Chapter 8
DeepGender2: A Generative Approach Toward Occlusion and Low-Resolution Robust Facial Gender Classification via Progressively Trained Attention Shift Convolutional Neural Networks (PTAS-CNN) and Deep Convolutional Generative Adversarial Networks (DCGAN)

Felix Juefei-Xu, Eshan Verma and Marios Savvides

Abstract In this work, we have undertaken the task of occlusion and low-resolution robust facial gender classification. Inspired by the trainable attention model via deep architecture, and the fact that the periocular region is proven to be the most salient region for gender classification purposes, we are able to design a progressive convolutional neural network training paradigm to enforce the attention shift during the learning process. The hope is to enable the network to attend to particular high-profile regions (e.g., the periocular region) without the need to change the network architecture itself. The network benefits from this attention shift and becomes more robust toward occlusions and low-resolution degradations. With the progressively trained attention shift convolutional neural networks (PTAS-CNN) models, we have achieved better gender classification results on the large-scale PCSO mugshot database with 400 K images under occlusion and low-resolution settings, compared to the one undergone traditional training. In addition, our progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations, as well as low resolution. One way to further improve the robustness of the proposed gender classification algorithm is to invoke a generative approach for occluded image recovery, such as using the deep convolutional

F. Juefei-Xu (✉) · E. Verma · M. Savvides
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: felixu@cmu.edu

E. Verma
e-mail: everma@cmu.edu

M. Savvides
e-mail: msavvid@ri.cmu.edu

generative adversarial networks (DCGAN). The facial occlusions degradation studied in this work is a missing data challenge. For the occlusion problems, the missing data locations are known whether they are randomly scattered, or in a contiguous fashion. We have shown, on the PCSO mugshot database, that a deep generative model can very effectively recover the aforementioned degradation, and the recovered images show significant performance improvement on gender classification tasks.

8.1 Introduction

Facial gender classification has always been one of the most studied soft-biometric topics. Over the past decade, gender classification on constrained faces has almost been perfected. However, challenges still remain on less-constrained faces such as faces with occlusions, of low resolution, and off-angle poses. Traditional methods such as the support vector machines (SVMs) and its kernel extension can work pretty well on this classic two-class problem as listed in Table 8.8. In this work, we approach this problem from a very different angle. We are inspired by the booming deep convolutional neural network (CNN) and the attention model to achieve occlusion and low-resolution robust facial gender classification via progressively training the CNN with attention shift. From an orthogonal direction, when images are under severe degradations such as contiguous occlusions, we utilize a deep generative approach to recover the missing pixels such that the facial gender classification performance is further improved on occluded facial images. On one hand, we aim at building a robust gender classifier that is tolerant to image degradations such as occlusions and low resolution, and on the other hand, we aim at mitigating and eliminating the degradations through a generative approach. Together, we stride toward pushing the boundary of unconstrained facial gender classification.

8.1.1 Motivation

Xu et al. [70] proposed an attention-based model that automatically learns to describe the content of images which has been inspired by recent work in machine translation [3] and object detection [2, 57]. In their work, two attention-based image caption generators were introduced under a common framework: (1) a 'soft' deterministic attention mechanism which can be trained by standard back-propagation method and (2) a 'hard' stochastic attention mechanism which can be trained by maximizing an approximate variational lower bound. The encoder of the models uses a convolutional neural network as a feature extractor, and the decoder is composed of a recurrent neural network (RNN) with long short-term memory (LSTM) architecture where the attention mechanism is learned. The authors can then visualize that the network can automatically fix its gaze on the salient objects (regions) in the image while generating the image caption word by word.
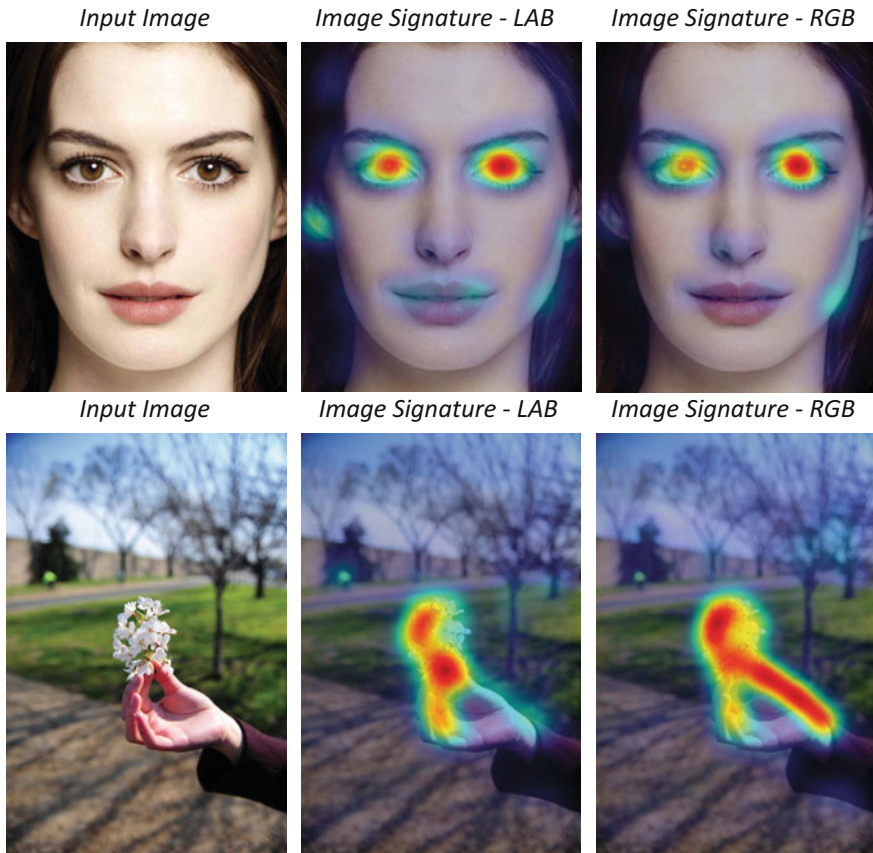
Fig. 8.1 (*Top*) Periocular region on human faces exhibits the highest saliency. (*Bottom*) Foreground object in focus exhibits the highest saliency. Background is blurred with less high-frequency details preserved

For facial gender classification, we know from previous work [16, 56] that the periocular region provides the most important cues for determining the gender information. The periocular region is also the most salient region on human faces, such as shown in the top part of Fig. 8.1, using a general purpose saliency detection algorithm [14]. Similar results can also be obtained using other saliency detection algorithms such as [15, 17]. We can observe from the saliency heat map that the periocular region does fire the most strongly compared to the remainder of the face.

Now we come to think about the following question:

> *Q: How can we let the CNN shift its attention toward the periocular region, where gender classification has been proven to be the most effective?*

The answer comes from our day-to-day experience with photography. If you are using a DSLR camera with a big aperture lens, and fixing the focal point onto an object in the foreground, all background beyond the object in focus will become out of focus and blurred. This is illustrated in the bottom part of Fig. 8.1 and as can be seen, the sharp foreground object (cherry blossom in hand) attracts the most attention in the saliency heat map.

Thus, we can control the attention region by specifying where the image is blurred or remains sharp. In the context of gender classification, we know that we can benefit from fixing the attention onto the periocular region. Therefore, we are 'forcing' what part of the image the network weighs the most, by progressively training the CNN using images with increasing blur levels, zooming into the periocular region, as shown in Table 8.1. Since we still want to use a full-face model, we hope that by employing the mentioned strategy, the learned deep model can be at least on par with other full-face deep models, while harnessing gender cues in the periocular region.

Q: Why not just use the periocular region crop?

Table 8.1 Blurred images with increasing levels of blur

| 13.33% | 27.62% | 41.90% |
| 56.19% | 68.57% | 73.33% |

Although experimentally, periocular is the best facial region for gender classification, we still want to resort to other facial parts (beard/mustache) for providing valuable gender cues. This is especially true when the periocular region is less ideal. For example, some occlusion like sunglasses could be blocking the eye region, and we want our network to still be able to generalize well and perform robustly, even when the periocular region is corrupted.

To strike a good balance between full face-only and periocular-only models, we carry out a progressive training paradigm for CNN that starts with the full face, and progressively zooms into the periocular region by leaving other facial regions blurred. In addition, we hope that the progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations.

Q: Why blurring instead of blackening out?

We just want to steer the focus, rather than completely eliminating the background, like the DSLR photo example shown in the bottom part of Fig. 8.1. Blackening would create abrupt edges that confuse the filters during the training. When blurred, low-frequency information is still well preserved. One can still recognize the content of the image, e.g., dog, human face, objects, etc. from a blurred image.

Blurring outside the periocular region, and leaving the high-frequency details at the periocular region will both help providing global and structural context of the image, as well as keeping the minute details intact at the region of interest, which will help the gender classification, and fine-grained categorization in general.

Q: Why not let CNN directly learn the blurring step?

We know that CNN filters operate on the entire image, and blurring only part of the image is a pixel location dependent operation and thus is difficult to emulate in the CNN framework. Therefore, we carry out the proposed progressive training paradigm to enforce where the network attention should be.

8.2 Related Work

In this section, we provide relevant background on facial gender classification and attention models.

The periocular region is shown to be the best facial region for recognition purposes [20–22, 24, 27, 31–36, 38, 39, 39, 40, 59, 62, 63], especially for gender classification tasks [16, 56]. A few recent work also applies CNN for gender classification [4, 50]. More related work on gender classification is consolidated in Table 8.8.

Attention models such as the one used for image captioning [70] have gained much popularity only very recently. Rather than compressing an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image. It also helps gaining insight and interpreting the results by visualizing where the model attends to for certain tasks. This mechanism can be viewed as a learnable saliency detector that can be tailored to various tasks, as opposed to the traditional ones such as [8, 14, 15, 17].

It is worth mentioning the key difference between the soft attention and the hard attention. The soft attention is very easy to implement. It produces distribution over input locations, reweights features, and feeds them as input. It can attend to arbitrary input locations using spatial transformer networks [19]. On the other hand, the hard attention can only attend to a single input location, and the optimization cannot utilize gradient descent. The common practice is to use reinforcement learning.

Other applications involving attention models may include machine translation which applies attention over input [54]; speech recognition which applies attention over input sounds [6, 9]; video captioning with attention over input frames [72]; image, question to answer with attention over image itself [69, 75]; and many more [67, 68].

8.3 Proposed Method

Our proposed method involves two major components: a progressively trained attention shift convolutional neural networks (PTAS-CNN) framework for training the unconstrained gender classifier, as well as a deep convolutional generative adversarial networks (DCGAN) for missing pixel recovery on the facial images so that the gender recognition performance can be further improved on images with occlusions.

8.3.1 Progressively Trained Attention Shift Convolutional Neural Networks (PTAS-CNN)

In this section we detail our proposed method on progressively training the CNN with attention shift. The entire training procedure involves $(k+1)$ epoch groups from epoch group 0 to k, where each epoch group corresponds to one particular blur level.

8.3.1.1 Enforcing Attention in the Training Images

In our experiment, we heuristically choose 7 blur levels, including the one with no blur at all. The example images with increasing blur levels are illustrated in Table 8.1. We use a Gaussian blur kernel with $\sigma = 7$ to blur the corresponding image regions.

Doing this is conceptually enforcing the network attention in the training images without the need of changing the network architecture.

8.3.1.2 Progressive CNN Training with Attention

We employ the AlexNet [46] architecture for our progressive CNN training. The AlexNet has 60 million parameters and 650,000 neuron, consisting of 5 convolution layers and 3 fully connected layers with a final 1000-way softmax. To reduce overfitting in the fully connected layers, AlexNet employs "dropout" and data augmentation, both of which are preserved in our training. The main difference is that we only need a 2-way softmax due to the nature of gender classification problems.

As illustrated in Fig. 8.2, the progressive CNN training begins with the first epoch group (Epoch Group 0, images with no blur), and the first CNN model $\mathcal{M}_0$ is obtained and frozen after convergence. Then, we input the next epoch group for tuning the $\mathcal{M}_0$ and in the end produce the second model $\mathcal{M}_1$, with attention enforced through training images. The procedure is carried out sequentially until the final model $\mathcal{M}_k$ is obtained. Each $\mathcal{M}_j (j = 0, \ldots, k)$ is trained with 1000 epochs and with a batchsize of 128. At the end of the training for step j, the model corresponding to best validation accuracy is taken ahead to the next iteration $(j + 1)$.

8.3.1.3 Implicit Low-Rank Regularization in CNN

Blurring the training images in our paradigm may have more implications. Here we want to show the similarities between low-pass Fourier analysis and low-rank approximation in SVD. Through the analysis, we hope to make connections to the low-rank regularization procedure in the CNN. We have learned from a recent work [65] that enforcing a low-rank regularization and removing the redundancy in the convolution kernels is important and can help improve both the classification accuracy and the computation speed. Fourier analysis involves expansion of the original data x_{ij} (taken from the data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$) in an orthogonal basis, which is the inverse Fourier transform:

$$x_{ij} = \frac{1}{m} \sum_{k=0}^{m-1} c_k e^{\mathbf{i}2\pi jk/m}. \tag{8.1}$$

The connection with SVD can be explicitly illustrated by normalizing the vector $\{e^{\mathbf{i}2\pi jk/m}\}$ and by naming it $\mathbf{v}'_k$:

$$x_{ij} = \sum_{k=0}^{m-1} b_{ik} v'_{jk} = \sum_{k=0}^{m-1} u'_{ik} s'_k v'_{jk}, \tag{8.2}$$
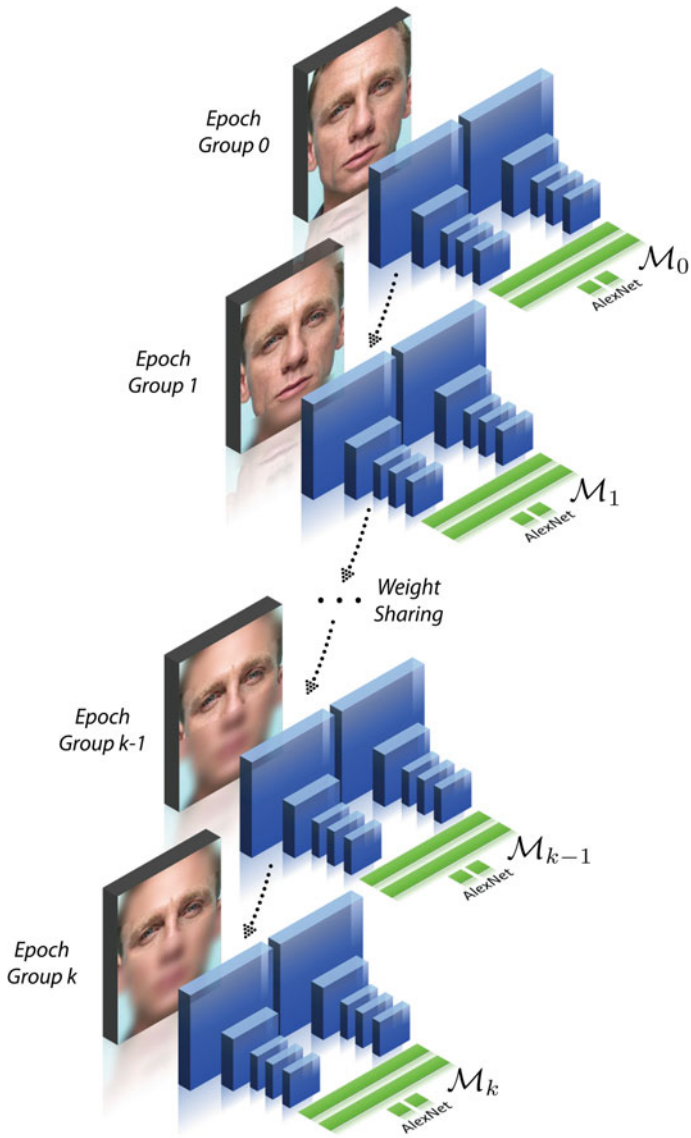
Fig. 8.2 Progressive CNN training with attention

which generates the matrix equation $\mathbf{X} = \mathbf{U}' \mathbf{\Sigma}' \mathbf{V}'^\top$. However, unlike the SVD, even though the $\{\mathbf{v}'_k\}$ are an orthonormal basis, the $\{\mathbf{u}'_k\}$ are not in general orthogonal. Nevertheless this demonstrates how the SVD is similar to a Fourier transform. Next, we will show that the **low-pass filtering in Fourier analysis** is closely related to the **low-rank approximation in SVD**.

Suppose we have N image data samples of dimension d in the original two-dimensional form $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Let matrix $\hat{\mathbf{X}}$ contain all the data samples undergone 2D Fourier transform $\mathcal{F}(\cdot)$, in the vectorized form:

$$\hat{\mathbf{X}} = \begin{bmatrix} | & | & & | \\ \text{vec}(\mathcal{F}(\mathbf{x}_1)) & \text{vec}(\mathcal{F}(\mathbf{x}_2)) & \ldots & \text{vec}(\mathcal{F}(\mathbf{x}_N)) \\ | & | & & | \end{bmatrix}_{d \times N} .$$

Matrix $\hat{\mathbf{X}}$ can be decomposed using SVD: $\hat{\mathbf{X}} = \hat{\mathbf{U}} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^\top$. Without loss of generality, let us assume that $N = d$ for brevity. Let $\mathbf{g}$ and $\hat{\mathbf{g}}$ be the Gaussian filter in spatial domain and frequency domain, respectively, namely $\hat{\mathbf{g}} = \mathcal{F}(\mathbf{g})$. Let $\hat{\mathbf{G}}$ be a diagonal matrix with $\hat{\mathbf{g}}$ on its diagonal. The convolution operation becomes dot product in frequency domain, so the blurring operation becomes

$$\hat{\mathbf{X}}_{\text{blur}} = \hat{\mathbf{G}} \cdot \hat{\mathbf{X}} = \hat{\mathbf{G}} \cdot \hat{\mathbf{U}} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^\top, \tag{8.3}$$

where $\hat{\mathbf{\Sigma}} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$ contains the singular values of $\hat{\mathbf{X}}_{\text{blur}}$, already sorted in descending order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$. Suppose we can find a permutation matrix $\mathbf{P}$ such that when applied on the diagonal matrix $\hat{\mathbf{G}}$, the diagonal elements are sorted in descending order according to the magnitude: $\hat{\mathbf{G}}' = \mathbf{P}\hat{\mathbf{G}} = \text{diag}(\hat{g}'_1, \hat{g}'_2, \ldots, \hat{g}'_d)$. Now, let us apply the same permutation operation on $\hat{\mathbf{X}}_{\text{blur}}$, we can thus have the following relationship:

$$\mathbf{P} \cdot \hat{\mathbf{X}}_{\text{blur}} = \mathbf{P} \cdot \hat{\mathbf{G}} \cdot \hat{\mathbf{U}} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^\top \tag{8.4}$$

$$\hat{\mathbf{X}}'_{\text{blur}} = \hat{\mathbf{G}}' \cdot \hat{\mathbf{U}} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^\top = \hat{\mathbf{U}} \cdot (\hat{\mathbf{G}}' \hat{\mathbf{\Sigma}}) \cdot \hat{\mathbf{V}}^\top \tag{8.5}$$

$$= \hat{\mathbf{U}} \cdot \text{diag}(\hat{g}'_1 \sigma_1, \hat{g}'_2 \sigma_2, \ldots, \hat{g}'_d \sigma_d) \cdot \hat{\mathbf{V}}^\top. \tag{8.6}$$

Due to the fact that Gaussian distribution is not a heavy-tailed distribution, the already smaller singular values will be brought down to 0 by the Gaussian weights. Therefore, $\hat{\mathbf{X}}_{\text{blur}}$ actually becomes low-rank after Gaussian low-pass filtering. To this end, we can say that low-pass filtering in Fourier analysis is equivalent to the low-rank approximation in SVD up to a permutation.

This phenomenon is loosely observed through the visualization of the trained filters, as shown in Fig. 8.15, which will be further analyzed and studied in future work.

8.3.2 Occlusion Removal via Deep Convolutional Generative Adversarial Networks (DCGAN)

In this section, we first review the basics of DCGAN and then show how DCGAN can be utilized for occlusion removal, or missing pixel recovery on face images.

8.3.2.1 Deep Convolutional Generative Adversarial Networks

The generative adversarial network (GAN) [12] is capable of generating high-quality images. The framework trains two networks, a generator $\mathcal{G}_\theta(\mathbf{z}) : \mathbf{z} \to \mathbf{x}$, and a discriminator $\mathcal{D}_\omega(\mathbf{x}) : \mathbf{x} \to [0, 1]$. $\mathcal{G}$ maps a random vector $\mathbf{z}$, sampled from a prior distribution $p_\mathbf{z}(\mathbf{z})$, to the image space. $\mathcal{D}$ maps an input image to a likelihood. The purpose of $\mathcal{G}$ is to generate realistic images, while $\mathcal{D}$ plays an adversarial role to discriminate between the image generated from $\mathcal{G}$, and the image sampled from data distribution p_{data}. The networks are trained by optimizing the following minimax loss function:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \Big[\log(\mathcal{D}(\mathbf{x})) \Big] + \mathbb{E}_{\mathbf{z} \sim p_\mathbf{z}(\mathbf{z})} \Big[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}))) \Big],$$

where $\mathbf{x}$ is the sample from the p_{data} distribution; $\mathbf{z}$ is randomly generated and lies in some latent space. There are many ways to structure $\mathcal{G}(\mathbf{z})$. The deep convolutional generative adversarial network (DCGAN) [60] uses fractionally strided convolutions to upsample images instead of fully connected neurons as shown in Fig. 8.3.

The generator $\mathcal{G}$ is updated to fool the discriminator $\mathcal{D}$ into wrongly classifying the generated sample, $\mathcal{G}(\mathbf{z})$, while the discriminator $\mathcal{D}$ tries not to be fooled. In this work, both $\mathcal{G}$ and $\mathcal{D}$ are deep convolutional neural networks and are trained with an alternating gradient descent algorithm. After convergence, $\mathcal{D}$ is able to reject images that are too fake, and $\mathcal{G}$ can produce high-quality images faithful to the training distribution (true distribution p_{data}).

8.3.2.2 Occlusion Removal via DCGAN

To take on the missing data challenge, we need to utilize both the $\mathcal{G}$ and $\mathcal{D}$ networks from DCGAN, pre-trained with uncorrupted data. After training, $\mathcal{G}$ is able to embed the images from p_{data} onto some nonlinear manifold of $\mathbf{z}$. An image that is not from p_{data} (e.g., corrupted face image with missing pixels) should not lie on the learned manifold. Therefore, we seek to recover the "closest" image on the manifold to the corrupted image as the proper reconstruction.

Let us denote the corrupted image as $\mathbf{y}$. To quantify the "closest" mapping from $\mathbf{y}$ to the reconstruction, we define a function consisting of contextual loss and perceptual loss, following the work of Yeh et al. [73].

In order to incorporate the information from the uncorrupted portion of the given image, the **contextual loss** is used to measure the fidelity between the reconstructed image portion and the uncorrupted image portion, which is defined as
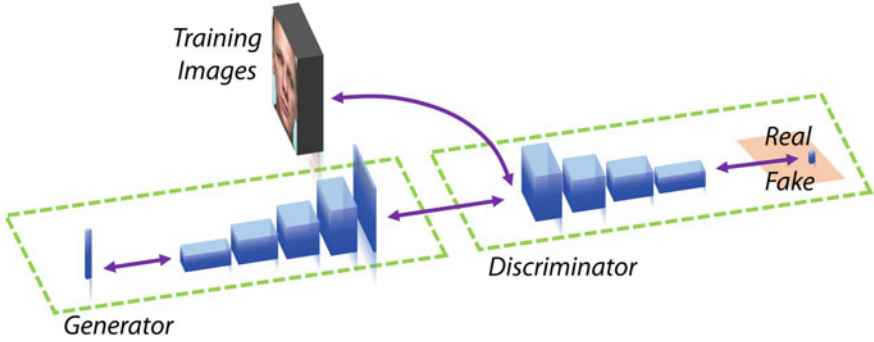
Fig. 8.3 Pipeline of a standard DCGAN with the generator $\mathcal{G}$ mapping a random vector $\mathbf{z}$ to an image and the discriminator $\mathcal{D}$ mapping the image (from true distribution or generated) to a probability value

$$\mathcal{L}_{\text{contextual}}(\mathbf{z}) = \|\mathbf{M} \odot \mathcal{G}(\mathbf{z}) - \mathbf{M} \odot \mathbf{y}\|_1, \tag{8.7}$$

where $\mathbf{M}$ denotes the binary mask of the uncorrupted region and $\odot$ denotes the element-wise Hadamard product operation. The corrupted portion, i.e., $(1 - \mathbf{M}) \odot \mathbf{y}$, is not used in the loss. The choice of ℓ_1-norm is empirical. From the experiments carried out in [73], images recovered with ℓ_1-norm loss tend to be sharper and with higher quality compared to ones reconstructed with ℓ_2-norm.

The **perceptual loss** encourages the reconstructed image to be similar to the samples drawn from the training set (true distribution p_{data}). This is achieved by updating $\mathbf{z}$ to fool $\mathcal{D}$, or equivalently to have a high value of $\mathcal{D}(\mathcal{G}(\mathbf{z}))$. As a result, $\mathcal{D}$ will predict $\mathcal{G}(\mathbf{z})$ to be from the data with a high probability. The same loss for fooling $\mathcal{D}$ as in DCGAN is used:

$$\mathcal{L}_{\text{perceptual}}(\mathbf{z}) = \log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z}))). \tag{8.8}$$

The corrupted image with missing pixels can now be mapped to the closest $\mathbf{z}$ in the latent representation space with the defined perceptual and contextual losses. We follow the training procedure in [60] and use Adam [45] for optimization. $\mathbf{z}$ is updated using back-propagation with the total loss:

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}}(\mathcal{L}_{\text{contextual}}(\mathbf{z}) + \lambda\mathcal{L}_{\text{perceptual}}(\mathbf{z})) \tag{8.9}$$

where λ is a weighting parameter. After finding the optimal solution $\hat{\mathbf{z}}$, the hallucinated full-face image can be obtained by

$$\mathbf{x}_{\text{hallucinated}} = \mathbf{M} \odot \mathbf{y} + (1 - \mathbf{M}) \odot \mathcal{G}(\hat{\mathbf{z}}). \tag{8.10}$$

Examples of face recovery from contiguous occlusions are shown in Fig. 8.11 using DCGAN. Applying this deep generative approach for occluded image recovery is expected to improve the performance of unconstrained gender classification.

8.4   Experiments: Part A

In this section we detail the training and testing protocols employed and various occlusions and low resolutions modeled in the testing set. Accompanying figures and tables for each subsection encompass the results and observations and are elaborated in each section.[1]

8.4.1   Database and Preprocessing

Training set: We source images from 5 different datasets, each containing samples of both classes. The datasets are J-Mugshot, O-Mugshot, M-Mugshot, P-Mugshot, and Pinellas. All the datasets, except Pinellas, are evenly separated into males and females of different ethnicities. The images are centered, by which, we mean that we have landmarked certain points on the face, which are then anchored to fixed points in the resulting training image. For example, the eyes are anchored at the same coordinates in every image. All of our input images have the same dimension 168×210. The details of the training datasets are listed in Table 8.2. The images are partitioned into training and validation and the progressive blur is applied to each image as explained in the previous section. Hence, for a given model iteration, the training set consists of $\sim$90 k images.

Testing set: The testing set was built primarily from the following two datasets: (1) The AR Face database [55] is one of the most widely used face databases with occlusions. It contains 3,288 color images from 135 subjects (76 male subjects + 59 female subjects). Typical occlusions include sunglasses and scarves. The database also captures expression variations and lighting changes. (2) Pinellas County Sherrif's Office (PCSO) mugshot database is a large-scale database of over 1.4 million images. We took a subset of around 400 K images from this dataset. These images are not seen during training.

The testing images are centered and cropped in the same way as the training images, though other preprocessing like the progressive blur are not applied. Instead, to model real world occlusions we have conducted the following experiments to be discussed in Sect. 8.4.2.

---

[1]**A note on legend**: (1) Symbols $\mathcal{M}$ correspond to each model trained, with $\mathcal{M}_F$ corresponding to the model trained on full face (equivalent to $\mathcal{M}_0$), $\mathcal{M}_P$ to one with just periocular images and $\mathcal{M}_k$, $k \subseteq (1, \dots, 6)$ to the incremental models trained. (2) The tabular results show model performance on the original images in column 1 and corrupted images in other columns.

Table 8.2 Datasets used for progressive CNN training

Database	Males	Females
J-Mugshot	1900	1371
M-Mugshot	1772	805
Pinellas subset	13215	3394
P-Mugshot	46346	12402
O-Mugshot	4164	3634
	67397	21606
Total		89003

8.4.2   Experiment I: Occlusion Robustness

In Experiment I, we carry out occlusion robust gender classification on both the AR Face database and the PCSO mugshot database. We manually add artificial occlusions to test the efficacy of our method on the PCSO database and test on the various images sets in the AR Face dataset.

Fig. 8.4 Overall classification accuracy on the PCSO (400 K). Images are not corrupted
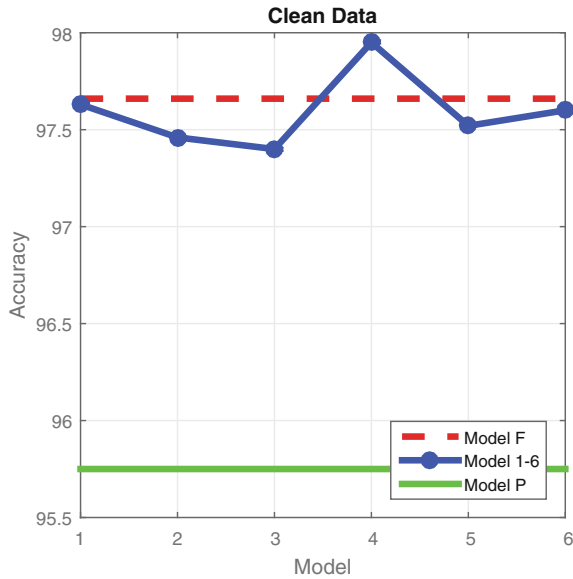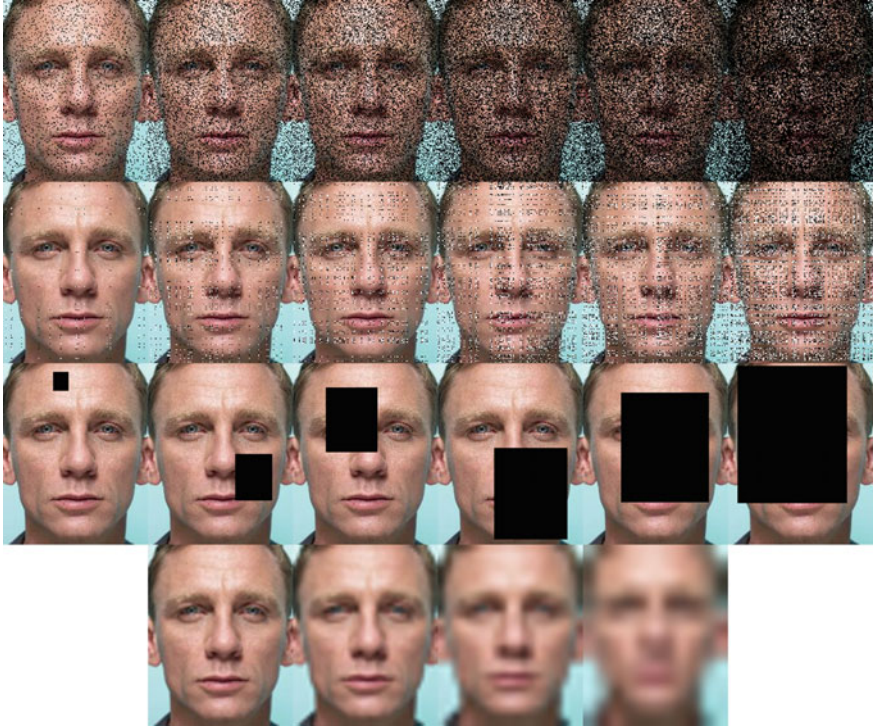
Fig. 8.5 Various degradations applied on the testing images for Experiment I and II. Row 1 random missing pixel occlusions; Row 2 random additive Gaussian noise occlusions; Row 3 random contiguous occlusions. Percentage of degradation for Row 1–3 10, 25, 35, 50, 65 and 75%. Row 4 various zooming factors (2x, 4x, 8x, 16x) for low-resolution degradations

8.4.2.1 Experiments on the PCSO Mugshot Database

To begin with, the performance of various models on the clean PCSO data is shown in Fig. 8.4. As expected, if the testing images are clean, it should be preferable to use $\mathcal{M}_F$, rather than $\mathcal{M}_P$. We can see that the progressively trained models $\mathcal{M}_1 - \mathcal{M}_6$ are on par with $\mathcal{M}_F$.

We corrupt the testing images (400 K) with three types of facial occlusions. These are visualized in Fig. 8.5 with each row corresponding to some modeled occlusions.

(1) Random Missing Pixels Occlusions

Table 8.3 Overall classification accuracy on the PCSO (400 K). Images are corrupted with **random missing pixels** of various percentages

Corruption	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	97.06	93.61	89.15	82.39	79.46	77.40
$\mathcal{M}_1$	97.63	96.93	92.68	87.99	81.57	78.97	77.20
$\mathcal{M}_2$	97.46	96.83	93.19	89.17	83.03	80.06	77.68
$\mathcal{M}_3$	97.40	96.98	94.06	90.65	84.79	81.59	78.56
$\mathcal{M}_4$	97.95	97.63	95.63	93.10	87.96	84.41	80.22
$\mathcal{M}_5$	97.52	97.26	95.80	94.07	90.40	87.39	83.04
$\mathcal{M}_6$	97.60	97.29	95.50	93.27	88.80	85.57	81.42
$\mathcal{M}_P$	95.75	95.45	93.84	92.02	88.87	86.59	83.18

Varying factors of the image pixels (10, 25, 35, 50, 65 and 75%) were dropped to model lost information and grainy images.[2] This is corresponding to the first row in Fig. 8.5. From Table 8.3 and Fig. 8.6, $\mathcal{M}_5$ performs the best with $\mathcal{M}_6$ showing a dip in accuracy suggesting a tighter periocular region is not well suited for such applications, i.e., a limit on the periocular region needs to be maintained in the blur set. There is a flip in performance of the models $\mathcal{M}_P$ and $\mathcal{M}_F$ going from the original to 25% with the periocular model generalizing better for higher corruptions. As the percentage of missing pixels increases, the performance gap between $\mathcal{M}_P$ and $\mathcal{M}_F$ increases. As hypothesized, the trend of improving performance between progressively trained models is maintained across factors indicating a better learned model toward noise.

(2) Random Additive Gaussian Noise Occlusions

Gaussian white noise ($\sigma = 6$) was added to image pixels in varying factors (10, 25, 35, 50, 65 and 75%). This is corresponding to the second row in Fig. 8.5 and is done to model high noise data and bad compression. From Table 8.4 and Fig. 8.7, $\mathcal{M}_4 - \mathcal{M}_6$ perform best for medium noise. For high noise, $\mathcal{M}_5$ is the most robust. Just like before, as the noise increases, the trend undertaken by the performance of $\mathcal{M}_P$ & $\mathcal{M}_F$ and $\mathcal{M}_5$ & $\mathcal{M}_6$ is maintained and so is the performance trend of the progressively trained models.

(3) Random Contiguous Occlusions

---

[2]This can also model the dead pixel/shot noise of a sensor and these results can be used to accelerate inline gender detection by using partially demosaiced images.
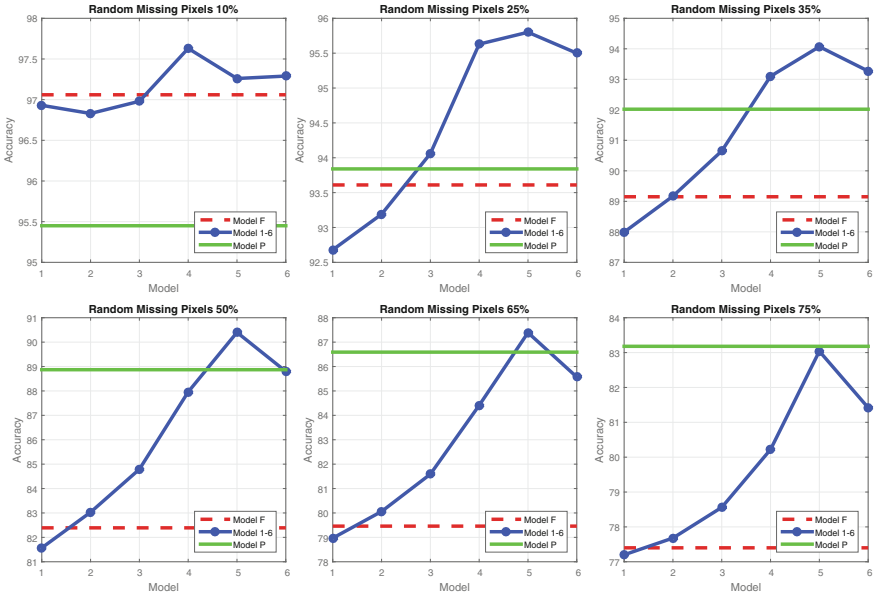
Fig. 8.6 Overall classification accuracy on the PCSO (400 K). Images are corrupted with **random missing pixels** of various percentages

Table 8.4 Overall classification accuracy on the PCSO (400 K). Images are corrupted with **additive Gaussian random noise** of various percentages

Corruption	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	97.00	94.03	91.19	86.47	83.43	79.94
$\mathcal{M}_1$	97.63	96.93	94.00	91.26	87.00	84.27	81.15
$\mathcal{M}_2$	97.46	96.87	94.43	92.19	88.75	86.44	83.33
$\mathcal{M}_3$	97.40	97.00	95.18	93.27	89.93	87.55	84.16
$\mathcal{M}_4$	97.95	97.67	96.45	95.11	92.43	90.28	87.06
$\mathcal{M}_5$	97.52	97.29	96.25	95.21	93.21	91.65	89.12
$\mathcal{M}_6$	97.60	97.32	96.04	94.77	92.46	90.80	88.08
$\mathcal{M}_P$	95.75	95.59	94.85	94.00	92.43	91.15	88.74

To model big occlusions like sunglasses or other contiguous elements, continuous patches of pixels (10, 25, 35, 50, 65 and 75%) were dropped from the image as seen in the third row of Fig. 8.5. The most realistic occlusion corresponds to the first few patches, and other patches are extreme cases. For the former cases, $\mathcal{M}_1 - \mathcal{M}_3$ are able to predict the classes with the highest accuracy. From Table 8.5 and Fig. 8.8, for such large occlusions and missing data, more contextual information is needed for correct classification since $\mathcal{M}_1 - \mathcal{M}_3$ perform better than other models.
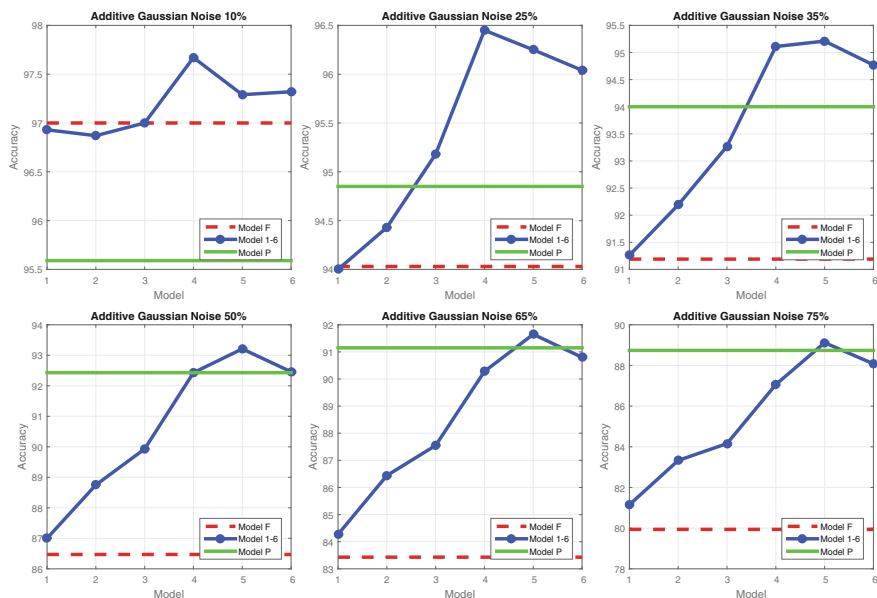
Fig. 8.7 Overall classification accuracy on the PCSO (400 K). Images are corrupted with **additive Gaussian random noise** of various percentages
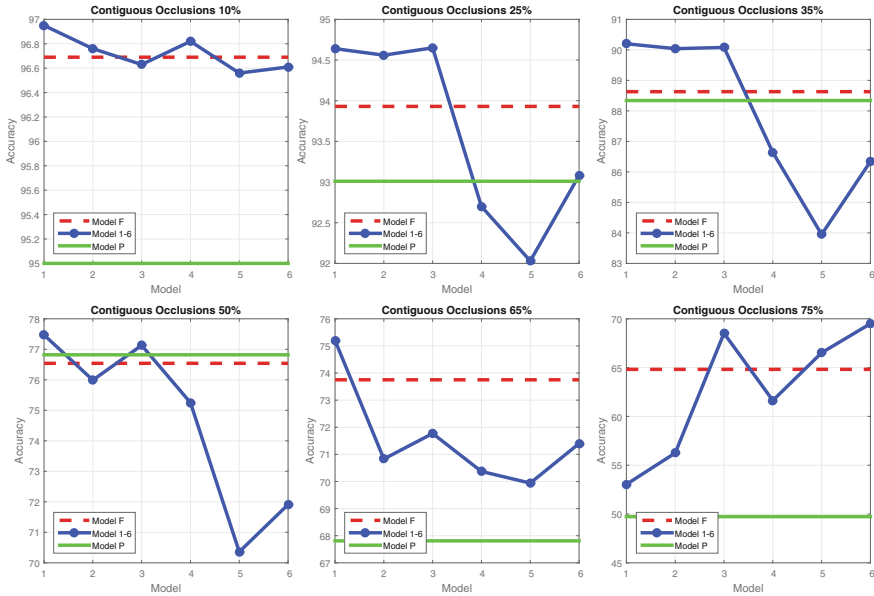
Table 8.5 Overall classification accuracy on the PCSO (400 K). Images are corrupted with **random contiguous occlusions** of various percentages

Corruption	0%	10%	25%	35%	50%	65%	75%
$\mathcal{M}_F$	97.66	96.69	93.93	88.63	76.54	73.75	64.82
$\mathcal{M}_1$	97.63	96.95	94.64	90.20	77.47	75.20	53.04
$\mathcal{M}_2$	97.46	96.76	94.56	90.04	75.99	70.83	56.25
$\mathcal{M}_3$	97.40	96.63	94.65	90.08	77.13	71.77	68.52
$\mathcal{M}_4$	97.95	96.82	92.70	86.64	75.25	70.37	61.63
$\mathcal{M}_5$	97.52	96.56	92.03	83.95	70.36	69.94	66.52
$\mathcal{M}_6$	97.60	96.61	93.08	86.34	71.91	71.40	69.50
$\mathcal{M}_P$	95.75	95.00	93.01	88.34	76.82	67.81	49.73

However, since they perform better than $\mathcal{M}_F$, our scheme of focused saliency helps generalizing over occlusions.

8.4.2.2   Experiments on the AR Face Database

We partitioned the original set to smaller subsets to better understand our methodology's performance under different conditions. Set 1 consists of neutral expression,

Fig. 8.8 Overall classification accuracy on the PCSO (400 K). Images are corrupted with **random contiguous occlusions** of various percentages

full-face subjects. Set 2 has full face but varied expressions. Set 3 includes periocular occlusions such as sunglasses and Set 4 includes these and other occlusions like clothing, etc. Set 5 is the entire dataset including illumination variations.

Referring to Table 8.6 and Fig. 8.9, for Set 1, the full-face model performs the best and this is expected as this model was trained on images very similar to this. Set 2 suggests that the models need more contextual information when expressions are introduced. Thus, $\mathcal{M}_4$ which has focus on periocular but has face information too performs best. For Set 3, we can see two things: one, $\mathcal{M}_P$ performs better than $\mathcal{M}_F$ indicative of its robustness to periocular occlusions. Two, $\mathcal{M}_5$ is the best as it combines periocular focus with contextual information gained from incremental training.

Set 4 performance brings out why periocular region is preferred for occluded faces. We ascertained that some texture and loss of face contour is throwing off the models $\mathcal{M}_1 - \mathcal{M}_6$. The performance of the models on Set 5 reiterates previously stated observations of the combined importance of contextual information about face contours and the importance of periocular region. This is the reason for the best accuracy reported by $\mathcal{M}_3$.

Table 8.6 Gender classification accuracy on the AR Face database

Sets	Set 1	Set 2	Set 3	Set 4	Set 5 (Full Set)
$\mathcal{M}_F$	98.44	93.23	89.06	83.04	81.65
$\mathcal{M}_1$	97.66	92.71	86.72	81.70	82.82
$\mathcal{M}_2$	97.66	92.71	90.62	82.14	85.10
$\mathcal{M}_3$	97.66	93.23	91.41	80.80	85.62
$\mathcal{M}_4$	98.44	95.31	92.97	77.23	84.61
$\mathcal{M}_5$	96.88	93.49	94.53	80.36	84.67
$\mathcal{M}_6$	96.09	92.71	92.97	79.02	83.90
$\mathcal{M}_P$	96.09	90.62	91.41	86.61	83.44
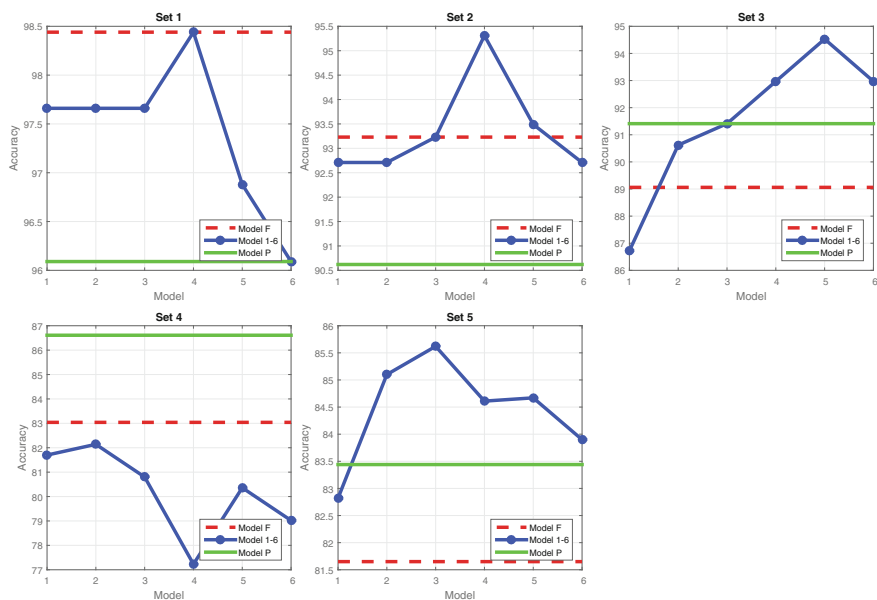
Fig. 8.9 Gender classification accuracy on the AR Face database

8.4.3 Experiment II: Low-Resolution Robustness

Our scheme of training on Gaussian blurred images should generalize well to low-resolution images. To test this hypothesis, we tested our models on images from the PCSO mugshot dataset by first down-sampling them by a factor and then blowing them back up (zooming factor for example: 2x, 4x, 8x, 16x).[3] This inculcates the loss of edge information and other higher order information and is captured in the

---

[3]Effective pixel for 16x zooming factor is around 10×13, which is a quite challenging low-resolution setting.

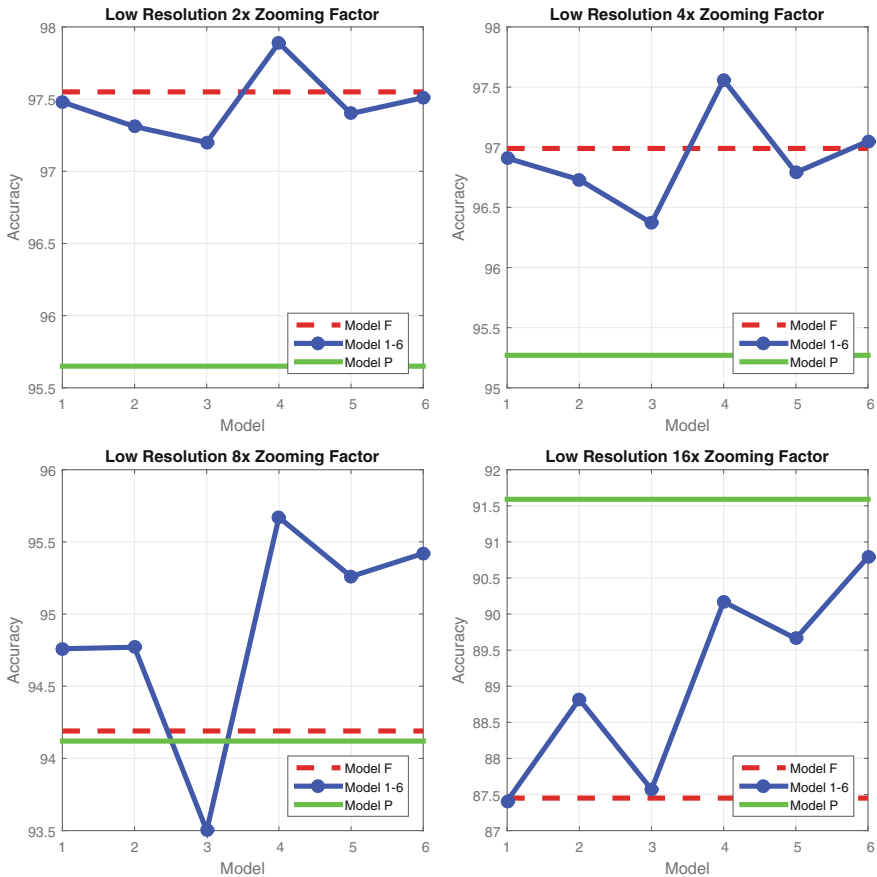**Table 8.7** Overall classification accuracy on the PCSO (400 K). Images are down-sampled to a **lower resolution** with various zooming factors

Zooming factor	1x	2x	4x	8x	16x
$\mathcal{M}_F$	97.66	97.55	96.99	94.19	87.45
$\mathcal{M}_1$	97.63	97.48	96.91	94.76	87.41
$\mathcal{M}_2$	97.46	97.31	96.73	94.77	88.82
$\mathcal{M}_3$	97.40	97.20	96.37	93.50	87.57
$\mathcal{M}_4$	97.95	97.89	97.56	95.67	90.17
$\mathcal{M}_5$	97.52	97.40	96.79	95.26	89.66
$\mathcal{M}_6$	97.60	97.51	97.05	95.42	90.79
$\mathcal{M}_P$	95.75	95.65	95.27	94.12	91.59

Fig. 8.10 Overall classification accuracy on the PCSO (400 K). Images are down-sampled to a **lower resolution** with various zooming factors

last row of Fig. 8.5. As seen in Table 8.7 and Fig. 8.10 for cases, 2x, 4x, 8x, the trend between $\mathcal{M}_1 - \mathcal{M}_6$ and their performance with respect to $\mathcal{M}_F$ is maintained. As mentioned before, $\mathcal{M}_4$ performs well due to the balance between focus on periocular region and saving the contextual information of a face.

8.4.4   Discussion

We have proposed a methodology for building a gender recognition system which is robust to occlusions. It involves training a deep model incrementally over several batches of input data preprocessed with progressive blur. The intuition and intent is twofold, one to have the network focus on periocular regions of the face for gender recognition. And two, to preserve contextual information of facial contours to generalize better over occlusions.

Through various experiments we have observed that our hypothesis is indeed true and that for a given occlusion set, it is possible to have high accuracy from a model that encompasses both of above-stated properties. Irrespective of the fact that we did not train on any occluded data, or optimize for a particular set of occlusions, our models are able to generalize well over synthetic data and real-life facial occlusion images.

We have summarized the overall experiments and consolidated the results in Table 8.8. For PCSO large-scale experiments, we believe that 35% occlusion is the right amount of degradations, on which accuracies should be reported. Therefore, we average the accuracy from our best model on three types of occlusions (missing pixel, additive Gaussian noise, and contiguous occlusions) which gives 93.12% in Table 8.8. For low-resolution experiments, we believe 8x zooming factor is the right amount of degradations, so we report the accuracy 95.67% in Table 8.8. Many other related work on gender classification are also listed for a quick comparison. This table is based on [16].

8.5   Experiments: Part B

In order to boost the classification accuracy and support the models trained in the previous section, we trained a GAN to hallucinate occlusions and missing information in the input image. The next two sections detail our curation of the input data and the selection of the testing sets for our evaluation. The gender model ($\mathcal{M}_k$) definitions remain the same as in the previous section. And we use $\mathcal{G}_z$ and $\mathcal{D}_x$ to denote the Generator and Discriminator models from GAN.

Table 8.8 Summary of many related work on gender classification. The proposed method is shown in the top rows

Authors	Methods	Dataset	Variation	Unique Subj.	Resolution	# of Subjects (Male/Female)	Accuracy
Proposed	Progressive CNN training w/ attention	Mugshots	Frontal-only, mugshot	Yes	168×210	89k total tr, 400k total te	97.95% te
			Occlusion	Yes		89k total tr, 400k total te	93.12% te
			Low-resolution	Yes		89k total tr, 400k total te	95.67% te
		AR Face	Expr x4, occl x2	No		89k total tr, 76/59 te	85.62% te
Hu et al. [16]	Region-based MR-8 filter bank w/ fusion of linear SVMs	Flickr	Li,exp,pos,bkgd,occl	Yes	128×170	10037/10037 tr, 3346/3346 te	90.1% te
		FERET	Frontal-only, studio	Yes		320/320 tr, 80/80 te	92.8% te
Chen and Lin [7]	Color and edge features w/ Adaboost+weak classifiers	Web img	Lighting, expression background	Yes	N/A	1948 total tr, 210/259 te	87.6% te

Table 8.8 (continued)

Authors	Methods	Dataset	Variation	Unique Subj.	Resolution	# of Subjects (Male/Female)	Accuracy
Wang et al. [10]	Gabor filters w/ polynomial-SVM	BioID	Lighting and expression	Yes	286×384	976/544 tr, 120 total te	92.5% te
Golomb et al. [11]	Raw pixel w/ neural network	SexNet	Frontal-only, studio	Yes	30×30	40/40 tr, 5/5 te	91.9% tr
Gutta et al. [13]	Raw pixel w/ mix of neural net RBF-SVM & decision tree	FERET	Frontal-only, studio	No	64×72	1906/1100 tr, 47/30 te	96.0% te
Jabid et al. [18]	Local directional patterns w/ SVM	FERET	Frontal-only, studio	No	100×100	1100/900 tr, unknown te	95.1% te
Lee et al. [48]	Region-based w/ linear regression fused w/ SVM	FERET	Frontal-only, studio	N/A	N/A	1158/615 tr, unknown te	98.8% te
		Web img	Unknown			1500/1500 tr, 1500/1500 te	88.1% te

Table 8.8 (continued)

Authors	Methods	Dataset	Variation	Unique Subj.	Resolution	# of Subjects (Male/Female)	Accuracy
Leng and Wang [49]	Gabor filters w/ fuzzy-SVM	FERET	Frontal-only, studio	Yes	256×384	160/140 total, 80% tr, 20% te	98.1% te
		CAS-PEAL	Studio	N/A	140×120	400/400 total, 80% tr, 20% te	93.0% te
		BUAA-IRIP	Frontal-only, studio	No	56×46	150/150 total, 80% tr, 20% te	89.0% te
Lin et al. [51]	Gabor filters w/ linear SVM	FERET	Frontal-only, studio	N/A	48×48	Unknown	92.8% te
Lu & Lin [52]	Gabor filters w/ Adaboost + linear SVM	FERET	Frontal-only, studio	N/A	48×48	150/150 tr, 518 total te	90.0% te
Lu et al. [53]	Region-based w/ RBF-SVM fused w/ majority vote	CAS-PEAL	Studio	Yes	90×72	320/320 tr, 80/80 te	92.6% te
Moghaddam and Yang [58]	Raw pixel w/ RBF-SVM	FERET	Frontal-only, studio	N/A	80×40	793/715 tr, 133/126 te	96.6% te
Yang et al. [71]	Texture normalization w/ RBF-SVM	Snapshots	Unknown	N/A	N/A	5600/3600 tr, unknown te	97.2% te
		FERET	Frontal-only, studio			1400/900 tr, 3529 total te	92.2% te

8.5.1 Network Architecture

The network architectures and approach that we use are similar to the work of [60]. The input size of the image is 64×64. While training we used Adam [45] optimization method because it does not require hand-tuning of the learning rate, momentum, and other hyper-parameters.

For learning z we use the same hyper-parameters as learned in the previous model. In our experiments, running 2000 epochs helped converge to a low loss.

8.5.2 Database and Preprocessing

In order to train $\mathcal{G}_z$ and $\mathcal{D}_x$ our approach initially was to use the same input images as used to train $\mathcal{M}_k$. However, this resulted in the network not converging to a low loss. In other words, we were not able to learn a generative distribution that the generator could sample from. Our analysis and intuition suggested that in order for the adversarial loss to work, the task had to be a challenge for both $\mathcal{G}_z$ and $\mathcal{D}_x$. Our fully frontal pose images were ill-posed for this model.

Hence to train the GAN, we used the Labeled Faces in the Wild (LFW) database [47] and aligned the faces using dLib as provided by the OpenFace [1]. We trained on the entire dataset comprising around 13,000 images with 128 images held out for the purpose of qualitatively showing the image recovery results as in Fig. 8.11. In this case, by visual analysis as well by analytical analysis the $\mathcal{G}_z$ was better able to learn a distribution of z, p_g, that was a strong representation of the data, p_{data}. That is symbolically, $p_g = p_{\text{data}}$. The results and detailed evaluation of the model are done later in Sect. 8.5.3. In this part of the experiment, we use a subset of the PCSO database containing 10,000 images (5,000 male and 5,000 female) for testing the gender classification accuracy. The reason we did not test on the entire 400 K PCSO is simply because the occlusion removal step involves an iterative solver which is time consuming.

8.5.3 Experiment III: Further Occlusion Robustness via Deep Generative Approach

As shown in Fig. 8.11, we chose to model occlusions based on percentage of pixels missing from the center of the image. The shown images are from the held-out portion of the LFW dataset. We show recovered images in Fig. 8.11 as a visual confirmation that the DCGAN is able to recover unseen images with high fidelity, even under pretty heavy occlusions as much as 75% in the face center. The recovery results on the PCSO datasets to be tested for gender classification are comparable to that of the LFW, but we are not allowed to display mugshot images in any published work.
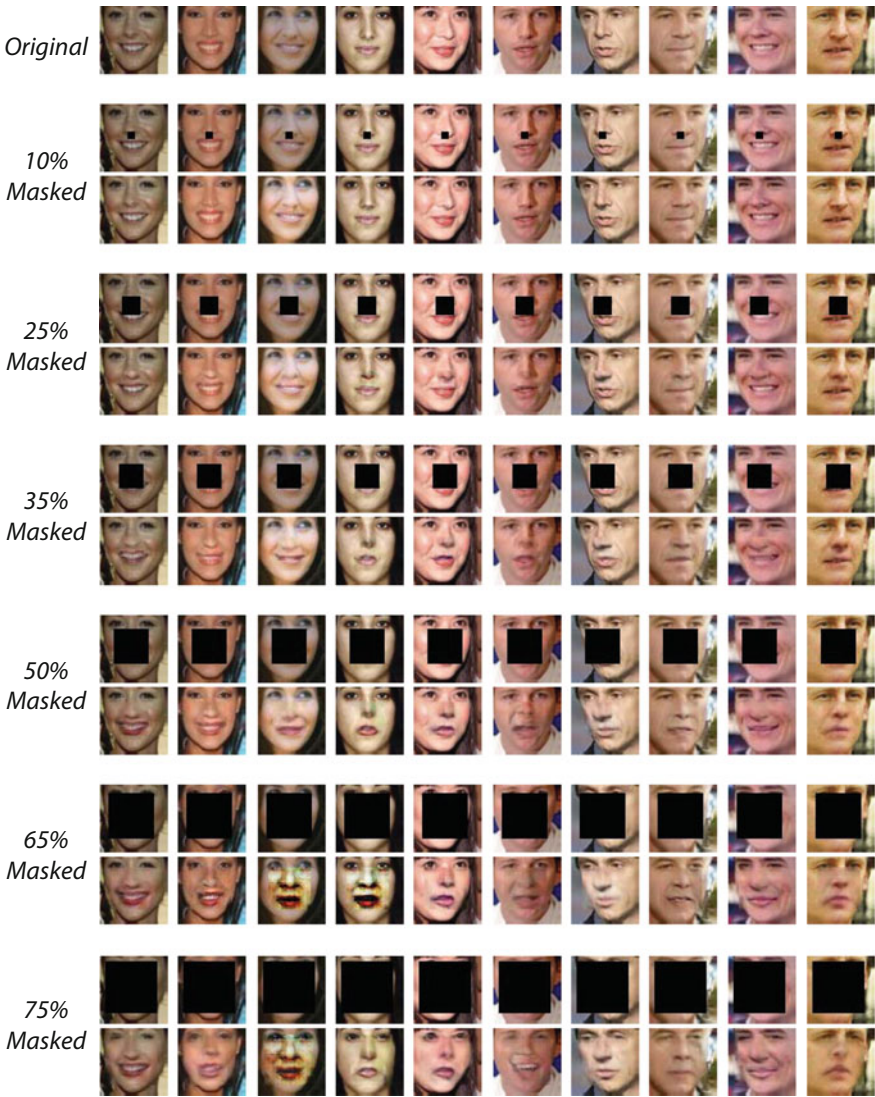
Fig. 8.11 Qualitative results of occlusion removal using DCGAN on images from LFW dataset

For training the $\mathcal{G}_z$ and $\mathcal{D}_x$, we used the LFW data that captures a high variance of poses, illumination, and faces. We found this was critical in helping especially the $\mathcal{G}_z$ converge to a stable weights. The model was able to generalize well to various faces and poses. As can be seen in Fig. 8.11, the model is able to generate missing information effectively.

Fig. 8.12 Quality measure of occlusion removal on the PCSO subset
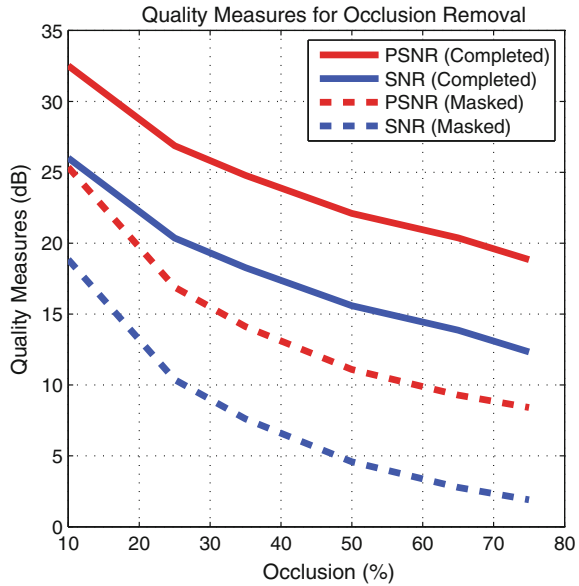
Table 8.9 Image quality measures (in dB) on the masked and occlusion removed images

Occlusion (%)	😷 PSNR	😀 PSNR	😷 SNR	😀 SNR
10	25.3607	32.5194	18.8488	26.0101
25	16.8941	26.8720	10.3822	20.3627
35	14.112	24.7825	7.6001	18.2732
50	11.0937	22.1043	4.5819	15.5950
65	9.2849	20.3719	2.773	13.8627
75	8.4169	18.8392	1.905	12.3299

Not relying on visual inspection, we plotted the PSNR and SNR of the recovered (occlusion removed) faces from the 10K PCSO subset in Fig. 8.12. This is our quality measure of occlusion removal using GAN. As can be seen in Table 8.9, the PSNR/SNR is better for completed images and as expected is higher for images with lesser occlusions.

The primary motivation behind training a GAN was to improve the classification of $\mathcal{M}_k$ models. This is covered in Table 8.10. The first column of the table is our baseline case. This was constructed using upsampled images from the resolution as needed by $\mathcal{G}_z$ and $\mathcal{D}_x$ to the resolution expected by $\mathcal{M}_k$. All other accuracies should be evaluated with respect to this case. (Figure 8.13)

Table 8.10 Overall classification accuracy for Experiment III on the PCSO subset. Images are corrupted with **centered contiguous missing pixels** of various percentages

Corrup.	0%	10%	10%	25%	25%	35%	35%	50%	50%	65%	65%	75%	75%
$\mathcal{M}_F$	96.64	85.39	96.15	71.12	95.29	56.08	93.60	54.05	87.88	57.82	78.42	50.19	78.56
$\mathcal{M}_1$	96.13	87.09	95.58	62.25	94.39	53.77	93.18	54.17	87.92	59.85	79.69	51.20	79.68
$\mathcal{M}_2$	95.74	85.73	94.75	68.30	93.46	53.59	92.47	55.03	86.45	51.26	78.19	50.13	78.81
$\mathcal{M}_3$	95.44	85.43	94.94	73.14	93.36	51.65	92.18	53.19	86.82	50.81	78.97	50.09	79.43
$\mathcal{M}_4$	97.52	83.35	96.97	56.34	94.78	50.24	92.82	53.58	86.93	52.72	78.96	50.12	79.70
$\mathcal{M}_5$	95.78	86.19	94.97	53.98	92.47	50.33	91.05	52.44	86.62	52.20	78.24	50.23	79.03
$\mathcal{M}_6$	96.21	86.41	95.59	58.47	92.84	50.42	91.30	52.02	87.15	51.02	78.74	50.40	79.77
$\mathcal{M}_P$	93.46	87.55	93.25	77.56	93.21	76.44	91.73	69.28	88.39	76.33	79.42	54.39	80.96

Figure 8.14 corresponds to the other columns of Table 8.10. The accuracies on completed images using $\mathcal{G}_z$ are significantly better than the accuracies on masked images. This suggests that the hallucinations are able to preserve the gender sensitivity of the original images.

The above statement can also be verified through visual inspection of Fig. 8.11. Even for high-percentage occlusions, $50 - 75\%$, the contours and features of the original face are pretty accurately reproduced by $\mathcal{G}_z$.

8.5.4 Discussion

In the current generative model for image occlusion removal, we assume that the occlusion mask is known to the algorithm, which is the **M** in Eq. 8.7. Although it is beyond the scope of this work to study how an algorithm can automatically determine the occlusion region, it will be an interesting research direction. For example, [61] is able to automatically tell which part is the face or the non-face region.

One big advantage of the DCGAN or GAN in general is that it is entirely unsupervised. The loss function is based off essentially measuring the similarity of two distributions (the true image distribution and the generated image distribution), rather than image-wise comparisons, which may require labeled ground-truth images be provided. The unsupervised nature of the GAN has made the training process much easier by not needing careful curating of the labeled data.

Fig. 8.13 Overall classification accuracy for Experiment III on the PCSO subset. Images are not corrupted
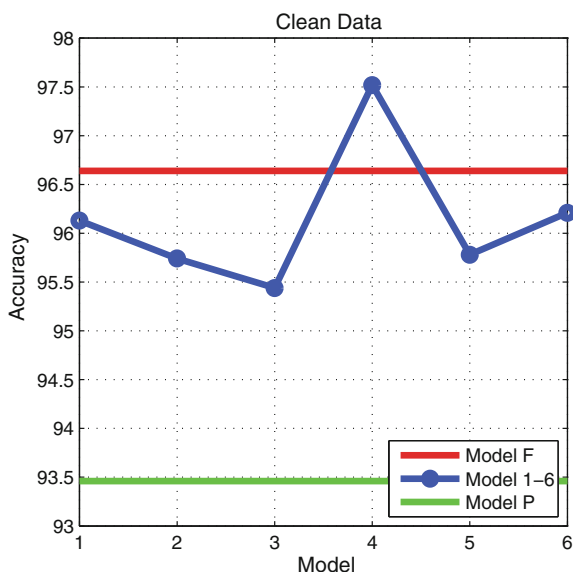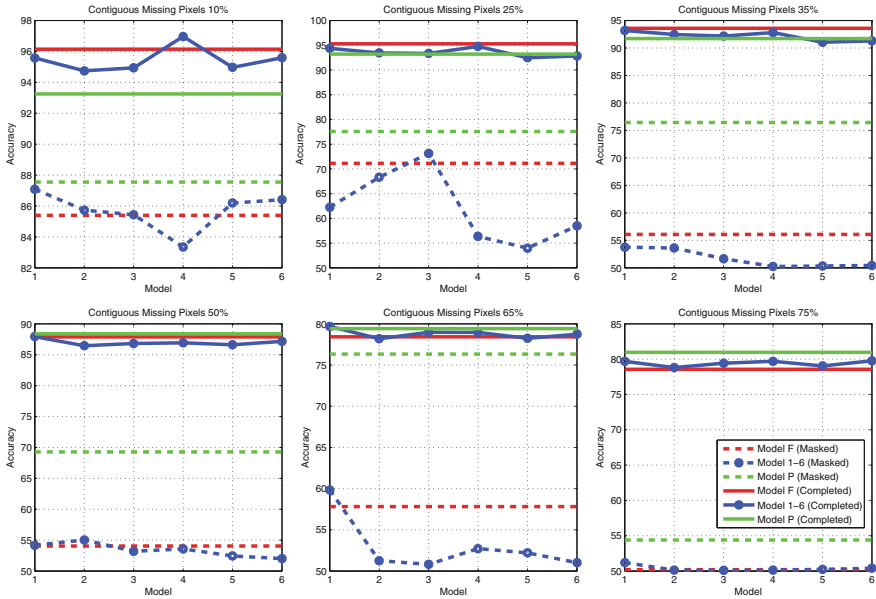
Fig. 8.14 Overall classification accuracy on the PCSO subset. Images are corrupted with **centered contiguous occlusions** of various percentages

The effectiveness of the DCGAN is also salient. It not only can recover the high-percentage missing data with high fidelity, which translates to significant improvement on the gender classification tasks, but also can be used for future data augmentation in a total unsupervised manner. We can essentially generate as much gender-specific data as needed, which will be an asset for training an even larger model.

During our experimentation, we find that training the DCGAN using more constrained faces (less pose variations, less lighting variations, less expression variations, etc.) such as the PCSO mugshot images actually degrades the recovery performance. The reason could be as follows. When the training data is less diversified, let us use one extreme case, which is a training dataset comprising thousands of images from only one subject. In this case, the 'true distribution' becomes a very densely clustered mass, which means whatever images the generator $\mathcal{G}$ tries to come up with, the discriminator $\mathcal{D}$ will (almost) always say that the generated ones are not from the true distribution, because the generated image distribution can hardly hit that densely clustered mass. This way, we are essentially giving a too easy task for the discriminator to solve, which prevents the discriminator to become a strong one, which leads to poorly performing generator as well in this adversarial setting. In a nutshell, during the DCGAN training, we definitely need more variations in the training corpus.

8.6 Conclusions and Future Work

In this work, we have undertaken the task of occlusion and low-resolution robust facial gender classification. Inspired by the trainable attention model via deep architecture, and the fact that the periocular region is proven to be the most salient region for gender classification purposes, we are able to design a progressive convolutional neural network training paradigm to enforce the attention shift during the learning process. The hope is to enable the network to attend to particular high-profile regions (e.g., the periocular region) without the need to change the network architecture itself. The network benefits from this attention shift and becomes more robust toward occlusions and low-resolution degradations. With the progressively trained CNN models, we have achieved better gender classification results on the large-scale PCSO mugshot database with 400 K images under occlusion and low-resolution settings, compared to the one undergone traditional training. In addition, our progressively trained network is sufficiently generalized so that it can be robust to occlusions of arbitrary types and at arbitrary locations, as well as low resolution.

To further improve the gender classification performance on occluded facial images, we invoke a deep generative approach via deep convolutional generative adversarial networks (DCGAN) to fill in the missing pixels for the occluded facial regions. The recovered images not only show high fidelity as compared to the original un-occluded images but also significantly improve the gender classification performance.

In summary, on one hand, we aim at building a robust gender classifier that is tolerant to image degradations such as occlusions and low resolution, and on the other hand, we aim at mitigating and eliminating the degradations through a generative approach. Together, we are able to push the boundary of unconstrained facial gender classification.

Future work: We have carried out a set of large-scale testing experiments on the PCSO mugshot database with 400 K images, shown in the experimental section. We have noticed that, under the same testing environment, the amount of time it takes to test on the entire 400 K images various dramatically for different progressively trained models ($\mathcal{M}_0 - \mathcal{M}_6$). As shown in Fig. 8.15, we can observe a trend of testing time decrease when testing using $\mathcal{M}_0$ all the way to $\mathcal{M}_6$, where the curves correspond to the additive Gaussian noise occlusion robust experiments. This same trend is observed across the board for all the large-scale experiments on PCSO. The time difference is stunning. For example, if we look at the green curve, $\mathcal{M}_0$ takes over 5000 seconds while $\mathcal{M}_6$ only around 500. One of the future directions is to study the cause of this phenomenon. One possible direction is to study the sparsity or the smoothness of the learned filters.

Shown in our visualization (Fig. 8.15) of the 64 first-layer filters in AlexNet for models $\mathcal{M}_0$, $\mathcal{M}_3$, and $\mathcal{M}_6$, respectively, we can observe that the progressively trained filters seem to be smoother and this may be due to the implicit low-rank regularization phenomenon discussed in Sect. 8.3.1.3. Other future work may include studying how the ensemble of models [43, 44] can further improve the performance and how various
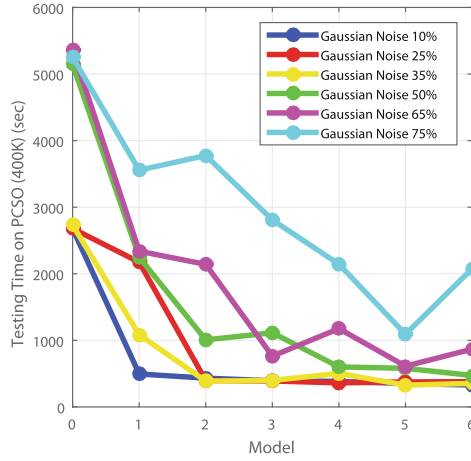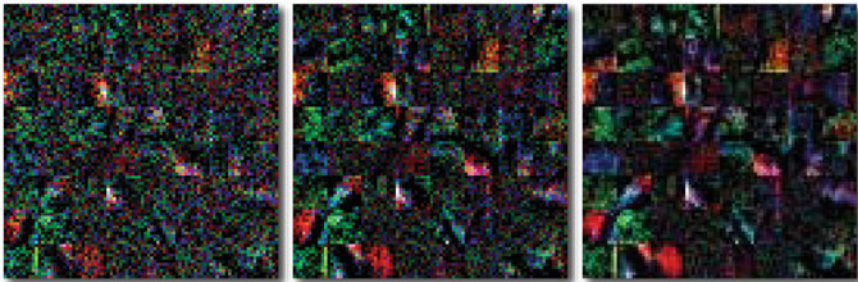
Fig. 8.15 (*Top*) Testing time for the additive Gaussian noise occlusion experiments on various models. (*Bottom*) Visualization of the 64 first-layer filters for models $\mathcal{M}_0$, $\mathcal{M}_3$, and $\mathcal{M}_6$, respectively

multimodal soft-biometrics traits [5, 23, 25, 26, 28–30, 37, 41, 42, 64, 66, 74] can be fused for improved gender classification, especially under more unconstrained scenarios.

References

1. B. Amos, B. Ludwiczuk, M. Satyanarayanan, Openface: a general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science (2016)
2. J. Ba, V. Mnih, K. Kavukcuoglu, Multiple object recognition with visual attention (2014), arXiv:1412.7755
3. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate (2014), arXiv:1409.0473
4. A. Bartle, J. Zheng, Gender classification with deep learning. Stanford cs224d Course Project Report (2015)

5. P. Buchana, I. Cazan, M. Diaz-Granados, F. Juefei-Xu, M. Savvides, Simultaneous forgery identification and localization in paintings using advanced correlation filters, in *IEEE International Conference on Image Processing (ICIP)* (2016), pp. 1–5
6. W. Chan, N. Jaitly, Q.V. Le, O. Vinyals, Listen, attend and spell (2015), arXiv:1508.01211
7. D.Y. Chen, K.Y. Lin, Robust gender recognition for uncontrolled environment of real-life images. IEEE Trans. Consum. Electron. **56**(3), 1586–1592 (2010)
8. M.M. Cheng, Z. Zhang, W.Y. Lin, P. Torr, BING: binarized normed gradients for objectness estimation at 300fps, in *IEEE CVPR* (2014), pp. 3286–3293
9. J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in *Advances in Neural Information Processing Systems* (2015), pp. 577–585
10. W. Chuan-xu, L. Yun, L. Zuo-yong, Algorithm research of face image gender classification based on 2-D gabor wavelet transform and SVM, in *ISCSCT'08. International Symposium on Computer Science and Computational Technology, 2008*, vol. 1 (IEEE, 2008), pp. 312–315
11. B.A. Golomb, D.T. Lawrence, T.J. Sejnowski, Sexnet: a neural network identifies sex from human faces, in *NIPS*, vol. 1 (1990), p. 2
12. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems* (2014), pp. 2672–2680
13. S. Gutta, J.R. Huang, P. Jonathon, H. Wechsler, Mixture of experts for classification of gender, ethnic origin, and pose of human faces. IEEE Trans. Neural Netw. **11**(4), 948–960 (2000)
14. J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in *NIPS* (2006), pp. 545–552
15. X. Hou, J. Harel, C. Koch, Image signature: highlighting sparse salient regions. IEEE TPAMI **34**(1), 194–201 (2012)
16. S.Y.D. Hu, B. Jou, A. Jaech, M. Savvides, Fusion of region-based representations for gender identification, in *IEEE/IAPR IJCB* (2011), pp. 1–7
17. L. Itti, C. Koch, E. Niebur et al., A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
18. T. Jabid, M.H. Kabir, O. Chae, Gender classification using local directional pattern (LDP), in *IEEE ICPR* (IEEE, 2010), pp. 2162–2165
19. M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in *NIPS* (2015), pp. 2008–2016
20. F. Juefei-Xu, M. Savvides, Can your eyebrows tell me who you are?, in *2011 5th International Conference on Signal Processing and Communication Systems (ICSPCS)* (2011), pp. 1–8
21. F. Juefei-Xu, M. Savvides, Unconstrained periocular biometric acquisition and recognition using COTS PTZ camera for uncooperative and non-cooperative subjects, in *2012 IEEE Workshop on Applications of Computer Vision (WACV)* (2012), pp. 201–208
22. F. Juefei-Xu, M. Savvides, An augmented linear discriminant analysis approach for identifying identical twins with the aid of facial asymmetry features, in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2013), pp. 56–63
23. F. Juefei-Xu, M. Savvides, An image statistics approach towards efficient and robust refinement for landmarks on facial boundary, in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2013), pp. 1–8
24. F. Juefei-Xu, M. Savvides, Subspace based discrete transform encoded local binary patterns representations for robust periocular matching on NIST's face recognition grand challenge. IEEE Trans. Image Process. **23**(8), 3490–3505 (2014)
25. F. Juefei-Xu, M. Savvides, Encoding and decoding local binary patterns for harsh face illumination normalization, in *IEEE International Conference on Image Processing (ICIP)* (2015), pp. 3220–3224
26. F. Juefei-Xu, M. Savvides, Facial ethnic appearance synthesis, in *Computer Vision - ECCV 2014 Workshops*. Lecture Notes in Computer Science, vol. 8926 (Springer International Publishing, Berlin, 2015), pp. 825–840
27. F. Juefei-Xu, M. Savvides, Pareto-optimal discriminant analysis, in *IEEE International Conference on Image Processing (ICIP)* (2015), pp. 611–615

28. F. Juefei-Xu, M. Savvides, Pokerface: partial order keeping and energy repressing method for extreme face illumination normalization, in *2015 IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2015), pp. 1–8

29. F. Juefei-Xu, M. Savvides, Single face image super-resolution via solo dictionary learning, in *IEEE International Conference on Image Processing (ICIP)* (2015), pp. 2239–2243

30. F. Juefei-Xu, M. Savvides, Weight-optimal local binary patterns, in *Computer Vision - ECCV 2014 Workshops*. Lecture Notes in Computer Science, vol. 8926 (Springer International Publishing, Berlin, 2015), pp. 148–159

31. F. Juefei-Xu, M. Savvides, Fastfood dictionary learning for periocular-based full face hallucination, in *2016 IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2016), pp. 1–8

32. F. Juefei-Xu, M. Savvides, learning to invert local binary patterns, in *27th British Machine Vision Conference (BMVC)* (2016)

33. F. Juefei-Xu, M. Savvides, Multi-class Fukunaga Koontz discriminant analysis for enhanced face recognition. Pattern Recognit. **52**, 186–205 (2016)

34. F. Juefei-Xu, M. Cha, J.L. Heyman, S. Venugopalan, R. Abiantun, M. Savvides, Robust local binary pattern feature sets for periocular biometric identification, in *4th IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)* (2010), pp. 1–8

35. F. Juefei-Xu, M. Cha, M. Savvides, S. Bedros, J. Trojanova, Robust periocular biometric recognition using multi-level fusion of various local feature extraction techniques, in *IEEE 17th International Conference on Digital Signal Processing (DSP)* (2011), pp. 1–7

36. F. Juefei-Xu, K. Luu, M. Savvides, T. Bui, C. Suen, Investigating age invariant face recognition based on periocular biometrics, in *2011 International Joint Conference on Biometrics (IJCB)* (2011), pp. 1–7

37. F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, M. Savvides, Gait-ID on the move: pace independent human identification using cell phone accelerometer dynamics, in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2012), pp. 8–15

38. F. Juefei-Xu, D.K. Pal, M. Savvides, Hallucinating the full face from the periocular region via dimensionally weighted K-SVD, in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2014), pp. 1–8

39. F. Juefei-Xu, D.K. Pal, M. Savvides, Methods and software for hallucinating facial features by prioritizing reconstruction errors (2014). U.S. Provisional Patent Application Serial No. 61/998,043, 17 Jun 2014

40. F. Juefei-Xu, K. Luu, M. Savvides, Spartans: single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios. IEEE Trans. Image Process. **24**(12), 4780–4795 (2015)

41. F. Juefei-Xu, D.K. Pal, M. Savvides, NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2015), pp. 141–150

42. F. Juefei-Xu, D.K. Pal, K. Singh, M. Savvides, A preliminary investigation on the sensitivity of COTS face recognition systems to forensic analyst-style face processing for occlusions, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2015), pp. 25–33

43. F. Juefei-Xu, V.N. Boddeti, M. Savvides, Local binary convolutional neural networks (2016), arXiv:1608.06049

44. F. Juefei-Xu, V.N. Boddeti, M. Savvides, Local binary convolutional neural networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)

45. D. Kingma, J. Ba, Adam: a method for stochastic optimization (2014), arXiv:1412.6980

46. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in *NIPS* (2012), pp. 1097–1105

47. G.B.H.E. Learned-Miller, Labeled faces in the wild: updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst (2014)

48. P.H. Lee, J.Y. Hung, Y.P. Hung, Automatic gender recognition using fusion of facial strips, in *IEEE ICPR* (IEEE, 2010), pp. 1140–1143
49. X. Leng, Y. Wang, Improving generalization for gender classification, in *IEEE ICIP* (IEEE, 2008), pp. 1656–1659
50. G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in *IEEE CVPRW* (2015)
51. H. Lin, H. Lu, L. Zhang, A new automatic recognition system of gender, age and ethnicity, in *WCICA 2006, The Sixth World Congress on Intelligent Control and Automation, 2006*, vol. 2 (IEEE, 2006), pp. 9988–9991
52. H. Lu, H. Lin, Gender recognition using adaboosted feature, in *ICNC 2007, Third International Conference on Natural Computation, 2007*, vol. 2 (IEEE, 2007), pp. 646–650
53. L. Lu, Z. Xu, P. Shi, Gender classification of facial images based on multiple facial regions, in *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 6 (IEEE, 2009), pp. 48–52
54. M.T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2015)
55. A. Martinez, R. Benavente, The AR Face Database. CVC Technical report No. 24 (1998)
56. J. Merkow, B. Jou, M. Savvides, An exploration of gender identification using only the periocular region, in *IEEE BTAS* (2010), pp. 1–5
57. V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in *NIPS* (2014), pp. 2204–2212
58. B. Moghaddam, M.H. Yang, Gender classification with support vector machines, in *IEEE FG* (IEEE, 2000), pp. 306–311
59. D.K. Pal, F. Juefei-Xu, M. Savvides, Discriminative invariant kernel features: a bells-and-whistles-free approach to unsupervised face recognition and pose estimation in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
60. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks (2015), arXiv:1511.06434
61. S. Saito, T. Li, H. Li, Real-time facial segmentation and performance capture from RGB input (2016), arXiv:1604.02647
62. M. Savvides, F. Juefei-Xu, Image Matching Using Subspace-Based Discrete Transform Encoded Local Binary Patterns (2013). US Patent US 2014/0212044 A1
63. M. Savvides, F. Juefei-Xu, U. Prabhu, C. Bhagavatula, Unconstrained biometric identification in real world environments, in *Advances in Human Factors and System Interactions*. Advances in Intelligent Systems and Computing, vol. 497 (2016), pp. 231–244
64. K. Seshadri, F. Juefei-Xu, D.K. Pal, M. Savvides, Driver cell phone usage detection on strategic highway research program (SHRP2) face view videos, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2015), pp. 35–43
65. C. Tai, T. Xiao, X. Wang, W. E, Convolutional neural networks with low-rank regularization. ICLR (2016), http://arxiv.org/abs/1511.06067
66. S. Venugopalan, F. Juefei-Xu, B. Cowley, M. Savvides, Electromyograph and keystroke dynamics for spoof-resistant biometric authentication, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2015), pp. 109–118
67. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in *IEEE CVPR* (2015), pp. 3156–3164
68. T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in *IEEE CVPR* (2015), pp. 842–850
69. H. Xu, K. Saenko, Ask, attend and answer: exploring question-guided spatial attention for visual question answering (2015), arXiv:1511.05234
70. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell - neural image caption generation with visual attention, in *ICML* (2015), pp. 2048–2057

71. Z. Yang, M. Li, H. Ai, An experimental study on automatic face gender classification, in *IEEE ICPR*, vol. 3 (IEEE, 2006), pp. 1099–1102
72. L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in *IEEE ICCV* (2015), pp. 4507–4515
73. R. Yeh, C. Chen, T.Y. Lim, M. Hasegawa-Johnson, M.N. Do, Semantic image inpainting with perceptual and contextual losses (2016), arXiv:1607.07539
74. N. Zehngut, F. Juefei-Xu, R. Bardia, D.K. Pal, C. Bhagavatula, M. Savvides, Investigating the feasibility of image-based nose biometrics, in *IEEE International Conference on Image Processing (ICIP)* (2015), pp. 522–526
75. Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7W: grounded question answering in images (2015), arXiv:1511.03416