# Deep Facial Expression Recognition with Occlusion Regularization

Nikul Pandya, Philipp Werner(✉), and Ayoub Al-Hamadi

Neuro-Information Technology, Otto von Guericke University Magdeburg,
Magdeburg, Germany
{nikulbhai.pandya,philipp.werner,ayoub.al-hamadi}@ovgu.de,
http://www.iikt.ovgu.de/nit.html

**Abstract.** In computer vision, occlusions are mainly known as a challenge to cope with. For instance, partial occlusions of the face may lower the performance of facial expression recognition systems. However, when incorporated into the training, occlusions can be also helpful in improving the overall performance. In this paper, we propose and evaluate occlusion augmentation as a simple but effective regularizing tool for improving the general performance of deep learning based facial expression and action unit recognition systems, even if no occlusion is present in the test data. In our experiments we consistently found significant performance improvements on three databases (Bosphorus, RAF-DB, and AffectNet) and three CNN architectures (Xception, MobileNet, and a custom model), suggesting that occlusion regularization works independently of the dataset and architecture. Based on our clear results, we strongly recommend to integrate occlusion regularization into the training of all CNN-based facial expression recognition systems, because it promises performance gains at very low cost.

**Keywords:** Facial expression recognition · Facial action unit intensity estimation · Occlusion regularization · Data augmentation · CNN

## 1 Introduction

Deep learning methods, especially CNNs, outperform previous state of the art in nearly all computer vision tasks, e.g. object detection, image classification, and facial expression recognition. Facial expression recognition attracts researchers' attention because of its many applications in human-machine interaction, social robotics, medical diagnosis and treatment, and semi-automated driving.

Regularization is one of the key elements of deep learning, allowing to generalize well to unseen data, even when training on a limited training set or with an imperfect optimization procedure [8]. Some widely and successfully used regularization techniques are data augmentation, drop-out, batch normalization, and weight decay, which are also common in expression recognition [11,22]. In addition to these methods, this paper proposes an occlusion-based regularization

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12510, pp. 410–420, 2020.
https://doi.org/10.1007/978-3-030-64559-5_32

technique, which consistently improves performance in facial expression recognition and can be combined with any existing regularization technique and network architecture. Occlusion regularization is a specific form of data augmentation and very easy to implement: Training images are synthetically occluded by random black bars or objects at random locations.

The work's main contributions are:

1. We propose to apply occlusion augmentation in facial expression recognition tasks, such as recognition of emotion categories and recognition of facial action unit intensities. Occlusion augmentation is a simple and effective regularizer, which can be applied with any CNN approach. It is beneficial even if the test data does *not* contain occlusions.
2. We experimentally show the resulting performance improvements using three datasets with different expression recognition tasks (RAF [12], AffectNet [16], and Bosphorus [18] databases) and three CNN architectures (pre-trained Xception [2] and MobileNet [6] as well as a custom architecture).
3. We compare our results with state-of-the-art results. We clearly outperform prior work on the Bosphorus dataset. On the RAF dataset we reach comparable results with our simple approach, which may also be applied to further improve results of more sophisticated state-of-the-art approaches.

## 2 Related Work

Most of the work related to occlusion in facial expression recognition intended to improve performance on partially occluded images. In contrast, our work addresses performance improvements on all face images, including occlusion-free images. For a general overview on facial expression recognition and on expression recognition under partial occlusion, the reader is referred to Li et al. [11] and Zhang et al. [24], respectively. A recent approach on occlusion-aware expression recognition is the CNN network with an attention mechanism proposed by Li et al. [13]. They combined multiple representations from facial regions of interest by weighting via a proposed gate unit, which computes an adaptive weight from the region itself according to the unobstructedness and importance. This way they improved performance on both occluded and occlusion-free face images.

Kukačka et al. [8] review and classify the literature on regularization. Among the most widely used methods are data augmentation, batch normalization, and drop-out. There are lots of works on using data augmentation to improve the performance of a deep learning network in general, which includes facial expression recognition tasks. Bengio et al. [1] showed that the performance of a deep neural network can be improved by data augmentation in the image classification problem. Even before, back in 1998, LeCun et al. [9] used various affine transformation for data augmentation for training LeNet. Lemley et al. [10] proposed a smart data augmentation technique to optimize data augmentation during training. Lin et al. [14] used data augmentation and compact feature learning to improve the performance of the facial expression recognition model. Sarandi

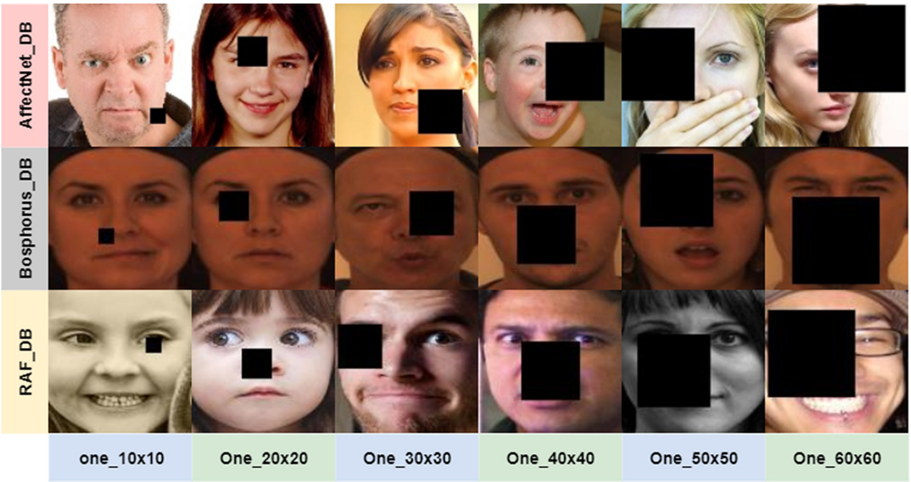

**Fig. 1.** Example images of the used databases with synthetic random black-bar occlusions of the used sizes. The occlusions are only augmented during training. Testing is done with the original (mostly occlusion-free) images.

et al. [17] used synthetic object occlusion for 3D body pose estimation performance improvements, which inspired our work on occlusion-based regularization in the facial expression domain.

Ioffe and Szegedy [7] showed how batch normalization can improve training time and the performance of deep learning networks. Batch normalization has a regularizing effect, because mean and standard deviation used for normalization vary between the randomly composed mini-batches. This introduces additional variation and teaches the layers to be robust to a lot of variation in their input. Another widely used concept of regularization in deep learning networks is drop-out, which was introduced by Hinton et al. [5]. Drop-out randomly removes hidden neurons during the training of a deep network. By doing so, the network does not depends on a specific activation during training, which reduces overfitting.

## 3 Approach

We propose to augment synthetic occlusions on the images that are used to train expression recognition models. The position of the occlusion mask in pixel coordinates is randomly selected for each sample (and epoch) in a way that it is always completely within the image. Two types of occlusions are considered in this work:

**Black-Bar Occlusions:** We use square occlusion masks of the sizes $10 \times 10$ to $60 \times 60$ pixels for all the experiments and set all pixels to zero. Some examples of random black bar occlusion can be seen in Fig. 1.

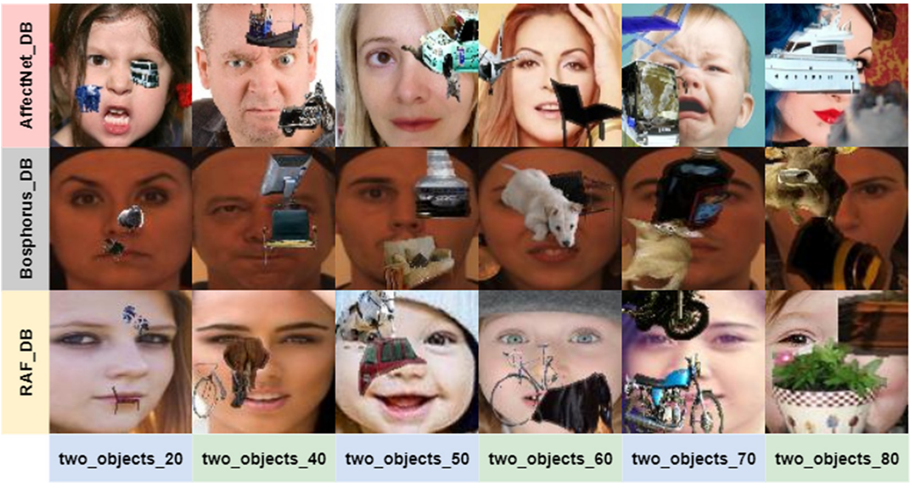

**Fig. 2.** Example images of the used databases with synthetic random object occlusions of the used sizes. The occlusions are only augmented during training. Testing is done with the original (mostly occlusion-free) images.

**Object Occlusions:** The PASCAL VOC 2011 [4] dataset is used to augment real objects (excluding faces) on face images. After several experiments, we selected to occlude each training image with two objects, because this resulted in better performance than using one object. Some exemplary occlusion masks and how much occlusion they create on training images can be seen in Fig. 2.

The occlusions do *not* resemble realistic occlusions, such as occlusions by hands, glasses, or other objects, because our goal is providing a simple regularization technique. Synthesizing realistic occlusions is a complex task, hard to implement, and – as our experiments show – not necessary for improving the performance.

Similar to other data augmentation techniques and batch normalization, occlusion augmentation increases the variance of the input and teaches the network to be more robust to variations. It encourages the network not to base its decisions exclusively on few local activations, but to combine multiple indicating activations globally. The size of the occlusion is a critical parameter: Occluding more pixels increases the variation of the training images and thus the regularization effect. However, occluding more pixels also hides more information that may be needed for a correct prediction.

The occlusion augmentation can be used with any CNN architecture and training loss. Further, it can be combined with any other regularization technique and be implemented as an extension of an arbitrary data augmentation pipeline. Although it is a simple approach, it is effective in improving the performance, as we will see in the following experiments.

# 4 Experiments and Results

To show the regularization effect of occlusion augmentation we use three facial expression datasets: Bosphorus [18], RAF [12], and AffectNet [16]. We present experiments on varying the degree of occlusion augmentation using three CNN models: Xception [2], MobileNet [6], and a custom architecture. This way we verify occlusion regularization with both, standard models pre-trained on ImageNet [3] and a custom model with random (Xavier) weight initialization. All three models are trained using both black-bar and object occlusion augmentation.

The custom model architecture contains six convolution layers (kernel $3 \times 3$, 16/32/.../512 channels, ReLU), all except the first followed by MaxPoolig2D (pool size $2 \times 2$). After the convolution part, we flatten the features, apply dropout ($p = 0.2$), and append the final dense layer, using softmax activation for classification and linear activation for regression. The image size used for Xception and MobileNet is $128 \times 128 \times 3$ and $100 \times 100 \times 3$ for the custom model. We conduct the experiments with the Keras deep learning framework. The occlusions are augmented with custom Python source code (using OpenCV).

## 4.1 Black-Bar Occlusion Regularization

The **Bosphorus Database** [18] contains 2,902 images, each with 26 facial action unit (AU) intensity labels. The images of 87 subjects (2,470 samples) are used for training and 17 subjects' faces (432 samples) are used as test images. We align all the training and test images with a similarity transform using facial landmarks provided with the database. The Xception and MobileNet networks are trained with classification loss (categorical cross-entropy) and the custom model with regression loss (mean squared error), because we want to verify that occlusion regularization works with both classification and regression. Figure 3 illustrates the performance improvements using black-bar occlusion regularization. The baselines (occlusion size of zero, i.e. no occlusion) are the average of three runs for each of the models. The y-axis in the plot presents the average of the 26 AUs' ICC(3,1) values [19] of the models on the test data (0 corresponds to chance level and 1 to error-free prediction) and the x-axis presents different occlusion sizes used in the respective training runs. It can be clearly seen from the plots that the performance of the models improves as the size of the occlusion mask increases until some range. Then, it starts to decline again. This is as to be expected, because at some occlusion size the negative impact of hiding information starts to outweigh the positive effect of regularization.

The **Real-World Affective Faces** Database (RAF-DB) [12] is a large database with around 30,000 diverse real-world face images downloaded from the internet. All images were annotated with basic or compound emotions by 40 trained annotators. Images with basic emotion expressions were used for experiments (including 12,271 training images and 3,068 test images). The RAF database provides both original images and aligned images; we use aligned face images for our experiments. Figure 4 shows the model performance improvements

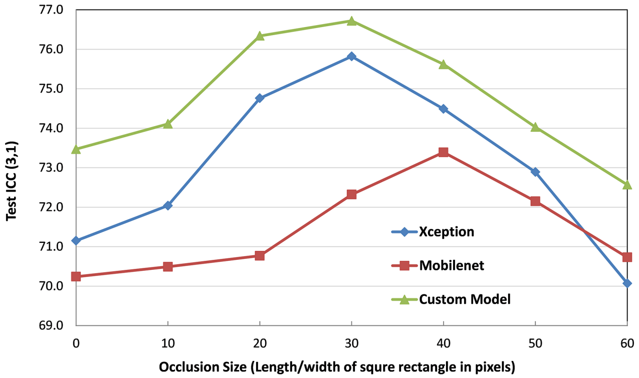

**Fig. 3.** Performance on Bosphorus (black-bar occlusion). An occlusion size of 0 corresponds to baseline results, which are outperformed by augmenting mid-size occlusions.

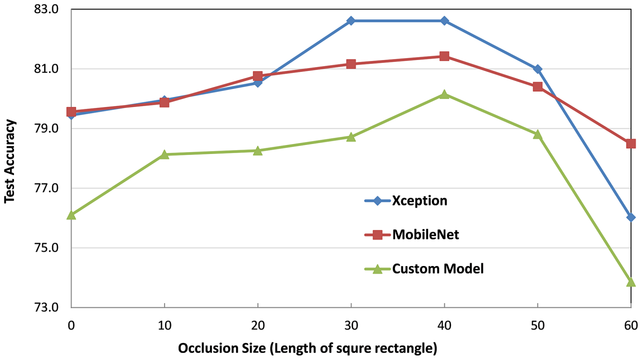

**Fig. 4.** Performance on RAF (black-bar occlusion). An occlusion size of 0 corresponds to baseline results, which are outperformed by augmenting mid-size occlusions.

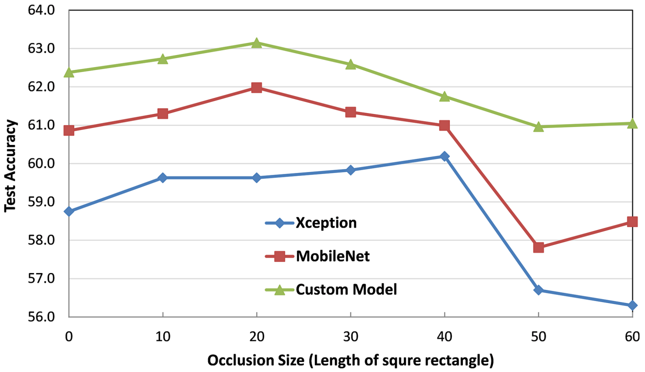

**Fig. 5.** Performance on AffectNet (black-bar occlusion). An occlusion size of 0 corresponds to baseline results, which are outperformed by augmenting mid-size occlusions.

using black bar occlusion regularization. The y-axis shows the test accuracy of the model on non-occluded test images and the x-axis shows the different occlusion sizes used for occlusion regularization. The curves of all three models are qualitatively similar to those obtained with the Bosphorus database, i.e. the performance increases with the occlusion size up to a certain point and decreases if the occlusion size is increased further.

**AffectNet** [16] is the largest labeled expression recognition database by far: It contains around 400,000 manually annotated facial images. For this work, 99,852 training images and 2,549 test images were selected randomly to reduce the required computational effort. The AffectNet database provides aligned faces that we directly used in our experiments. The performance plots are depicted in Fig. 5. Again, we observed performance improvements for all models up to a tipping point, after which performance decreases again.

### 4.2 Object Occlusion Regularization

We repeat all the above experiments using the object occlusion augmentation on the training images. Then, model performances are tested on the test data, which are mainly free of occlusions. We present results with two occlusion masks per image unlike a single mask in black bar occlusion augmentation, because we found a better regularization effect compared to a single mask. The performance plots are qualitatively similar to black-bar occlusion regularization. So we show the numbers in Tables 1, 2, and 3. Again, occlusion-based regularization outperforms the training without occlusions, at least up to a certain size of occlusion (see the bold numbers in the tables).

On the Bosphorus dataset the best object augmentation resulted in performance improvements of 3.3%, 3.7%, and 0.7% for the Xception, MobileNet, and custom architectures, respectively. These are lower than the improvements of the best black-bar augmentation, which are 5.6%, 4.3%, and 3.2%.

On RAF the performance improvements of the best object augmentation are 2.3%, 3%, and 0.7% for the Xception, MobileNet, and custom architectures, respectively. With black-bar augmentation we get improvements of 4.7%, 1.7%, and 1.6%, which is better on average.

In contrast to the Bosphorus and RAF databases, object occlusion augmentation performs better than black-bar augmentation on the AffectNet database, with performance improvements of 2.7%, 1.7%, and 2.9% (Xception, MobileNet, and custom architecture), compared to 2.6%, 1.1%, and 0.8%.

### 4.3 Comparison with State of the Art

We compare our results with the state of the art, although beating it is not the focus of our work. Table 4 shows the comparison on the Bosphorus database (mean of ICC measures of 26 AUs' intensity outputs). It can be seen from the table that we outperform the previous state of the art on this database clearly.

We also compare our best test accuracy on the RAF dataset with the existing state of the art. Table 5 shows that we obtain comparable results. Since occlusion

**Table 1.** Test performance achieved on Bosphorus datasets with different CNN models (columns) by augmenting the training data with synthetic **object occlusions** (rows).

| Training | Bosphorus Database (ICC(3,1)) | | |
|---|---|---|---|
| Occlusions | Xception | MobileNet | Custom |
| No Occlusions | 0.712 | 0.702 | 0.735 |
| Two 10 × 10 | **0.745** | **0.724** | **0.736** |
| Two 20 × 20 | **0.715** | **0.739** | **0.739** |
| Two 30 × 30 | **0.721** | **0.703** | **0.742** |
| Two 40 × 40 | 0.675 | 0.660 | 0.721 |
| Two 50 × 50 | 0.608 | 0.569 | 0.698 |

**Table 2.** Test performance achieved on RAF datasets with different CNN models (columns) by augmenting the training data with synthetic **object occlusions** (rows).

| Training | RAF Database (Accuracy in %) | | |
|---|---|---|---|
| Occlusions | Xception | MobileNet | Custom |
| No Occlusions | 79.5 | 79.6 | 76.1 |
| Two 10 × 10 | **80.0** | **78.8** | **76.3** |
| Two 20 × 20 | **80.4** | **81.3** | **76.2** |
| Two 30 × 30 | **81.0** | **81.2** | **76.8** |
| Two 40 × 40 | **81.1** | **82.6** | **76.6** |
| Two 50 × 50 | **81.8** | **80.9** | 75.4 |

**Table 3.** Test performance achieved on AffectNet datasets with different CNN models (columns) by augmenting the training data with synthetic **object occlusions** (rows).

| Training | AffectNet Database (Accuracy in %) | | |
|---|---|---|---|
| Occlusions | Xception | MobileNet | Custom |
| No Occlusions | 58.8 | 60.9 | 62.4 |
| Two 10 × 10 | **61.2** | **61.8** | **63.8** |
| Two 20 × 20 | **60.3** | **61.2** | **63.3** |
| Two 30 × 30 | **63.5** | **61.1** | **63.4** |
| Two 40 × 40 | **63.3** | **62.6** | **64.1** |
| Two 50 × 50 | 58.4 | **61.3** | **62.1** |

**Table 4.** State of the art on Bosphorus database.

| Model | Test ICC(3,1) |
|---|---|
| Easy Ensemble [15,21] | 0.340 |
| SVR Ensemble imbalanced [21] | 0.553 |
| SVR Ensemble balanced [21] | 0.533 |
| SVR Ensemble MIDRUS [21] | 0.603 |
| **Custom model with black-bar occlusion regularization** | **0.767** |

**Table 5.** State of the art on RAF database.

| Model | Test Accuracy in % |
|---|---|
| VGG16 [20] | 80.96 |
| DLP-CNN [25] | 80.89 |
| pCNN [13] | 81.64 |
| GAN-Inpainting [23] | 81.87 |
| Xception [2] | 79.45 |
| **Xception with black-bar occlusion regularization** | **82.61** |
| gCNN [13] | 83.05 |
| gACNN [13] | 85.07 |

regularization is working for all different datasets and models used in this work, we think that gCNN and gACNN can be further improved if these models are trained with additional occlusion regularization. For the AffectNet we did not use the full dataset, so comparison with other works is not fair.

## 5 Conclusion

We proposed and evaluated the idea of using occlusion augmentation for regularization in order to improve performance in facial expression recognition. Two types of occlusion augmentation were considered: black-bar occlusion and object occlusion. With both we found significant performance improvements (compared to not using occlusion augmentation) on three databases, three CNN architectures, two recognition tasks (basic emotions and AU intensities), and two loss functions (softmax cross-entropy and mean squared error). On the Bosphorus and RAF databases we observe greater performance improvements using black-bar than object occlusion regularization. On the AffectNet database it was vice versa. Due to the consistent improvements, we strongly recommend to integrate occlusion regularization into the training of all CNN-based facial expression recognition systems. We propose to use black-bar regularization (which is easy to implement and yields good results) with a square size in range of about 20–40% of the CNN input image size.

Future work should investigate occlusion augmentation and occlusion regularization in further experiments. Adding more randomization (e.g. regarding size and aspect ratio) may be a promising direction. Another approach is to randomly select for each image, whether the occlusion augmentation should be applied (leaving a subset of the images occlusion-free). Moreover, an algorithm may be developed which can automatically find the best occlusion augmentation for a particular database by searching the parameter space.

**Acknowledgments.** This work was funded by the German Federal Ministry of Education and Research (BMBF), project HuBA (03ZZ0470). The sole responsibility for the content lies with the authors.

## References

1. Bengio, Y., et al.: Deep learners benefit more from out-of-distribution examples. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 164–172 (2011)
2. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
4. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
5. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv: 1207.0580 (2012)
6. Howard, A.G., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861 [cs.CV] (2017)
7. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv: 1502.03167 (2015)
8. Kukacka, J., Golkov, V., Cremers, D.: Regularization for deep learning: A taxonomy. arXiv: 1710.10686 [cs.LG] (2017)
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
10. Lemley, J., Bazrafkan, S., Corcoran, P.: Smart augmentation learning an optimal data augmentation strategy. IEEE Access **5**, 5858–5869 (2017)
11. Li, S., Deng, W.: Deep facial expression recognition: A survey. arXiv: 1804.08348 [cs.CV] (2018)
12. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2852–2861 (2017)
13. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. **28**(5), 2439–2450 (2018)

14. Lin, F., Hong, R., Zhou, W., Li, H.: Facial expression recognition with data augmentation and compact feature learning. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1957–1961. IEEE (2018)
15. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. **39**(2), 539–550 (2009)
16. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affective Comput. **10**(1), 18–31 (2017)
17. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: Synthetic occlusion augmentation with volumetric heatmaps for the 2018 ECCV posetrack challenge on 3D human pose estimation. arXiv: 1809.04987 (2018)
18. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3D face analysis. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BioID 2008. LNCS, vol. 5372, pp. 47–56. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89991-4_6
19. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. **86**(2), 420 (1979)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556 (2014)
21. Werner, P., Saxen, F., Al-Hamadi, A.: Handling data imbalance in automatic facial action intensity estimation. In: British Machine Vision Conference (BMVC), pp. 124.1–124.12 (2015)
22. Werner, P., Saxen, F., Al-Hamadi, A., Yu, H.: Generalizing to unseen head poses in facial expression recognition and action unit intensity estimation. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2019)
23. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514 (2018)
24. Zhang, L., Verma, B., Tjondronegoro, D., Chandran, V.: Facial expression analysis under partial occlusion: a survey. ACM Comput. Surv. **51**(2) (2018). https://doi.org/10.1145/3158369
25. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3219–3228 (2017)