Priming Deep Neural Networks with Synthetic Faces for Enhanced Performance

Adam Kortylewski Andreas Schneider Thomas Gerig Clemens Blumer Bernhard Egger Corius Reyneke Andreas Morel-Forster Thomas Vetter

Department of Mathematics and Computer Science University of Basel

Abstract

Today's most successful facial image analysis systems are based on deep neural networks. However, a major limitation of such deep learning approaches is that their performance depends strongly on the availability of large annotated datasets. In this work, we prime deep neural networks by pre-training them with synthetic face images for specific facial analysis tasks. We demonstrate that this approach enhances both the generalization performance as well as the dataset efficiency of deep neural networks. Using a 3D morphable face model, we generate arbitrary amounts of annotated data with full control over image characteristics such as facial shape and color, pose, illumination, and background. With a series of experiments, we extensively test the effect of priming deep neural networks with synthetic face examples for two popular facial image analysis tasks: face recognition and facial landmark detection. We observed the following positive effects for both tasks: 1) Priming with synthetic face images improves the generalization performance consistently across all benchmark datasets. 2) The amount of real-world data needed to achieve competitive performance is reduced by 75% for face recognition and by 50% for facial landmark detection. 3) Priming with synthetic faces is consistently superior at enhancing the performance of deep neural networks than data augmentation techniques. Furthermore, our experiments provide evidence that priming with synthetic faces is able to enhance performance because it reduces the negative effects of biases present in real-world training data.

1. Introduction

Facial image analysis tasks such as face recognition and facial landmark detection have gained a lot of attention from the computer vision community. In recent years, advances in deep learning [33] and the availability of large-scale datasets have enabled a large increase in the performance of facial image analysis systems [57, 51, 47]. However, relying on large-scale datasets to enable successful facial image analysis is a limiting factor for several reasons: 1) Current large-scale datasets were mostly collected from the web and therefore have biases regarding several facial properties such as e.g. the head pose or the illumination. It is well known that such biases have a strong negative effect on the generalization performance of machine learning systems [59, 32, 58]. 2) The manual annotation of additional data, which prevents unbiasing of available datasets as well as the collection of data for novel facial image analysis tasks is expensive. 3) Large datasets of faces often cannot be released publicly due to privacy issues [51].

A promising approach for overcoming the data collection problem in computer vision is to use simulated datasets [46, 26, 6, 9, 16, 43]. Data synthesis allows massive amount of training samples with fine-grained annotation to be generated. In a recent special issue of the IJCV [15] several works demonstrated the "reasonable effectiveness of synthetic visual data" e.g. for autonomous driving [42], indoor scene understanding [28] or 3D object detection [36]. Although the approach of using synthetic data for training has been widely applied, it has not received much attention in the context of facial image analysis. Only Kim et al. [30] have collected the first evidence that synthetic data is useful for facial image analysis in the context of 3D face reconstruction, where no real-world ground truth data is available for training. In this work, we address the question: Can we use synthetic face images to enhance the performance of deep neural networks when performing facial image analy-
sis tasks?

We present an approach for generating synthetic face images with detailed annotations. We evaluate the effects of priming deep neural networks with synthetic faces for a specific facial image analysis task, before training them with real-world data. Our experiments are carried out on two popular facial analysis tasks: face recognition and facial landmark detection.

For the synthesis of face images we use a 3D Morphable Face Model (3DMM) [5, 17]. The 3DMM is a statistical model of 3D faces which has been widely applied to facial image analysis in an analysis-by-synthesis setting. The generation of novel synthetic face images is a basic capability of the 3DMM, which allows us to create possibly infinite datasets with any desired type of annotation with arbitrary depth (number of samples per identity) and width (number of identities). Additionally, synthetic face images can be released publicly for research purposes. The advantage of using 3DMMs for data synthesis over related generative face models such as e.g. GANs [4, 19] is that the 3DMM provides full control over disentangled parameters that change the facial identity in the terms of shape and albedo texture as well as pose, illumination and facial expression.

In our experiments, we pre-train deep neural networks with synthetic face images and subsequently fine-tune them with varying amounts of real-world data. In our analysis we observed the following positive effects when performing face recognition and facial landmark detection tasks:

1. **Enhanced data-efficiency.** The amount of real-world data needed to achieve competitive performance is reduced considerably.

2. **Enhanced generalization performance.** The generalization performance of deep neural networks is enhanced consistently across all benchmark datasets.

3. **Superiority when compared to data augmentation and transfer learning approaches.** Priming deep neural networks with synthetic faces has a considerably more positive effect on performance when compared to general techniques such as data augmentation and transfer learning from classification tasks.

Finally, we demonstrate that deep learning systems for facial image analysis strongly benefit from priming with synthetic face images in terms of generalization performance and real-world data efficiency. Since our approach proves to be effective for two very different facial image analysis tasks, we conclude that other facial analysis tasks might also benefit from priming. Curiously, despite the success of 3D Morphable Face Models at facial image analysis, we are not aware of any previous work that uses this effective and easily accessible approach for training face recognition or facial landmark detection systems.

2. Related Work

**Deep face recognition.** The performance of face recognition systems significantly increased with the introduction of deep convolutional neural networks [44, 34]. In the last years, major performance gains could be achieved by increasing the amount, quality and availability of training data. Taigman et al. [57] used four million images to train a Siamese network architecture. Further architectural changes have since steadily increased face recognition performances such as DeepID 1-3 [56, 53, 54] or the pose-aware model of Masi et al. [38].

A particularly successful network architecture was proposed by Schroff et al. [51], who introduced an innovative triplet-loss to their FaceNet model. After training with a large-scale private dataset, they achieved a considerable increase in state-of-the-art face recognition performance. For our face recognition experiments we use a publicly available implementation of this FaceNet network ⁶.

**Deep facial landmark detection.** The detection of fiducial points is one of the most important tasks in facial image analysis. Common approaches include regression-based [63, 7, 29] and model-based [12, 40] methods. The former learns local updates given an initial alignment and the latter is based on adapting a holistic face model to the target image. In recent years, facial landmark detection approaches based on deep learning have achieved remarkable levels of performance [55]. Trigeorgis et al. [60] combined a CNN with an RNN for cascaded regression and thereby jointly learned a feature representation for each fiducial point as well as the local update procedure. Multi-task learning approaches have gained recent popularity [67, 48]. They train a single deep network for the prediction of multiple facial image analysis tasks, thus enforcing the development of a holistic face representation which improves robustness and generalization across all individual tasks. In this work, we investigate whether deep learning approaches to facial landmark detection benefit from training with synthetic data. For our facial landmark detection experiments we use a publicly available implementation of the multi-task HyperFace network ⁷.

**Facial image analysis with augmented data.** Massive training datasets have been a critical component for deep learning systems to achieve ever higher performance at facial image analysis tasks. However, the collection, annotation, and publication of millions of face images have proven to be practically unfeasible. Therefore, a major branch of research has evolved around techniques to augment available data in order to increase dataset sizes artificially. Rudimentary approaches to data augmentation are geometric transformations such as mirroring [8, 64], translational shift.

---

⁶https://github.com/takiyu/hyperface

⁷https://github.com/takiyu/hyperface
Patel et al. [45] performed face image relighting in order to make illumination changes between images of the same person more robust. Hu et al. [23] composed novel face images by blending parts of different donor face images into a novel image. Masi et al. [39] used domain-specific data augmentation by augmenting the Casia dataset [65] in 3D. They aligned a 3D face shape to landmarks in each training image and subsequently modified the face with regard to shape deformation, 3D pose, and expression neutralization. This artificial augmentation process does not follow any statistical distribution and thus can lead to unrealistic training data. Furthermore, they did not change the illumination of the images, thus relying on the illumination distribution of the Casia dataset. In this work, we follow a different approach by studying the effect of using fully synthetic face images which were generated with 3DMMs, a statistical illumination model, a camera model and computer graphics.

Deep learning with fully synthetic data. Fully synthetic datasets which are generated with computer graphics have been widely used for the evaluation and training of computer vision tasks such as optical flow [6], autonomous driving systems [9], object detection [22], pose estimation [52, 43, 25] and text detection [21]. Recently, Qiu and Yuille [61] developed UnrealCV, a computer graphics engine for the assessment of computer vision algorithms for scene analysis. Following this approach, Kortylewski et al. [32] recently proposed to use synthetically generated face images for analyzing the generalization performance of different neural network architectures when performing face recognition in the virtual domain. Gaidon et al. presented Virtual KITTI [16], where they used synthetically generated data to pre-train a deep convolutional neural network for the tasks of object detection, tracking and scene segmentation in the context of automated driving systems. They showed that Deep Learning systems behave similarly when trained in the synthetic domain and evaluated in the real domain and vice-versa. In the context of facial analysis, Abbasnejad et al. [1] trained a deep convolutional neural network for expression analysis on synthetic data and achieved state-of-the-art results in action unit classification on real data. Kim et al. [30] trained an AlexNet architecture for the regression of 3D morphable model parameters and achieved competitive results when compared to approaches that learn from real-world data. Curiously, despite the widespread use of synthetic data for deep learning, we are not aware of any work that studied how synthetic data can be leveraged to enhance face recognition and facial landmark detection in real-world images.

In this work, we use a statistical model of 3D face shape, texture, expression and illumination to generate synthetic face images, which we use to train deep neural networks for face recognition and facial landmark detection.

3. Face Image Generator

We synthesize face images by sampling from a statistical 3D Morphable Model [5] of face shape, color and expression. For data synthesis, we use the generator proposed by Kortylewski et al. [32] and extend it with facial expressions as well as a natural illumination prior. In the following, we describe the most important parameters of the model and their influence on the facial appearance in the image.

Facial identity. In order to simulate different facial identities we use the Basel Face Model 2017 [18] (BFM). The BFM describes a statistical distribution of face shape, color and expression which was learned from 200 neutral high-resolution 3D face scans. The parameters follow a Gaussian distribution. By drawing random samples from this distribution we generate novel 3D face meshes with unique color, shape and facial expression.

Illumination. We assume the Lambertian reflectance model and approximate the environment map with 27 spherical harmonics coefficients (9 for each color channel). In order to generate natural illuminations, the spherical harmonics illumination parameters are sampled from the Basel Illumination Prior (BIP) [13]. The BIP describes an empirical distribution of spherical harmonics coefficients estimated from 14’348 real-world face images.

Pose and camera. The simulated faces are viewed with a fixed pinhole camera. The facial orientation with respect to the camera is controlled by six pose parameters (yaw, pitch, roll, and 3D translation). Throughout our experiments, we vary the head position and pose angles, while normalizing the head position to the center of the image frame.
Background. We simulate changes in the background with a non-parametric background model by sampling randomly from a set of background images of the describable texture database [11]. The purpose of these random structured background changes is to aid deep learning systems in discovering the irrelevance of background structures for facial image analysis tasks.

The synthesized images (Fig. 1) are fully specified by the aforementioned parameter distributions. Note that our data generation process follows a statistical distribution of face shapes and textures which is in contrast to, for example, the data augmentation in [39], where shape deformation between a few fixed 3D shapes is performed. The major benefit of the generator is that by sampling from its parameters we are able to synthesize an arbitrary amount of face images with any desired number of identities, with expressions, in different head poses and natural illumination conditions.

Software. The synthetic face image generator is publicly available.

**4. Experiments**

In this section, we empirically analyze the effect of priming deep neural networks with synthetic face images. After introducing our experimental setup (Section 4.1), we study the performance of deep networks when performing face recognition and facial landmark detection trained using a combination of real and synthetic data (Section 4.2). We compare our approach to traditional data augmentation methods and to transfer learning from a pre-trained network in Section 4.3. Finally, we study how changing the pose distribution and the number of identities in the synthetic data affects the generalization performance of the primed models (Section 4.4).

**4.1. Experimental Setup**

**Face recognition.** Our face recognition experiments are based on the publicly available OpenFace framework [2]. For face detection and alignment we use a publicly available multi-task CNN [46]. In case the face detection fails, we use the face boxes as defined in the individual datasets [5]. We train the FaceNet-NN4 architecture that was originally proposed by Schroff et al. [51] with the vanilla setting, as provided in the OpenFace framework. The aligned images are scaled to 96 × 96 pixels. The triplet loss is trained with batches of 20 identities and 15 sample images per identity for 200 epochs.

The real-world training data for face recognition is sampled from the cleaned Casia WebFace dataset [65], which comprises 455,594 images of 10,575 different identities.

From this dataset, we remove the 27 identities which overlap with the IJB-A dataset. For testing the generalization performance we use: 1) CMU-Multi-PIE [20], which was recorded under controlled illumination and background conditions. We use the neutral identities from session one with the frontal illumination setting. Images from the two overhead cameras are excluded. 2) LFW [24] has been the de facto standard face recognition benchmark for many years. Face images in this dataset are subject to a complex illumination, partial occlusion, and background clutter. 3) IJB-A [31] was proposed to further push the frontiers of face recognition. The conditions regarding pose, illumination and partial occlusion are more complex than LFW. In addition, subjects are possibly described by multiple gallery images, these image sets are commonly referred to as templates. We evaluate face recognition networks at the task of face verification. We measure the distance between two face images as the cosine distance between their 128-dimensional feature embeddings from the last layer of the FaceNet model.

\[ s(a, b) = \frac{a \cdot b}{\|a\|_2 \|b\|_2}. \] (1)

For comparing the templates in the IJB-A dataset, we perform softmax averaging of the similarity scores between each image pair as proposed by Masi et al. [39]. We do not perform any dataset adaptation, thus we test the most challenging face recognition protocol with only unrestricted, labeled outside data. For LFW and IJB-A, the pairwise comparisons are provided by their respective protocols. For the Multi-PIE dataset, we follow the LFW protocol. Thus, we perform 10 fold cross validation with 300 pairs of positive and negative samples.

**Facial landmark detection.** For facial landmarking we use a publicly available implementation of the HyperFace network [6] [47]. HyperFace is a Multi-Task CNN that learns a common representation in order to perform several facial image analysis tasks jointly: Face detection, pose and gender estimation, and facial landmark detection. For training the HyperFace model, we use the exact training setup as proposed in its open source implementation. Unless otherwise stated, the Neural Networks are randomly initialized with Gaussian noise of zero mean and unit variance.

The real-world data used for training is sampled from the AFLW [37] dataset. We randomly select 1K images from the AFLW set for testing and use the rest for training. For benchmarking we use three facial landmarking datasets, all of which contain large variations in pose, illumination, facial occlusion and expression: 1) Our test split from the AFLW set. 2) The test set of the LFPW [3] dataset. 3) The 300-W [50] dataset which has been introduced as part of a challenge for “in the wild” facial landmark detection and is

---

4https://github.com/kpzhang93/mtcnn_face_detection_alignment
5For LFW and IJB-A these face boxes are provided in the dataset, for Multi-PIE we use the annotations provided in [14].

6https://github.com/takiyu/hyperface
current one of the most challenging datasets available for facial landmark detection. We combine all 600 images from the indoor and outdoor scenes into our test set.

In our experiments, we compare the 21 AFLW-landmarks available in AFLW and 300-W. For LFPW we use the 20 landmarks which overlap with the AFLW set. We compute the detection accuracy as the mean euclidean distance between the prediction and the ground truth annotation, normalized according to the face size.

**Synthetic face image generation.** The synthetic face images used for training are generated by randomly sampling the parameters of our face image generator (Section 3). The shape, color and expression parameters are sampled according to the Gaussian distribution defined by the BFM [18], while the parameters for the illumination are sampled from the empirical BIP [13]. The head pose is sampled according to a uniform pose distribution on the yaw, pitch and roll angles in the respective ranges $r_{yaw} = [-90^\circ, 90^\circ]$, $r_{pitch} = [-30^\circ, 30^\circ]$ and $r_{roll} = [-15^\circ, 15^\circ]$. Background clutter is simulated by randomly sampling a background texture as explained in Section 3. For face recognition, we generate one million face images with $10K$ different identities and 100 example images per identity. For landmark detection we use $50K$ training images with $10K$ identities and five sample per identity. The number of synthetic images per task were determined empirically. In Section 4.4 we present an additional experimental analysis of the effect of changing these dataset characteristics.

### 4.2. Effectiveness of Priming with Synthetic Faces

In the following experiments, we study how priming deep neural networks with synthetic faces affects the generalization performance and data efficiency.

**Learning from synthetic data only.** When training with synthetic data only (SYN-only), we observe that for the CMU-Multi-PIE benchmark the performance is similar to that of a network trained with real-world data (Real-100%). This suggests that our synthetic face image generator can sufficiently model the facial appearance in constrained visual environments. However, on the benchmarks of LFW and IJB-A the SYN-only model performs worse when compared to a network trained with the full real-world dataset. Hence, a prominent real-to-virtual performance gap can be observed when testing with “in the wild” images.

A similarly prominent real-to-virtual performance gap can be observed for the task of facial landmark detection (Table 2). The SYN-only model performs worse when compared to the one trained with real-world data (Real-100%)
Table 3: Comparison of the face recognition performance when initializing the weights of our model randomly to training with augmented data, and to priming with synthetic face images. We measure the performance when using 50% and 100% of the cleaned Casia dataset for training. We test using the CMU Multi-PIE, LFW and IJB-A benchmarks. While the effect of data augmentation is marginal, priming with synthetic face images leads to a consistently enhanced generalization performance.

| Datasets | Multi-PIE | LFW | IJB-A |
|----------|-----------|-----|-------|
| # real data | 50% 100% | 50% 100% | 50% 100% |
| Random | 86.0 91.2 | 92.7 94.1 | 81.9 86.8 |
| Augmented | 87.7 91.3 | 91.3 94.6 | 82.8 86.7 |
| Primed | 93.0 93.3 | 94.8 95.8 | 88.2 90.6 |

Table 4: Comparison of the facial landmark detection performance when initializing the weights of our model randomly to training with augmented data, and to priming with synthetic face images. We measure the performance when using 50% and 100% of the AFLW training data. For the AFLW and LFPW datasets we show the detection accuracy at 3% error and for 300-W at 5% error. Priming with synthetic data leads to a consistent superior performance.

| Datasets | AFLW | LFPW | 300-W |
|----------|------|------|-------|
| # real data | 50% 100% | 50% 100% | 50% 100% |
| Random | 43.8 47.4 | 57.5 63.1 | 75.6 78.2 |
| Augmented | 44.9 49.7 | 61.2 68.8 | 74.3 84.3 |
| ImageNet | 46.3 47.6 | 65.2 65.3 | 79.8 85.1 |
| Primed | 47.8 50.9 | 68.5 70.5 | 82.2 87.8 |

Across all benchmarks.

These measurements complement a number of previous works which studied the effectiveness of using synthetic data for training computer vision models. These works show that for image analysis tasks such as e.g. object detection [22] and optical flow [6] synthetic data could very well replace real-world data, whereas, for example, for semantic segmentation [41] a real-to-virtual performance gaps exists, that can be closed by fine-tuning with real-world data.

For the context of facial image analysis, our findings demonstrate that so far it is not possible to fully replace real-world data with synthetic face images. A main reason for the persisting performance gap between models trained with real-world and synthetic face images is that some important characteristics of faces are currently not modeled with the parametric model. These include e.g. facial hair, partial occlusion, the mouth interior and detailed skin textures.

**Fine-tuning with real-world data.** In order to bridge the real-to-virtual performance gap we fine-tune the SYN-only model with different subsets (10%, 25%, 50%, 100%) of the real-world training data. In this way, the synthetic data will prime the model towards the target facial image analysis task, enabling the model to leverage the information in the real-world data more efficiently during the fine-tuning process. The performance of the primed models is denoted as “+ Primed” in Tables 1 & 2. The primed models considerably outperform the unprimed models at both tasks: face recognition as well as facial landmark detection. Interestingly, even when fine-tuning with the full real-world datasets the models still have an enhanced performance compared to the unprimed models. Importantly, when less real-world data is available the overall increase in generalization performance becomes even more apparent.

At the task of face recognition, the real-to-virtual gap can almost be closed with 25% of the real data, already outperforming the Real-100% model with 50% of the real-world dataset. Remarkably, priming with synthetic data leads to a performance increase across all benchmarks, even though the individual datasets have very different imaging characteristics. Note that we do not perform any dataset adaptation in our experiments.

For facial landmark detection we observe similar effects when priming with synthetic data. The real-to-virtual gap can mostly be closed with 50% of the real data. Large performance gains are achieved at small error rates of 3% compared to the unprimed models. Similarly, to the task of face recognition we observe a strong generalization performance across datasets without performing dataset adaptation. This can be observed best from the results on the challenging 300-W benchmark. Note that there is a domain gap between the AFLW training data and the 300-W test data, which manifests in the worse performance of the Real-100% model compared to the results of the LFPW benchmark. Using our approach this performance gap is constantly improved by about 7−9% at an error rate of 5% for all dataset sizes.

In summary, our results in Tables 1 & 2 show that priming with synthetic faces has the following positive effects for both face recognition and facial landmark detection tasks:

1. A compelling increase in the generalization performance is achieved when priming with synthetic data before training with 100% of the real data.
2. After priming, only 25% − 50% of the real-world data is needed to achieve the same performance as is obtained when training an unprimed model with the full
real-world dataset.

3. Priming with synthetic data enhances the generalization performance consistently across all benchmark datasets.

### 4.3. Comparison to data augmentation and transfer learning

In this Section, we compare our priming approach to related techniques for enhancing the generalization ability of deep neural networks when performing facial image analysis. A popular approach to increase the training data (and thus the performance) is data augmentation. Thereby, the original training data is manipulated with general image transformations such as color changes, 2D rotations and mirroring [8, 64, 62, 10]. Another widely applied tool for performance increase is transfer learning from image classification, where a model is pre-trained on a general object classification task and subsequently fine-tuned on the desired facial image analysis task. In this way, the model can re-use the knowledge acquired for discriminating objects and transfer it to facial image analysis.

In our experimental setup, we augment training data by mirroring across the vertical axis and by rotating a face image twice around the center of the face randomly in a range of $[-30^\circ, 30^\circ]$. In this way, we augment the head pose distribution in the training data. For pre-training, we use an AlexNet [33, 27] that was trained for image classification on the ImageNet dataset [49]. We subsequently fine-tune this model for facial landmark detection. We could not test this setup for face recognition, as there are no pre-trained FaceNet architectures available. As a general performance baseline, we use a model with randomly initialized weights.

In Tables 3 & 4, we summarize the performance results for all tested approaches when using 50% of the real-world training data and when using all of the available data for fine-tuning. We observe that data augmentation with 2D image transformations has only a marginal effect on face recognition. Apparently, the large-scale face recognition dataset already offers enough data to generalize well across the different head poses that were simulated during data augmentation. For facial landmark detection we observe a positive effect, due to the smaller scale of the training dataset. At this task, pre-training seems to have a more positive effect than data augmentation, when only 50% of the training data is used.

Our approach of priming the models with synthetic face images is superior to data augmentation with 2D image transformations and transfer learning from a classification task. In our results the performance of face recognition and facial landmark detection is enhanced with any amount of available real-world training data. The major advantage of our approach is that it offers face-specific knowledge that goes beyond the information offered by general techniques such as data augmentation and transfer learning. It is also noteworthy that all three approaches are to some extent complementary, and can thus be combined.

### 4.4. Effect of Changing the Synthetic Data Distribution

In this section, we change the pose-distribution and the number of identities in the synthetic dataset and study how
this affects the generalization performance when performing face recognition and facial landmark detection tasks. Note that we change one dataset characteristic at a time while keeping all other parameters fixed to the original values of our previous experiments. The experimental results are summarized in Tables 5 & 6.

**Bias to frontal pose.** We bias the yaw pose in the synthetic data to a range of $r_{yaw} = [-35^\circ, 35^\circ]$, while keeping the amount of data the same as in the original SYN dataset. In our experiments, we prime the models with the biased dataset and subsequently fine-tune with the full real world datasets (“primed-frontal” in Tables 5 & 6). From the results, we can observe a performance decrease at both facial image analysis tasks compared to priming with the original unbiased dataset (“primed”). This negative effect is particularly prominent for the most challenging datasets, IJB-A and 300-W. This decrease in the generalization performance demonstrates that deep neural networks do benefit from the large pose variation in the synthetic data, suggesting that conceptual knowledge about 3D head pose variation is learned and transferred to the real-world data.

**Increasing the number of training identities.** We double the number of identities in the synthetic data to $20K$, thus generating double the amount of training images for each facial image analysis task. It is important to note that such an increase in the training data would require a great effort for real-world data, whereas for synthetic data it is essentially cost-free. After priming with the new dataset and subsequently fine-tuning on the full real-world datasets, we observe an additional increase in the generalization performance when performing either facial image analysis task (“primed-double” in Tables 5 & 6). The performance increase is particularly prominent for face recognition, whereas for facial landmark detection it is marginal, suggesting that more complex facial image analysis tasks benefit more from additional synthetic data. Nevertheless, these results confirm our observations that priming with synthetic face images increases the generalization performance of facial image analysis tasks. Additionally, the results suggest that the generalization performance could be increased even more when additional synthetic training data is used.

In summary, the experiments in this section demonstrate that the additional variability in the synthetic data, in terms of pose and facial identities, is an important factor for the overall increase in the generalization performance of real-world data.

5. Discussion

In this work, we demonstrated the positive effects of priming deep neural networks with synthetic face images for face recognition and facial landmark detection. In our experiments, we observed the following phenomena for both facial image analysis tasks:

**Enhanced generalization performance.** Priming with synthetic data followed by fine-tuning with real-world data enhances the generalization performance consistently across all benchmark datasets compared to training with real-world data only. Our experiments in Section 4.4 show that the additional variability in the synthetic data, in terms of head pose and facial identity, is a crucial factor for the measured increase in generalization performance. Thus, we provide evidence that the negative effects of biases in the real-world data are alleviated when priming with synthetic data that is unbiased in these variables.

**Enhanced data efficiency.** The improvement in generalization performance becomes more prominent, as we reduce the amount of real-world data available at training time. More importantly, the smaller the amount of real-world training data, the larger is the improvement in performance, compared to models that were not primed. This result is intuitive since with increasing number of real-world data, any algorithm will eventually reach the optimal accuracy which is possible under the model class. Hence, the role of priming with synthetic face images is to provide knowledge about those facial appearance variations that are simulatable by our data generator. Only relatively few real-world data are then needed to learn the image variation which currently cannot be simulated well. Using our priming approach the number of real-world data needed to achieve competitive performance was reduced by 75% for face recognition, and by 50% for facial landmark detection.

**Superiority to data augmentation and transfer learning from classification tasks.** Common ways of improving the sample efficiency and generalization performance of deep neural networks are data augmentation and transfer learning from large-scale object classification tasks. Our experiments show that the proposed method of priming with synthetic face images has a considerably more positive effect on the performance of deep neural networks than data augmentation and transfer learning from a classification task. This observation is intuitive because our approach provides additional information that is specific to the target facial image analysis tasks. This kind of knowledge cannot be learned from a general object classification task nor when the training data is augmented with basic image transformations.

Notably, our approach of priming with synthetic face images is beneficial for two very different facial image analysis tasks: face recognition and facial landmark detection. We conclude from this observation that other facial image analysis tasks, such as face segmentation or pose estimation, could also potentially benefit from priming. In order to support future research in this direction our synthetic data generator has been made publicly available.
Acknowledgment

Adam Kortylewski is supported by a Novartis University of Basel Excellence Scholarship for Life Sciences. We gratefully acknowledge the support of NVIDIA with the donation of a Titan Xp and a Quadro P5000.

References

[1] Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., Lucey, S.: Using synthetic data to improve facial expression analysis with 3d convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1609–1618 (2017) 3

[2] Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science (2016) 4

[3] Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE transactions on pattern analysis and machine intelligence (2013) 4

[4] Berthelot, D., Schumm, T., Metz, L.: Began: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017) 2

[5] Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1995) 2, 3

[6] Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European Conference on Computer Vision, pp. 611–625. Springer (2012) 1, 3, 6

[7] Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision (2014) 2

[8] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014) 2, 7

[9] Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2722–2730 (2015) 1, 3

[10] Chen, Y.C., Patel, V.M., Phillips, P.J., Chellappa, R.: Dictionary-based face recognition from video. In: European Conference on Computer Vision, pp. 766–779. Springer (2012) 3, 7

[11] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014) 4

[12] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models—their training and application. Computer vision and image understanding (1995) 2

[13] Egger Bernhard Schoenborn Sandro, S.A.K.A.A.M.F.C.B., Vetter, T.: Occlusion-aware 3d morphable models and an illumination prior for face image analysis. International Journal of Computer Vision (2018) 3, 5

[14] El Shafey, L., McCool, C., Wallace, R., Marcel, S.: A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition. IEEE transactions on pattern analysis and machine intelligence (2013) 4

[15] Gaidon, A., Lopez, A., Perronnin, F.: The reasonable effectiveness of synthetic visual data. International Journal of Computer Vision (2018) 1

[16] Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4340–4349 (2016) 1, 3

[17] Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models—an open framework. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, pp. 75–82. IEEE (2018) 2

[18] Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models—an open framework. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, pp. 75–82. IEEE (2018) 3, 5

[19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014) 2

[20] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing (2010) 4
[21] Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016) 3

[22] Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: European Conference on Computer Vision, pp. 345–360. Springer (2014) 3, 6

[23] Hu, G., Peng, X., Yang, Y., Hospedales, T.M., Verbeek, J.: Frankenstein: Learning deep face representations using small data. IEEE Transactions on Image Processing (2018) 3

[24] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007) 4

[25] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence (2014) 3

[26] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014) 1

[27] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014) 7

[28] Jiang, C., Qi, S., Zhu, Y., Huang, S., Lin, J., Yu, L.F., Terzopoulos, D., Zhu, S.C.: Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars. International Journal of Computer Vision (2018) 1

[29] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014) 2

[30] Kim, H., Zollhöfer, M., Tewari, A., Thies, J., Richardt, C., Theobalt, C.: Inversefacenet: Deep single-shot inverse face rendering from a single image. arXiv preprint arXiv:1703.10956 (2017) 1, 3

[31] Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1931–1939 (2015) 4

[32] Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Forster, A., Vetter, T.: Empirically analyzing the effect of dataset biases on deep face recognition systems. CVPRW (2017) 1, 3

[33] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012) 1, 7

[34] LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks (1995) 2

[35] Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–42 (2015) 3

[36] Loing, V., Marlet, R., Aubry, M.: Virtual training for a real application: Accurate object-robot relative localization without calibration. International Journal of Computer Vision (2018) 1

[37] Martin Koestinger Paul Wohlhart, P.M.R., Bischof, H.: Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In: Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011) 4

[38] Masi, I., Rawls, S., Medioni, G., Natarajan, P.: Pose-aware face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4838–4846 (2016) 2

[39] Masi, I., Tran, A.T., Hassner, T., Leksut, J.T., Medioni, G.: Do we really need to collect millions of faces for effective face recognition? In: European Conference on Computer Vision, pp. 579–596. Springer (2016) 3, 4

[40] Matthews, I., Baker, S.: Active appearance models revisited. International journal of computer vision (2004) 2

[41] McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation. In: Proceedings of the International Conference on Computer Vision (ICCV), vol. 4 (2017) 6
[42] Müller, M., Casser, V., Lahoud, J., Smith, N., Ghanem, B.: Sim4cv: A photo-realistic simulator for computer vision applications. International Journal of Computer Vision pp. 1–18 (2018) 1

[43] Park, D., Ramanan, D.: Articulated pose estimation with tiny synthetic videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 58–66 (2015) 1, 3

[44] Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC (2015) 2

[45] Patel, V.M., Wu, T., Biswas, S., Phillips, P.J., Chellappa, R.: Illumination robust dictionary-based face recognition. In: Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 777–780. IEEE (2011) 3

[46] Qiu, W., Yuille, A.: Unrealcv: Connecting computer vision to unreal engine. In: European Conference on Computer Vision, pp. 909–916. Springer (2016) 1

[47] Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017) 1, 4

[48] Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pp. 17–24. IEEE (2017) 2

[49] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (2015) 7

[50] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2013) 4

[51] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823 (2015) 1, 2, 4

[52] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 1297–1304. Ieee (2011) 3

[53] Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems, pp. 1988–1996 (2014) 2

[54] Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015) 2

[55] Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3476–3483 (2013) 2

[56] Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014) 2

[57] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701–1708 (2014) 1, 2

[58] Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias. In: Domain Adaptation in Computer Vision Applications, pp. 37–55. Springer (2017) 1

[59] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 1521–1528. IEEE (2011) 1

[60] Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4177–4187 (2016) 2

[61] Weichao Qiu Fangwei Zhong, Y.Z.S.Q.Z.X.T.S.K.Y.W.A.Y.: Unrealcv: Virtual worlds for computer vision. ACM Multimedia Open Source Software Competition (2017) 3

[62] Wolf, L., Hassner, T., Taigman, Y.: Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. IEEE transactions on pattern analysis and machine intelligence (2011) 3, 7
[63] Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 532–539 (2013) 2

[64] Yang, H., Patras, I.: Mirror, mirror on the wall, tell me, is the error small? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4685–4693 (2015) 2, 7

[65] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014) 3, 4

[66] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters (2017). DOI 10.1109/LSP.2016.2603342 4

[67] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, pp. 94–108. Springer (2014) 2