. This MLP predicts the projection of the positive example returned from the second network and the loss function is L2 loss.

The contrastive learning approach is applied quite commonly in the computer vision domain, and it has gained good performances in popular datasets like ImageNet. Compared to the other two approaches, only a few studies used the SSL approach for person re-id problem as described in the next section.

2.3 SSL for person re-identification problem

SSL approach is also useful for person re-id problem since data annotation for this problem is even more difficult than other computer vision tasks. Besides, learning good representations from large unlabeled datasets could be helpful for the downstream task to achieve better results. Therefore, we focused on this approach to tackle person re-id problem. For the transfer learning phase, we are trying to improve supervised person re-id models rather than unsupervised models. In this context, the first study [6] that makes the first attempt to apply SSL techniques to person re-id problem also presents a large unlabeled dataset namely LUPerson. They state the need for such a large scale dataset for person re-id problem and they actually pave the way for other studies [7][8] to use SSL methods for this problem. The LUPerson is an unlabeled person dataset of 4M images and over 200K identities, which, they state, is 30 times larger than the largest available RE-ID dataset.

As shown in Table 1, the number of images and identities are far beyond the existing datasets. Also, it has many diverse environments, including much more scenes and different camera views as indicated in their studies. On the other hand, this study also shows that using datasets like ImageNet for pretraining is not that helpful due to the domain gap between ImageNet and person datasets.

For their SSL framework, this study uses MoCoV2 [9], a modified and better version of original MoCo framework. However, they indicates that directly applying MoCoV2 to person re-id problem is not possible, that is, it requires special attention. The first observation they made is the default augmentations used in original framework don't fit this problem well. Although cropping, resizing and flipping are still useful for person re-id, the color jitter is harmful to the problem as they indicated. This is, they indicates, because of the fact that the color information is very necessary to identify different identities. Besides, they say random erasing improves the performance of the model. Another observation they show is as the augmentation becomes stronger, e.g. larger erasing area for random erasing, the SSL task becomes harder, so better representations are obtained from pretraining stage.

Another key factor that they studied is the temperature parameter in the contrastive loss. As they indicate, a too large value for temperature will make the model to hardly distinguish the positive and negative samples. On the contrary, a too large temperature value will cause the model not to learn from hard negative examples. Therefore, they studied carefully to find the optimal temperature value. As a result, they argue that a relatively small temperature is better for person re-id since interclass variation is small compared to other tasks.

They studied to improve both supervised and unsupervised re-id methods using SSL frameworks. For supervised learning, they use different existing models, but they achieved best results with MGN,

which is one of the state-of-the-art models in supervised person re-id. They improved the MGN model performance by benefiting from pretraining on the LUPerson dataset using MoCoV2 framework. The baseline MGN model they compared uses supervised ImageNet pretraining.

The second study that makes use of SSL methods for person re-id [7], is a vision transformer (ViT) based model and they are the first that uses ViT based SSL for person re-id problem. For the pretraining part, they used the LUPerson dataset which was created by the previous study that we mentioned. They also highlight the domain gap issue when using non-person datasets like ImageNet for pretraining. They focused on ViT based SSL methods to choose the best one for person re-id. After some experiments, they concluded that DINO outperforms other ViT based SSL methods for person re-id task. The second topic they investigated is how to reduce the large LUPerson dataset because of the need for high computational resources to train such a large dataset. To that hand, they proposed a metric called Catastrophic Forgetting Score (CFS) to evaluate the gap between pre-training and downstream data, and thus to reduce the pretraining data based on similarities of samples. A subset of the images with high CFS score are selected for the pretraining stage. In fact, this score is computed by training some lightweight models for a small number of epochs. On the model side, they proposed a ReID-specific patch embedding called IBN-based convolution stem to improve the peak performance. For supervised fine tuning model, they used TransReID, a ViT based supervised person re-id model, because of its significant performance boost over CNN based models. Using DINO as their SSL framework and with this modification on the ViT model, they conducted experiments with supervised learning, unsupervised domain adaptation and unsupervised learning. They achieved better results than the supervised pretrained counterpart that uses ImageNet. They state that pretraining on 50% filtered LUPerson dataset with 8 x V100 GPUs takes 51 hours ,which shows the computational requirements of training on such large datasets.

3 Method

For pretraining and fine tuning stages, we used the same networks as our baseline [6] i.e. MoCoV2 as our SSL framework and MGN as our supervised fine-tuning model. MoCoV2 is an improved version of MoCo which is based on the idea of contrastive learning using a memory bank of past projections as negative examples. MoCo tries to keep a large dictionary since it makes the problem harder to match a query from a large dictionary and thus improve learning representations. For that reason, they create the dictionary as a queue of data samples in which the encoded representations of the current mini-batch are added and the oldest are removed. This, in fact, makes it possible to keep a large dictionary independent from minibatch size as opposed to other SSL frameworks that must keep a large mini-batch size to make the dictionary larger. So, the dictionary size is a hyperparameter in MoCo framework which makes it flexible. Although using a queue provides a large dictionary, it makes the gradient update much difficult as the gradient must propagate through all samples in the queue. Likely, they have a solution to this by updating the key network, which is the network that negative examples go through, with momentum of the query network. Thus, only the query network is updated online and requires backpropagation as shown in Figure 2. MoCoV2 improves this framework in two ways by benefiting from the improvements that are obtained in SimCLR[8]. SimCLR shows the benefits of adding an MLP projection head and more data augmentation for the representation learning ability of the model. So, MoCoV2 directly applies these ideas to MoCo and obtains remarkable improvements.

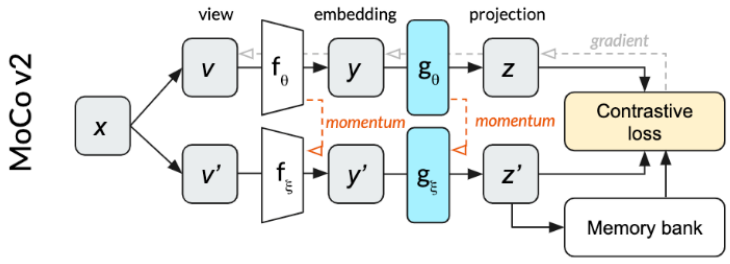

Figure 2: The architecture of MoCoV2 framework [9]

The loss function that MoCoV2 uses is the same as the other contrastive learning approaches. The contrastive learning could be imagined as a dictionary look-up problem in which the matching key of a query is to be found from a bunch of keys in the dictionary. The contrastive loss function is as follows.

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum_{k^-} \exp\left(q \cdot k^-/\tau\right)} \tag{1}$$

Here q is a query embedding, k+ is an embedding of the positive key sample, and {k-} are embeddings of the negative key samples. T is a temperature hyper-parameter. Generally, a query and a key creates a positive pair if they are augmentations of the same image, and negative pair if not.

For the supervised person re-id model, we used MGN, same as our baseline model. MGN extracts both global features and local features of the person images and uses both classification loss and triplet loss functions. It has a multi-branch deep network architecture, one branch for global feature representations and two branches for local feature representations. In MGN architecture, horizontal stripes are used for local feature learning. It uses a ResNet-50 backbone that feeds all three branches as seen in Figure 3.

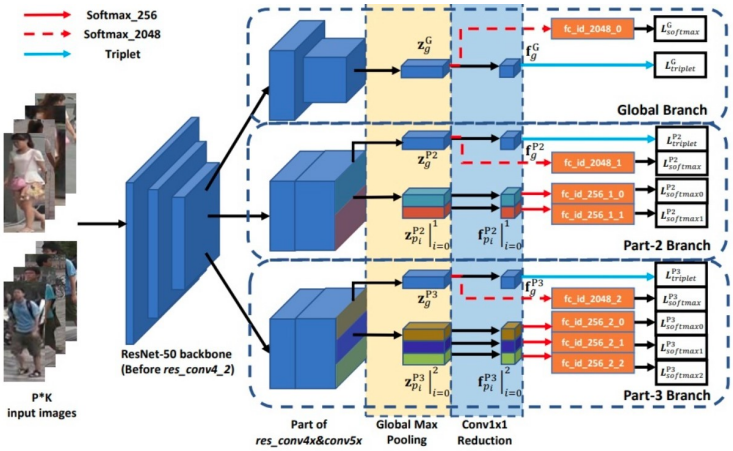

Figure 3: Multiple Granularity Network(MGN) architecture. [1]

They divide the part after res_conv4_1 block of ResNET-50 into three independent branches. In upper brach, they first apply stride-2 convolution and global max-pooling. After that, a 1x1 convolution layer with batch normalization and ReLU activation is applied and a 256-dim features obtained from 2048-dim features. This branch is called global branch and it does not include any partitioning. The other two branches have similar architecture to global one but in this case no down-sampling is applied in res_conv5_1 block to keep proper areas of receptive fields, as they states. The output features are divided into two and three identical stripes in horizontal direction. After that, the operations are similar to global branch. For testing stages, features coming from all branches are concatenated as the final feature that represents the image for similarity check.

In this project, at the beginning, we want to observe the difference between using a large scale dataset and using a relatively small dataset in pretraining tasks. After showing the gap between the datasets, we plan to explore the effects of various data augmentation methods used in training.

Fu et al. [6] performs their pretrain tasks on the LUPerson dataset that is also proposed by them. As mentioned in previous sections, the pretraining process is data hungry methods for this domain and it requires a large scale dataset. Considering this, Fu et al. created a dataset named “LUPerson” for person re-identification unsupervised pretraining tasks instead of using ImageNet. This LUPerson dataset consists of more than 4M bounding boxes collected from more than 200K different identities. In this study, we used a smaller dataset, Market1501[11] dataset, for our experiments. The main

Setting	Default	+RE	−GS	−GB	−CJ	−CJ + RE
mAP	73.4	74.2	73.2	73.3	74.0	74.7
$cmc1$	74.0	74.8	73.9	74.1	74.6	75.4

Table 2: Transfer performance on the CUHK03 dataset with different data augmentations. "+, -"mean with and without, and "RE, GS, GB, CJ" mean RandomErasing, GrayScale, GaussianBlurring, and ColorJitter respectively.

motivation for why we use the Market1501 dataset instead of LUPerson is that dealing with large datasets like LUPerson is not feasible because we do not have access to enough resources for processing the LUPerson. The Market1501 dataset has only 13K training images while LUPerson contains over 4M unlabeled training images. Another motivation is that the LUPerson dataset is only available on Baidu which is inaccessible from outside China. Therefore, we perform our pretraining task on the Market1501 dataset.

Augmentation is crucial in machine learning models created for computer vision tasks to make the model more robust to changes in dataset distribution. It is of higher importance when it comes to SSL approach since the contrastive learning, a type of SSL methods that we prefered for our problem, is constructed based on the augmentations of training set images. That is why, we wanted to pay special attention to augmentations in our study to improve the performance of the baseline model. We also studied the impact of augmentation in the supervised finetuning stage which is not explored in the SSL person re-id literature, to the best of our knowledge.

The first study that performs SSL for person re-id problem [6] studies carefully on the effect of data augmentation in their works. Actually, their findings in data augmentaion stage constitutes one of their three main contributions, as they state. They study several data augmentation techniques in their SSL pretraining stage, i.e. augmentations in MoCo V2 framework with respect to transfer learning performance on CUHK03 dataset. They maintain that the common augmentation such as cropping, resizing and flipping are essential also for person re-id problem. So, they studied more on other augmentations like color distorting,random erasing etc.

As shown in Table 2, they conducted several experiments by including and excluding augmentations. Removing Gaussian blurring and random gray-scale does not affect performance much, so they didn't study them in their work. As for color jitter, exluding it boosted performance by 0.6% in terms of both mAP and cmc-1. Based on these experiments, they concluded that color information is crucial for person re-id problem and so color jitter should not included in pretraining augmentations. As we mention in the following sections, some type of color jitters like changing brightness or contrast may be useful also for person re-id, so totally ignoring color jitter may not be the best option. They also included another augmentation method, random erasing, since it is mostly adopted in the person re-id tasks. As we can see in Table 2, using random erasing improves both mAP and cmc1 by 0.8%. Furthermore, with experiments done by increasing maximum erasing area, they showed that stronger data augmentation gives better results for person re-id pretraining. By underlying the importance of task-specific augmentations, this study has led us to increase our efforts in this direction.

The seconds study that uses SSL for person re-id [7], does not mention any pretraining augmentation techniques specifically designed or chosen for person-re id problem. Yet, they use some data augmentations to analyze the feature invariance of pretrained models. They synthesize new images by six different augmentations applied on each image. Then, calcualtes a Centered Kernel Alignment (CKA) score to measure the feature invariance of pretrained models between original and simulated datasets. The higher CKA score a model has, the more invariant features they generate, thus the better it performs. The augmentations they use are shown in Figure 4. As seen in this figure, color jitter like changing brightness and contrast are used to measure the feature invariance, which also encourage us in direction of using them in SSL pretraining stage.

We studied two groups of augmentations: color distortions and affine transformations. Color distortions includes altering the brightness and the contrast values of images. We exclude hue in color jitter not to change the appearance of the color itself, as shown to be harmful in our baseline study. Affine transformations consists of rotation, translation, scale and shear components. The

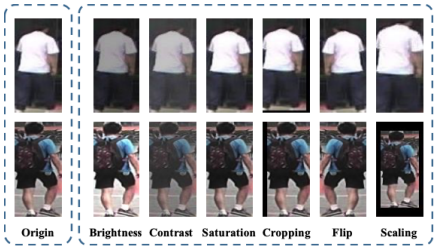

Figure 4: Examples of generated images in trans re-id paper [7].

motivation behind using affine transformation is based on the idea of simulating different camera angles and also varying distances to camera with the help of rotation, translation, scale and shear. That is, if our pretrained model learns to generate invariant representations for affine transformation augmented images, then the supervised model may learn better to recognize different images of an identity obtained from several cameras. Example images generated from these augmentations can be seen in Figure 5.

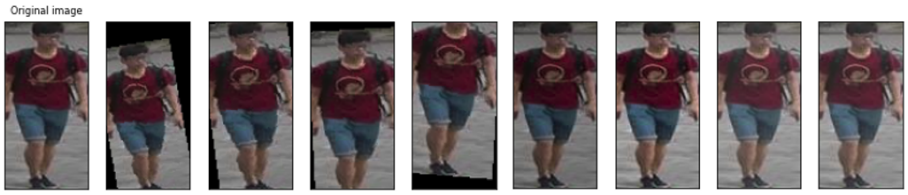

Figure 5: Examples of color and affine augmentations in our study.The first four belong to affine transformations, others belong to color distortion.

At the beginning, we intended to use a novel approach for random crop, contrastive crop, which is proposed by a recent study [11]. It is based on the idea of eliminating crops that include purely the background patches without the object of interest. Although this augmentation is shown to outperform the default random crop for a common computer vision problem, we think, it is not likely to improve the performance of person re-id problem since the images in person images are already cropped from bounding boxes so there is no such "irrelevant crops" problem.

We study aforementioned augmentations in both pretraining and fine tuning stages. In pretraining, augmentations are used to optimize the contrastive loss function. That is, the distance between augmentations of the same image are minimized while distance between augmentations from different images are maximized. So, adding augmentations in pretraining is the default choice. Yet, we think it is also helpful to include these augmentations in the supervised finetuning stage to make the model more robust against changes in dataset, which is not considered in SSL person re-id studies, as far as we know. We use MGN for our baseline model in supervised finetuning stage. In the default configuration of this model only random horizontal flipping is used to augment the data. So, we enhance this with our proposed augmentations mentioned above.

4 Experiments

In order to create a baseline for our data augmentation research, we first train the ResNet50 model with the Market1501 dataset in an unsupervised manner. The pretraining process is performed for 100 epochs with 64 mini-batch size on Nvidia RTX2060 GPU. Then, pretrained backbone ResNet50 model with MGN supervised architecture is finetuned using 10%, 30%, 50%, and 70% of the Market1501 dataset, separately. The samples selected at the identity level. For instance, while using the 10% of the dataset, we take all images from randomly selected 75 identities out of 750. The achieved results are shown in Table 3.

Pretrain	10%	30%	50%	70%
IN Sup.	53.1/76.9	75.2/90.8	81.5/93.5	84.8/94.5
IN unsup.	58.4/81.7	76.6/91.9	82.0/94.1	85.4/94.5
LUP unsup.	64.6/85.5	81.9/93.7	85.8/94.9	88.8/95.9
Market Unsup. (Ours)	31.4/55.0	57.4/80.1	67.2/86.1	73.1/89.7

Table 3: The mAP and CMC-1 scores of the model which is finetuned with different portions of Market-1501.

Setting	Def_SSL+ft	CJ_SSL+ft	AF_SSL+ft	PT+Def_ft	PT+CJ_ft	PT+AF_ft
mAP	31.4	32.0	31.1	48.6	48.2	46.4
$cmc1$	55.0	56.1	55.0	74.2	73.5	72.4

Table 4: The table showing the results of the effects of the augmentation methods used in the pretraining and the finetuning stages. Def: Default, SSL: Self-Supervised Learning, ft: Finetuning, CJ: Color Jitter, AF: Affine Transforms, PT: Pretrained.

It seems that our model pretrained on Market1501 is far behind the others. There are some possible reasons to mention. Firstly, in the pretraining task, the scale of the dataset is crucial and the number of images Market1501 contains less than 1% of those LUPerson has. Although the gap between the numbers is huge, we pretrained the model on Market1501 for the same number of epochs as used in LUPerson. Secondly, due to the scarce resource, we take mini-batch size as 64, which is defined as 2560 in the related paper. They benefit from the distributed training opportunities with multiple GPUs. Even though small batch size causes a longer training process, we keep the learning rate and the number of epochs the same. Because of these decisions, our training process did not converge, and it ended earlier than needed.

As a next step, we tried to research the effects of different augmentation methods on pretraining and finetuning stages. Basically, color-based augmentations and affine transforms are implemented. In color augmentation, we only change contrast and brightness because the re-id models are highly dependent to color information. In this direction, we build 6 different training cases, 3 of them are for pretraining and the other 3 of them are processed for finetuning. To investigate the impact of the implemented augmentations, we firstly create a baseline by training the models with default augmentation methods. Then, an ablation study is conducted by adding the augmentations to trainings one by one. The results are shown in Table 4. The code is available at github.com/nezihkasim/CMPE597_DL_LUPerson

5 Conclusion

Firstly, the model pretrained on Market1501 is far behind the model pretrained on LUPerson . The first reason explaining the gap is that the scale of the dataset is crucial in the pretraining task. The number of images Market1501 contains less than 1% of those LUPerson has. Secondly, due to the scarce resources we have, we had to take mini-batch size as 64, which is chosen as 2560 in the related paper. Additionally, they benefit from the distributed training opportunities with multiple GPUs. After observing the performance degradation due to insufficient resources and small dataset, we focused on improving our baseline performance using enhanced augmentations. Applying specifically designed color distortions like contrast, brightness on pretraining obtains 1% improvements in terms of cmc-1 score. By considering the fact that our baseline study obtained only 1.5% performance increase in CUHK03 dataset using their optimal choice of augmentations, we can conclude that we have achieved noteworthy results. We have used only 10% of the dataset in fune tuning due to high training durations, which means the improvement would be higher if large portion of the dataset could have beed used. Adding affine transformations has not changed the performance in contrast to what we have expected. We couldn't conducted sufficient amount of experiments to find the optimal parameter settings of the affine augmentations. Hence, we cannot say affine transformation does not help getting improvements in person re-id problem and we think it is a good point to study for future works. On the other hand, we tried applying augmentations in supervised fine tuning stage but it didnt improved performance, rather it degraded it by 1% to 2%. Again, adjusting augmentation parameters may help to compensate the deterioration but it requires many experiments to perform.

References

[1] Wang, G., Yuan, Y., Chen, X., Li, J., & Zhou, X. (2018, October). Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM international conference on Multimedia (pp. 274-282).

[2] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[3] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

[4] Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271-21284.

[5] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9650-9660).

[6] Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., ... & Chen, D. (2021). Unsupervised pre-training for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14750-14759).

[7] Luo, H., Wang, P., Xu, Y., Ding, F., Zhou, Y., Wang, F., ... & Jin, R. (2021). Self-Supervised Pre-Training for Transformer-Based Person Re-Identification. arXiv preprint arXiv:2111.12084.

[8] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[9] Zhu, K., Guo, H., Yan, T., Zhu, Y., Wang, J., & Tang, M. (2022). Part-Aware Self-Supervised Pre-Training for Person Re-Identification. arXiv preprint arXiv:2203.03931.

[10] Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

[11] Zheng, Liang, et al. "Scalable person re-identification: A benchmark." Proceedings of the IEEE international conference on computer vision. 2015..

[12] Peng, X., Wang, K., Zhu, Z., & You, Y. (2022). Crafting Better Contrastive Views for Siamese Representation Learning. arXiv preprint arXiv:2202.03278.

[13] Papers with code - market-1501 benchmark (person re-identification). The latest in Machine Learning. (n.d.). Retrieved June 6, 2022, from https://paperswithcode.com/sota/person-re-identification-on-market-1501