Synthetic Data Augmentation for Facial Re-identification

Synthetic Data Augmentation for Facial Re-identification

Anonymous Authors

Anonymous Submission

May 2019

Abstract

Facial Re-identification datasets which facilitate the training of Deep Neural Networks (DNNs), tend to be high quality images of celebrities harvested from the internet. There is however a domain gap between these datasets, and the low quality samples used in real-world systems and scenarios such as surveillance footage. In this work we describe a novel process of data augmentation using synthetically generated images, which aids cross-domain generalisability, without the need to acquire large amounts of real data in the target domain. We also contribute a new dataset derived from this process: *syn-Face*. Our approach is validated by training with standard high quality datasets with synthetic augmentation and testing in 2 different realistic sets.

Keywords: Facial Re-identification; Data Augmentation; Dataset Generation; Deep Learning

1 Introduction

Facial Re-identification is the task whereby an image of a person's face used as a query / probe is compared to known identities in a gallery. This results in a similarity metric between the probe and each corresponding identity in the gallery, which can then be ranked from most-similar to least. This is very useful, for example, in security applications whereby a suspect or individuals on a watch-list may be detected in surveillance footage.

The field has exploded in recent years with the advent of practical Deep Neural Networks (DNNs), the computational capacity to train them, and the availability of large datasets such as *VGGFace2* [1], *MS-Celeb-1M* [2], and the *Diversity in Faces Dataset* [3].

Different deep neural approaches has been proposed to perform metric learning. In [4], the authors introduced the triplet-loss, a metric learning loss function that aims to embed images into Euclidean space with similar images pulled towards each other and dissimilar images pushed apart. The VGG16 architecture [5] combined the triplet-loss with a Convolutional Neural Network (CNN) and a large unconstrained facial dataset. This was further extended in [1] with an even larger dataset, and a network based on Resnet50 [6]. Also, DNNs designed specifically for facial recognition / re-identification have been developed, such as *HaarNet*[7] which uses a trunk-and-branch DNN architecture based on Haar-like features (as described in [8]).

However, although there has been plenty of success with facial re-identification under lab conditions [1], [9], [10], this has not necessarily translated to real-world performance when applied to everyday scenarios, particularly in the security / surveillance space. The apparent inability of systems to successfully transfer between controlled and real-world scenarios, is at least partly due to the difference in domains those scenarios present. The large datasets used to train facial re-identification systems are primarily harvested from the internet (see [1], [2]) and aimed at different purposes to re-identification, such as face recognition. This means that these datasets [1], [2] mainly consist of high-quality images of celebrities, taken with professional equipment in good lighting conditions, with a bias for frontal images or people facing the camera. This is in contrast to the low-quality images that are typical of surveillance footage. It is a very common problem in re-identification to have

a gallery consisting of high-quality stills, with probe images being low-quality frames of surveillance footage: this is made a very difficult task if there are no images with similar properties in the training data.

In addition, although many available datasets are unconstrained (in that they contain a wide variety of poses, where the subject is not necessarily under the instruction of the person capturing the image), these datasets still struggle to encapsulate the full complexity of real-world data such as subjects being unaware of the camera's presence or actively avoiding it. Only a few small datasets aiming to reflect this complexity are available [11], [12] but their size makes them insufficient for training DNN architectures.

Given the scarcity of real-world / domain-specific images, and the amount of effort needed to acquire them, it is imperative that we find methods to utilise the existing datasets in ways that promote cross-domain generalisability. One solution is to employ advanced data augmentation techniques that are able to introduce the missing variability of the existing dataset.

Traditional data augmentation techniques include random crops, mirroring, and in-plane rotations [13], [14], with the goal being to create many unique samples as is feasible from a single image.

The authors of [13] uses affine transformations (translations and rotations), which [14] expands on by introducing patches, as well as mirroring and lighting noise. Focusing on colour, [15] manipulates the brightness, hue and contrast, whereas [16] performs colour-casting: randomly manipulating each channel in an RGB image by a random value across the entire image.

More innovative approaches to data augmentation, able to introduce a fully new variability mode not present in the original set, have been proposed although more scarcely. Research in [17] introduced a number of face specific augmentations such as landmark perturbation, as well as hair, glasses, and illumination synthesis, while [18] used background substitution. In addition, [17] also uses 3D image reconstruction for synthesising poses, although the reconstruction requires high-quality near-frontal images to function correctly. More recently, Generative Adversarial Networks (GANs) have been used to perform style transfers on datasets [19].

While these techniques compensate for the lack of data, they barely introduce transformations such as high-to-low resolution changes or out-of-plane 3D rotations. In this paper we propose a new type of data augmentation for facial re-identification: adding synthetically generated data into a system's training dataset which compensates for it's lack of specific variability. Furthermore, we generate and make available a new dataset, *syn-Face* (detailed in sec. 2.1) and use it in conjunction with VGGFace2 [1] to train a neural network (sec. 2.2) on a facial re-identification task. Our approach is then validated on a number of datasets with different intrinsic properties to gauge cross-domain generalisability.

2 Methodology

2.1 syn-Face: The Synthetic Dataset

The generation of the *syn-Face* is a multi-stage progress. Fist, 3D models of humanoids are constructed. Then, images are generated by placing each 3D model in a virtual scene, and rendering to set parameters.

The models are generated in the Blender [20] 3D creation suite using the MB-Lab plug-in, which constructs highly detailed and realistic humanoids. MB-Lab has 3 base models (*afro*, *asian*, and *caucasian*) which are subdivided further into 18 base *phenotypes*, each of which have a male and female version. Each phenotype was assigned probability distributions for eye and skin colour based on the Fitzpatrick scale [21]. For each model the sex and base phenotype were chosen at random. The apparent age, height, and mass were then sampled from a normal distribution. Finally, the physical proportions of the model were randomly adjusted according to a normal distribution. This technique allows us to create as many unique 3D humanoids as we desire. The images are then rendered in Blender itself with set lighting, camera positions, and resolution.

2.1.1 Degrees of Freedom

The rendered images in the dataset have 4 degrees of freedom: *Camera Zenith*, *Camera Azimuth*, *Lighting Scenario*, and *Resolution*.

Out-Of-Plane Rotation (OOPR) Synthetic data generation allows us to create complementary images for training, effectively augmenting the training set in ways that are outside the ability of traditional techniques. An example of this is the Out-of-Plane Rotation, where the head is rotated in 3D space, instead of the standard 2D rotation used in tradition augmentation.

While existing images can only be rotated around a point in the image plane, we can generate many images of the exact same scene, rendered from different viewpoints as shown in fig. 1 and fig. 2. We propose that this is an aid to learning, as it may help the model more readily learn frontalization: the process of transforming images containing unconstrained poses into full frontal portraits.

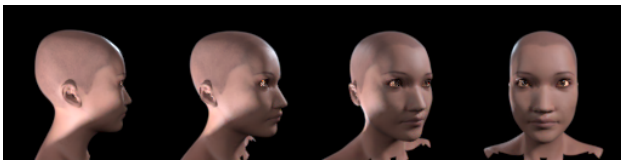
Camera Zenith and Azimuth: We use a spherical co-ordinate system for virtual camera placement with r a set constant, and the origin centred on the model's head. 4 values are used for the zenith (φ), and 7 values for azimuth (θ) as illustrated in fig. 1 and fig. 2 (only 4 angles shown).

Lighting Scenarios: We use 4 lighting scenarios when rendering the images (fig. 3). Scenario *A* uses portrait lighting consisting of 4 lights at set angles and intensities. Scenarios *B*, *C*, and *D* use two light sources: a dominant point lamp positioned at $\varphi = 60°$ with $\theta = 180°$, $\theta = 225°$, and $\theta = 270°$ respectively, and a weak sun lamp to provide a minimal level of ambient lighting.

Resolution: Finally, we rendered each image at 4 different resolutions (fig. 4): 256×256, 128×128, 64×64, and 32×32.

(a) $45°$ (b) $60°$ (c) $75°$ (d) $90°$

Figure 1: Camera placement at φ with $\theta = 270°$

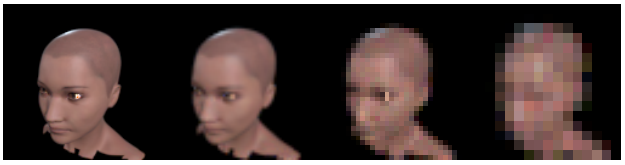


(a) $180°$ (b) $210°$ (c) $240°$ (d) $270°$

Figure 2: Camera placement at θ with $\varphi = 90°$

(a) A (b) B (c) C (d) D

Figure 3: Lighting Scenarios



(a) 256^2 (b) 128^2 (c) 64^2 (d) 32^2

Figure 4: Resolution Levels

2.2 Re-identification Network

The model we propose for face re-identification is a Deep Neural Network (DNN) based on ResNet50 [6], with a final densely connected layer of 128 units (as shown in fig. 5). We use the output of this layer as a point in a 128-dimensional embedding space. The network is trained using the triplet-loss [4] with semi-hard mining

as implemented in TensorFlow. The goal is to train the network so that the distance in the embedding space is decreased for similar images, and increased for dissimilar images. Classification / re-identification can then be performed in the embedding space by calculating the distance between embeddings for respective images. Triplet loss has been validated [4] as metric learning paradigm for re-identification.

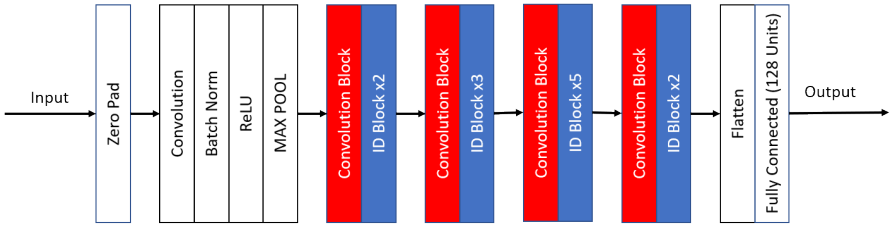


Figure 5: Simplified Resnet50 Architecture. *Convolution Blocks* reduce the output feature map dimensions using convolutions of stride 2, while doubling the number of filters. *ID Blocks* also contain convolutional layers, but preserve the number of filters and filter map dimensions. Both blocks have skip connections, as detailed in [6].

$$L = max(d(a,p)) - d(a,n) + margin, 0) \tag{1}$$

The triplet-loss function is shown in eq. 1, where L is the loss, d is a dissimilarity function (in our case Euclidean distance), and a is an anchor image. p and n are positive and negative image matches respectively. The $margin$ defines how separated the positive and negative examples should be.

2.2.1 Unitary Mini-batches

Training using mini-batches is a common technique to speed up the convergence of DNN training. However, one possible concern in using a metric-learning loss function, such as the triplet loss, is that including both real and synthetic data in mini-batches together may collapse the corresponding clusters in the embedding space. To address this, we include experiments that use *unitary mini-batches*: mini-batches that include samples from a single dataset only. Training samples are grouped by dataset, and then segmented into mini-batches. The mini-batches are then shuffled into a random order.

3 Datasets and Experimental Design

3.1 Non Synthetic (Real) Datasets

(a) (b) (c) (d) (e) (f)

Figure 6: Real Datasets used in this paper for re-identification. a) and b) *VGGFace2*, c) and d) *ChokePoint*, e) and f) *MMF*.

VGGFace2 [1]: The VGGFace2 dataset contains 3.31 million images of 9131 subsets, split into *train* and *test* sets consisting of 8631 and 500 subjects respectively. The images are harvested from Google Images. See

fig. 6a and fig. 6b. As they are images of celebrities, the quality tends to be high with the pictures taken with good lightening and professional-grade equipment. The dataset is however unconstrained, with various poses in a wide variety of environments. We will use this dataset for both training out model and evaluation.

Multi-Modal Faces (MMF) [12]: Multi-Modal Faces is a small asymmetric dataset that can be used to evaluate models in still-to-video and video-to-still scenarios. It consists of 2 subsets: *A* - High quality frontal stills, and *B* - Unconstrained stills mined from lower quality video.See fig. 6e for a sample from *A*, and fig. 6f for a sample from *B*. The dataset contains 77 subjects, all of which are represented in both *A* and *B* subsets. We use this dataset for evaluation only.

ChokePoint [11]: ChokePoint is a video-to-video (v2v) dataset that attempts to mimic the characteristics of real-world surveillance video. The dataset consists of 48 video sequences, taken from an array of 3 cameras placed above two portals (pedestrian choke-points). The sequences recorded at the first portal have 25 subjects, whereas the second portal has 29 subjects. See fig. 6c and fig. 6d. We re-purpose this dataset to fit an S2V scenario, constructing a gallery and probe for each portal.

3.2 Experiment Overview

The metric we will use to assess model performance is the Cumulative Matching Characteristic (CMC), which states the percentage of times a correct match was returned in the top n images, where n is referred to as the "rank".

For the sake of computation time, we use 49984 non-synthetic ("real") images of VGGFace2's *train* set for training. We use this number of real images as an anchor value for 100% across all experiments. In certain experiments, synthetic data is added to the training set. When this is the case, the number of synthetic images added is expressed as an adding percentage to the number of real images $+x\%$.

Testing follows existing protocols set for each dataset where applicable. To test with VGGFace2 in a re-identification scenario, a single image was removed from each identity in the *test* set, forming a gallery in aggregate. Every other image in the *test* set was used as a probe. The same testing protocol was used for ChokePoint, with the exception that it's existing protocol (specified in [11]) mandated testing separately on two defined subsets and combining the results. For MMF, we used the same re-identification protocol as specified in [12] for 10 rounds.

In order to analyse individual aspects of our proposed synthetic data augmentation such as the use of unitary mini-batches (UM) and the effect of OOPR, individual experiments are devised. When we use the synthetic images to train for OOPR only, we apply the following constraints when sampling images from the synthetic dataset: $\varphi \in \{60, 75, 90\}$, $\theta \in \{210, 240, 270, 300, 330\}$, Lighting $\in \{A\}$, and Resolution $\in \{256 \times 256\}$.

4 Results

4.1 Synthetic Data vs Traditional Data Augmentation Techniques

In these experiments, we perform a comparative analysis between the inclusion of synthetic data while training, and several traditional data augmentation techniques in isolation, and the expected gains in testing for the same training scenario as well as more realistic scenarios.

Each of the techniques, when used, are applied according to each's corresponding probability distribution, per-image, per-epoch. Therefore, the same input image will have different potential augmentations applied across multiple epochs. The traditional techniques are:

- **Random Mirror**: There is a 50% chance that an image is mirrored when input into the model.

- **Random Rotate**: Images are rotated to a uniformly random degree between $\pm 5^{\circ}$ when input into the network. The rotated images have to maintain the dimensions of the originals, so areas of images that are now "unoccupied" due the rotation are set to black.
- **Random Down-sample**: There is a 50% chance that an image is down-sampled using bilinear interpolation when input into the model. This results in the resolution decreasing to $1/4$ of the original image, as the height and width are reduced by $1/2$ each.

Table 1: Synthetic Data vs Traditional Data Augmentation Techniques - CMC Scores (%) at Rank 1

Experiment Name	VGGFace2	ChokePoint	MMF
Baseline	3.26	7.25	10.51
Real (Mirror)	**4.75**	**10.59**	12.64
Real (Rotate)	4.04	10.55	**15.83**
Real (Downsample)	3.42	6.82	11.71
Synth +20%	2.67	9.42	9.34
Synth +20% UM	2.95	8.44	11.78
Synth +20% (OOPR Only)	3.28	8.67	8.19
Synth +20% (OOPR Only) UM	2.49	9.99	7.59

Tbl. 1 shows the Rank 1 results for these experiments across the testing datasets. Note however that results for a single experiment across different datasets are not comparable, as each dataset contains a different number of identities. In general, all data augmentation techniques seems to improve learning. When comparing traditional versus synthetic DA, when used in isolation, the traditional augmentation techniques provide a larger increase in performance over the baseline. This is very clear on the same dataset that was used for training (VGGFace2) but less obvious in the more realistic testing sets. The poor performance of Synth on VGG was expected since the introduced variability is not exhibited in this dataset, so any improvement in generalisation to real scenarios will not be reflected in accuracy. More importantly, since DA techniques are not applied in isolation but combined, our synthetic DA will demonstrate complementarity to the traditional techniques as we will show in later results.

As for the variants of *syn-Face* (OOPR only and/or UM), there is no obvious pattern showing which combination is intrinsically superior. However, the differences between the variants are not negligible. Therefore, we surmise that the relative performance of one variant against another is primarily due to domain differences between the datasets tested on.

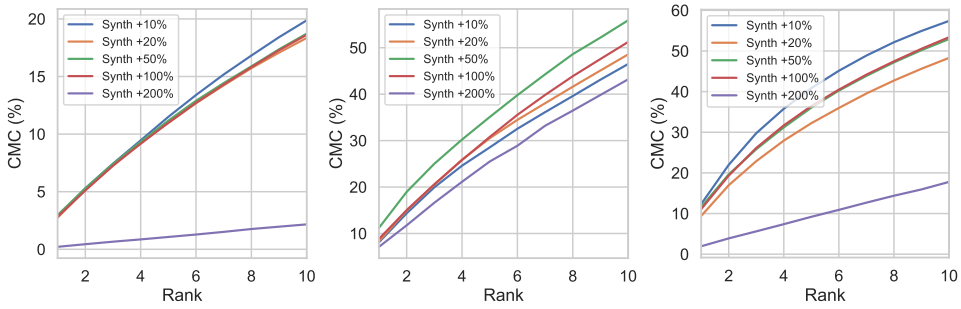
4.2 Optimisation of Real-Synthetic Data Ratio

Here, we aim to find the best ratio of real to synthetic data. Excluding the amount of synthetic data added, all other experimental variables remain fixed: neither *OOPR Only* nor UM are used in these experiments.

Table 2: Optimisation of Real-Synthetic Data Ratio - CMC Scores (%) at Rank 1

Experiment Name	VGGFace2	ChokePoint	MMF
Synth+10%	2.82	8.12	**12.33**
Synth+20%	2.74	8.34	9.40
Synth+50%	**2.97**	**11.21**	11.58
Synth+100%	2.83	8.84	11.09
Synth+200%	0.21	7.13	1.94

Fig. 7 shows the CMC curves and tbl. 2 shows the Rank 1 results. It's clear that for performance, the best ratio of non-synthetic to synthetic data is 100:50 (2:1), with Synth+50% achieving the highest Rank 1 result on 3 datasets, and second highest on *MMF*.



(a) VGGFace2 (b) ChokePoint (c) MMF

Figure 7: Optimisation of Real-Synthetic Data Ratio - CMC Rank 1 to 10 per dataset.

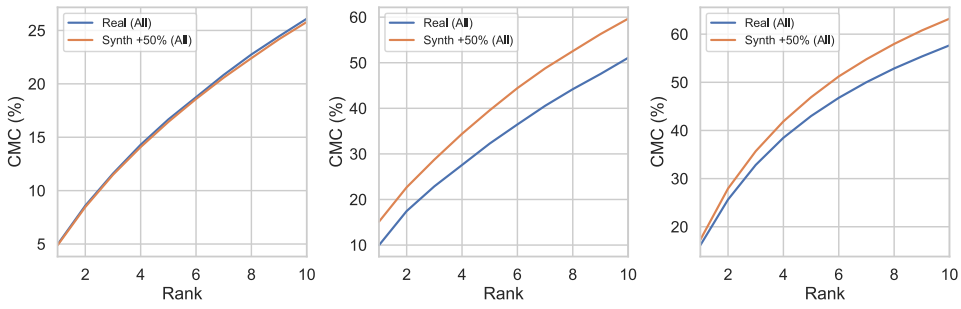
Another notable result here, is that a ratio of 100:200 (1:2) destroys the network's ability to perform re-identification on *VGGFace2* and *MMF*: the two datasets with the higher quality images. As might be expected, there appears to be a limit to how much synthetic data can be added to the training set, before the DNN overfits to synthetic images.

4.3 Combining Traditional Techniques with Synthetic Data

Here, we combine all of the previously explored techniques to maximise the performance over our baseline result, and demonstrate the complementarity of our proposed synthetic data augmentation to the traditional techniques. Here, we use Synth+50% with all traditional techniques as it proved to be the best ratio from the previous experiments. Baseline is included in tbl. 3 for comparative purposes.

Table 3: Combining Traditional Techniques with Synthetic Data - CMC Scores (%) at Rank 1.

Experiment Name	VGGFace2	ChokePoint	MMF
Baseline	3.26	7.25	10.5
Real (All)	**4.96**	9.97	16.2
Synth+50% (All)	4.89	**15.12**	**17.3**



(a) VGGFace2 (b) ChokePoint (c) MMF

Figure 8: Combining Traditional Techniques with Synthetic Data - CMC Rank 1 to 10 per dataset.

Tbl. 3 shows the Rank 1 results, with fig. 8 showing the corresponding CMS curves up to Rank 10. By these results it can be seen how our Synthetic DA seems complementary to traditional techniques, improving the result for realistic challenging sets.

As expected, for *VGGFace2*'s set, applying all of the traditional augmentations improves performance more that also adding synthetic data, since this is the same dataset used for training and non of the synthetic modalities are shown in test. However, as can be seen in fig. 8a the difference in performance appears to be negligible. Meanwhile, the model gained a sizeable performance boost on *ChokePoint* and *MMF* when using synthetic data, in conjunction with traditional augmentation techniques.

5 Conclusion

In this work, we have shown how synthetically generated images can be used to augment datasets when training deep neural networks on facial re-identification scenarios. In particular, this type of augmentation is most beneficial when there are notable differences between the training and testing domains.

We have provided one general example of this augmentation, the *syn-Face* dataset, but countless other variants can be produced. The absolute control we have during data generation, allows for the creation of datasets designed to complement the performance of models, in specific target domains. We will be investigating this, as well as other ways to utilise synthetically generated data for cross-domain training, in future work.

6 References

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *CoRR*, vol. abs/1710.08092, 2017.

[2] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-celeb-1M: A dataset and benchmark for large-scale face recognition," *CoRR*, vol. abs/1607.08221, 2016.

[3] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," *arXiv:1901.10436 [cs]*, Jan. 2019.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 815–823.

[5] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British machine vision conference*, 2015.

[6] 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)*, 2016, pp. 770–778.

[7] M. Parchami, S. Bashbaghi, and E. Granger, "Video-based face recognition using ensemble of haar-like deep convolutional neural networks," in *2017 international joint conference on neural networks (IJCNN)*, 2017, pp. 4625–4632.

[8] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *2014 IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[10] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[11] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *CVPR 2011 WORKSHOPS*, 2011, pp. 74–81.

[12] G. Brown, J. M. del Rincon, and P. Miller, "A comparative study of face re-identification systems under real-world conditions," in *Irish machine vision and image processing conference proceedings 2018: Proceedings*, 2018, pp. 137–144.

[13] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh international conference on document analysis and recognition, 2003. Proceedings.*, 2003, vol. 1, pp. 958–963.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[15] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," p. 6.

[16] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," p. 11.

[17] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, Mar. 2017.

[18] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *2015 12th IEEE international conference on advanced video and signal based surveillance (AVSS)*, 2015, pp. 1–6.

[19] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[20] Blender Online Community, *Blender - a 3D modelling and rendering package*. Blender Foundation.

[21] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through VI," *Arch Dermatol*, vol. 124, no. 6, pp. 869–871, Jun. 1988.