Intelligent Systems with Applications 16 (2022) 200137

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

# Deep Residual Network and Transfer Learning-based Person Re-Identification

Arpita Gupta [a,*], Pratik Pawade [b], Ramadoss Balakrishnan [b]

[a] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India
[b] Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India-620015

ARTICLE INFO

*Keywords:*
Transfer learning
Person Re-Identification
Deep residual network
Deep learning

ABSTRACT

Person re-identification (person re-id) is a field in which a person is recognized from different views, which plays a significant role in surveillance. One of the major problems in-person re-id is the unavailability of large labeled datasets, which has bounded the performance of the deep learning models. In this study, different variations were experimented with of deep residual networks trained with transfer learning, and then fine-tuning was done on the pre-trained model. These models are pre-trained on a larger dataset, ImageNet, to learn the visual features better. The proposed model is based on deep residual networks because of its better understanding and performance of the visual features, which leads to better classification. The proposed model was evaluated on the Market-1501 dataset, which consists of 1501 identities collected from 6 cameras. In this paper, the effect of hyperparameters is studied for better accuracy. The model has achieved the highest mAP score of 68.4%, with an improvement of 5.3%, and 96.0% of Rank-1 with 12.4%. The proposed model has outperformed all the existing models trained in a supervised manner on the Market 1501 dataset. The results have proved that this model could help better person re-id classification problems even when there is no availability of large labeled datasets and could be used for better security and surveillance.

## Introduction

The recent technological advancement has led to the camera network in all public places like airports, schools, colleges, and other places. This allowed us to cover a broad area span and have a non-overlapping field view for better coverage (Peng, Xiang, Wang, Pontil, Gong, Huang, & Tian, 2016). Collecting data from these networks has led to a massive amount of data that could be analyzed for safety or other purposes. The fields like deep learning, pattern recognition, and computer vision play a vital role in analyzing this data to get an outfit that could help humans in daily life, but this collected data is not properly labeled, which leads us to transfer learning (Grigorev, Tian, Rho, Xiong, Liu & Jiang, 2019). Analyzing this data could help find crime patterns, search for stolen vehicles, person re-identification, and other fields. (Peng, Xiang, Wang, Pontil, Gong, Huang, & Tian, 2016; Bai, Yang, Huang, Dou, Yu, & Xu, 2020)

The motivation behind person identification is to match a person in the probe example with the gallery data collected from the cameras (Zhao, Ouyang, & Wang, 2013). The main challenge is the changes across camera views, lighting, the pose of human angle, obstruction, resolution, and the background (Zhao, Ouyang & Wang, 2013). The identification is made by matching the probe image with the gallery dataset to find the person's location or track it for other applications (Zheng, Shen, Tian, Wang, Wang & Tian, 2015). Two methods exist: first, open-world matching in this the probe image may or may not exist in the gallery dataset, so the new class will be created for future classification of identification (Bedagkar-Gala, & Shah, 2014; Zheng, Yang, & Hauptmann, 2016; it is not always possible that a person will be a known person. The open-world setting is more realistic and useful but hard to implement. Open-world identification has been explored (Bedagkar-Gala & Shah, 2014); they tell the usefulness of the open world, its applications, and work done. There are multiple models which concentrate either on feature representation; in one of the works Harmonious Attention CNN (HA-CNN) model for simultaneous feature representation optimization and combined learning of soft pixel attention and hard regional attention is used to improve person recognition in uncontrolled (misaligned) pictures (Li, Zhu, & Gong, 2018), in another study local feature learning, which conducts an alignment/matching by figuring out the shortest path between two sets of local features, enhance global feature learning substantially and does so without the need for

* Corresponding author: Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India. Tel.: +91-9407885161.
*E-mail addresses:* arpitagupta2993@gmail.com (A. Gupta), ppratikpawade7@gmail.com (P. Pawade), brama@nitt.edu (R. Balakrishnan).

https://doi.org/10.1016/j.iswa.2022.200137
Received 30 April 2022; Received in revised form 27 September 2022; Accepted 5 October 2022
Available online 12 October 2022

additional supervision. In this study, only the global feature to calculate picture similarities after joint learning was used(Zhang, Luo, Fan, Xiang, Sun, Xiao, Jiang, Zhang, & Sun, 2017), and in one study, it was the first time that a CNN architecture takes into account human body shape information to speed up feature learning (Zhao, Tian, Sun, Shao, Yan, Yi, Wang, & Tang, 2017). Then there are studies working on the distance metric; one study demonstrates employing a variation of the triplet loss to achieve end-to-end deep metric learning beats most other reported approaches by a significant margin for models learned from scratch and ones that have been pretrained. (Hermans, Beyer & Leibe, 2017), In another study, Margin Sample Mining Loss (MSML), a novel metric learning loss with hard sample mining, may achieve superior accuracy compared to existing metric learning losses, such as triplet loss (Xiao, Luo & Zhang, 2017); on some frequently used datasets, better re-ID results than the state-of-the-art ones are obtained by integrating SN loss on top of Resnet50 in one study (Li, Ding, Li, Zhang & Fu, 2018); based on the quadruplet loss for the individual ReID, a quadruplet deep network with a margin-based online hard negative mining is suggested (Chen, Chen, Zhang & Huang, 2017); in contrast to conventional metric learning losses, such as triplet loss, present a novel metric learning loss with hard sample mining termed margin sample mining loss (MSML) (Xiao, Li, Wang, Lin & Wang, 2017) or both. Recent advancements in deep learning have led to deep re-id, which has shown significant improvement in the field of vision (Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018); in one study, a realistic unsupervised learning environment is taken into account, and it is demonstrated that our model can generalize to a fresh re-id dataset without any further adjustment. (Qian, Fu, Xiang, Wang, Qiu, Wu, Jiang, & Xue, 2018); in another study, a network of many sources for cross-domain person re-identification (Wei, Zhang, Gao, & Tian, 2018).

Many problems are having a broad set of labeled datasets with multiple classes, while in-person re-id lacks large labeled datasets, and collecting the dataset is a difficult task (Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018). The dataset, the largest person re-id, is just adequate in size. So, to overcome this disadvantage of the small training set, transfer learning is used in which the model is trained on a larger dataset and then tested on the person re-id datasets. The training is done on a significantly larger dataset, and then the model is fine-tuned and tested, which helps overcome the gap created by the non-existence of a large dataset in person re-id. One of the significantly large datasets is ImageNet, which is very popular and successfully used in transfer learning. ImageNet (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2010) is a dataset containing millions of images of object categories of thousands and has been proven to be competent in pre-training the model for a specific task. There are some problems in using ImageNet for pre-training the model for person re-id. ImageNet object categorization is very different from the person re-id, and the person re-id dataset has a lower resolution than the ImageNet images as it is collected from a network of CCTV cameras. However, the model could be designed to transfer the knowledge from the ImageNet and be used in person re-id.

The following paper has been organized: Section 2 explains the related work for a person re-id based on deep networks, transfer learning, and other techniques. Section 3 describes the proposed model based on transfer learning and fine-tuning. The proposed model section explains the working of the model and its architecture. Section 4 describes the performance evaluation, results achieved by the proposed model, and the details of the datasets used, followed by section 5 conclusion.

## Related Work

Existing methods in person re-id have used distance metric learning (Hirzer, Beleznai, Roth, & Bischof, 2011; instead of manually creating a single feature to address the issue, they build a feature space using our knowledge of the situation and then allow a machine learning algorithm choose which representation works best. (Gray& Tao, 2008); another study developed utilizing a single code package that contains 22 metric learning and ranking approaches and 11 feature extraction algorithms. (Karanam, Gou, Wu, Rates-Borras, Camps, & Radke, 2019), discriminative subspace learning (Liao, Hu, Zhu, & Li, 2015), learning to rank (Paisitkriangkrai, Shen, & Van Den Hengel, 2015), or deep learning in which the similarity metrics and learning features is done simultaneously (Ahmed, Jones & Marks, 2015). Person Re-id is taking images from the probe set and comparing the distance or making feature comparison with the gallery image and finding the nearest or ranking the commonality of the image with the other class images. In this study, better feature understanding is concentrated on, which is the key to transfer learning. Training different models, ResNet-18, ResNet-34, ResNet-50 (He, Zhang, Ren & Sun, 2016), and ResNeXt-50 32 × 4d (Hu, Shen, & Sun, 2018) on a larger dataset to overcome the problem of a smaller dataset in person re-id field and in this study used SoftMax loss in our architecture. The study by (Zheng, Yang & Zheng et al., 2017) has explored the combination of two models for better discriminative pedestrian descriptors using the Siamese network.

### *Transfer Learning*

Transfer learning technique in deep learning is getting much attention in all the fields where there is no availability of large labeled datasets (Peng, Xiang, Wang, Pontil, Gong, Huang, & Tian, 2016; Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018). A most common design used in transfer learning is fine-tuning in which the network is firstly pre-trained on a larger dataset, then from the pre-trained model, some of the layers are taken in regards to the target network, and then either all or some of the layers are fine-tuned to get the desired network. A study on transfer learning proves that an increase in source and target discrepancy can lead to less generalization, which is one of the obstacles in transfer learning (Peng, Xiang, Wang, Pontil, Gong, Huang, & Tian, 2016). There is no use in increasing the depth of networks, while changes in transformation in layers could lead to better performance. One of the studies has explored the dynamic nature of cameras (Ahmed, Lejbolle, Panda, & Roy-Chowdhury, 2020). Another study proposed Classification and Latent Commonality for unsupervised person re-id (Tian, Teng, Zhang, Wang, & Fan, 2021), a totally different approach from our work.

Several other studies show that a multi-task joint approach for training is designed for a person re-id whose aim is to decrease the inconsistency between the source and target datasets, which helps in reducing the boundary difference of the area (Peng, Xiang, Wang, Pontil, Gong, Huang, & Tian, 2016; Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018). If the source dataset is similar to the area of the target dataset, the performance is better as there is lesser inconsistency in the area boundary, while it is a bit unsuitable to use a different area dataset for training and testing, which leads us to use it fine-tuning for better results. Studies have shown that combining joint learning with multi-task learning and fine-tuning is incompatible with different area dataset training and testing tasks (Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018).

### *Loss function*

In order to determine how effectively a neural network represents the training data, a loss function is used to compare the target and anticipated output values. This loss between the expected and goal outputs is something we want to reduce throughout training. In one study, in order for convolutional neural networks (CNNs) to learn angularly discriminative features, they suggest the angular softmax (A-Softmax) loss. A-Softmax loss may be seen geometrically as imposing discriminative constraints on a hypersphere manifold, which coincides with the assumption that faces also lie on a manifold (Liu, Wen, Yu, Li, Raj, & Song, 2017). One demonstrates that employing a variation of the triplet loss to achieve end-to-end deep metric learning beats most other

reported approaches by a significant margin for models learned from scratch as well as ones that have been pretrained(Hermans, Beyer, & Leibe, 2017). Another study's suggested strategy for learning the embedding with the structured loss goal on the lifted issue works noticeably better than existing approaches on all tried embedding dimensions with the Google Neural Network (Song, Xiang, Jegelka, & Savarese, 2016). According to one study, the instance loss provides a superior weight initialization for the ranking loss, allowing for the learning of more discriminative embeddings (Zheng, Zheng, Garrett, Yang, Xu, & Shen, 2020). In another study, two fundamental deep feature learning methods, namely learning with class-level labels and pair-wise labels, both have a unified formula for the Circle loss. Comparing the Circle loss to the loss functions optimising technique, we demonstrate analytically that the Circle loss offers a more flexible optimization strategy towards a more specific convergence objective (Sun, Cheng, Zhang, Zhang, Zheng, Wang, & Wei 2020). One of the key parts of neural networks is the loss function. Loss is nothing more than a neural network's prediction mistake. Loss Function is the name of the procedure used to compute the loss. To put it simply, the gradients are calculated using the Loss.

*Supervised Domain Adaption Learning*

Supervised domain adaption is the transfer learning technique in which all the training examples are labeled. Domain adaption is the discipline in transfer learning in which a model learns from different domain datasets and performs the same task on the various domain of the dataset. The adaption is done from one domain dataset and applied to a different dataset. Our model uses this task of learning from one and laying on another supervised.

This study compared deep residual networks based on transfer learning, pre-trained on a larger dataset and ImageNet for person re-id. Our proposed models have used SoftMax loss function and fine-tuning. This study has proposed a model based on supervised domain adaption using deep residual networks and compared it with different variations possible. Also, evaluated the network on the benchmark Market1501 (Zheng, Shen, Tian, Wang, Wang & Tian, 2015) dataset.

## Proposed Model

*Overview*

The network architecture is primarily composed of the deep residual network (ResNet), and different types of ResNet are compared to find which network will fulfill the requirements of a person's re-id. This network aims to distinguish the images of visually different classes. These networks take the labeled domain datasets as input for training the network and then use this other domain knowledge on the benchmark dataset to do the needful. This network architecture uses SoftMax loss and fine-tuning.

Our base network is composed of ResNet, and its different variations are compared, which learn the features or representation from the input labeled training set. In this study, the ResNet-18, ResNet-34, ResNet-50, and ResNeXt-50 32 × 4D are experimented on. ResNet is widely used in person re-id models, and it has been selected as this network has shown great results in other computer vision tasks. Employing ResNet has helped, as it could easily generalize the features of the images from our large training dataset, ImageNet, and has shown promising results. Also, two-stepped tuning is used, as suggested by (Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018), which changes the SoftMax layer totally of the pre-trained model with random initialization and then freezing layer by layer to understand which layer should not be changed since the one step fine-tuning was not helpful in our model. This fine-tuning technique has helped us to achieve our results.

*Transfer Learning*

In this study, using a transfer learning technique for training deep models is employed. This study aims to find the model which could learn the features from the large training dataset and evaluate the results on the person re-id task. In this study, using the pre-trained layers of the networks (ResNet-18, ResNet-34, ResNet-50 (Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018), and ResNeXt-50 32 × 4D (Hu, Shen, & Sun, 2018)), which extract the feature from the large training dataset and fine-tuning to get the desired network. Also, compared four variants of ResNet which are ResNet-18, ResNet-34, ResNet-50, and ResNeXt-50 32 × 4D. This methodology has proven effective as it gives a better result than any other supervised model trained on a different dataset.

*ResNet*

ResNet also is known as a deep residual network because of its residual connections, which help better information transfer in the network. Residual connections, similar to skip connections, provide a way to propagate the gradients directly through the network (Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018). These skip connections help take the global feature into account and skip the layers, which helps in training by reducing time and optimal tuning of the network layers. There is a backpropagation in ResNet, which helps remove the vanishing gradient problem. ResNet was proposed in 2015 and won first prize in the image classification task in ILSVRC (Image Large Scale Visual Recognition Challenge) 2015 on ImageNet (Thenmozhi, & Srinivasulu Reddy,2019).

In ResNet, the degradation of the gradient problem is addressed in deeper models. ResNet introduces the skip connection in the layers by residual mapping. Let x denote the inputs to the first of these layers and h(x) underlying mapping to be fit by a few stacked layers. As shown in fig. 1, the mapping is modified by f(x)+x from f(x)=h(x)-x, which hypothesizes that instead of optimizing the original mapping, it is easier to optimize the residual mapping. ResNet has proved that it can solve the degradation problem by simply driving the weights of the multiple nonlinear layers toward zero to approach identity mappings (Xiao, Li, Wang, Lin & Wang, 2017).

*ResNet-18*

ResNet-18, as shown in fig. 2, consists of 18 layers in which the architecture could be represented composition of 4 residual blocks. In each of them, there are 3 × 3 filters, and downsampling is done in further layers. ResNet-18 gives accurate results, but it just converges faster and has a higher error rate. ResNet-18 is better than an 18-layer network and performs well in a person re-id task. The SoftMax layer in the network is replaced after training in ImageNet, as it should not overlap. Here two-step fine-tuning in this network is followed.

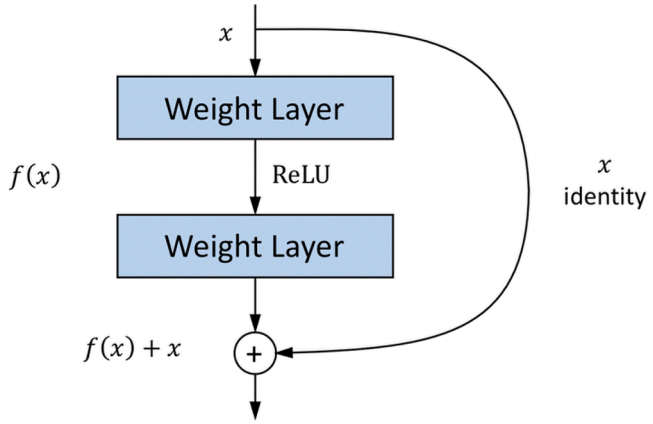


**Fig. 1.** Residual Connection.

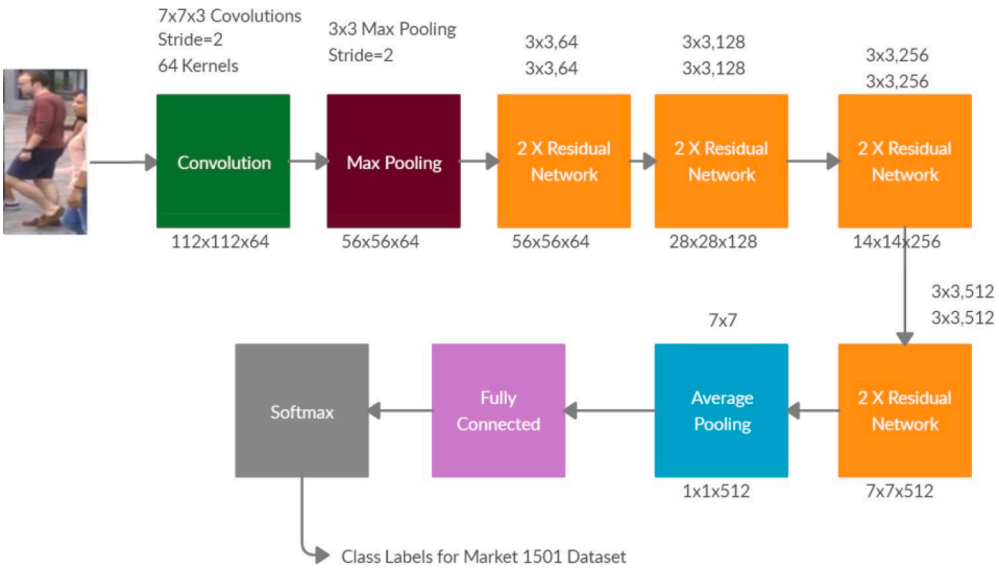


**Fig. 2.** ResNet-18 using Transfer learning architecture with fine-tuning.

*ResNet 34*

ResNet-34 contains 34 layers in which there are four residual blocks, convolution layer, which extracts features; max-pooling layer objective is to down-sample an input representation, average pooling layer, fully connected, and SoftMax layer as shown in fig. 3. This network also consists of 3 × 3 filters grouped in 3, 4, 6, and 3 layers. ResNet-34 performs better than ResNet-18, showing lower training error and better generalizing ability. ResNet- 18 and ResNet- 34 perform mostly identically in our transfer learning study. Here two-step fine-tuning in this network is followed.

*ResNet 50*

As the name suggests, this network consists of four residual blocks and has 50 layers, as shown in fig. 4. This network has replaced the two filters block as in ResNet-34 with three-layer blocks. The filters used in this network follow the bottleneck technique: firstly, there is a 1 × 1 filter followed by 3 × 3 followed by a 1 × 1 filter, where the 1 × 1 layers are responsible for reducing and restoring, leaving the 3 × 3 layer a bottleneck with smaller input/output dimensions. Here two-step fine-tuning in this network is followed.

*ResNeXt-50 32 × 4d*

This network is a modified version of ResNet- 50 inspired by VGG networks, as shown in fig. 5. In this network, the stack of the residual blocks is formed, which follows two rules:- a) the block shares the same hyperparameter if they produce spatial maps of the same size, and b) the width of the blocks is multiplied by a factor of 2 when the spatial map is downsampled by a factor of 2. In this network, a new dimension was introduced, cardinality.

$$y = x + \sum_{i=1}^{C} T_i(x) \tag{1}$$

In equation 1 above, C is the cardinality, the size of the set of transformations to be aggregated (Hu, Shen, & Sun, 2018). The above equation represents the aggregated transformations, where c controls the modifications. Experiments have shown that cardinality is more effective than traditional depth and width. In the above equation, the aggregation of transforms acts as the residual connections; the whole y is the output. Here two-step fine-tuning in this network is followed.

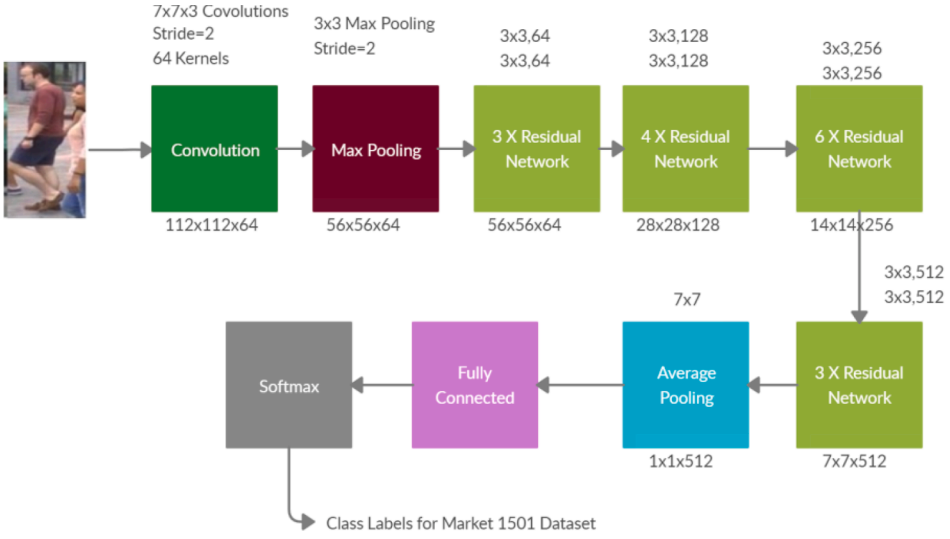


**Fig. 3.** ResNet-34 using Transfer learning architecture with fine-tuning.

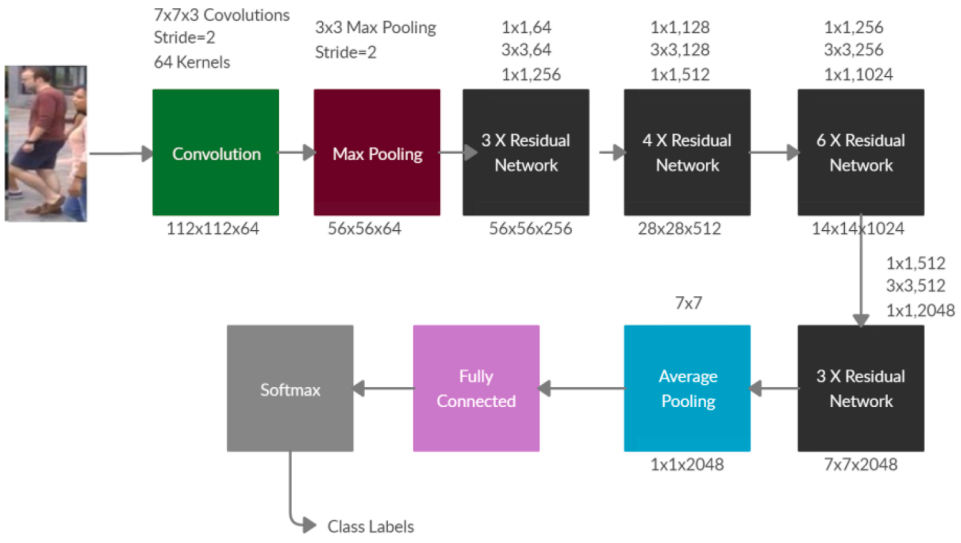


**Fig. 4.** ResNet-50 using Transfer learning architecture with fine-tuning.

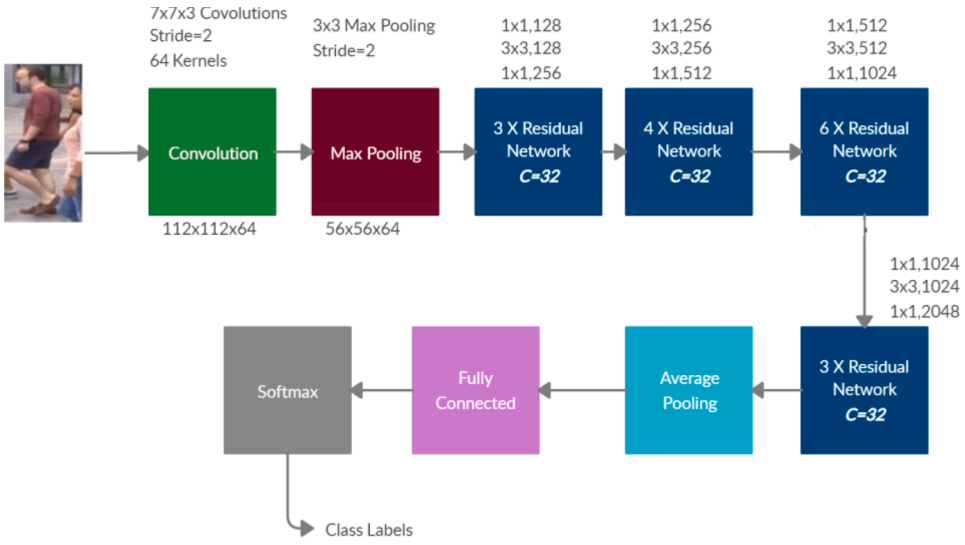


**Fig. 5.** ResNeXt-50 using Transfer learning architecture with fine-tuning.

*Loss Function*

SoftMax loss is widely used in computer vision in every subfield. SoftMax function helps better learn the model and can easily find any outliers in the data. Equation 2 explains the working of the SoftMax loss.

$$\mathscr{L}_{CE} = -\frac{1}{N}\sum log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{2}$$

In equation 2, $y_i$ is the class label of the training sample at i; $x_i$ is the training sample at I, N represents the number of labels, W is the weight matrix, and bias is shown by b.

## Results and Discussion

In this study, experimentation on the Market-1501 dataset for person re-id has been conducted. In this study, a transfer learning-based ResNet model pre-trained on ImageNet and later fine-tuned is proposed, outperforming the supervised domain adaption models. The following subsections describe the details of the datasets and the results achieved by the proposed methods.

*Datasets and Settings*

For experimenting, ImageNet is used (Deng, Dong, Socher, Li, Li, & Fei-Fei,2010) and Market 1501 (Zheng, Shen, Tian, Wang, Wang & Tian, 2015) datasets. Pre-training the models on ImageNet. ImageNet dataset consists of 1000 images that illustrate the synonym set as it is in the WordNet hierarchy. The images of ImageNet are used for pre-training because this is a large dataset; as explained in the introduction section, there is a need for a larger training dataset for training which helps in better training. Pretraining on ImageNet has solved the problem of the non-existence of high-quality images with annotations for training the model.

The Market-1501 dataset is used (sample shown in fig. 6) for further training and testing in our experimentation. This dataset is collected from 6 cameras, one low-resolution and six high-resolution cameras, placed in the supermarket of Tsinghua University. There is an overlap in field-of-view in different cameras. The dataset contains 32,668 annotated boxes of 1,501 identities. Table 1 provides more details about the datasets.

The supervised transfer learning method is used for 751 identities for training and gallery training settings consisting of 12936 and 15913

**Fig. 6.** Market-1501 Sample

**Table. 1**
Details of Market 1501 dataset

| Dataset | Market-1501 |
|---|---|
| **#identities** | 1,501 |
| **#BBoxes** | 32,668 |
| **#distractors** | 2,793 |
| **#cam. Per ID** | 6 |
| **DPM/Hand** | DPM |
| **Evaluation** | mAP |

images, respectively while using 750 identities for the query, which includes 3368 images, all collected from 6 cameras. Also, using the same train, gallery, and query settings for all the models.

*Evaluation Metrics*

Rank-1, 5, 10, and 20 are used to evaluate the performance of the person re-id methods. Rank-1 is to check the percentage of how many times the predicted label is the same as the ground label. Rank-5 and further are calculated if there are many numbers of classes. Like Rank-5 determines if the ground truth label is in the top 5 most probable ones or not, and so on for 10 and 20. Mean average precision (mAP) is evaluated as suggested by (Zheng, Shen, Tian, Wang, Wang & Tian, 2015) to evaluate Market-1501.

*Results*

On the large dataset person re-id dataset, namely Market 1501, two-step fine-tuning is used on the variations of ResNet. The results of our model are shown in table 2, and the comparison with state-of-the-art supervised trained models on person re-id in table 3. TransLearn in the table represents transfer learning, and FT represents fine-tuning. As shown in table 1, the highest mAP score and rank-1 are achieved by ResNeXt-50_32 × 4d with transfer learning, and fine-tuning has outperformed all the existing models. Drawing the following conclusions from the model trained with transfer learning and fine-tuning: (1) Both our models, ResNet-18 + TransLearn + FT and ResNet-34 + TransLearn + FT, perform quite similarly (2) As you become more lenient on the model's performance evaluation all give an approximately same result which is not practical (3) ResNeXt-50_32 × 4d + TransLearn + FT outperform all the models with the mAP score of 68.40% and Rank-1 of 96%.

**Table 3**
Performance evaluation comparison with existing models on Market-1501 dataset supervised

| **Model** | **Rank-1** | **mAP** |
|---|---|---|
| XQDA(Liao, Hu, Zhu, & Li, 2015) | 54.1 | 28.4 |
| SCSP (Liao, Hu, Zhu, & Li, 2015) | - | - |
| DNS(Zhang, Xiang, & Gong, 2016) | 71.5 | 46.0 |
| Siamese LSTM(Varior, Shuai, Lu, Xu, & Wang, 2016) | 61.6 | 35.3 |
| Gated S-CNN(Varior, Haloi, & Wang, 2016) | 76.0 | 48.4 |
| CAN (Liu, Feng, Qi, Jiang & Yan, 2017) | - | - |
| GAN(Zheng et al., 2017) | 83.97 | 66.07 |
| Re-rank (Zhong, Zeng, Cao & Li, 2017) | 77.11 | 63.63 |
| CNN+ SID(Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018) | 83.6 | 62.2 |
| CNN+ PV(Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018) | 81.5 | 63.1 |
| CNN+ TL(Chen, Wang, Shi, Yan, Geng, Tian & Xiang, 2018) | 72.4 | 49.7 |
| STNET(Wang, Hu & Zhang,2020) | 73.5 | 48.2 |
| STNET(Wang, Hu & Zhang,2020) | 74.6 | 49.7 |
| MSTNet-pixel(Wang, Hu & Zhang, 2020) | 75.8 | 50.4 |
| MSTNet-task(Wang, Hu & Zhang, 2020) | 76.2 | 51.1 |
| MSTNet(Wang, Hu & Zhang,2020) | 80.9 | 55.2 |
| OGNet(Zheng, Zheng & Yang, 2020) | 87.71 | 69.52 |
| Ours | **96.0** | **68.4** |

The following observations could be made from table 3:- (1) Our model significantly outperformed the state-of-the-art Market 1501 dataset with a gap of 12.4% in rank-1 and 5.3% in mAP, which has proved to be the best performing model(ResNeXt-50_32 × 4d) in our four models. (2) All the deep learning-based are performing better than the non-deep learning model because of the advantage of hand-crafted feature-based models.

Our model has outperformed the existing one because of the feature representation learning from a larger dataset ImageNet and fine-tuning. Our model has outperformed a deep network-based model (Chen, Wang,

**Table 2**
Performance evaluation on the Market-1501 dataset

| **Parameters** | **ResNet-18+TransLearn +FT** | **ResNet-34 + TransLearn +FT** | **ResNet-50 + TransLearn +FT** | **ResNeXt-50_32 × 4d +TransLearn+FT** |
|---|---|---|---|---|
| **mAP Score** | 65.60 | 66.20 | 67.70 | **68.40** |
| **Rank-1** | 84.30 | 83.60 | 84.50 | **96** |
| **Rank-5** | 93.30 | 93.10 | **94** | 93.50 |
| **Rank-10** | 95.60 | 95.70 | **96** | 95.50 |
| **Rank-20** | **97.40** | **97.40** | 97 | 97.10 |

Shi, Yan, Geng, Tian & Xiang, 2018) and (Wang, Hu & Zhang, 2020) mainly because of better feature representation from ImageNet. In this study, all the models with a single loss only are considered.

Below fig. 7 shows the classification accuracy of different variations of ResNet with depth when trained with transfer learning and fine-tuned. Fig. 8 indicates that the deepest variant of ResNet with aggregated transformation creates a better network connection and feature learning connection.

Fig. 7 shows the comparison of results in graphical form with regards to mAP and Rank 1. The comparison is with the existing and related studies.

From fig. 8, the overall performance of the models could be concluded. This figure shows that after rank-10, all models perform similarly. It indicates that the ResNeXt-50_32 × 4d performs well in most cases.

The graph in fig. 9 shows the variation in the rank metric in all the models from the graph; the rank-1 could be seen, and all the graphs give approximately the same result, which proves that going deeper with the models does not lead to better results always.

*Effects of epochs*

The number of epochs was decided based on the best accuracy achieved in the number of which epochs. All the models were trained on 40 epochs. Starting to train our model with 10 epochs then increasing the epochs to 40 where achieving the highest mAP score of 68.40% in ResNeXt-50_32 × 4d + TransLearn + FT model, 67.70% in ResNet-50 + TransLearn + FT model, 66.20% in ResNet-34 + TransLearn + FT model and 65.60% in ResNet-18 + TransLearn + FT model.

From Figs. 10 and 11, we can understand that mAP score variation with epochs shows that ResNet-18, ResNet-34, and ResNet-50 start with low performance and later achieve a better mAP score. In Fig. 12, variation of rank, the score is shown, which shows that the accuracy is achieved highest at the epochs of 40, where the rank of all the models except ResNeXt-

50_32 × 4d varies from lower to higher, while there is no significant variation in the other model.

Fig. 12 and 13 show the variation of epochs in different models trained on transfer learning and fine-tuning. The variation indicates that all the models after certain epochs are performing similarly. The highest accuracy achieved is 96.0 for the Market 1501 dataset at 40 epochs, compared with other existing models, and our models have outperformed all the existing models.

Visualizing the results to understand how transfer learning and fine-tuning contribute and how the model varies with training. Overall classification performance of the proposed network is the highest accuracy achieved by any network in supervised learning, 96.0% with Rank-1score and a gap of 12.4% from the existing one, while the mAP score of 68.4% with a difference of 5.3%.

The best performance metric achieved is by ResNeXt-50_32 × 4d with transfer learning and fine-tuning. This study experimented with different models of varying layers from 18 to 50, showing that the deeper model does not always perform better than the shallower model. The model performs better than any existing model because of feature representation learning from a larger dataset and fine-tuning. This detailed learning of features from ImageNet and then Market 1501 must lead to better performance of the models, proving that the network is better aware of the visual features. This better awareness of the feature is the key to the performance of the proposed model.

*Contributions*

*This study proposed deep residual network training via transfer learning in a variety of configurations, testing, and pre-training the model, also fine-tuning. These models were previously trained on ImageNet, a bigger dataset, to understand the visual characteristics more effectively.*

- Without having access to large source data, we present a reliable and effective multiple metric hypothesis transfer learning approaches to effectively adapt a newly introduced camera to an existing person re-id framework.
- To demonstrate the superiority of our suggested strategy over current options, we conduct rigorous tests on a number of benchmark datasets.
- We have considered benchmark datasets leading to results attained with no geographic boundaries.

## Conclusion

This study proposed a transfer learning-based model which has outperformed the existing models and has proved to be helping in solving the problem of availability of larger labeled datasets. The models are fine-tuned to achieve better accuracy. Our study has shown that if models are combined with transfer learning and fine-tuning technique, they will perform better in the classification problem of person re-

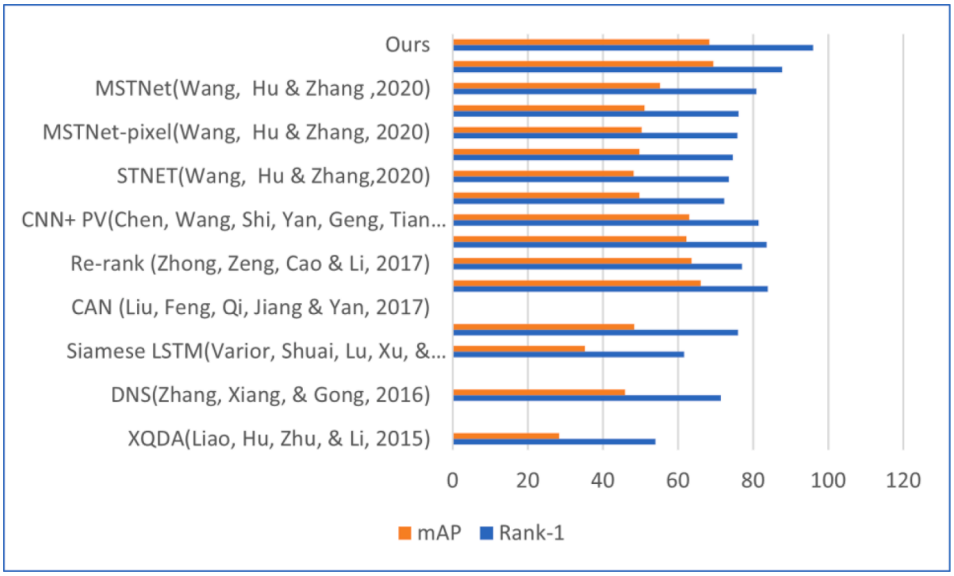


**Fig. 7.** Classification results comparison.

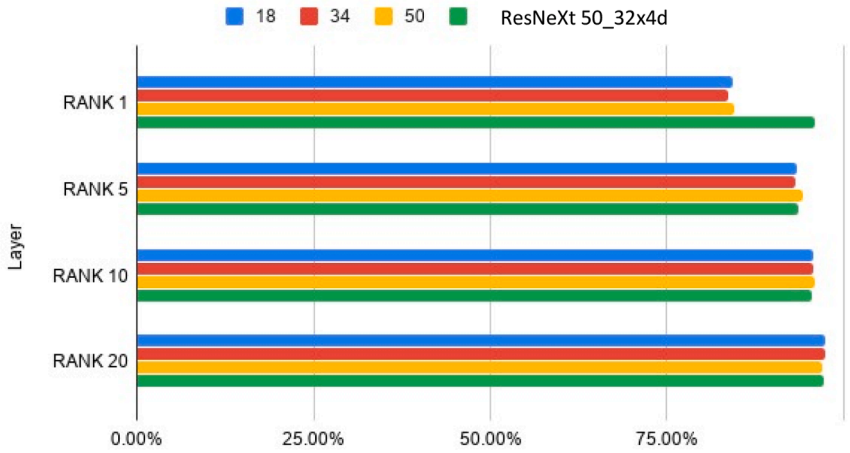


**Fig. 8.** Classification accuracy in different proposed with transfer learning and fine-tuning.

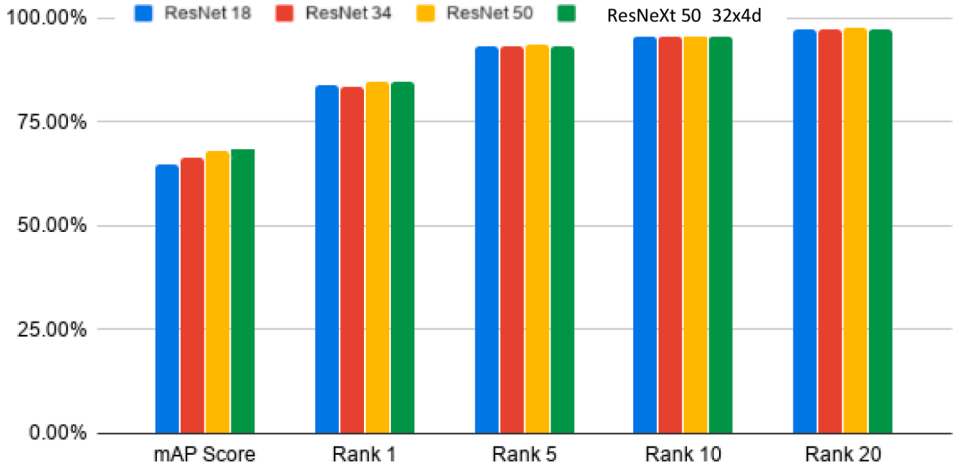


**Fig. 9.** Overall performance of different proposed models with transfer learning and fine-tuning.

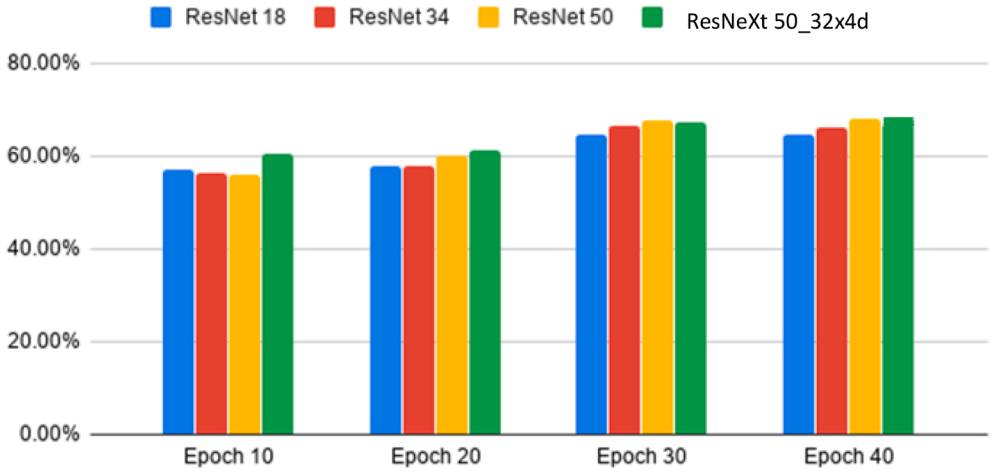


**Fig. 10.** Variation in mAP score with epochs.

identification. The experimental results have also shown the effect of epochs on performance. This study has performed experiments on variants of ResNet from 18 layers to 50 because going deeper resulted in more time complexity and not better results, which could not be achieved with better training of these shallower models. Also, compared the models with existing ones based on deep networks and other techniques. The results show that our proposed model of ResNeXt+TransLearn+Ft has achieved performance with the highest mAP score of 68.4%, with an improvement of 5..3%, and 96.0% of Rank-1 with an improvement of 12.4%, which is the suggested evaluation metric for Market 1501 dataset in supervised settings. In the future, our work will be focused more on real-time application and making it unsupervised. This model

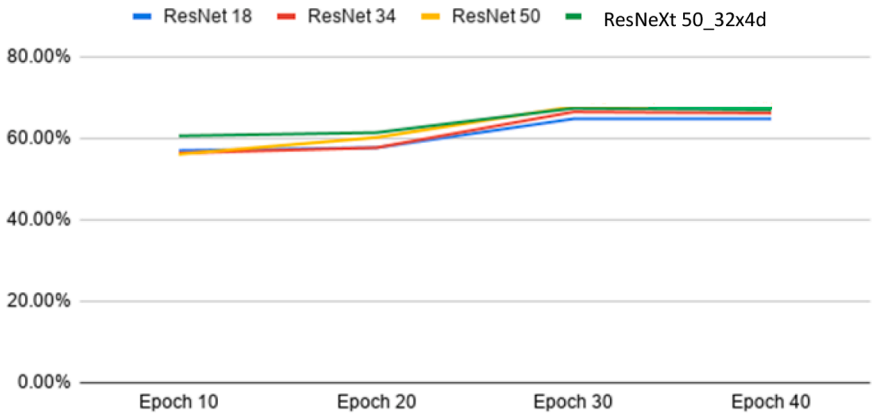


**Fig. 11.** Variation in mAP score.

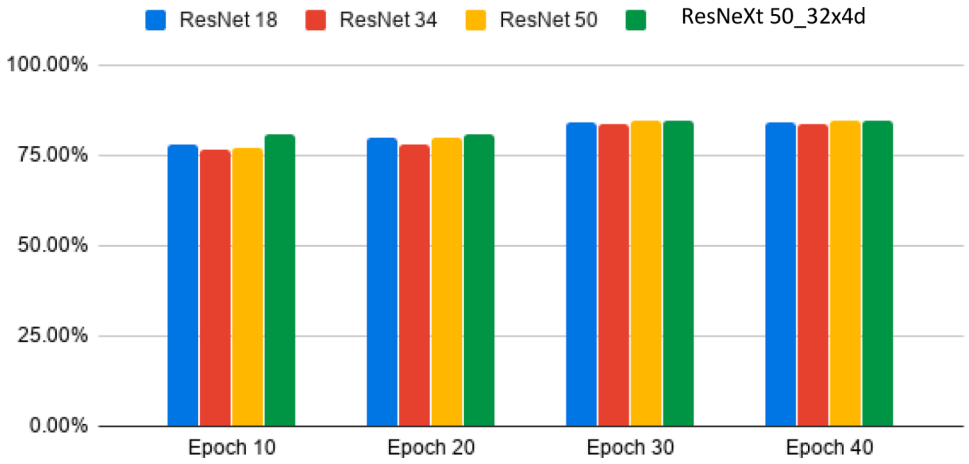


**Fig. 12.** Variation in Rank score with epochs.

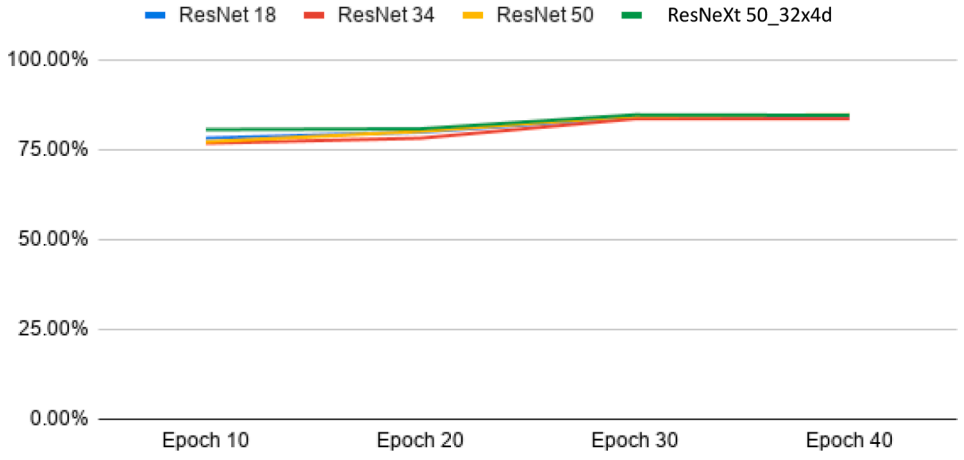


**Fig. 13.** Variation in the Rank score.

will help classify the person's re-id problem better when there is no large training dataset. Our results imply that this strategy also delivers competitive performance and can provide re-id-specific domain knowledge and insights to guide future research. In future the person re-id could be used in public places for security management and for locating a missing person or a child. We provide a thorough discussion on obstacles and potential in learning more powerful structures.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data Availability

Data will be made available on request.

## References


Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015* (pp. 3908–3916). https://doi.org/10.1109/CVPR.2015.7299016

Ahmed, S. M., Lejbolle, A. R., Panda, R., & Roy-Chowdhury, A. K. (2020). Camera on-boarding for person re-identification using hypothesis transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12144–12153). https://doi.org/10.1109/CVPR42600.2020.01216

Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., & Xu, Y. (2020). Deep-Person: Learning discriminative deep features for person Re-Identification. *Pattern Recognition, 98*. https://doi.org/10.1016/j.patcog.2019.107036

Bedagkar-Gala, A., & Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing, 32*(4), 270–286. https://doi.org/10.1016/j.imavis.2014.02.001

Chen, H., Wang, Y., Shi, Y., Yan, K., Geng, M., Tian, Y., & Xiang, T. (2018). Deep Transfer Learning for Person Re-Identification. In *2018 IEEE 4th International Conference on Multimedia Big Data, BigMM 2018*. https://doi.org/10.1109/BigMM.2018.8499067

Chen, W., Chen, X., Zhang, J., & Huang, K. (2017). Beyond triplet loss: A deep quadruplet network for person re-identification. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January* (pp. 1320–1329). https://doi.org/10.1109/CVPR.2017.145

Gray, D., & Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features – *ECCV 2008. 5302*(October 2008). https://doi.org/10.1007/978-3-540-88682-2.

Grigorev, A., Tian, Z., Rho, S., Xiong, J., Liu, S., & Jiang, F. (2019). Deep person re-identification in UAV images. *Eurasip Journal on Advances in Signal Processing*, (1), 2019. https://doi.org/10.1186/s13634-019-0647-z

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hermans, A., Beyer, L., & Leibe, B. (2017). *In Defense of the Triplet Loss for Person Re-Identification*. http://arxiv.org/abs/1703.07737.

Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6688 LNCS*, 91–102. https://doi.org/10.1007/978-3-642-21227-7_9.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141). https://doi.org/10.1109/CVPR.2018.00745

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., & Radke, R. J. (2019). A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(3), 523–536. https://doi.org/10.1109/TPAMI.2018.2807450

Li, K., Ding, Z., Li, K., Zhang, Y., & Fu, Y. (2018). Support neighbor loss for person re-identification. In *, 3. MM 2018 - Proceedings of the 2018 ACM Multimedia Conference* (pp. 1492–1500). https://doi.org/10.1145/3240508.3240674

Li, W., Zhu, X., & Gong, S. (2018). Harmonious Attention Network for Person Re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Ii* (pp. 2285–2294). https://doi.org/10.1109/CVPR.2018.00243

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by Local Maximal Occurrence representation and metric learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015* (pp. 2197–2206). https://doi.org/10.1109/CVPR.2015.7298832

Liu, H., Feng, J., Qi, M., Jiang, J., & Yan, S. (2017). End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing, 26*(7), 3492–3506. https://doi.org/10.1109/TIP.2017

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 212–220).

Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4004–4012).

Paisitkriangkrai, S., Shen, C., & Van Den Hengel, A (2015). Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015* (pp. 1846–1855). https://doi.org/10.1109/CVPR.2015.7298794

Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., & Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December* (pp. 1306–1315). https://doi.org/10.1109/CVPR.2016.146

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.-G., & Xue, X (2018). Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 650–667). https://github.com/naiq/PN_GAN.

Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., & Wei, Y. (2020). Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6398–6407).

Thenmozhi, K., & Srinivasulu Reddy, U. (2019). Crop pest classification based on deep convolutional neural network and transfer learning. *Computers and Electronics in Agriculture, 164*(September), Article 104906. https://doi.org/10.1016/j.compag.2019.104906

Tian, J., Teng, Z., Zhang, B., Wang, Y., & Fan, J. (2021). Imitating targets from all sides: An unsupervised transfer learning method for person re-identification. *International Journal of Machine Learning and Cybernetics, 12*(8), 2281–2295. https://doi.org/10.1007/s13042-021-01308-6

Varior, R. R., Haloi, M., & Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9912 LNCS* (pp. 791–808). https://doi.org/10.1007/978-3-319-46484-8_48

Varior, R. R., Shuai, B., Lu, J., Xu, D., & Wang, G. (2016). A siamese long short-term memory architecture for human re-identification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9911 LNCS* (pp. 135–153). https://doi.org/10.1007/978-3-319-46478-7_9

Wang, H., Hu, J., & Zhang, G. (2020). Multi-source transfer network for cross domain person re-identification. *IEEE Access, 8*. https://doi.org/10.1109/ACCESS.2020.2991440

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person Transfer GAN to Bridge Domain Gap for Person Re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 79–88). https://doi.org/10.1109/CVPR.2018.00016

Xiao, Q., Luo, H., & Zhang, C. (2017). *Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification*. http://arxiv.org/abs/1710.00478.

Xiao, T., Li, S., Wang, B., Lin, L., & Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3415–3424).

Zhang, L., Xiang, T., & Gong, S. (2016). Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December* (pp. 1239–1248). https://doi.org/10.1109/CVPR.2016.139

Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., & Sun, J. (2017). *AlignedReID: Surpassing Human-Level Performance in Person Re-Identification*. http://arxiv.org/abs/1711.08184.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., & Tang, X. (2017). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January* (pp. 907–915). https://doi.org/10.1109/CVPR.2017.103

Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3586–3593). https://doi.org/10.1109/CVPR.2013.460

Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). *Person Re-identification: Past, Present and Future, 14*(8), 1–20. http://arxiv.org/abs/1610.02984.

Zheng, Z., Zheng, L., & Yang, Y. (2017). A discriminatively learned cnn embedding for person re-identification. In *, 14. ACM transactions on multimedia computing, communications, and applications (TOMM)* (pp. 1–20). https://doi.org/10.1145/3159171

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y. D. (2020). Dual-path convolutional image-text embeddings with instance loss. In *, 16. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (pp. 1–23).

Zheng, Z., Zheng, N., & Yang, Y. (2020). Parameter-efficient person re-identification in the 3d space. *arXiv preprint arXiv:2006.04569*. https://doi.org/10.48550/arXiv.2006.04569.

Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1318–1327). https://doi.org/10.48550/arXiv.1701.08398

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, Kai, & Fei-Fei, Li (2010). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/cvpr.2009.5206848 [Dataset].

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable Person Re-identification : A Benchmark University of Texas at San Antonio. *Iccv*, 1116–1124. https://doi.org/10.1109/ICCV.2015.133 [Dataset].