Shape Constrained Network for Eye Segmentation in the Wild

Bingnan Luo[1], Jie Shen*[1,2], Shiyang Cheng[2], Yujiang Wang[1], and Maja Pantic[1,2]

[1]Department of Computing, Imperial College London, UK
[2]Samsung AI Centre, Cambridge, UK
{bingnan.luo16, jie.shen07}@imperial.ac.uk, shiyang.c@samsung.com, {yujiang.wang14, m.pantic}@imperial.ac.uk

Abstract

Semantic segmentation of eyes has long been a vital pre-processing step in many biometric applications. Majority of the works focus only on high resolution eye images, while little has been done to segment the eyes from low quality images in the wild. However, this is a particularly interesting and meaningful topic, as eyes play a crucial role in conveying the emotional state and mental well-being of a person. In this work, we take two steps toward solving this problem: (1) We collect and annotate a challenging eye segmentation dataset containing 8882 eye patches from 4461 facial images of different resolutions, illumination conditions and head poses; (2) We develop a novel eye segmentation method, Shape Constrained Network (SCN), that incorporates shape prior into the segmentation network training procedure. Specifically, we learn the shape prior from our dataset using VAE-GAN, and leverage the pre-trained encoder and discriminator to regularise the training of SegNet. To improve the accuracy and quality of predicted masks, we replace the loss of SegNet with three new losses: Intersection-over-Union (IoU) loss, shape discriminator loss and shape embedding loss. Extensive experiments shows that our method outperforms state-of-the-art segmentation and landmark detection methods in terms of mean IoU (mIoU) accuracy and the quality of segmentation masks. The dataset is available at https://ibug.doc.ic.ac.uk/resources/ibug-eye-segmentation-dataset/

1. Introduction

Eyes not only are the most vital sensory organ but also play a crucial role in conveying a person's emotion state and mental well-being [16]. Recently, considering facial segmentation has gained good performacne, as [39], eye segmentation task is gradually highlighted. Although there have been numerous works on blink detection [29, 5, 38], we argue that accurate segmentation of sclera and iris can provide much more information than blinks alone, thus allowing us to study the finer details of eye movement such as saccade, fixation, and other gaze patterns. As a pre-processing step in iris recognition, iris segmentation in high resolution expression-less frontal face images have been well studied by the biometric community. However, the commonly used Hough-transform-based method does not work well on low-resolution images captured under normal Human-Computer Interaction (HCI) and/or video-chat scenarios. This is particularly evident when the boundary of eyes and iris are blurry, and the shape of the eye can differ greatly due to pose variation and facial expression. To our knowledge, this work presents the first effort in solving the eye segmentation problem under such challenging conditions.

To investigate the topic of eye segmentation in the wild, the first problem we need to address is the lack of data. Albeit both biometric community and facial analysis community published an abundance of eye datasets over the years, none can be used as is for our purpose, because the former category only contains high resolution eye scans while the latter category lacks annotation of segmentation masks for sclera and iris. In fact, existing databases were collected in controlled environment (and mainly in high resolution), while there is no in-the-wild eye database that contains eye images in a wide range of resolutions. As a step towards the solution, we create a sizable eye segmentation dataset of 8882 eye patches by manually annotating 4461 face images selected from HELEN [23], 300VW [35], 300W [33], CVL [28], IMDB [32], Utdallas Face database [26], and Columbia Gaze database [37].

To solve the segmentation problem, we propose a novel method, Shape Constrained Network (SCN), that incorpo-

*Corresponding author

rates shape prior into the segmentation model. Specifically, we first pre-train a VAE-GAN [22] on the ground truth segmentation masks to learn the latent distribution of eye shapes. The encoder and discriminator are then utilised to regularise the training of the base segmentation network through the introduction of shape embedding loss and shape discriminator loss. This approach not only enables the model to produce accurate eye segmentation masks, but also helps it suppress artifacts, especially on low-quality images where the fine details are missing. In addition, since the regularisation is applied during the training, SCN does not incur additional computational cost to the base segmentation network during inference. Through extensive experiments, we demonstrate that SCN outperforms state-of-the-art segmentation and landmark localisation methods in terms of mean mIoU metric.

The main contribution of this work are as follows:

- We collect and annotate a large eye segmentation dataset consisting of 8882 eye patches from 4461 face images in the wild, this is the first of its kind and a significant step towards solving the problem of eye segmentation.
- We propose Shape Constrained Network (SCN), a novel segmentation method that utilises shape prior to increase accuracy on low quality images and to suppress artifacts.
- We redesign the objective function of SegNet with three new losses: Intersection-over-Union (IoU) loss, shape discriminator loss and shape embedding loss.

2. Related Works

Eyes localisation. Early methods [12, 11] often rely on edge information of the original image or handcrafted feature map when locating eyes and iris. In [11], the eye can be modelled as two parabolic curves (lids) and an ellipse (iris) respectively, whose parameters are determined by Hough transformation. Even though this method has been widely used in many iris recognition systems, it is very sensitive to image noises and pose changes. On a separate note, these algorithms are designed to work on eye scans of high quality (i.e. minimum of 70 pixels in iris radius), whereas for an in-the-wild image captured with consumer-grade camera, they do not perform well.

Everingham and Zisserman [15] attempted to solve this problem with 3 different approaches: (a) ridge regression that minimizes errors in the predicted eye positions; (b) a Bayesian model of eye and non-eye appearance; (c) a discriminative detector trained using AdaBoost. This is one of the earliest detectors that achieved some degrees of success in detecting eyes from the low resolution images. However, it still felt short of detecting eyes in extreme poses and illumination conditions, partly because it utilized image intensities rather than robust image feature (e.g., HoG [10]). Needless to say, they merely detected two landmarks, which were not sufficient for dense segmentation.

As a matter of fact, many existing 2D/3D facial landmarks detection methods [40, 19, 4, 1, 2] are able to provide significantly better localisation of eyes than the aforementioned methods, owing to the tremendous efforts in collecting and annotating large facial image databases [33, 35, 23, 13]. Unfortunately, the majority of these works only provide a small number of landmarks for one single eye (e.g., 6 landmarks in 68-point markup [33]), which is barely enough for describing the full structure of eye (i.e., iris, pupil and sclera) in a 2D image. Moreover, a significant portion of those annotated images do not display clear structure of eyes. To the best of our knowledge, there is no large scale database for dense eye landmarks localisation or eye segmentation. In this paper, we take a step forward by collecting the first in-the-wild eye database that is annotated with landmarks and fine-grained segmentation mask.

Deep semantic segmentation of image. The above methods are all condition-sensitive algorithms, as they are meticulously designed based on the predefined setting (e.g., the number of points, shapes or curves), thus may not suit our specific purpose. More recently, various deep learning techniques have achieved impressive results in semantic segmentation of images. Fully Convolutional Networks (FCN) [24] is one of the most influential deep learning methods for image segmentation. FCN is indeed an encoder-decoder network that predicts the segmentation mask in an end-to-end manner. It adopts VGG-16 [36] as the backbone of encoder, and utilises the transposed convolution for upsampling and generating the mask. SegNet [3] also adopted VGG-16 in the encoder network, however, comparing with FCN, it removed the fully connected layers and led to a more light-weight model. Additionally, inspired by unsupervised feature learning [17], the decoder of SegNet employed the max-unpooling layers, which reuse indices of the corresponding max-pooling operations of the encoder. The reuse of indices not only improves boundary delineation but also helps reduce the number of training parameters. DeepLab [7] proposed to use Atrous Convolutional Neural Network (Atrous-CNN) to generate the segmentation mask directly from the input image. The mask is further refined by a fully-connected Conditional Random Field (CRF) layer with mean-field approximation for fast inference.

One drawback of these methods is that they need to learn the shape prior from input image from scratch, which is often an inefficient procedure. Since the shapes of sclera and iris are highly regular, shape information can be exploited for eye segmentation. On the other hand, in low resolution images that do not display many details (such as prominent

edges), neglecting a shape prior can lead to sub-optimal performance for this task because the pixel intensity alone does not provide sufficient contextual information.

Deep generative models with shape constraint. Several deep generative models that take advantage of shape prior have been developed. Shape Boltzmann Machine (ShapeBM) [14] provided a good way to construct a strong model of binary shape using Deep Boltzmann Machines (DBMs) [34]. ShapeBM is an inference-based generative model that can generate realistic and different examples from the training data. Nonetheless, ShapeBM is quite sensitive to the appearance changes of object in different views, thus it is less appealing for the task of eye segmentation in-the-wild. More recently, Anatomically Constrained Neural Networks (ACNN) [31] incorporated shape prior knowledge into semantic segmentation or super-resolution models. Since the shape prior of ACNN were learned by auto-encoder, the reconstructed segmentation masks were often blurry and lack sharp edges. Shape prior can also be modelled in Variational Auto-Encoder (VAE) [21]. VAE tries to learn latent representation of training examples by mapping them to a posterior distribution. Unfortunately, VAE still fails to produce clear and sharp segmentation mask. To address this problem, Larsen et al. [22] presented VAE-GAN that combined VAE and GAN with a shared generator. The element-wise reconstruction error of VAE is replaced by feature-wise errors to better capture data distributions. VAE-GAN can optimally balance the similarity and variation between the inputs and outputs.

3. Dataset

Due to the lack of available data for eye segmentation in-the-wild, we create a new dataset by annotating 4461 facial images found in HELEN [23], 300VW [35], 300W [33], CVL [28], IMDB-WIki [32], Utdallas Face database [26], and Columbia Gaze database [37]. The particular images were selected to ensure a variety of head poses, image qualities, resolutions, eye shapes and gaze directions are represented in this dataset.

Once the facial images are collected, we use an facial landmark detector [19] to find an approximate location of the eyes in each image. For each eye patch, we manually annotate the segmentation mask. Each pixel in the patch is labelled as either background, sclera, or iris. Based on the annotated segmentation mask, the bounding box of the eye patch is then adjusted accordingly so that it is always centred on the eye with a fixed aspect ratio of 2:1. Some examples of the eye patches and their corresponding segmentation masks are illustrated in Figure 1.

Each eye patch is further tagged with 3 discrete attributes: head pose (near-frontal or non-frontal), resolution (high resolution or low resolution), and occlusion. The *head pose* attribute is manually annotated following the guideline that a head-yaw within 30 degree is considered *near-frontal* while the rest being considered *non-frontal*. The *resolution* tag is derived by comparing the eye patch's area to a fixed threshold of 4900 pixels, which is typically the number of pixels one can expect from a face image captured by 720P HD webcam during video chat. Distribution of the eye patch size in our dataset is shown in Figure 2. The *occlusion* attribute labels whether the image contains hairs, glasses, or profile view of the face (namely, self-occlusion). Detailed statistics of the dataset is given in Table 1.

Name	Value
Total number of faces	4461
Total number of eye patches	8882
Non-frontal faces proportion	18.35%
Low-resolution eye patches proportion	57.58%
Proportion of images with occlusions	16.05%

Table 1. Dataset statistics.

4. Shape Constrained Network

In this section, we illustrate the proposed Shape Constrained Network (SCN). SCN mainly contains a segmentation network and a shape regularization network, we design the loss functions for each part of network and explain the training of SCN in details.

4.1. Overview

We adapt SegNet [3] for our front-end segmentation network, and employ VAE-GAN [22] to regularise the predicted shape as well as to discriminate between real and fake examples. Our network is depicted in Figure 3. The training of SCN is divided into two steps: (1) First, we pretrain shape regularisation network (i.e., VAE-GAN) using the ground truth eye segmentation masks; (2) We borrow its encoder $E(.)$ and discriminator $D(.)$ for training our main segmentation network $S(.)$. The inference of SCN is indeed the same as that of SegNet, as we do not alter its encoder-decoder structure, we mainly reformulate the losses and improve the training by adding shape regularization.

4.2. Modeling shape prior

We utilise VAE-GAN [22] to learn the shape prior from ground truth segmentation masks. Simply put, VAE-GAN is a combination of Variational Auto-Encoder (VAE) and Generative Adversarial Networks (GANs), where they share a common decoder/generator. Specifically, in VAE, encoder tries to learn the parameters that map segmentation masks to the latent space of $\mathcal{N}(0, I)$, while the generator decodes the latent vector $z \sim \mathcal{N}(\mu, \sigma)$ to synthesise segmentation mask. In the part of GANs, the discriminator takes the generated mask and ground truth mask, and learns

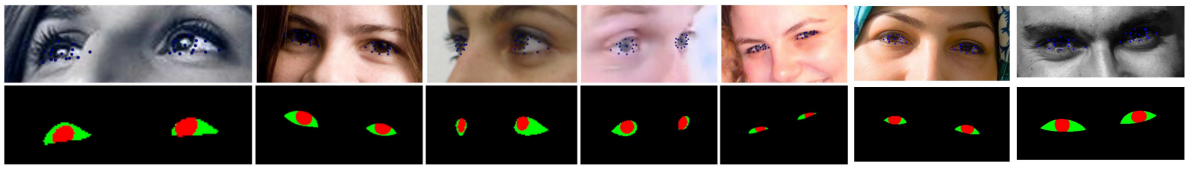

Figure 1. Examples of the eye patches (top row) and their corresponding segmentation masks (bottom row). Control points used to generate the segmentation masks are also made visible.

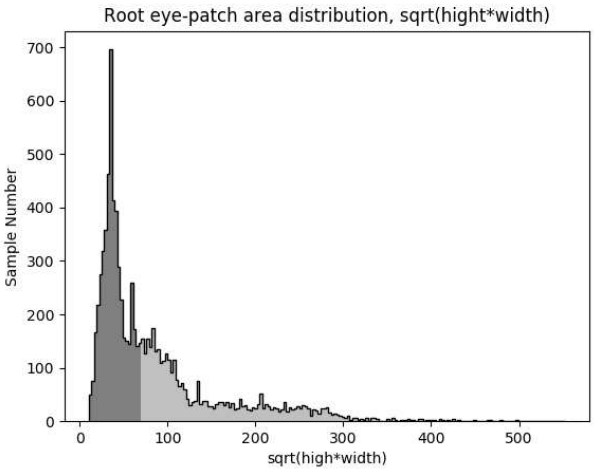

Figure 2. Distribution of eye-patch size (measured by the square root of area) in our dataset. The shaded part at the lower-end of the histogram indicates the samples tagged as 'low resolution'.

to judge between *real* and *fake*. Given a training example y, the training losses of VAE-GAN can be written as:

$$\begin{aligned} \mathcal{L}_{prior} &= D_{KL}(q(z|y)\|p(z)), \\ \mathcal{L}_{rec} &= \mathbb{E}_{q(z|y)}[\log p(D_l(y)\|z))] \\ \mathcal{L}_{gan} &= D(y) + \log(1 - D(\hat{y})) + \log(1 - D(\hat{y}_p)) \\ \mathcal{L}_{total} &= \mathcal{L}_{prior} + \mathcal{L}_{rec} + \mathcal{L}_{gan}, \end{aligned} \tag{1}$$

where $\hat{y}$ and $\hat{y}_p$ are the masks generated from the feature embedding z of ground truth data and randomly sample latent vector $z_p \sim \mathcal{N}(0, I)$ correspondingly. $q(z|y)$ presents the distribution of latent vector z given the input y, $p(z)$ is the normal distribution; $D_{KL}(.)$ is the KL divergence, and $\mathcal{L}_{prior}$ constrains the latent distribution to Gaussian. $D(.)$ and $D_l(.)$ denotes the discriminator and its feature from the l^{th} hidden layer respectively. $\mathcal{L}_{rec}$ is the reconstruction loss measuring the Euclidean distance of l^{th} hidden layer's output in the discriminator between the original image and the image reconstructed by auto-encoder. In VAE-GAN, the similarity of the ground truth and the reconstructed image is not evaluated directly. Instead, they are first fed into the discriminator and the distance between their l^{th} feature maps is used to measure the similarity. $\mathcal{L}_{gan}$ is an adversarial loss to play the minimax game between three candidates: original images, reconstructed images and images randomly sampled from latent space. The original images provide the discriminator with real examples, while the other two candidates aim at fooling the discriminator. The authors of VAE-GAN did not indicate any method to choose the l^{th} hidden layer. Theoretically, l can be any hidden convolutional layer in the discriminator. In this paper, we empirically chose l=1.

4.3. Eye segmentation network

We borrow the architecture of SegNet [3] for our eye segmentation network, but reformulate the loss function to improve the segmentation accuracy and robustness. As mentioned previously, SegNet is indeed an encoder-decoder network without fully connected layers, this is achieved by reusing pooling indices calculated in the max-pooling step of the encoder to perform non-linear upsampling in the corresponding decoder. Owing to this, our segmentation network has less trainable parameters while maintaining a good performance.

4.3.1 Network loss design

Shape reconstruction loss. Based on VGG-16 [36], SegNet employs softmax cross entropy as the loss function, however, as Intersection-over-Union (IoU) is more effective in evaluating the segmentation accuracy, we replace the original loss with the differentiable IoU loss [30]. Moreover, comparing with cross entropy loss, IoU loss can better balance the contribution from different regions, thus avoiding the domination of one particular category (i.e., the background pixels, especially when the eye is nearly closed). This loss is defined as:

$$\mathcal{L}_{iou} = \frac{\hat{y} * y}{\hat{y} + y - \hat{y} * y + \epsilon}, \tag{2}$$

where $\hat{y}$ and y indicate reconstructed mask and ground truth mask respectively, both variables are in the region of $[0, 1]$. ϵ is a very small number to avoid division by zero.

Shape embedding loss. Regularisation of the eye shape is important for producing a good segmentation mask. Inspired by ACNN [27], we regularise the shape prediction in the latent space of pre-trained VAE-GAN. Given a training image I, the segmentation network predicts the mask $\hat{G}$, which can be encoded to $\hat{z}$ such that $\hat{z} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$ by VAE.

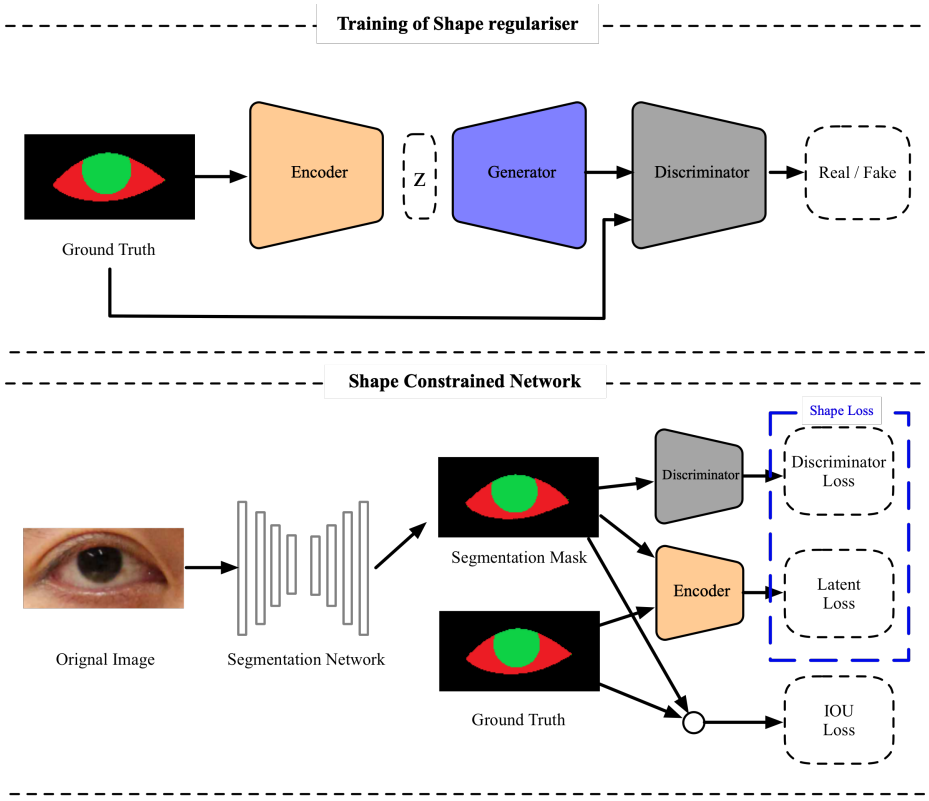

Figure 3. Overview of proposed Shape Constrained Network. SCN is constructed by VGG-16 based SegNet and VAE-GAN. We first use ground truth segmentation masks to train a VAE-GAN and reuse its encoder and discriminator. After that, we use a segmentation network to predict the segmentation mask (i.e., $S(x)$). which is fed into the pre-trained encoder to obtain the latent code, z. Meanwhile, the ground truth segmentation mask is also encoded into the latent space to obtain a ground truth latent code, $\hat{z}$. Therefore, we can use these two latent representations to formulate a shape embedding loss (see Eq. 3). We use the pre-trained discriminator to judge the realness of the predicted mask. A differentiable IoU loss is also employed to ensure the accuracy of reconstruction.

Similarly, the ground truth mask G can also be encoded, i.e., $z \sim \mathcal{N}(\mu, \sigma)$. Assume the distance between two latent vectors is $d = z - \hat{z}$, where $d \sim \mathcal{N}(\mu - \hat{\mu}, \sqrt{\sigma^2 + \hat{\sigma}^2})$, to ensure that feature embedding of predicted mask lies close to that of ground truth, we need to minimise the expectation $\mathbb{E}[d^2]$ of error distance d in terms of L2-norm. Therefore, the latent loss can be computed as:

$$\mathcal{L}_z = \mathbb{E}[d^2] = \mathbb{E}^2[d] + Cov[d] = (\mu - \hat{\mu})^2 + \sigma^2 + \hat{\sigma}^2,$$

since the variance σ of ground truth mask feature embedding is not related to any segmentation model parameters, it can be left out. Our shape embedding loss function becomes:

$$\mathcal{L}_z = (\mu - \hat{\mu})^2 + \lambda_z \hat{\sigma}^2, \tag{3}$$

where λ_z is used to balance the precision and error tolerance.

Shape discriminator loss. The discriminative power of VAE is usually not strong enough to single out hard negative examples, hence, we propose a discriminator loss to further regularise the generated mask. This loss is defined as follows:

$$\mathcal{L}_{disc} = \mathbb{E}[\log(1 - D(\hat{y}))]. \tag{4}$$

Although the discriminator loss can improve the quality of the segmentation result, it might also prolong the convergence of training. Therefore, it is important to weight the contribution of this loss.

4.3.2 Objective function

Combining Eq. 2, 3 and 4, we formulate the final objective function as follows:

$$\mathcal{L} = \mathcal{L}_{iou} + \lambda_1 \mathcal{L}_z + \lambda_2 \mathcal{L}_{disc}, \tag{5}$$

where λ_1 and λ_2 are two hyper parameters for trade-off between two shape regularisation losses, viz. shape embedding loss and shape discriminator loss.

4.4. Training of Shape Constrained Network

The segmentation network and shape regularisation network need to be trained separately. First, we train the VAE-

Algorithm 1 Training of Shape Constrained Network

Require: $\theta_s, \theta_e, \theta_g, \theta_d \leftarrow$ initialise network parameters.
repeat
  $y \leftarrow$ sample mini-batch from ground truth masks.
  $z \leftarrow E(y)$
  $z_p \leftarrow \mathcal{N}(0, I)$
  $\mathcal{L}_{prior} \leftarrow D_{KL}(q(z|y)\|p(z))$
  $\hat{y}_p \leftarrow G(z_p)$
  $\hat{y} \leftarrow G(z)$
  $\mathcal{L}_{rec} \leftarrow -(D_l(y) - D_l(\hat{y}))^2$
  $\mathcal{L}_{gan} \leftarrow D(y) + \log(1 - D(\hat{y})) + \log(1 - D(\hat{y}_p))$
  Updating parameters:
  $\theta_e \leftarrow \theta_e - \nabla_{\theta_e}(\mathcal{L}_{prior} + \mathcal{L}_{rec})$
  $\theta_g \leftarrow \theta_g - \nabla_{\theta_g}(\alpha\mathcal{L}_{rec} - \mathcal{L}_{gan})$
  $\theta_d \leftarrow \theta_d - \nabla_{\theta_d}\mathcal{L}_{gan}$
until Converged
Freeze θ_e and θ_d.
repeat
  $x, y \leftarrow$ sample mini-batch from the dataset.
  $\hat{y} \leftarrow S(x)$
  $\hat{z} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}) \leftarrow E(\hat{y})$
  $z \sim \mathcal{N}(\mu, \sigma) \leftarrow E(y)$
  $\mathcal{L}_z \leftarrow (\mu - \hat{\mu})^2 + \lambda_z\hat{\sigma}^2$
  $\mathcal{L}_{iou} \leftarrow \frac{\hat{y}*y}{\hat{y}+y-\hat{y}*y+\epsilon}$, $\epsilon = 1e^{-8}$
  $\mathcal{L}_{disc} \leftarrow \log(1 - D(\hat{y}))$
  Updating parameters:
  $\theta_s \leftarrow \theta_s - \nabla_{\theta_s}(\mathcal{L}_{iou} + \lambda_1\mathcal{L}_z + \lambda_2\mathcal{L}_{disc})$
until Converged

GAN using only ground truth segmentation masks. Our objective is to obtain a discriminative latent space to represent the underlying shape distribution $p(s|z; \theta_e, \theta_g, \theta_d)$, where s indicates the shape, θ_e denotes the parameters of encoder, θ_g describes the parameters of generator, θ_d are discriminator parameters and z is the latent vector.

Next, we freeze all the parameters of VAE-GAN, and connect the encoder and discriminator to the end of an untrained segmentation network. These two modules are only used to compute the shape embedding and discriminator losses as defined in Eq. 3 and 4, whilst their parameters will not be altered. Last, we train the segmentation network using the loss function Eq. 5.

Algorithm 1 shows the step-by-step training procedure of SCN. In that, $S(.)$ describes the segmentation network, $E(.)$ is the encoder, and $G(.)$ is the generator. θ_s indicates the parameters of segmentation network.

5. Experiments

All experiments were performed on the aforementioned eye dataset, which is further divided into separate train, validation, and test sets with the ratio of 8:1:1. The three subsets were constructed in a subject-independent manner such that images of the same subject (as extracted from the meta data) are always put into the same subset.

During the experiments, mean IoU metric is used to to evaluate segmentation accuracy on sclera (S-mIoU), iris (I-mIoU), and the combined foreground classes (Mean mIoU). To ensure a fair comparison, all methods under comparison were re-trained on the same training set as ours using their publicly available implementation. Paired T-test with Bonferroni correction were applied to all results to test whether the performance difference between our proposed approach and the compared method is statistically significant.

5.1. Implementation details

Our method is implemented using TensorFlow. Batch normalization [18] is used before each weight layer in the network. During training, data augmentation was performed by random horizontal flipping of the images. Adam optimizer [20] with a learning rate of 0.0002 was used for training the networks. For the shape regulariser, since it is difficult to test the convergence of GAN [9], the network was trained Figfor a fixed number of 100 epochs. For the segmentation network, early stopping [6] was used to prevent over-fitting, with the number of tolerance steps set to 50. The weights λ_1 and λ_2 for the two shape loss terms were both set to 0.3.

5.2. Ablation study

An ablation study was performed to verify that both the shape embedding loss and the shape discriminator loss helped to significantly improve segmentation accuracy in terms of Mean mIoU. The results are shown in Table 2. As can be seen, adding shape embedding loss increased the Mean mIoU by 2%, while further adding the shape discriminator loss brought an additional 1.5% improvement.

Method	S-mIoU	I-mIoU	Mean mIoU
SCN (full loss)	**71.86%**	**86.18%**	**79.02%**
SCN (only with $\mathcal{L}_z$)	70.26%	84.69%	77.47%†
SegNet[3]	66.06%	82.92%	74.49%†

Table 2. mIoU accuracy of the baseline segmentation network as compared to SCN with full loss and SCN with only the shape embedding loss. † indicates significant difference (0.95 confidence) between the performance of our method and that of the compared method.

5.3. Comparison with state-of-the-arts

We compared SCN to a number of state-of-the-art segmentation method [41, 42, 25, 8, 3, 7], as well as three landmark localisation methods [4, 40, 19]. All compared methods were re-trained on the same training set during this experiment. The segmentation methods were trained

Method	S-mIoU	I-mIoU	Mean mIoU	Inference Time
SCN(ours)	**71.86%**	**86.18%**	**79.02%**	0.033s
FAN [4]	71.41%	85.95%	78.68%†	0.111s
PSPNet [42]	70.44%	85.40%	77.92%†	0.070s
DeepLab V3+ [8]	69.78%	85.46%	77.62%†	0.041s
DenseASPP [41]	68.34%	83.94%	76.14%†	0.137s
ERT[1] [19]	66.42%	83.57%	74.99%†	**0.003s**
SegNet [3]	66.06%	82.92%	74.49%†	0.033s
FCN [24]	63.91%	82.79%	73.35%†	0.033s
DeepLab V2 [7]	63.41%	82.01%	72.71%†	0.110s
SDM[2] [40]	61.37%	78.70%	70.03%†	0.037s

Table 3. mIoU and average inference speed achieved by SCN and other segmentation and landmark detection methods. The rows are sorted in descending order with respect to Mean mIoU. † indicates significant difference (0.95 confidence) between the performance of our method and that of the compared method. The experiment was performed on a machine with Intel Core(R) i7-6700 3.4GHz CPU, 32GB memory, and a single Nvidia GeForce GTX 1080 Ti GPU. Inference time is recorded for a single prediction.

and tested in the same setting as SCN. For the landmark localisation methods, the control points created during the annotation process were used as the training targets. During testing, we interpolated (cubic-spline for eyelids and ellipse for iris) the predicted landmark positions to create the segmentation mask for comparison. Result of this experiment is shown in Table 3. SCN achieved higher Mean mIoU than all other methods. Through paired T-test with Bonferroni correction, we further found that the differences are all statistically significant (95% confidence). Visualisation of some random examples for the best-performing methods are shown in Figure 4. It can be clearly seen that SCN is quite robust and less likely to produce artifacts, which is attributed to the shape constraint.

In addition to accuracy, we also report the inference time of each method in Table 3. Although ERT [19] has the shortest inference time, it is less accurate than most deep methods. Among all deep methods, SCN runs the fastest (0.033s per image), achieving the same speed as that of SegNet [3]. This is because the VAE-GAN is only used during training, thus does not incur additional computational cost during inference.

5.4. Cross-resolution comparison

In this experiments, we wanted to investigate how the change of image resolution might affect segmentation performance of our method. Different from previous experiments, we ensure that the train set only contains high-resolution images ($\sqrt{s_{eye}} \geqslant 70$, where s_{eye} is the area of eye patch in pixels), while the test set only contains low-resolution images. The ratio is roughly 5:1. All samples are resized to 160×80 for training and testing. We compared with six state-of-the-art segmentation methods in this experiment, the result is shown on Table 4. It is clear that SCN is consistently better than the other methods in S-mIoU and I-mIoU (at least 0.7% better in Mean mIoU), despite of the fact that fewer details are presented in the low-resolution image. Thereinto, S-mIoU and I-mIoU denote the intersection over union metric for sclera and iris, respectively. We attribute this to show that the shape prior knowledge learnt by VAE-GAN from only high-resolution data can also benefit low-resolution eye segmentation.

Method	S-mIoU	I-mIoU	Mean mIoU
Ours	**63.91%**	**80.95%**	**72.46%**
PSPNet [42]	63.31%	80.20%	71.76%†
DenseASPP [41]	61.09%	79.03%	70.06%†
DeepLab v3+ [8]	61.59%	78.54%	70.07%†
DeepLab V2 [7]	57.57%	76.79%	67.18%†
SegNet [3]	59.47%	76.62%	68.05%†
FCN [24]	57.71%	76.04%	66.88%†

Table 4. Model accuracy of cross-resolution comparism. SCN is significantly better than the other models' performance. The table shows SCN can be robust to adapt different image resolution conditions. † indicates significant difference (0.95 confidence) between the performance of our method and that of the compared method.

6. Conclusion

In this paper, we aimed at solving the problem of low-resolution eye segmentation. First, we proposed an in-the-wild eye dataset that includes 8882 eye patches from frontal and profile faces, the majority of which are captured in low resolution. We collected a significant number of samples that exhibit occlusion, weak/strong illumination and glasses. Then, we developed the Shape Constrained Network (SCN) that employs SegNet as the backend segmen-

[1]Using the implementation available at `https://github.com/davisking/dlib`

[2]Using the implementation available at `https://github.com/FengZhenhua/Supervised-Descent-Method`

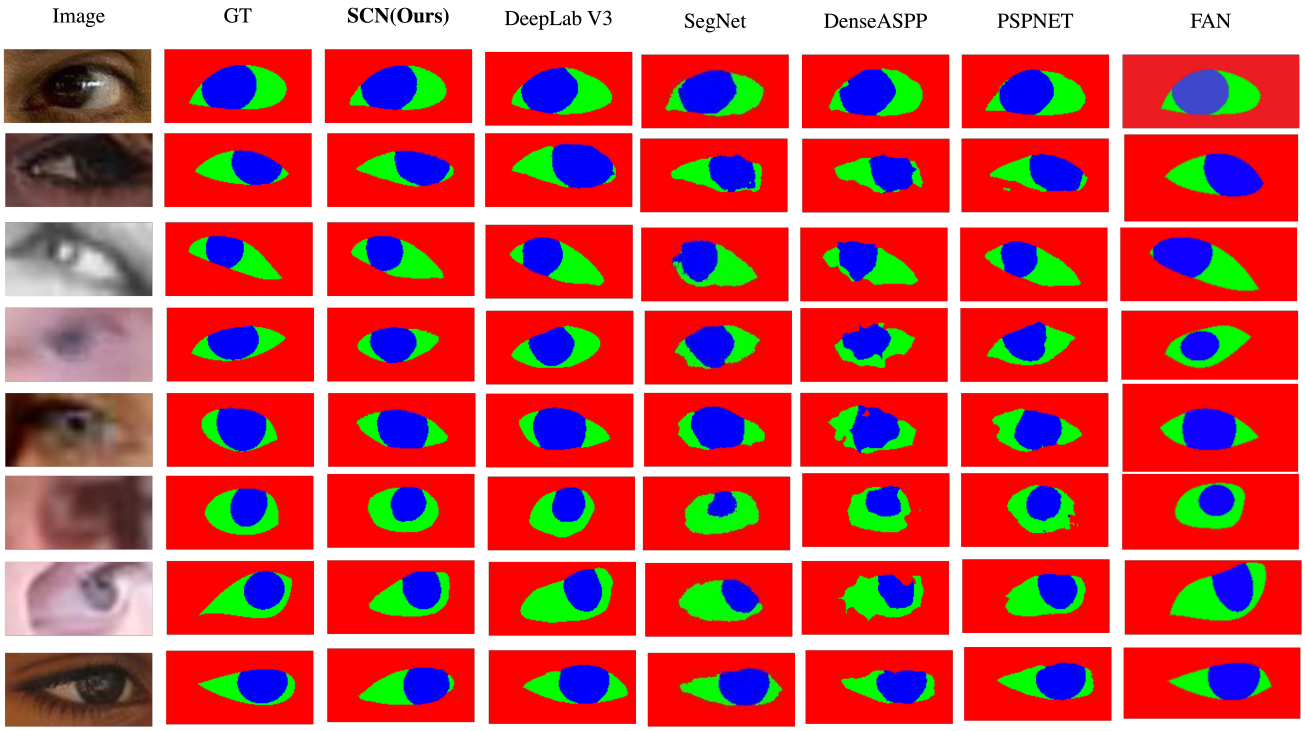

Figure 4. Qualitative visualisation of segmentation results based the eye segmentation dataset from SCN, SCNDeepLab V3SegNet, DenseASPP, PSPNET and FAN. Please check our supplementary materials for the visualisation results of Deeplab V2, FCN, SDM and ERT.

tation network, and we introduced shape prior to the training of SegNet by integrating the pre-trained encoder and discriminator from VAE-GAN. Based on the new training paradigm, we design three new losses: Intersection-over-Union (IoU) loss, shape discriminator loss and shape embedding loss.

We demonstrated in ablation studies that adding shape prior information is beneficial in training segmentation network. We outperformed several state-of-the-art segmentation methods as well as landmark alignment methods in subject-independent experiments. Last, we evaluate SCNs performance in low-resolution images, with a cross dataset experiment in which the model is trained on high-resolution data and tested on low-resolution data. The results show that SCN can generalise well to variations in image resolution.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451, 2013.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR*, pages 1859–1866, 2014.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *T-PAMI*, 39(12):2481–2495, 2017.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[5] H. Cai, B. Liu, Z. Ju, S. Thill, T. Belpaeme, B. Vanderborght, and H. Liu. Accurate eye center localization via hierarchical adaptive convolution. In *British Machine Vision Conference*. British Machine Vision Association, 2018.

[6] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pages 402–408, 2001.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE Computer Society, 2005.

[11] J. Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009.

[12] J. G. Daugman. High confidence visual recognition of persons by a test of statistical independence. *T-PAMI*, 15(11):1148–1161, 1993.

[13] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, pages 1–26, 2018.

[14] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn. The shape boltzmann machine: a strong model of object shape. *International Journal of Computer Vision*, 107(2):155–176, 2014.

[15] M. Everingham and A. Zisserman. Regression and classification approaches to eye localization in face images. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 441–446. IEEE, 2006.

[16] B. M. Hood, J. D. Willen, and J. Driver. Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, 9(2):131–134, 1998.

[17] F. J. Huang, Y.-L. Boureau, Y. LeCun, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[18] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[19] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[23] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692. Springer, 2012.

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[26] M. Minear and D. C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633, 2004.

[27] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. ORegan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2018.

[28] P. Peer. Cvl face database. *Computer vision lab., faculty of computer and information science, University of Ljubljana, Slovenia. Available at http://www. lrv. fri. uni-lj. si/facedb. html*, 2005.

[29] P. Polatsek. Eye blink detection. *Slovak University of Technology in Bratislava. Faculty of Informatics and Information Technologies. IIT. SRC*, 18, 2013.

[30] M. A. Rahman and Y. Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016.

[31] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya. Learning and incorporating shape models for semantic segmentation. In *MICCAI*, pages 203–211. Springer, 2017.

[32] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[33] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[34] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *AISTATS*, pages 448–455, 2009.

[35] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV-W*, pages 50–58, 2015.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280. ACM, 2013.

[38] F. Timm and E. Barth. Accurate eye centre localisation by means of gradients. *Visapp*, 11:125–130, 2011.

[39] Y. Wang, B. Luo, J. Shen, and M. Pantic. Face mask extraction in video sequence. *International Journal of Computer Vision*, 127(6-7):625–641, 2019.

[40] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[41] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.