

# Effect of Data Augmentation Methods on Face Image Classification Results

Ingrid Hrga<sup>1</sup><sup>a</sup> and Marina Ivasic-Kos<sup>2</sup><sup>b</sup>

<sup>1</sup>Faculty of Informatics, Juraj Dobrila University of Pula, Rovinjska 14, Pula, Croatia

<sup>2</sup>Department of Informatics, University of Rijeka, Radmile Matejčić 2, Rijeka, Croatia

**Keywords:** Data Augmentation, Image Classification, Transfer Learning.

**Abstract:** Data augmentation encompasses a set of techniques to increase the size of a dataset artificially. Insufficient training data means that the network will be susceptible to the problem of overfitting, leading to a poor generalization capability of the network. Therefore, research efforts are focused on developing various augmentation strategies. Simple affine transformations are commonly used to expand a set. However, more advanced methods, such as information dropping or random mixing, are becoming increasingly popular. We analyze different data augmentation techniques suitable for the image classification task in this paper. We investigate how the choice of a particular approach affects the classification results depending on the size of the training dataset, the type of transfer learning applied, and the task's difficulty, which we determine based on the objectivity or subjectivity of the target attribute. Our results show that the choice of augmentation method becomes crucial in the case of more challenging tasks, especially when using a pre-trained model as a feature extractor. Moreover, the methods that showed above-average results on smaller sets may not be the optimal choice on a larger set and vice versa.


## 1 INTRODUCTION


Data augmentation has become an integral part of almost every machine learning pipeline because modern deep neural networks require many training examples. Larger training sets enable them to generalize better, i.e., successfully applying what has been learned to new data, as the ultimate learning goal. Although approaches based on unsupervised, semi-supervised, or self-supervised learning are increasingly being developed, most tasks are still based on supervised learning, which requires training samples of data and associated labels.

Various large data sets, such as ImageNet (Deng, J. et al., 2009), of labeled images depicting everyday scenes, have been collected and made available to the public, which ultimately enabled advances in computer vision. Still, such data sets often do not meet the real-world needs of many domains with specific data requirements. Moreover, collecting and labeling new data is often difficult or impossible, sometimes due to various restrictions, e.g., GDPR

(Regulation (EU) 2016/679) or because the labeling process requires expert knowledge. And the lack of data in sufficient quantity and quality sometimes sets a barrier to greater adoption of deep models.

Even when there are no such constraints, large amounts of data are needed to cover variations of objects depicted in a scene. For example, in an image recognition task, the problem of different lighting, occlusion, size, or relationships is often present. However, the human perception is much less susceptible to such disturbances allowing them to recognize a scene in just a fraction of a second (Fei-Fei, L. et al., 2007; Hrga, I., & Ivašić-Kos, M., 2019). In contrast, computers have only pixel values at their disposal. Nonetheless, an image recognition system should recognize objects correctly regardless of their size, color, or where they are in the image. And with a more extensive set of data, the likelihood that the model will capture the necessary variation during training also increases. Therefore, it is common for more learning data to provide a better model (Shorten, C., & Khoshgoftaar, T. M., 2019).

<sup>a</sup> <https://orcid.org/0000-0001-5118-4282>

<sup>b</sup> <https://orcid.org/0000-0002-1940-5089>

This paper analyzes different data augmentation techniques suitable for the image classification task. We investigate how the choice of a particular method affects the classification results depending on the size of the training set, the type of transfer learning applied, and the difficulty of the task. Furthermore, we determine the difficulty of the task based on the objectivity or subjectivity of the target attribute.

The issue of subjectivity often comes to the fore when dealing with images of people, often used in a wide range of different systems, such as identification, surveillance, or smart home systems, and areas such as medicine, sports, or fashion. Sometimes classification is facilitated because an explicit part of the image can be referenced when deciding on a label. However, in the case of more subjective attributes, such as beauty, youth, or emotional states, it is usually necessary to consider multiple aspects of the image when making a decision. It is not easy to reach a consensus even among humans when it comes to such attributes.

The paper is structured as follows: after the introductory part, we outline some of the most used augmentation techniques for image classification in the second section. In parallel, we provide an overview of the related work. The third section presents the methods employed in our research and the experimental setup. The classification results are shown in the fourth section, where we also interpret the observed outcomes. Finally, in the last section, we conclude and propose guidelines for future work.

## 2 BACKGROUND AND RELATED WORK

Data augmentation encompasses a set of techniques to increase the size of a data set artificially. However, insufficient training data means the model will be susceptible to overfitting, leading to a poor generalization capability. Furthermore, because of many degrees of freedom, the model can memorize the training data instead of learning how to solve a task. High-capacity neural networks with millions of parameters, such as modern Convolutional Neural Networks (CNNs), are particularly prone to this problem.

There are various ways to alleviate overfitting. One is to constrain the model with regularization techniques, such as dropout (Srivastava, N. et al., 2014). Transfer learning is also one way to take advantage of small learning set by using a pre-trained network. However, it is common for more learning

data to give a better model (Shorten, C., & Khoshgoftaar, T. M., 2019), so the first choice to reduce overfitting is to collect additional data. When it is impossible to collect new data, it is necessary to use the existing data more effectively, e.g., by applying some of the data augmentation techniques.

The data augmentation techniques commonly used in computer vision can be categorized into those that apply transformations to existing images and those that create new images (Shorten, C., & Khoshgoftaar, T. M., 2019). A special group consists of augmentation methods derived from Neural Architecture Search (NAS) to learn an optimal augmentation policy. The following sections explain selected methods in more detail.

### 2.1 Data Augmentation Techniques That Transform Existing Images

Some of the popular choices in this group are geometric or color transformations and information-dropping techniques.

a) Simple geometric and color transformations have become an integral part of most data processing pipelines because they are effective (Perez, L., & Wang, J., 2017), computationally rather simple, and relatively safe to apply. In addition, if used in a limited range, they generally do not affect the class label. However, such transformations do not produce a completely new image, so the diversity they add to the data set is relatively small.

The standard set of operations involves, for example, random cropping, scaling, translation, or rotation. Color transformations that adjust brightness, contrast, hue, or saturation are slightly less common, while spatial transformations that deform objects are rarer in applications. One of the earlier significant applications of simple image augmentations is (Krizhevsky, A. et al., 2012).

b) Information dropping techniques transform images by removing some of the information they contain. They can encompass simple operations, such as adding noise, removing colors, or dropping color channels, and various methods that involve masking parts of the image, which have recently become increasingly popular. The dropout technique for regularization inspires those methods. Dropout stochastically drops network units during training to prevent co-adaptation while augmentation approaches for information dropping act only on the input layer to remove one or more image areas (Shorten, C., & Khoshgoftaar, T. M., 2019).

Information dropping techniques contribute to the increase in robustness (Chen, P. et al., 2020) because

the model is forced to take into consideration also the context. Therefore, they are especially suitable for tasks such as classification or object detection when occlusion occurs (Chen, P. et al., 2020).

Cutout (DeVries, T., & Taylor, G. W., 2017) masks a randomly selected square area of the image with random values, while Random erasing (Zhong, Z. et al., 2020) randomly masks a rectangular area of the image. Different from them, Hide and seek (Singh, K. K. et al., 2018) randomly selects multiple parts to hide, and then the network is forced to search for relevant information throughout the image. Finally, a somewhat different approach takes Grid Mask (Chen, P. et al., 2020). The method hides evenly spaced square regions whose size can be adjusted to help preserve important information crucial for the final decision, making the task more difficult but not impossible.

## 2.2 Data Augmentation Techniques That Create New Images

The techniques that create new images can be divided into those that involve learning (e.g., GANs) and those without learning (e.g., mixing images).

a) Image mixing techniques have proven successful in various tasks, although the resulting images are less understandable to humans (Shorten, C., & Khoshgoftaar, T. M., 2019) and is more challenging to ensure the preservation of class labels. In essence, they randomly select two or more images of the same or different classes and then blend them or cut and combine parts of different images into a new one.

Sample pairing (Inoue, H., 2018) mixes two randomly selected images by averaging their pixel values, and the new image simply gets the class label of the first image. The technique has been shown to improve accuracy, especially in the case of a small training set. Mixup (Zhang, H. et al., 2017) creates a new image based on two images with different labels and performs the linear interpolation of both of them. Interpolation weights are randomly chosen from the beta distribution. Finally, random image cropping and patching (RICAP) (Takahashi, R. et al., 2019) creates a new image by randomly taking four images, cutting a patch from each one, and transferring it to the new image. In addition to images, the method also mixes class labels.

b) Generative adversarial networks (GANs) learn the distribution of a dataset. By sampling from the learned distribution, new synthetic examples can be generated that are very similar to the original dataset. Numerous network variations have been proposed,

and in the context of data augmentation, GANs can be used to generate new data based on a specific dataset (Frid-Adar, M. et al., 2018) or to generate transformed data directly (Antoniou, A. et al., 2017). GANs can also be trained to transfer the style of one image to another image (Isola, P. et al., 2017; Karras, T. et al., 2019).

## 2.3 Data Augmentation Techniques Inspired by Neural Architectural Search

A special group of augmentation techniques automates finding the optimal data augmentation policy. They are inspired by Neural architecture search (NAS), where reinforcement learning and evolution strategies have been employed to learn optimal network topology for a given task (Cubuk, E. D. et al., 2020). Similarly, learning an optimal augmentation policy improves accuracy and robustness (Cubuk, E. D. et al., 2020), although additional steps increase complexity and computational costs.

Smart augmentation (Lemley, J. et al., 2017) uses two networks. The augmentation network learns to generate augmented images by merging several examples of the same class. Those examples are then used to train a second, task-specific network. Autoaugment (Cubuk, E. D. et al., 2019) uses reinforcement learning to find an optimal augmentation policy. For each image from a training batch, a sub-policy determines the transformations to apply and their settings. A search algorithm selects among many sub-policies that will result in the best validation accuracy for the selected data set. RandAugment (Cubuk, E. D. et al., 2020) simplifies automation of the augmentation process by reducing the search space using a simple grid-search strategy. In addition, two parameters can be used to adjust the number of augmentations and their magnitude.

# 3 METHODS AND EXPERIMENTAL SETUP

## 3.1 Dataset

We base our experiments on the CelebA (Liu, Z. et al., 2015) dataset, consisting of more than 200k images of celebrities. Every image is annotated with 40 attributes. Of these, we selected two attributes for two different binary classification tasks.

We chose "Mouth Slightly Open" as an example of an objective attribute, i.e., with a clear representation in the image, and "Attractive" as a highly subjective attribute because it can hardly be associated with just one part of the image. The attributes are roughly equally represented in the dataset.

The set is already split into training, validation, and test sets. We did not use all the training data as this would significantly slow down the experiments, for which a standard laptop with a single GPU was used. Instead, we extracted three training subsets from the original training set to compare the impact of each augmentation technique with the change in the data set size. The first training set consists of 50,000 images selected by random selection. From them, we created two additional subsets of 10,000 and 1,000 images so that the smaller subset is always the proper subset.

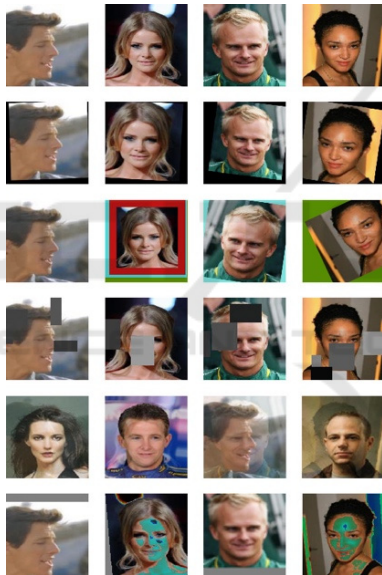


Figure 1: Augmentations used in our experiments. From the top row to the bottom row: no augmentation, simple affine transformations, more aggressive affine transformations, information dropping, Mixup, RandAugment.

### 3.2 Augmentation Techniques

All image transformations were performed using the Imgaug library (Jung, A. B. et al., 2020). The original images have a resolution of 178x218 pixels. Since the pre-trained CNN model requires inputs to be 224x224 pixels in size, all images were resized first. For the experiments, we chose the following augmentations and their settings (Figure 1):

1. Simple affine transformations (we refer to them as AF(s)) with a limited range of settings and applied

in a deterministic order: scale (0.85, 1.15), translate px (-20, 20), rotate (-15, 15), flip left-right with probability  $p=0.5$ .

2. Affine transformations (AF) with more aggressive settings. For each image, between 0 and 4 transformations were selected and applied in a random order: scale (0.5, 1.5), translate px (-20, 20), rotate (-25, 25), flip left-right with probability  $p=0.5$ . Additionally, a fill color for the background is also randomly chosen when needed. Although images in the original dataset are cropped and aligned, the background is still visible and shows strong variability, predominantly in color.

3. As a representative of the information-dropping technique (CUT), we used a variation of the Hide and seek and Cutout. Every image was masked with a randomly selected number of rectangles, from one to five, filled with a color chosen from black, white, or shades of gray. The size of rectangles varied between 0.1 and 0.4 relative to the image size.

4. Mixup (MIX) interpolates two images and their labels by sampling from the mixup vicinal distribution (Zhang, H. et al., 2017):

$$\begin{aligned} \mu(x', y' | x_i, y_i) &= \\ &= \frac{1}{n} \sum_j \mathbb{E}[\delta(x' = \lambda x_i + (1 - \lambda)x_j, \\ & \quad y' = \lambda y_i + (1 - \lambda)y_j)], \\ \lambda &\sim \text{Beta}(\alpha, \alpha), \text{ for } \alpha \in (0, \infty) \end{aligned} \quad (1)$$

where  $x_i$  and  $x_j$  are original images,  $y_i$  and  $y_j$  are their respective labels,  $x'$  and  $y'$  are interpolated image and label,  $\lambda$  is the interpolation weight,  $\alpha$  is a mixup hyper-parameter, which controls the interpolation strength. For mixup, we set  $\alpha = 1$ .

5. RandAugment (RND) is applied with a group of settings suggested by the authors (Cubuk, E. D. et al., 2020). Parameter  $n$  represents the number of transformations applied to an image, and  $m$  represents the magnitude for all the transformations, with  $m = 0$  being the weakest. We set  $n = 2$ , and  $m = 9$ . When needed, the background is filled with medium gray.

### 3.3 Classification

We used MobileNet v3 small (Howard, A. et al., 2019) for the classification task. The network was pre-trained on the ImageNet dataset, so we adapted the classification layer to the binary classification task. We chose a smaller CNN network because the paper aims not to achieve the highest classification

accuracy but to compare augmentation techniques under controlled settings. Moreover, although various architectures have specialized for facial analysis (Schroff, Kalenichenko & Philbin, 2015; Han et al., 2017; Jang, Gunes, & Patras, 2019), we opted for a general-purpose network to leave more room to spot differences between augmentation methods.

Since ImageNet contains images of people, we did not train the models from scratch but used transfer learning in two ways. First, we trained only the classifier layer on the CelebA set when the model was used as a feature extractor. Then, we fine-tuned the entire model on the CelebA dataset in the second case. We used Adam optimizer (Kingma, D. P., & Ba, J., 2014) with a learning rate of 0.0005 in the first and 0.0002 in the second case. We trained the models for 20 epochs on datasets consisting of 1,000, 10,000, and 50,000 examples, with five different augmentation techniques and without augmentation. For every attribute combination, dataset size, type of transfer learning, and augmentation technique, we chose the model with the highest validation accuracy, resulting in 72 models. We tested on the entire original test set and calculated the accuracy and F1 for each model. We consider classification with the "Mouth Slightly Open" attribute as the target class to be an easier task.

## 4 RESULTS

The classification results are shown in Table 1. It can be observed that generally better results were achieved with the "Mouth Slightly Open" (MSO) attribute as the target class than with "Attractive" (ATT). The maximum accuracy and F1 scores were achieved on the 50k dataset and fine-tuning. In the case of MSO, the scores are 92.96% and 92.79%, respectively, while for ATT, the scores are 82.15% and 82.39%. Also, the entire range of accuracy or F1 values in the case of MSO is larger, which indicates that the size of the dataset contributed more significantly to the result of MSO than of ATT. For instance, the difference in percentage points between the total achieved minimum and maximum accuracy for MSO is 25.28, while for ATT, it is only 5.88.

If taking into account the size of the training set and the type of transfer learning, then the difference between the best and the worst augmentation results in the case of MSO and feature extraction (FE) is larger for the largest set (4.21 points difference in F1 and 3.61 in accuracy). At the same time, fine-tuning (FT) is the smallest for the largest set (0.88 points in both accuracy and F1). Similar can be observed for

the ATT attribute, although with a smaller range of values. That was expected because the influence of the pre-training set is significant in the case of feature extraction.

Suppose standardizing the results by considering the target attribute, transfer type, and dataset size to track how the relationship between individual augmentation techniques changes with a change in dataset size. It can be observed that this relationship remains fairly constant for the combination of ATT and FE. AF(s) lags behind other techniques, while RND scores above average. For the combination of ATT and FT, there is a greater change in the relationship between individual augmentation techniques. This change is more significant at the transition from a set size of 1k to 10k than 10k to 50k. In the case of the smallest set, AF stands out, and with the largest set, CUT and NOAUG produced the highest scores. It is noticeable that CUT and MIX improved and gained more importance with the increase in data set size.

In the case of MSO and FE, the situation is unclear. The results are quite uneven, and there is no single best option, as it varies with data set size and transfer learning type, except that AF(s) still produced the worst results. Although the relationship between individual techniques is somewhat similar to that found in ATT, when fine-tuning (FT), the augmentation techniques showed a change that differs from what was observed in ATT.

It is interesting to note that the best results in the case of MSO and FE are those achieved without augmentation. FT is the opposite, where the best results were achieved by somewhat more aggressive affine transformations (AF) and a variant of the information dropping technique (CUT). It is even more noticeable how MIX improved from the worst result on the smallest set to the best on the largest set.

In general, MIX improves with increasing set size and has shown the largest regularization effect among the tested techniques. With fine-tuning of other hyperparameters, MIX could bring additional improvements.

Although the relationship between individual augmentations changes with the change in dataset size, this change expressed in accuracy or F1 scores becomes smaller as the dataset increases, especially with fine-tuning. Increasing the size above 10k for MSO had virtually no significant impact on the results.

Table 1: Accuracy and F1scores for the selected augmentation techniques for combinations of transfer learning type and training dataset size. Best scores are marked with \*.

"Mouth Slightly Open"												
Augmentation	Accuracy						F1					
	FE 1k	FE 10k	FE 50k	FT 1k	FT 10k	FT 50k	FE 1k	FE 10k	FE 50k	FT 1k	FT 10k	FT 50k
AF(S)	0.6768	0.7038	0.7070	0.8295	0.9057	0.9228	0.6638	0.6898	0.6879	0.8099	0.9021	0.9217
AF	*0.7007	0.7264	0.7397	*0.8458	*0.9154	0.9272	0.6708	0.7083	0.7226	0.8302	*0.9128	0.9251
CUT	0.6887	0.7186	0.7318	0.8430	0.9104	0.9263	0.6815	0.6870	0.7190	*0.8319	0.9072	0.9243
MIX	0.6938	0.7210	0.7350	0.8279	0.9121	*0.9296	0.6697	0.6982	0.7097	0.8210	0.9082	*0.9279
NOAUG	0.6977	*0.7337	*0.7432	0.8401	0.9067	0.9208	*0.6825	*0.7123	0.7300	0.8288	0.9035	0.9191
RND	0.6961	0.7285	0.7383	0.8406	0.9094	0.9227	0.6865	0.7023	*0.7245	0.8272	0.9067	0.9209

"Attractive"												
Augmentation	Accuracy						F1					
	FE 1k	FE 10k	FE 50k	FT 1k	FT 10k	FT 50k	FE 1k	FE 10k	FE 50k	FT 1k	FT 10k	FT 50k
AF(S)	0.7626	0.7688	0.7688	0.7869	0.8036	0.8162	0.7651	0.7710	0.7803	0.7883	0.7954	0.8126
AF	0.7686	0.7780	0.7780	*0.7935	0.8111	0.8203	0.7736	0.7869	0.7865	*0.7979	0.8112	0.8192
CUT	*0.7688	0.7790	0.7790	0.7839	*0.8115	0.8193	0.7743	0.7867	*0.7901	0.7884	0.8133	*0.8239
MIX	0.7635	0.7769	0.7769	0.7793	0.8068	0.8207	0.7676	0.7774	0.7853	0.7793	*0.8142	0.8191
NOAUG	0.7672	0.7786	0.7786	0.7820	0.8078	*0.8215	0.7730	0.7850	0.7883	0.7837	0.8105	0.8237
RND	0.7662	*0.7813	*0.7813	0.7913	0.8099	0.8196	*0.7759	*0.7884	0.7900	0.7939	0.8140	0.8212

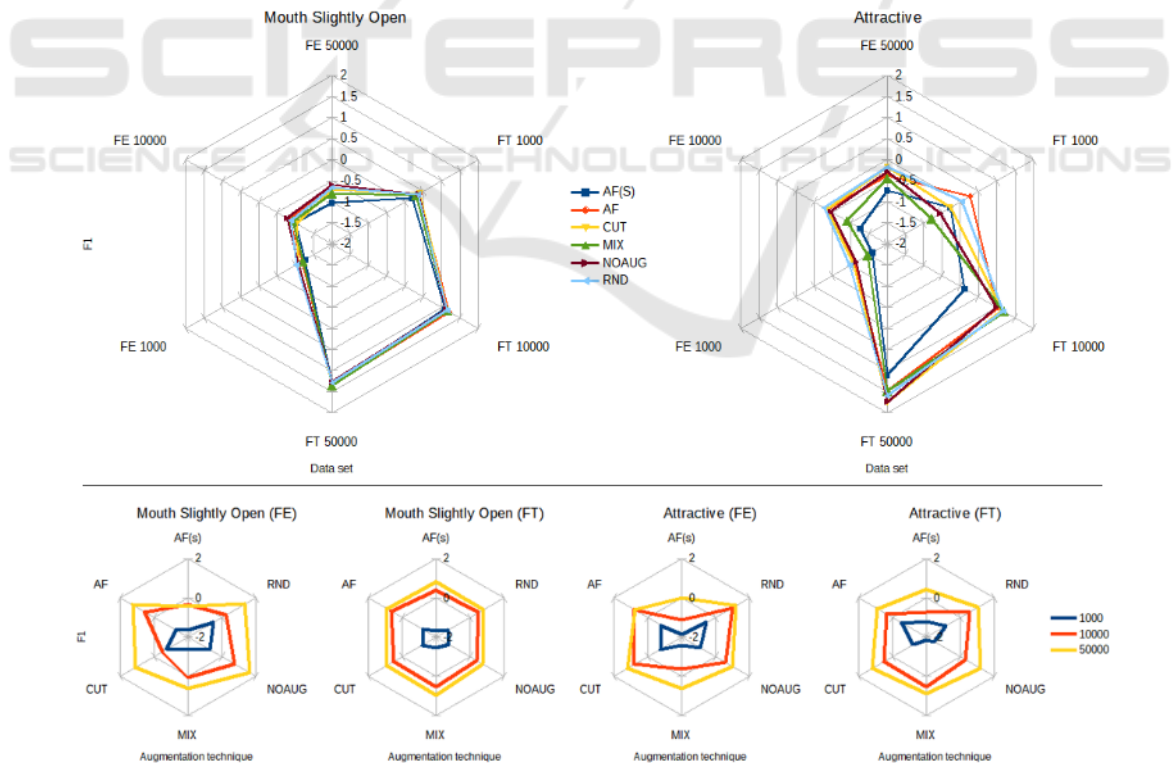


Figure 2: Standardized F1 scores. The top row shows the classification results with the target attributes "Mouth Slightly Open" (left) and "Attractive" (right) expressed in F1 values standardized by the target attribute. The bottom row shows F1 scores normalized by attribute and transfer learning type.

For comparison, Table 2 shows the results of selected state-of-the-art models tested on the CelebA dataset. These models specialize in various facial analysis tasks, and each of them uses certain architecture specifics, e.g., (Sharma & Foroosh, 2020) proposed special slim modules to achieve computational efficiency. Some models (Han et al., 2017; Jang, Gunes, & Patras, 2019) use additional sets of facial images for training, while others (Rudd, Günther, & Boulton, 2016; Hand, & Chellappa, 2017) use multitask learning to take advantage of all available attributes at the same time. In addition, some models use image augmentation extensively (Jang, Gunes, & Patras, 2019), while others do not use it at all (Rudd, Günther, & Boulton, 2016). It can be observed that the results achieved in our experiments are still comparable. However, a general-purpose network was used without any special adjustments. It was trained on a subset of the CelebA training set, while others are trained on the entire CelebA training set.

Table 2: Accuracy on the CelebA test set of selected models specialized for facial analysis. For comparison, our best results were achieved on the 50k dataset with fine-tuning and CUT augmentation for ATT and MIX for MSO attributes.

Method	Accuracy*	
	ATT	MSO
LNet + ANet (Liu et al., 2015)	0.8100	0.9200
Moon (Rudd, Günther, & Boulton, 2016)	0.8167	0.9354
MCNN + AUX (Hand, & Chellappa, 2017)	0.8306	0.9374
DMTL (Han et al., 2017)	0.8500	0.9400
Face-SSD (Jang, Gunes, & Patras, 2019)	0.8130	0.9190
Slim-CNN (Sharma & Foroosh, 2020)	0.8285	0.9379
Our (FT 50k CUT/MIX)	0.8239	0.9279

\* All results are from original papers.

## 5 CONCLUSIONS

Facial image analysis is often performed automatically in systems used in various application domains, so it is important to be aware of all the factors influencing the outcome.

In this paper, we analyzed the influence of augmentation techniques on the results of face image classification by comparing five augmentation techniques with training without image augmentation on datasets of three sizes and using two variants of

transfer learning. In addition, we took into account the objectivity or subjectivity of the target attribute.

The results show that simple affine transformations applied with a small magnitude did not prove useful enough. At the same time, mixing images gained importance with the increase of dataset size, especially with fine-tuning. Mixing images also showed the strongest regularization effect. However, due to the uniformity of the scenes and the lack of typical problems, such as occlusion, the information dropping technique has not come to the fore enough, so similar research should also be conducted on images of more complex scenes.

Noticeably, with the change in training set size, especially with fine-tuning, the relationship between individual augmentation techniques changes, indicating that those that showed above-average results on smaller sets may not be the optimal choice on a larger set and vice versa.

Also, there is a noticeable difference in the results concerning the target attribute. In the case of an easier task, there is less variation of the classification scores than in the case of a more subjective target attribute, suggesting that the augmentations should be chosen more carefully in the latter case.

However, in total, with the increase in the set size, the difference in the achieved accuracy or F1 scores between the best and worst augmentation techniques decreases, especially with fine-tuning.

The experiments showed that the augmentation techniques should be chosen carefully and require additional research. Therefore, we intend to expand the study to different models and images of various scenes. Moreover, we want to analyze whether the classifier "looks" at the image changes due to the chosen augmentation technique.

## ACKNOWLEDGEMENTS

Croatian Science Foundation supported this research under the project IP-2016-06-8345, "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS), and the University of Rijeka under the project number 18-222-1385.

## REFERENCES

- Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

- Chen, P., Liu, S., Zhao, H., & Jia, J. (2020). Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*.
- Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 113-123).
- Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 702-703).
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE.
- DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
- Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance at a real-world scene? *Journal of vision*, 7(1), 10-10.
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018, April). Synthetic data augmentation using GAN for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)* (pp. 289-293). IEEE.
- Han, H., Jain, A. K., Wang, F., Shan, S., & Chen, X. (2017). Heterogeneous face attribute estimation: A deep multitask learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40(11), 2597-2609.
- Hand, E. M., & Chellappa, R. (2017). Attributes for improved attributes: A multitask network utilizing implicit and explicit relationships for facial attribute classification. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314-1324).
- Hrga, I., & Ivašić-Kos, M. (2019, May). Deep image captioning: An overview. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 995-1000). IEEE.
- Inoue, H. (2018). Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*.
- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).
- Jang, Y., Gunes, H., & Patras, I. (2019). Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. *Computer Vision and Image Understanding*, 182, 17-29.
- Jung, A. B., Wada, K., Crall, J., Tanaka, S., Graving, J., Yadav, S., ... & Laporte, M. (2020). Imaging. *GitHub: San Francisco, CA, USA*.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp.4401-4410).
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097-1105.
- Lemley, J., Bazrafkan, S., & Corcoran, P. (2017). Smart augmentation learning is an optimal data augmentation strategy. *Ieee Access*, 5, 5858-5869.
- Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730-3738).
- Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.
- Regulation (EU) 2016/679 of the European Parliament and of the Council [online] <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=HR#d1e40-1-1> [Accessed: 20<sup>th</sup> October 2021.]
- Rudd, E. M., Günther, M., & Boulton, T. E. (2016). Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision* (pp. 19-35). Springer, Cham.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).
- Sharma, A. K., & Foroosh, H. (2020). Slim-CNN: A lightweight CNN for face attribute prediction. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 329-335). IEEE.
- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1-48.
- Singh, K. K., Yu, H., Sarmasi, A., Pradeep, G., & Lee, Y. J. (2018). Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
- Takahashi, R., Matsubara, T., & Uehara, K. (2019). Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9), 2917-2931.
- Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.
- Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 13001-13008).