BiOcularGAN: Bimodal Synthesis and Annotation of Ocular Images

Darian Tomasevic1,*, Peter Peer1,†, Vitomir Štruc2,‡

1 Faculty of Computer and Information Science, University of Ljubljana, Slovenia
2 Faculty of Electrical Engineering, University of Ljubljana, Slovenia
*darian.tomasevic@fri.uni-lj.si, †peter.peer@fri.uni-lj.si, ‡vitomir.struc@fe.uni-lj.si

Abstract

Current state-of-the-art segmentation techniques for ocular images are critically dependent on large-scale annotated datasets, which are labor-intensive to gather and often raise privacy concerns. In this paper, we present a novel framework, called BiOcularGAN, capable of generating synthetic large-scale datasets of photorealistic (visible light and near-infrared) ocular images, together with corresponding segmentation labels to address these issues. At its core, the framework relies on a novel Dual-Branch StyleGAN2 (DB-StyleGAN2) model that facilitates bimodal image generation, and a Semantic Mask Generator (SMG) component that produces semantic annotations by exploiting latent features of the DB-StyleGAN2 model. We evaluate BiOcularGAN through extensive experiments across five diverse ocular datasets and analyze the effects of bimodal data generation on image quality and the produced annotations. Our experimental results show that BiOcularGAN is able to produce high-quality matching bimodal images and annotations (with minimal manual intervention) that can be used to train highly competitive (deep) segmentation models (in a privacy aware-manner) that perform well across multiple real-world datasets. The source code for the BiOcularGAN framework is publicly available at https://github.com/dariant/BiOcularGAN.

1. Introduction

Modern biometric systems are predominantly based on convolutional neural networks (CNNs) and transformer models, which rely on massive annotated (training) datasets to achieve competitive performance [29]. While large-scale datasets can today easily be collected from the web for many biometric modalities, such collection procedures often raise privacy and copyright-related concerns [12, 25]. Additionally, the annotation of such large-scale datasets is today (in most cases) still a manual, labor-intensive, and time-consuming task. These points are especially true for datasets dedicated to the segmentation of ocular images (in various imaging domains), where, next to the data collection, the generation of high-quality (multi-class) semantic annotations is known to be a costly endeavor [40, 42].

Researchers are, therefore, increasingly looking into automatic techniques that allow for the generation of synthetic datasets that require no (or minimal) human intervention during the annotation process [8, 24, 30, 44]. However, several challenges are associated with such an approach: (i) the synthetic (training) samples need to be as close as possible to the expected real-world data to allow for the trained model to perform well during deployment, (ii) the synthesis procedure must allow for the generation of large and diverse datasets that can cater to the data needs of modern deep learning models, and (iii) data annotations need to be produced automatically, without (or with minimal) supervision. To meet these challenges, existing solutions often resort to Generative Adversarial Networks (GANs) [10, 13] due to their ability to generate highly photorealistic and detailed synthetic data and the fact that the model’s internal representations can be exploited to generate semantic segmentation labels alongside the generated images [44].

Motivated by the needs for large-scale synthetic datasets and the capabilities of recent generative models, we present in this paper a novel data generation framework, called BiOcularGAN, capable of generating aligned photorealistic (bimodal) ocular images in the visible (VIS) and near-infrared (NIR) spectra along with corresponding segmentation masks, as illustrated in Figure 1. The key components of the framework are (i) a novel dual-branch StyleGAN2 (DB-StyleGAN2) model, which extends the capabilities of previous StyleGAN versions to bimodal data synthesis, and (ii) a data annotation procedure, inspired by [44], that exploits the semantic information encoded by the bimodal synthesis network for segmentation mask generation. We evaluate the proposed approach in experiments with five diverse datasets and investigate the impact of the bimodal (VIS and NIR) generation process on the quality of the synthesized images. Furthermore, we analyze the ability of BiOcularGAN to generate useful datasets by observing how well current semantic segmentation models, trained on syn-

Figure 1: Example data generated with BiOcularGAN. The proposed framework is based on a novel Dual-Branch StyleGAN2 model and can generate (synthetic) per-pixel aligned visible light (VIS) and near-infrared (NIR) ocular images as well as corresponding segmentation masks.
thetic labeled data, generalize to diverse real-world datasets. In summary, we make the following main contributions:

• We present BiOcularGAN, a powerful framework for generating large labeled datasets of ocular images based on bimodal data representations that can be used to train contemporary segmentation models.

• We design a novel bimodal generative model, i.e., the Dual-Branch StyleGAN2 (DB-StyleGAN2), capable of synthesizing visually convincing (aligned) ocular images in both the visible and near-infrared domains.

• We show that using bimodal information as the basis for generating ground truth segmentation masks leads to improvements in the quality of the generated annotations compared to solutions using only a single modality, e.g., the state-of-the-art DatasetGAN [44].

2. Related work

Image and Dataset Generation. Image synthesis techniques have experienced rapid development in the past decade, most notably due to the introduction of Generative Adversarial Networks (GANs) [10]. Over time, a myriad of improvements and iterations to the GAN model have been proposed, from manipulating latent space distributions [3] to using multiple discriminator networks [7]. Despite numerous advancements [13, 27], some of the inner workings of the generator networks remained poorly understood [1].

More recently, a powerful new generation model, called StyleGAN, was proposed by Karras et al. in [16]. With its high-resolution image synthesis capabilities, the model drastically outperformed other unconditional image generation techniques across a variety of datasets. Since then, the authors further iterated on the model (with StyleGAN2 and StyleGAN3) [17, 15] and addressed several of its characteristic artifacts with changes to model architecture and training procedures. Most notably, Karras et al. [14] also introduced various image augmentations to the discriminator, thus immensely lowering the amount of training data required to train the StyleGAN2 model.

Several approaches have also been proposed to enable the synthesis of segmentation masks alongside images generated by StyleGAN, either by using separate generator branches [24] or by exploiting the feature space of the generator [30, 44]. The latter approach showcased the ability to generate high-quality datasets of paired images and segmentation masks, with only a few annotated examples, and was aptly named DatasetGAN [44]. In this paper, we build on the outlined advances and present, to the best of our knowledge, the first StyleGAN2-based model for bimodal data synthesis. As we demonstrate in the experimental section, the model leads to visually convincing generation results and allows us to synthesize large datasets of matched ocular images in the VIS and NIR imaging domains with corresponding ground truth segmentation masks.

Ocular Synthesis. Despite the considerable progress in generative models, only a limited number of solutions capable of generating photorealistic high-quality ocular images have so far been presented in literature. Shrivastava et al. [37] presented one of the initial GAN-based models for ocular synthesis, capable of converting pre-rendered ocular images [41] into more realistic ones. Lee et al. [23] built on this approach with the use of CycleGAN [45]. However, the resulting images remained rather noisy and often did not match the original gaze direction. Concurrently, Kohli et al. [21] explored convolutional GAN models for iris generation. Despite significant artifacts, they successfully performed presentation attacks on the recognition systems of the time. Based on the need for large datasets, Facebook organized the OpenEDS Synthetic Eye Generation challenge [9]. Buhler et al. [4] emerged victorious with their Seg2Eyes model, a mix of StyleGAN [16] and Gaussian [31], capable of generating identity-preserving ocular images based on the desired style and input segmentation masks. Kaur et al. [18] introduced the EyeGAN model for the same task, and later upgraded it with a cyclic training mechanism [19] to ensure consistency of gaze direction and style. Boutros et al. [2] proposed an alternative solution to the problem with a novel D-ID-network solution. Nevertheless, the generated images still featured visible artifacts.

Despite significant improvements in ocular synthesis, all current approaches generate images of only a single modality. In addition, they are mostly focused on identity-preserving image generation and feature mechanisms that can limit the diversity of generated synthetic data. Different from these works, we focus in this paper on the generation of diverse and appearance-rich datasets of bimodal VIS and NIR data, along with matching synthetically-generated reference annotations. The bimodal aspect is especially useful from a segmentation aspect, since NIR images often contain important cues that are not present in VIS images, and vice versa. Furthermore, we base our work on insights from state-of-the-art image generation techniques, i.e., StyleGAN2 [14, 17], allowing us to learn highly successful models using a limited amount of training data.

3. Methodology

The main contribution of this work is the BiOcularGAN framework that allows for photo-realistic generation of bimodal ocular images and the corresponding reference segmentation masks. In this section, we describe BiOcularGAN in detail and elaborate on its main characteristics.

3.1. Overview of the BiOcularGAN framework

The proposed BiOcularGAN framework, depicted in Figure 2, consists of two key components. These being (i) the Dual-Branch StyleGAN2 (DB-StyleGAN2) generative model (§3.2), which generates pixel-aligned VIS and NIR ocular images (§3.3), and (ii) the Semantic Mask Generator (SMG) that produces corresponding semantic segmentation masks (§3.4). Jointly, these components allow for the generation of matching photo-realistic bimodal ocular images along with corresponding high-quality annotations and, consequently, for the creation of synthetic large-scale datasets that can be used for training data-hungry deep learning (segmentation) models in a privacy-aware manner, e.g., for semantic segmentation tasks.
Formally, the BioCularGAN generator \( G \) begins with an input latent code \( z \in \mathcal{Z} \) that is first transformed into an intermediate latent representation \( w \in \mathcal{W} \) and then fed to the DB-StyleGAN2 synthesis network \( g \), which produces the pixel-aligned VIS and NIR ocular images, \( \mathbf{x}_{\text{vis}} \in \mathbb{R}^{W \times H \times 3} \) and \( \mathbf{x}_{\text{nir}} \in \mathbb{R}^{W \times H \times 3} \), respectively, i.e.:
\[
\{\mathbf{x}_{\text{vis}}, \mathbf{x}_{\text{nir}}\} = G(z) = g(f(z)),
\]
where the latent-space transformation \( w = f(z) \) is implemented with a mapping network \( f \), as shown in Figure 2. To generate the semantic segmentation masks, the feature maps computed along the different layers of DB-StyleGAN2 are pooled and then fed to the semantic mask generator \( S \), similarly to [44], i.e.: \( \Omega = S(\phi_1(z), \phi_2(z), \ldots, \phi_k(z)) \), where \( \Omega \in \mathbb{R}^{W \times H} \) is the generated segmentation mask, \( \phi \) is a mapping implemented within the generator \( G \), and \( k \) is the number of feature maps used. Thus, given a latent code \( z \), drawn from a normal distribution, BioCularGAN generates a triplet of the following form: \( \{\mathbf{x}_{\text{vis}}, \mathbf{x}_{\text{nir}}, \Omega\} \).

3.2. Dual-Branch StyleGAN2

The key component of BioCularGAN is the novel Dual-Branch (DB) StyleGAN2 generator that extends the original StyleGAN2 [14, 17] for bimodal data generation. As illustrated in Figure 3, the generator consists of a mapping network \( f \) that follows a fully connected design, similarly to [17], as well as a dual-branch synthesis network, and is trained using two discriminators, \( D_{\text{VIS}} \) and \( D_{\text{NIR}} \), one for the VIS and one for the NIR images. Details on the generator and discriminators are given below.

The Generator \( (G) \) is responsible for producing the synthetic (NIR and VIS) ocular images and builds on recent insights and advancements in image generation [14, 17]. Similarly to the original StyleGAN2 design, it consists of a succession of synthesis blocks that produce images of progressively higher resolution, as shown on the left side of Figure 3. These consist of smaller style blocks (light gray boxes), which take the intermediate latent representation \( w \), transformed through \( k \) learned affine transformations \( A \), as the style input. Convolution weights \( w_x \) are then modulated based on the style input and later “demodulated” [17] – a procedure which mimics the effects of instance normalization. Style is thus incorporated into the convolution operation via the processed weights. The network starts from a constant input \( c \times 4 \times 4 \times 512 \). After each convolutional layer, the noise input (from the noise broadcast operation \( B \)) and bias \( b_x \) are applied to the signal, which is then passed through a leaky ReLU activation function. A unique feature of the proposed DB-StyleGAN2 model that enables bimodal image generation is the dedicated synthesis blocks that contain two output branches, one for generating VIS and the other for generating NIR data at a specific resolution. Here, each branch features a \( 1 \times 1 \) convolution layer, denoted \( f_{\text{VIS}} \) ("toVIS") and \( f_{\text{NIR}} \) ("toNIR") in Figure 3. The outputs of these branches are upsampled and merged with the output of the higher-resolution synthesis block to construct the final VIS and NIR images, thus, forming the DB-synthesis network, as seen on the right part of Figure 3.

The Discriminators \( (D_{\text{VIS}}, D_{\text{NIR}}) \) aim to determine, whether images are real or artificially generated, and help to ensure that the data generated by the DB generator is as close to the training data distribution as possible. For BioCularGAN, we utilize two discriminators, \( D_{\text{VIS}}, D_{\text{NIR}} \), one for each branch of the DB-synthesis network, corresponding to the VIS and NIR image modalities, as shown on the right side of Figure 3. The discriminators take a pair of real (or fake) bimodal images as input and first pass them through \( 1 \times 1 \) convolutional layers denoted \( f_{\text{VIS}} \) ("from VIS") and \( f_{\text{NIR}} \) ("from NIR"). The processed input is then passed through a ResNet-like [11] downsampling architecture, with each block consisting of two convolution layers and a separate skip connection. The output of each of the discriminators is a binary decision, i.e., real or fake. The two discriminators share the same architecture.

3.3. DB-StyleGAN2 training

Different from StyleGAN2, our dual-branch model produces two semantically similar output images in two distinct imaging domains. The training is, therefore, done with adversarial learning objectives involving two discriminators. Because of the (dual) bimodal output produced by DB-StyleGAN2, the training follows a multi-task learning regime, where the correlations between the two tasks (i.e., VIS and NIR image generation) help to efficiently
Table 1: Summary of the experimental dataset. We train (and validate) all components of BiOcularGAN on the cross-spectral datasets and evaluate segmentation performance on the visible spectrum datasets.

However, different from the procedure of Zhang et al. [44], we extract feature maps from each Leaky ReLU activation function in the dual-branch synthesis network (in Figure 3), related to a single style and resolution. This allows us to capture the semantic information of the bimodal ocular images before they are rendered in a certain imaging domain. We then upsample these feature maps to the output resolution and construct a $W \times H \times d$ tensor, from which $d$-dimensional feature vectors corresponding to each of the $W \times H$ image pixels can be obtained. Using the obtained high-dimensional feature vectors as input, we train an ensemble of 10 three-layer MLPs to classify pixels into the semantic classes. Here, manual annotations over an incredibly small set ($< 10$) of generated bimodal images are used as the ground truth for the training procedure. We note that a majority voting strategy is utilized over the predictions of the MLP ensemble to minimize the randomness of the learning stage. Once trained, the SMG can be used together with the DB-StyleGAN2 model to generate unlimited amounts of pixel-level aligned bimodal ocular images with corresponding semantic ground truth masks. Here, a single forward pass is needed to generate one triplet $\{x_{vis}, x_{nir}, \Omega\}$.

4. Experiments and result

4.1. Experimental setup

Datasets. We use five datasets for training and evaluation of BiOcularGAN, i.e., the PolyU cross-spectral Iris database (PolyU) [28], CrossEyed [35, 36], the Sclera Mobile Database (SMD) [6], SBVPI [34, 40] and MOBIUS [39]. The main characteristics of the datasets are summarized in Table 1, while the key details are provided below:

- **Cross-spectral datasets**: The PolyU and CrossEyed datasets contain ocular images captured in the near-infrared (NIR) and visible light (VIS) spectra. The acquisition procedure for both datasets was performed with custom sensors capable of simultaneous acquisition of the NIR and VIS images. The images in PolyU are aligned with pixel-level correspondences, while the CrossEyed data is loosely aligned, i.e., with small (random) perturbations in scale and position in the NIR-VIS image pairs. For our experiments, the image pairs of both datasets are split into subject disjoint training and evaluation parts in a ratio of 9 : 1. The training part

---

[Table 1: Summary of the experimental dataset.]

| Dataset          | # Images | # IDs | # Eyes | Resolution | Modality | Purpose |
|------------------|---------|------|-------|------------|----------|---------|
| PolyU [23]       | 12540   | 209  | 518   | 640 x 480  | NIR/VIS  | TR/SV   |
| CrossEyed [35, 36] | 1858    | 55   | 110   | 3264 x 2448| VIS      | SE      |
| SMD [6]          | 500     | 25   | 50    | 3264 x 2448| VIS/VIS  | SE      |
| MOBIUS [39]      | 3542    | 35   | 70    | 3000 x 1700| VIS      | SE      |
| SBVPI [34, 40]   | 1858    | 55   | 110   | 3000 x 1700| VIS      | SE      |

†TR – training, SV – synthesis validation, SE – segmentation experiments

---

Here, $d$ denotes the combined length of all extracted feature maps.
is used to learn the DB-StyleGAN2 model and SMG annotation procedure, whereas the (hold-out) evaluation part is reserved for the performance evaluation.

- **Visible spectrum datasets**: The SMD, SBVPI and MOBIUS datasets consist of high-resolution VIS ocular images captured primarily for research into sclera biometrics. All three datasets have manual annotations of some key regions of the ocular images, e.g., the sclera, iris or pupil, and are therefore used to evaluate the performance of the segmentation models trained with the annotated data generated by BiOcularGAN.

**Implementation Details.** All components of BiOcularGAN were implemented in PyTorch and are made publicly available from URL². The Dual-Branch StyleGAN2 is implemented based on the StyleGAN2-ADA variant [14]. The main part of the DB-StyleGAN2 is initialized with weights pretrained on the FFHQ dataset (of resolution 256 × 256) and then optimized further using the Adam optimizer [20] with a learning rate of 0.0025 and a batch size of 16. For the other hyperparameters, we use the recommended values $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$ for both, the generator and the two discriminators. We train all models for 2500 kimgs or until training diverges, due to the low amount of training data. To combat model divergence, we enable data augmentation in the form of horizontal image flipping and additionally employ the adaptive discriminator augmentation procedure proposed in [14]. For the Semantic Mask Generator (SMG), training is performed based on the cross-entropy loss and the Adam optimizer [20], with a learning rate of $10^{-4}$. Each MLP classifier is trained on randomly sampled image pixels in batches of 64. The training is stopped once no improvement is observed in the learning objective over 50 batches following the third epoch, similarly to [44]. Additional implementation details can be found in the publicly released source code.

**Experimental Hardware.** All experiments are conducted on a Desktop PC with an Intel i9-10900KF CPU with 64 GB of RAM and an Nvidia 3090 GPU with 24 GB of video RAM. Using this hardware, we trained two DB-StyleGAN2 models, one on PolyU and one on CrossEyed, denoted as **DB-StyleGAN2-P** and **DB-StyleGAN2-CE** hereafter. Once converged, the models are able to generate visually convincing bimodal ocular images of 256 × 256 pixels in size, as demonstrated in the following sections.

### 4.2. Synthesis evaluation

In the first set of experiments, we explore the capabilities of the trained DB-StyleGAN2 models.

**Visual Evaluation.** Figure 4 shows a selection of (real) VIS and NIR images from the PolyU and CrossEyed datasets, as well as a few examples generated by the two trained DB-StyleGAN2 models. As can be seen, both models are capable of generating high-quality and visually convincing images that well match the visual characteristics of the training data in the visual as well as near-infrared domain. The trained models are able to synthesize crisp image details, such as individual eyelashes, eyebrows, skin textures and even reproduce the specular reflections present in the training samples. Due to the dual-branch design of the DB-StyleGAN2 model, these fine image details are also consistent across the bimodal image pairs.

**VIS-NIR Pair Alignment.** While DB-StyleGAN2-P was trained on the per-pixel aligned data from PolyU, the training of DB-StyleGAN2-CE was performed with the loosely aligned images from CrossEyed. Nonetheless, both

---

²https://github.com/dariant/BiOcularGAN

---

Figure 4: **Visual examples of original and generated ocular images in both domains.** The first two columns show samples from the PolyU and CrossEyed datasets and the last two columns show examples of images generated by the DB-StyleGAN2 models trained on the PolyU (DB-StyleGAN2-P) and CrossEyed (DB-StyleGAN2-CE) datasets.
models produce well-aligned NIR and VIS images due to the shared style blocks in the StyleGAN2 model that capture the semantics of the ocular images, while the two branches generate the final output images within the specific imaging domains. To visualize the alignment of the original and synthesized image pairs, we generate composite images, where the RGB data from the VIS samples is first transformed into the YCbCr color space and the luma (Y) component is then replaced by the NIR image. This composition changes the color artifacts. Nevertheless, the trained DB-StyleGAN2 versions for the comparisons using the NIR and VIS images synthesized with the standard unimodal StyleGAN2 and the proposed bimodal DB-StyleGAN2.

### Image Diversity

Next, we qualitatively analyze the diversity of the images generated by the trained DB-StyleGAN2 models. Specifically, we are interested in the variations of ocular images the models are able to produce with respect to the data seen during training. To this end, we show in Figure 6 a randomly generated image pair produced by the DB-StyleGAN2-P and DB-StyleGAN2-CE models (left column of each presented example) as well as the most similar VIS-NIR pair from the training data – where the similarity is measured in terms of Mean Squared Error (MSE) between the VIS images. Several interesting observations can be made from the presented examples, i.e.: (i) the generated images share obvious similarities with the training data in terms of visual appearance, (ii) the models generate distinct data samples that differ from the training examples in terms of gaze direction, eye shape and color (for VIS), eyelash arrangement, eyelid appearance, pupil size, skin and iris texture, and other factors, and (iii) despite appearing similar in the VIS domain at first glance, considerable differences are present in the NIR domain in the presented examples, suggesting that the combined (bimodal) ocular images generated by the models are distinct.

### State-of-the-Art Comparison and Ablations

We compare the DB-StyleGAN2 models to the standard unimodal StyleGAN2 model from [16]. We note that StyleGAN2 represents a state-of-the-art model for image generation and while StyleGAN version 3 (StyleGAN3) was also introduced recently [15], it only offers superior performance (in terms of texture consistency) when generating sequences of images (or videos) but does not ensure improvements in the quality of the generated images. We train four StyleGAN2 versions for the comparisons using the NIR and VIS images from the two training datasets, i.e., PolyU and CrossEyed. The experiments presented in this section serve

![Image](image-url)
In Table 2 we show a comparison of the Learned Perceptual Image Patch Similarities (LPIPS) [43] between 5000 images generated by the unimodal models, whereas this is handled seamlessly in DB-StyleGAN2 through the dual-branch design. Using the unimodal models, whereas this is handled seamlessly in DB-StyleGAN2 through the dual-branch design. Because the generation process is based on latent space sampling, it is also challenging to produce matching samples in both domains. However, the DB-StyleGAN2 models are able to synthesize the bimodal images through a single generation step, whereas separate models need to be trained for the off-the-shelf StyleGAN2 generators. Because the generation process is based on latent space sampling, it is also challenging to produce matching samples in both domains using the unimodal models, whereas this is handled seamlessly in DB-StyleGAN2 through the dual-branch design. In Table 2 we show a comparison of the Learned Perceptual Image Patch Similarities (LPIPS) [43] between 5000 randomly generated images and the training (T) and hold-out validation (H) data from each dataset. As can be seen, on PolyU all models perform similarly (within the standard deviations), whereas our bimodal design has a slight advantage on CrossEyed, suggesting that the generated images are somewhat closer to the real data on average.

To get further insight into the synthesis capabilities of DB-StyleGAN2, we use t-distributed Stochastic Neighbor Embedding (t-SNE) [38] and visualize the distribution of features extracted from different types of images in Figure 8. For this purpose, we select a ResNet-101 model pretrained on ImageNet (from PyTorch) as a feature extractor and use the 2048-dimensional output of the penultimate model layer as the feature representation of the ocular images [11]. We generate 250 test images for the analysis by randomly sampling the latent space of the two DB-StyleGAN2 and all four unimodal StyleGAN2 models. As can be seen, the distributions corresponding to the generated images overlap reasonably well with the distributions of the original images (marked Real) for both types of models. However, in certain cases the unimodal models generate less overlap with the training-data distribution than the bimodal models – see results for CrossEyed VIS for example.

### Table 3: Training and run-time requirements.

The bimodal DB-StyleGAN2 model takes longer to train than the unimodal StyleGAN2, but is able to match the run-time performance of its unimodal counterpart.

| Model                  | Training time [hours]† | Run-time [ms]          |
|------------------------|------------------------|------------------------|
| DB-StyleGAN2           | ~ 20h                  | 13.994 ± 0.068         |
| StyleGAN2              | ~ 18h                  | 11.232 ± 0.071         |

†Approximate estimate

### Figure 9: Sample segmentation results.

The results were generated with two U-Net models, trained on artificial data generated by the DatasetGAN and BiOcularGAN frameworks, learned with the DB-StyleGAN2-P model.

### 4.3. Bimodal data annotation and segmentation

In the second set of experiments, we explore the advantages that bimodal information brings to the ground truth segmentation-mask generation process. To this end, we manually annotate 8 ocular images generated by each of the two DB-StyleGAN models using 4 target segmentation classes, i.e., the pupil, the iris, the sclera and the background. We use the NIR images as the basis for the manual annotation procedure (due to better contrast, distinct borders, etc.), but due to the alignment of the artificial bimodal images, these segmentation masks are also applicable to the...
Table 4: Cross-dataset segmentation performance comparison of models trained on artificially generated datasets. The segmentation models trained on 5000 images generated by BiOcularGAN outperform the ones trained on 5000 images generated by DatasetGAN across all datasets and performance measures (IoU and F1 scores along with pixel errors).

![DeepLab-V3](https://github.com/jkn1314/DeepLabV3) [FineTuning](https://github.com/milesial/Pytorch-UNet)  

VIS data. Using the generated annotations, we then train the mask generation procedure and synthesize a training dataset of 5000 pairs of VIS and NIR images with corresponding reference segmentation masks (and 500 for validation). Finally, we train a DeepLab-V3 [3] and U-Net [32] segmentation model using the synthetic datasets. Public implementations are used to foster reproducibility. To test the performance of the trained models, we use the (frontal gaze) VIS images from SMD, MOBIUS and SBVPi. Thus, segmentation performance with VIS images is used as a proxy for the quality of the generated segmentation masks.

**State-of-the-art Comparison.** In Table 4, we report the results of the segmentation experiments in terms of the Intersection-Over-Union (IoU), F1 score and overall Pixel error following established methodology [33, 39] and compare the performance ensured by the data generated by our BiOcularGAN to that produced by the unimodal DatasetGAN procedure from [44]. Here, the DatasetGAN approach is learned from the unimodal StyleGAN2 model trained on VIS images, and with 8 manually annotated images.

Interestingly, the segmentation models trained with the artificial dataset generated by the proposed BiOcularGAN framework clearly outperform the models trained with DatasetGAN on all three test datasets and across all three performance measures. This suggests that the joint bimodal supervision used to train the DB-StyleGAN model helps to better capture the semantic information of the images in the model layers and consequently leads to higher quality training data. This observation is further supported by the sample results in Figure 9, where we again see better segmentation performance following the use of the BiOcularGAN framework for data generation. Here, the examples were produced with U-Net and the BiOcularGAN and DatasetGAN frameworks trained using the PolyU data.

**Fine-grained Segmentation.** Because only a few manual annotations are needed to produce large amounts of training data for learning segmentation models, we manually annotate 2 images with a 10-class markup as shown on the left side of Figure 10. We then train a segmentation model (i.e., U-Net) with the dataset generated by BiOcularGAN using this fine-grained markup. The right part of Figure 10 shows some qualitative segmentation results generated with images from the SMD, MOBIUS and SBVPi datasets. Note that despite the fact the BiOcularGAN framework relied only on the DB-StyleGAN2-P model (that generates ocular images of mostly Asian subjects) and was learned with only 2 manually annotated images, the trained segmentation model still perform reasonably well on images from all three test datasets.

![Training data](https://github.com/MOBIUS) [Testing data](https://github.com/SBVPI)  

**Real-world Time Requirements.** The training of the data annotation procedure takes around 13 minutes on PolyU and 11 minutes on CrossEyed using 8 annotated images per dataset with our hardware setup. At run-time, a single segmentation mask is produced in 77.8 ms on average for an 256 × 256 image produced by DB-StyleGAN2.

5. Conclusion

In this paper, we presented BiOcularGAN, a framework for generating synthetic datasets of ocular images with corresponding ground truth segmentation masks. At the heart of the framework is a novel generative model, i.e., the dual-branch StyleGAN2 (DB-StyleGAN2), capable of generating photorealistic aligned bimodal (VIS and NIR) ocular images. Using the proposed BiOcularGAN framework, we showed that it is possible to generate large and representative synthetic datasets that can be used to train competition segmentation models that generalize well across a diverse set of ocular images. As part of our future work, we plan to further explore the DB-StyleGAN2 models for cross-modal recognition tasks and investigate image editing possibilities within the DB-StyleGAN2 latent space.
References

[1] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. Visualizing and understanding generative adversarial networks. In International Conference on Learning Representations (ICLR), pages 1–4, 2019.

[2] F. Bourtou, N. Damer, K. Raja, R. Ramachandhara, F. Kirchbuchner, and A. Kuijper. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. Image and Vision Computing, 104:104007, 2020.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In International Conference on Learning Representations (ICLR), pages 1–35, 2018.

[4] M. Buhler, S. Park, S. De Mello, X. Zhang, and O. Hilliges. Content-consistent generation of realistic eyes with style. In IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pages 1–5, 2019.

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Re-thinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[6] A. Das. Towards multi-modal sclera and iris biometric recognition with adaptive liveness detection. PhD thesis, School of Information and Communication Technology, Griffith University, 2017.

[7] I. Durugkar, I. Gemp, and S. Mahadevan. Generative multi-adversarial networks. In International Conference on Learning Representations (ICLR), pages 1–14, 2017.

[8] D. Galeev, K. Sofiiuk, D. Rukhovich, M. Romanov, O. Buchner, and A. Kuijper. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. In International Conference on Learning Representations (ICLR), pages 1–18, 2021.

[9] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. OpenEDS: Open eye dataset. arXiv preprint arXiv:1903.03702, 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS), pages 2672–2680, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[12] C. Jasserand. Massive facial databases and the GDPR: The new data protection rules applicable to research. In Data Protection and Privacy: The Internet of Bodies, pages 169–188. Bloomsbury Publishing, 2018.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In International Conference on Learning Representations (ICLR), pages 1–26, 2018.

[14] T. Karras, M. Aittala, J. Hellsten, S. Laine, and T. Aila. Training generative adversarial networks with limited data. In Advances in Neural Information Processing Systems (NeurIPS), pages 12104–12114, 2020.

[15] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. In Advances in Neural Information Processing Systems (NeurIPS), pages 852–863, 2021.

[16] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4401–4410, 2019.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8110–8119, 2020.

[18] H. Kaur and R. Manduchi. EyeGAN: Gaze-preserving, mask-mediated eye image synthesis. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 310–319, 2020.

[19] H. Kaur and R. Manduchi. Subject guided eye image synthesis with application to gaze redirection. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 11–20, 2021.

[20] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), pages 1–5, 2015.

[21] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore. Synthetic iris presentation attack using iDCGAN. In IEEE International Joint Conference on Biometrics (IJCB), pages 674–680, 2017.

[22] G. Kwon and J. C. Ye. Diagonal attention and style-based GAN for content-style disentanglement in image generation and translation. In IEEE/CVF International Conference on Computer Vision (ICCV), pages 13980–13989, 2021.

[23] K. Lee, H. Kim, and C. Suh. Simulated/unsupervised learning with adaptive data generation and bidirectional mappings. In International Conference on Learning Representations (ICLR), pages 1–15, 2018.

[24] D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8300–8311, 2021.

[25] B. Meden, P. Rot, P. Terhorst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Struc. Privacy-enhancing face biometrics: A comprehensive survey. IEEE Transactions on Information Forensics and Security (TIFS), 16:4147–4183, 2021.

[26] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for GANs do actually converge? In International Conference on Machine Learning (ICML), pages 3481–3490, 2018.

[27] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In International Conference on Learning Representations (ICLR), pages 1–26, 2018.

[28] P. R. Nalla and A. Kumar. Toward more accurate iris recognition using cross-spectral matching. IEEE Transactions on Image Processing (TIP), 26(1):208–221, 2016.

[29] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan. Iris recognition with off-the-shelf CNN features: A deep learning perspective. IEEE Access, 6:18848–18855, 2017.

[30] D. Pakhomov, S. Hira, N. Wagle, K. E. Green, and N. Navab. Segmentation in style: Unsupervised semantic image segmentation with application to gaze redirection. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 4401–4410, 2019.

[31] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2337–2346, 2019.
[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[33] P. Rot, Ž. Emeršič, V. Štruc, and P. Peer. Deep multi-class eye segmentation for ocular biometrics. In *IEEE International Work Conference on Bioinspired Intelligence (IWobi)*, pages 1–8, 2018.

[34] P. Rot, M. Vitek, K. Grm, Žiga Emeršič, P. Peer, and V. Štruc. Deep sclera segmentation and recognition. In *Handbook of Vascular Biometrics*, pages 395–432. Springer, 2020.

[35] A. Sequeira, L. Chen, P. Wild, J. Ferryman, F. Alonso-Fernandez, K. B. Raja, R. Raghavendra, C. Busch, and J. Bigun. Cross-eyed-cross-spectral iris/periocular recognition database and competition. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2016.

[36] A. F. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K. B. Raja, R. Raghavendra, C. Busch, T. de Freitas Pereira, et al. Cross-eyed 2017: Cross-spectral iris/periocular recognition competition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 725–732, 2017.

[37] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2107–2116, 2017.

[38] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008.

[39] M. Vitek, A. Das, Y. Pourcenoux, A. Missler, C. Paumier, S. Das, I. D. Ghosh, D. R. Lucio, L. A. Z. Jr., D. Menotti, F. Boutros, N. Damer, J. H. Grebe, A. Kuijper, J. Hu, Y. He, C. Wang, H. Liu, Y. Wang, Z. Sun, D. Osorio-Roig, C. Rathgeb, C. Busch, J. Tapia, A. Valenzuela, G. Žampoukis, L. Tsotchatzidis, I. Pratikakis, S. Nathan, R. Suganya, V. Mehta, A. Dhall, K. Raja, G. Gupta, J. N. Khiar, M. Akbari-Shahperi, F. Jaryani, M. Asgari-Chenaghlou, R. Vyas, S. Dakshit, S. Dakshit, P. Peer, U. Pal, and V. Štruc. SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment. In *International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.

[40] M. Vitek, P. Rot, V. Štruc, and P. Peer. A comprehensive investigation into sclera biometrics: A novel dataset and performance study. *Neural Computing & Applications*, 32(24):17941–17955, 2020.

[41] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3756–3764, 2015.

[42] L. A. Zanlorensi, R. Laroca, E. Luz, A. S. Britto, L. S. Oliveira, and D. Menotti. Ocular recognition databases and competitions: A survey. *Artificial Intelligence Review*, 55(1):129–180, 2022.

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[44] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler. DatasetGAN: Efficient labeled data factory with minimal human effort. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10145–10155, 2021.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.