Landmark-aware Self-supervised Eye Semantic Segmentation

Xin Cai[1,2], Jiabei Zeng[1], Shiguang Shan[1,2,3]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100090, China
[2]University of Chinese Academy of Sciences, Beijing, 100090, China
[3]Peng Cheng Laboratory, Shenzhen, 518055, China

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667031

***Abstract*— Learning an accurate and robust eye semantic segmentation model generally requires enormous training data with delicate segmentation annotations. However, labeling the data is time-consuming and manpower-consuming. To address this issue, we propose to segment the eyes using unlabelled eye images and a weak empirical prior on the eye shape. To make the segmentation interpretable, we leverage the prior knowledge of eye shape by converting the self-supervised learned landmarks of each eye component to the segmentation maps. Specifically, we design a symmetrical auto-encoder architecture to learn disentangled representations of eye appearance and eye shape in a self-supervised manner. The eye shape is represented as the landmarks on the eyes. The proposed method encodes the eye images into the eye shapes and appearance features and then it reconstructs the image according to the eye shape and the appearance feature of another image. Since the landmarks of the training images are unknown, we require the generated landmarks' pictorial representations to have the same distribution as a known prior by minimizing an adversarial loss. Experiments on TEyeD and UnitySeg datasets demonstrate that the proposed self-supervised method is comparable with supervised ones. When the labeled data is insufficient, the proposed self-supervised method provides a better pre-trained model than other initialization methods.**

I. INTRODUCTION

Understanding human eyes plays an important role in medical application, human-computer interaction, virtual reality, biometric security, and other areas. Explicitly parsing eye images into different eye components implies analyzing the semantic constituents (e.g., pupil, iris and sclera) of human eyes, and is useful for a variety of tasks, including gaze tracking, iris recognition, pupil diameter estimation, etc. All these applications require the eye parsing/segmentation methods to be robust to the various poses, illuminations, and other environments.

Efforts have been made in developing eye segmentation methods during the last decades. Some early works propose to segment sclera [7] or iris [1] using image processing methods, including edge detection [30] and ellipse fitting [8]. With the handcrafted features, they usually determine the sclera or iris regions by adjusting some thresholds according to the distribution of the data. These approaches highly depend on stable environments and are likely to fail in new distributed data. Recently, ascribe to the development of deep learning, appearance-based methods based on Convolution Neural Network (CNN) have gained popularity and achieved the state-of-the-art accuracy on eye semantic segmentation. However, a well-trained segmentation network, e.g., SegNet [3] and UNet [27], requires diverse training images with high-quality annotations about the eyes' regions. Labeling the segmentation needs the annotators to draw a fine edge of the target region. It takes minutes to complete the labeling of one image. Therefore, some works are proposed to segment the images in an unsupervised manner, e.g., IIC [18], PiCIE [6]. Although without annotations, most unsupervised segmentation methods cannot determine whether some small regions or super-pixels should be merged or not. It is also cumbersome to determine what the region stands for.

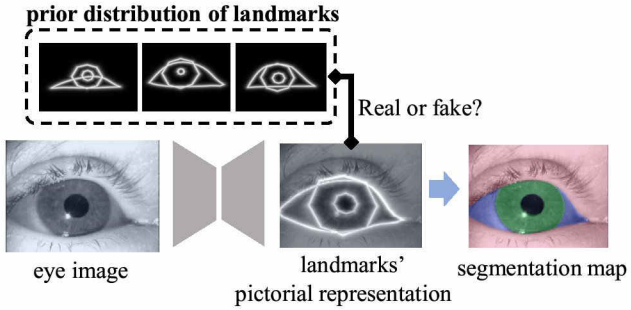

Fig. 1. Main idea of the proposed method (LS^{2}E-Seg). The input eye image is translated into its landmarks' pictorial representation. Then we convert the predicted landmarks to the segmentation map. Since the landmark annotations of the input images are unknown, we force the generated landmarks in-distinguishable among the real ones.

To this end, we propose a **L**andmark-aware **S**elf-**S**upervised **E**ye **Seg**mentation (LS^{2}E-Seg) method to segment pupil, iris, and sclera regions from eye images leveraging unlabelled images of eyes. To make the segmentation interpretable, we first learn the eyes landmarks and then convert the landmarks of each eye component to the segmentation maps. To learn the landmarks, inspired by the recent self-supervised landmarks learning framework [17], we train the landmark detector with unlabelled images and a set of landmarks' pictorial representation, which are not the labels to the training images but serve as a prior distribution of eye landmarks. Fig. 1 demonstrates the main idea of LS^{2}E-Seg. As shown in Fig. 1, given an eye image, an image-to-image translation network [15] is used to generate its landmarks' pictorial representation. To guarantee that the generated pictorial representation stands for the landmarks of eyes' components, we use a discriminator to judge if the

This work is partially supported by National Key R&D Program of China (No. 2017YFA0700800) and National Natural Science Foundation of China (No. 62176248, 61976203).

generated one is real or fake. Then, we convert the predicted landmarks of pupil, iris, and sclera to a segmentation map using ellipse fitting. Therefore, we obtain the segmentation of each eye component without annotations. Our contributions are summarised as follows:

1. We propose a self-supervised method to segment the pupil, iris, and sclera of the eye images according to the self-supervised detected landmarks. To the best of our knowledge, it is the first work to do eye semantic segmentation in a self-supervised manner.

2. The proposed method learns the landmarks by disentangling the shape (landmarks) and appearance of the eyes in an image reconstructing task. The eye segmentation is converted from the detected landmarks and thus it is interpretable.

3. Experiments on two datasets show the effectiveness of the proposed self-supervised method, which achieves comparable results with other supervised methods.

II. RELATED WORK

A. Eye semantic segmentation

Since the eye semantic segmentation is considered as a sub-problem of an image segmentation task, general image segmentation methods were applied to segment eyes. Lian *et al.* [22] proposed to use the attention mechanism on U-Net [27] to guide the model to learn discriminative features for iris segmentation. Naqvi *et al.* [24] presented ScleraNet, a residual encoder-decoder network based on SegNet[3]. To improve muti-class segmentation for eyes, Perry and Fernandez [26] proposed to leveraged dilated and asymmetric convolution, meanwhile Kansal *et al.* [19] chose to utilize squeeze-and excitation [14] block.

Besides applying the general segmentation method, some works utilized the unique characteristics of eye images. Kim *et al.* [20] proposed to add a heuristic filter after segmentation network because the sclera covered the iris, and iris wrapped pupil. Fuhl *et al.* [11] proposed a combined convolutional neural network architecture for eyelid landmark, pupil ellipse regression together with pupil area and eyelid area segmentation. Kothari *et al.* [21] proposed EllSeg framework for simultaneous segmentation and ellipse parameter prediction for both iris and pupil regions.

The deep learning based works rely on large, curated training datasets of eye images with well-annotated labels and have difficulty with generalizing unconstrained environments. Learning with limited or no external supervision for eye semantic segmentation is still a challenge.

B. Unsupervised segmentation methods

Unsupervised or self-supervised techniques have been explored recently to conduct image semantic segmentation without external supervision. A few works consider unsupervised semantic segmentation as a problem of clustering pixel-level features. Both Ji *et al.* [18] and Ouali *et al.* [25] leverage an end-to-end approach maximizing the discrete mutual information between augmented image pairs to learn a pixel-level clustering function and then obtain the probabilities of pixels over classes. PiCIE [6] conduct pixel-level feature clustering using invariance to photometric transformations and equivariance to geometric transformations. However, these methods can neither leverage the prior information of eye shape nor segment eye images for specific interpretable parts (pupil, iris and sclera) we want.

C. Unsupervised Keypoint Detection

There have been a few attempts in the literature to tackle keypoint detection under the unsupervised setting. Thewlis *et al.* [29] propose to learn sparse viewpoint invariant landmarks using the equivalence constraint and develop the method to a dense situation [28]. Zhang *et al.* [32] use an auto-encoder paradigm to learn explicit structural representations as landmarks. Jakab *et al.* [16] develop the auto-encoder formulation by using conditional image generation and a bottleneck to limit the geometric information flow. Based on [16], Jakab *et al.* [17], the most related work of ours, make use of an interpretable keypoint prior to learn 'semantically meaningful' keypoint directly. Inspired by [17], we extended the idea to the self-supervised eye semantic segmentation by combining the self-supervised keypoint detector with segmentation fitting.

III. METHODS

We aim to learn a function that maps an eye image to its semantic segmentation map without annotations. However, most general unsupervised segmentation methods can not produce interpretable segmentation maps directly. To conduct the interpretable self-supervised eye semantic segmentation, we first detect the landmarks of the iris, pupil and sclera and then using the landmarks to induce the segmentation. In the training procedure, we train a self-supervised landmark detector using a symmetric self-supervised learning framework with pairs of unlabeled eye images and a set of prior keypoints' pictorial representation. In the inference procedure, given an image, we first predict the landmarks of eyes and then convert the landmarks of each component (iris, pupil and sclera) into its corresponding segmentation map. Below, we introduce the two procedures in details.

A. Training the self-supervised landmark detector

We learn interpretable keypoints of eyes to identify the contours of iris, pupil and sclera. To avoid using the keypoint labels of images, we use a symmetrical auto-encoder architecture to learn disentangled representations between the appearance and shape of the eyes. The shape is depicted as the landmarks or keypoints of the eye.

Figure 2 (left) illustrates the self-supervised training framework. As can be seen, it takes two different images (I_1, I_2) of the same eye from a video clip as the inputs. Each of the image is fed into the appearance encoder E_a and the keypoint encoder E_k, respectively. E_k outputs the keypoint pictorial representation $\mathbf{K}$. $\mathbf{K}$ contains the information of the eye shape (keypoint) and is spatially aligned with I. To make $\mathbf{K}$ represent the shape (keypoint) information like $\mathbf{K}_{real}$ which are sampled from keypoint prior, we require the discriminator D to judge if the generated $\mathbf{K}$ is real or fake. The keypoint bottleneck compresses $\mathbf{K}$ into the coordinates of landmarks and reconstructs a purified keypoint pictorial

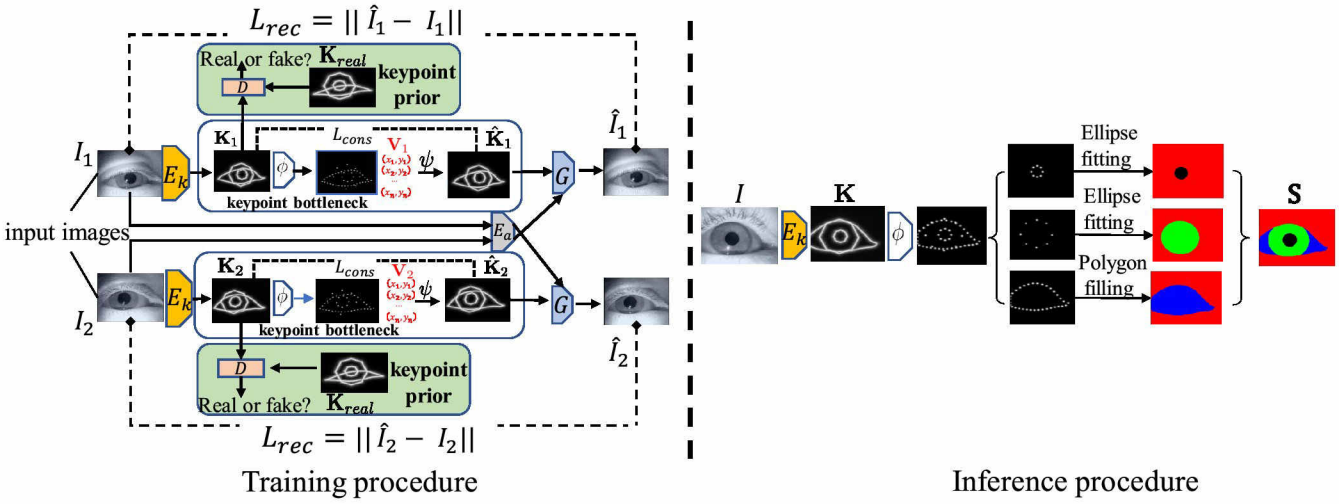

Fig. 2. Framework of the proposed symmetric landmark-aware self-supervised eye semantic segmentation method. In training procedure, the unlabelled input images (I_1, I_2) of the same eye are encoded by E_k and E_a to get the landmark and appearance features, respectively. E_k outputs the keypoint pictorial representation $\mathbf{K}$. The keypoint bottleneck compresses $\mathbf{K}$ into the coordinates of landmarks and reconstructs a purified keypoint pictorial representation $\hat{K}$. The inputs images are reconstructed by a generator G according to their own purified keypoint pictorial representation and the swapped appearance feature. To make $\mathbf{K}$ represent the shape (keypoint) information like $\mathbf{K}_{real}$, we require the discriminator D to judge if the generated $\mathbf{K}$ is real or fake. In inference procedure, the eye image I is translated to $\mathbf{K}$ by E_k. $\mathbf{K}$ is regressed to the points coordinates $\mathbf{V}$ by ϕ. Keypoints $\mathbf{V}$ is used for fitting pupil, iris and sclera to obtain the final eye semantic segmentation map $\mathbf{S}$.

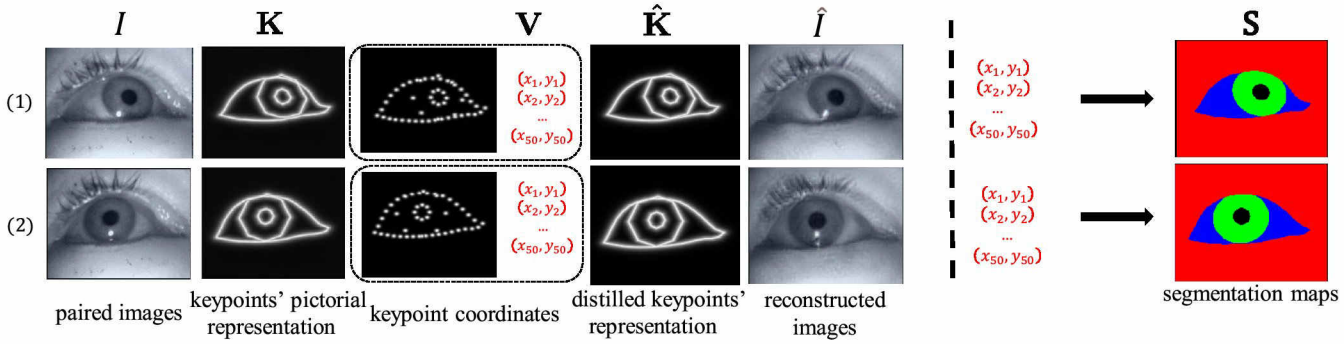

Fig. 3. Examples of the inputs or generated images during the training (left) and the converted eye semantic segmentation map in inference (right).

representation $\hat{\mathbf{K}}$. E_a outputs the appearance feature. According to their own purified keypoint pictorial representation and the swapped appearance feature, the inputs images are reconstructed by a generator G. This is because I_1 and I_2 are from the same eye under a similar environment, and should have the same appearance feature. If we swap their appearance feature, we can still reconstruct I_1 or I_2 .

Below, we present details of the components keypoint pictorial representation, keypoint representation prior, keypoint representation bottleneck, and the objective function for learning.

1) ***Keypoint Pictorial Representation****:* To learn the spatial structure of the eyes, we design the keypoint encoder E_k as an image translation network. E_k translates an eye image to a pictorial representation of the keypoints, which is an image spatially aligned to the input eye image. The generated pictorial representation represent the information of keypoints. It is composed of the edges that connect two keypoints and looks like an image of the sketched structure of an eye. In Fig. 2, $\mathbf{K}_1$ and $\mathbf{K}_2$ are two examples of the keypoint pictorial representations.

Mathematically speaking, we denote the keypoint pictorial representation as $\mathbf{K} \in \mathbb{R}^{W \times H}$, where W and H is the width and the height of the input image. The keypoint pictorial representation $\mathbf{K}$ could be generated according to the 2D keypoint coordinates $\mathbf{V} = [\mathbf{v}_1; \mathbf{v}_2; \cdots, \mathbf{v}_n]$, where n is the number of keypoints and $\mathbf{v}_i = (x_i, y_i) \in \mathbb{N}^2$ is the coordinate of the i-th keypoint. We denote the grid of pixel coordinates as $\Omega = \{1, 2, ..., H\} \times \{1, 2, ..., W\}$. The value at each position $\mathbf{u} \in \Omega$ of the keypoint pictorial representation $\mathbf{K}$ is computed as

$$k_{\mathbf{u}} = \exp(-\gamma \min_{\substack{(\mathbf{v}_i, \mathbf{v}_j) \in \mathcal{E}, \\ r \in [0,1]}} ||\mathbf{u} - r\mathbf{v}_i - (1-r)\mathbf{v}_j||_2), \quad (1)$$

where $k_{\mathbf{u}}$ is the element of $\mathbf{K}$ at position $\mathbf{u}$. $\mathcal{E}$ is the set of the pre-defined edges with the start-end keypoint pairs $(\mathbf{v}_i, \mathbf{v}_j)$ and γ is 0.2 in our experiments. When the pixel $\mathbf{u}$ is on the edge that connects $\mathbf{v}_i$ and $\mathbf{v}_j$, $\mathbf{u}$ can be represented as a linear combination $\mathbf{v}_i$ and $\mathbf{v}_j$. It means that there exist an $r \in [0, 1]$ such that $\mathbf{u} = r\mathbf{v}_i + (1-r)\mathbf{v}_j$, and $||\mathbf{u} - r\mathbf{v}_i - (1-r)\mathbf{v}_j||^2 = 0$. Then, $k_{\mathbf{u}} = 1$ and it looks bright in the keypoint pictorial representation. When the pixel $\mathbf{u}$ is far away from all the edges in $\mathcal{E}$, $k_{\mathbf{u}}$ is around 0. It looks dark in the pictorial representation. Therefore, $\mathbf{K}$ is visualized as smooth lines

with the ends as the keypoint pairs.

It is noted that we do not directly use the 2D keypoint coordinates $\mathbf{V}$ as the output of E_k to represent the keypoint information, because it is easier to train an image-to-image translator than to train an image-to-vector regressor without supervision. Using an image-to-image translator, we can take advantage of the inductive bias in CNN that assumes a certain type of spatial structure present in the input image [23].

*2) **Keypoint Representation Prior**:* Since we do not have the ground truth of the output of E_k, we leverage a keypoint representation prior which is a set of landmark pictorial representations computed from the landmarks of real eye images **other than** the training images. The keypoint representation prior is used to make the output of keypoint shape encoder E_k looks like the expected keypoint pictorial representation rather than other forms. We define a discriminator D to distinguish whether the keypoint pictorial representation $\mathbf{K}$ generated by E_k are from the keypoint representation prior to distribution $p(\mathbf{K})$ in order that the distribution of E_k's output can be closed to $p(\mathbf{K})$.

Mathematically speaking, let us assume the keypoint representation prior as $\mathcal{K}_{real} = \{\mathbf{K}_i\}_{i=1}^{N}$, where N is the size of the prior set. $\mathcal{D}_{tr} = \{(I_{i1}, I_{i2})\}_{i=1}^{M}$ denotes the training set of unlabelled image pairs, where M is the number of pairs. Then, similar to WGAN [2], we could optimize the parameters of the keypoint encoder E_k and discriminator D by minimize the maximum of an adversarial loss

$$\min_{E_k}\max_{D}\mathcal{L}_{adv}(D,E_k) = \min_{E_k}\max_{D}\frac{1}{N}\sum_{\mathbf{K}_i\in\mathcal{K}_{real}} D(\mathbf{K}_i) - \frac{1}{2M}\sum_{(I_{i1},I_{i2})\in\mathcal{D}_{tr}}\left(D(E_k(I_{i1})) + D(E_k(I_{i2}))\right), \quad (2)$$

where $E_k(I_{i1})$ and $E_k(I_{i2})$ are the generated keypoint pictorial representation of the input image I_{i1} and I_{i2}, respectively. $D(\mathbf{K}_i)$ denotes the output of the discriminator given keypoint pictorial representation $\mathbf{K}_i$.

*3) **Keypoint Representation Bottleneck**:* To purify the output of E_k as the information of landmarks, we adopt a keypoint representation bottle to prevent the keypoint pictorial representation containing appearance information. As can be seen in Fig. 2, in the keypoint bottleneck, we compress the keypoint pictorial representation $\mathbf{K}_1$ to the coordinates of the keypoints $\mathbf{V}_1$ through a mapping function ϕ. Then, we reconstruct a purified keypoint pictorial representation $\hat{\mathbf{K}}_1$ according to $\mathbf{V}_1$. For notation conveniency, we denote the operation of reconstruction as $\psi(\mathbf{V}_1) = \hat{\mathbf{K}}_1$, where the value of $\hat{\mathbf{K}}_1$ at the position $\mathbf{u}$ is computed using Eq. (1). Similarly, we obtain the purified keypoint pictorial representation $\hat{\mathbf{K}}_2$ of the other image $\mathbf{I}_2$.

We implement ϕ as a neural network regressor and pretrain its parameters using training pairs $\mathcal{V} = \{\mathbf{V}_{real}^i\}_{i=1}^{N}$, where $\mathbf{V}_{real}$ is the real eye keypoints. N is the size of the training set. We train the regressor network ϕ by

$$\min_{\phi}\mathcal{L}_{reg}(\mathcal{V}) = \min_{\phi}\sum_{i=1}^{N}||\mathbf{V}_{real}^i - \phi(\psi(\mathbf{V}_{real}^i))||_2, \quad (3)$$

where we compute the keypoint pictorial representation $\psi(\mathbf{V}_{real}^i)$ from $\mathbf{V}_{real}^i$ and re-generate the landmarks by $\phi(\psi(\mathbf{V}_{real}^i))$. ϕ is optimized by minimizing the discrepancy between the original landmarks and generated ones.

*4) **Objective Function for Learning**:* To train the self-supervised landmarks detector, we first pre-train the regressor ϕ by (3), then we alternatively update (1) the landmarks encoder E_k, appearance encoder E_a, and the generator G, (2) the discriminator D, (3) the regressor ϕ.

The landmarks encoder E_k, appearance encoder E_a, and the generator G are updated by the objective function using the unlabelled training image pairs and the keypoint prior:

$$\min_{E_k,E_a,G}\mathcal{L}_{rec} + \mathcal{L}_{cons} + \mathcal{L}_{adv}, \quad (4)$$

where $\mathcal{L}_{rec}$ is the image reconstruct loss. Given $\mathcal{D}_{tr} = \{(I_{i1}, I_{i2})\}_{i=1}^{M}$ as the training set of unlabelled image pairs, where M is the number of pairs, the reconstruct loss is formulated as

$$\mathcal{L}_{rec} = \frac{1}{2M}\sum_{i=1}^{M}||f(\hat{I}_{i1}) - f(I_{i1})||_2 + ||f(\hat{I}_{i2}) - f(I_{i2})||_2 + \frac{1}{2M}\sum_{i=1}^{M}||\hat{I}_{i1} - I_{i1}||_1 + ||\hat{I}_{i2} - I_{i2}||_1 \quad (5)$$

where $\hat{I}_{i1} = G(\psi(\phi(E_k(I_{i1}))), E_a(I_{i2}))$ is the reconstructed image of I_{i1} by generator D according to the purified landmark representation of I_{i1} and the appearance feature of I_{i2}. Similarly, the reconstructed image $\hat{I}_{i2} = G(\psi(\phi(E_k(I_{i2}))), E_a(I_{i1}))$. f is a pretrained VGG feature extractor network. The first summation item force reconstructed image features rather than pixels similar to original images, which make learning fast and robust. The second summation item keeps reconstructed images and original images spatially aligned, which is important to keypoint pictorial representation learning.

$\mathcal{L}_{cons}$ is the consistent loss that makes the output keypoint pictorial representation of E_k as purified as possible. It is formulated as

$$\mathcal{L}_{cons} = \frac{1}{2M}\sum_{i=1}^{M}\sum_{j=1}^{2}||E_k(I_{ij}) - \psi(\phi(E_k(I_{ij})))||_1. \quad (6)$$

$\mathcal{L}_{adv}$ is the adversarial loss described in Eq. (2) with fixed discriminated D.

The discriminator D is updated by maximizing the adversarial loss $\mathcal{L}_{adv}$ defined in Eq. (2) with fixed landmark encoder E_k by $\max_D \mathcal{L}_{adv}(D, E_k)$.

The regressor ϕ is updated by the objective function $\min_\phi \mathcal{L}_{cons} + \mathcal{L}_{reg}$, where $\mathcal{L}_{cons}$ is defined in Eq. (6) and is computed with the unlabelled training images and fixed landmarks encoder E_k. $\mathcal{L}_{reg}$ is defined in Eq. (3) and is computed with the prior real landmarks.

B. Segmentation with map fitting

In the inference procedure, we use the obtained interpretable keypoints $\mathbf{V}$ to create the semantic segmentation

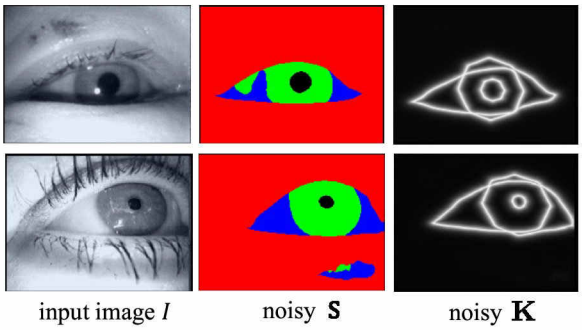

Fig. 4. Illustration of noisy S and noisy $\mathbf{K}$. E_k outputs the noisy S when we replace the keypoint representation prior to semantic segmentation map prior.

maps, which is illustrated in Fig.2 (right). We fit the iris landmarks and the pupil landmarks into ellipse by ellipse fitting method and get the ellipse parameters. According to the parameters, we draw the iris and pupil masks. We connect eyelid landmarks point-by-point to get the contour of sclera and draw the mask. Then, we integrate the masks of pupil, iris and sclera into the semantic segmentation maps $\mathbf{S}$.

Ellipse fitting finds the ellipse parameters $\mathbf{A}$ given the coordinates of a series of points $P = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$ on the ellipse edge. We denote an ellipse by a second-order polynomial:

$$F(\mathbf{p}; \mathbf{A}) = \mathbf{A} \cdot d_{\mathbf{p}} = ax^2 + bxy + cy^2 + dx + ey + f = 0 \quad (7)$$

where $\mathbf{A} = [a, b, c, d, e, f]^T$, $\mathbf{p} = [x, y]^T$ and $d_{\mathbf{p}} = [x^2, xy, y^2, x, y, 1]^T$. We minimize the sum of squared algebraic distances on P with regard to ellipse parameters $\mathbf{A}$:

$$\min_{\mathbf{A}} \sum_{i=1}^{N} F(\mathbf{p}_i; \mathbf{A})^2. \quad (8)$$

To avoid the trivial solution $\mathbf{A} = \mathbf{0}_6$, we add a quadratic constraint[4] on $\mathbf{A}$ as $\mathbf{A}^T\mathbf{C}\mathbf{A} = 1$, where $\mathbf{C}$ is a 6×6 constraint matrix. The optimization problem (8) are formulated as a generalized eigenvalue system and has a solution as

$$\mathbf{D}^T\mathbf{D}\mathbf{A} = \lambda\mathbf{C}\mathbf{A} \quad (9)$$

where $\mathbf{D} = [d_{\mathbf{p}_1}, d_{\mathbf{p}_2}, ..., d_{\mathbf{p}_n}]^T$. Here we use the constraint $4ac - b^2 = 1$ on $\mathbf{A}$ like [9]. Fitzgibbon *et al.* [9] proved that the system (9) produces exactly one positive eigenvalue λ^* which corresponds to an ellipse and we take its corresponding eigenvector $\mathbf{A}^*$ as our solution.

C. Discussion of the landmarks and segmentation

Instead of translating the eye images into the landmarks and then converting the landmarks into segmentation in our method, an alternative way is to translate the eye image I into the segmentation $\mathbf{S}$ directly in the self-supervised learning framework. The change degrades the semantic segmentation performance of our method, which will be shown in experimental results in IV-F. Below, we discuss the reasons and the advantages of our method.

First, we explicitly use the knowledge that pupils and iris are ellipse shapes by segmentation fitting. However, if we directly translate the eye image I to eye semantic segmentation map $\mathbf{S}$, the model can only learn this knowledge from the real segmentation representation prior implicitly.

Second, under the unsupervised manner, it introduces more noises by directly learning the semantic segmentation map $\mathbf{S}$ than learning the landmarks. Fig 4 shows some examples of the noisy $\mathbf{K}$ and the noisy $\mathbf{S}$ which are produced by E_k when using eye semantic segmentation map $\mathbf{S}$ and keypoint pictorial representation $\mathbf{K}$ as the representation prior form respectively.

Last, it's difficult to use a differentiable function $\psi(\mathbf{V}) = \mathbf{S}$ mapping keypoints to the segmentation maps like Eq. (1). Although we can use a neural network to simulate this, the introduction of a new neural network makes the back-propagation in training progress more difficult.

IV. Experiments

We compared our method with others under both unsupervised and supervised protocols on both real (TEyeD [10]) and synthetic (UnitySeg) eye datasets. We also conducted ablation experiments to validate the effectiveness of the key components of our method and illustrate the effectiveness of feature disentanglements.

A. Experimental settings

1) Dataset: **TEyeD** is the world's largest unified public data set of eye images collected by head-mounted devices. It contains more than 20 million carefully annotated images with 2D&3D landmarks, semantic segmentation, 3D eyeball annotation and the gaze vector and eye movement types. We randomly selected 126 videos containing 1.2M frames as training and validation set and 32 videos containing 500k frames as the test set. In our method, we did not use any of the labels of the training data but randomly selected another 500k ground truth landmarks as the landmarks' prior beyond the selected training and test set.

UnitySeg is a synthetic eye image dataset created by us using UnityEyes [31], a tool to generate labelled synthetic eye images. UnitySeg contains 200k images of 100 different subjects. In our method, the landmarks' prior was selected as the landmarks of 20 random subjects. The training set was selected as the unlabelled eyes images of another 60 random subjects. The test set was selected as another 20 subjects.

Only eyelid margin and iris landmark are accessible from UnityEyes, we created the ground truth of segmentation by using fit methods mentioned in Sec III-B according to the provided landmarks. In this dataset, only the iris and sclera segmentations were investigated.

2) Implementation Details: In our experiments, the detailed structures of the keypoint encoder E_k, appearance encoder E_a. the decoder G, the discriminator D and the regressor ϕ are presented in the supplemental materials. It is noted that other suitable networks can be substituted for the network we use. The proposed method was implemented using the deep learning toolbox PyTorch. The models are trained by optimizing the objective using RMSprop with a learning rate of $1 \cdot 10^{-4}$. The batch size is 32 and the values

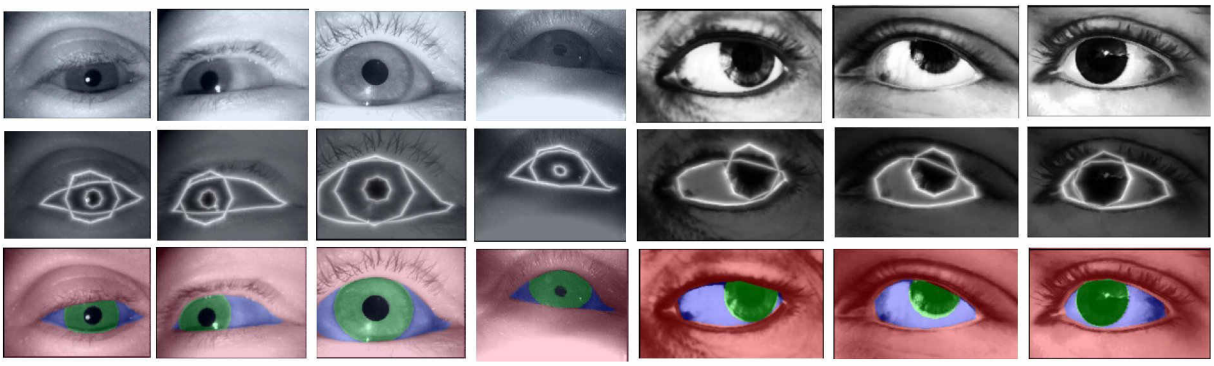

Fig. 5. **Visualized examples**. Keypoint pictorial representation and semantic segmentation results on the TeyeD (left four cols) and UnitySeg (right three cols).These results are produced by our method directly without any additional labelled data. This figure is best viewed in color.

TABLE I

QUANTITIES RESULTS OF SELF-SUPERVISED METHODS (TOP THREE ROWS) AND SUPERVISED METHODS (OTHER ROWS) ON TEYED AND UNITYSEG.

Dataset	TEyeD						UnitySeg				
Method	paras	mF1	mIoU	IoU(iris)	IoU(pupil)	IoU(sclera)	paras	mF1	mIoU	IoU(iris)	IoU(sclera)
self-supervised											
CycleGAN	3.125M	0.884	0.792	0.766	0.802	0.642	3.122M	0.905	0.827	0.835	0.703
LS^{2}E-Seg	3.125M	0.938	0.884	0.890	0.913	0.789	3.122M	0.939	0.886	0.887	0.797
LS^{2}E-Seg*	3.125M	**0.951**	**0.907**	**0.908**	**0.924**	**0.825**	3.122M	**0.949**	**0.897**	**0.903**	**0.809**
supervised											
RITnet	0.25M	0.934	0.877	0.882	0.891	0.782	——	——	——	——	——
E_k(Seg)	3.126M	0.955	0.916	0.919	0.933	0.832	3.122M	0.935	0.878	0.866	0.788
Resnet50	23.58M	0.960	0.923	0.922	0.925	0.859	23.57M	0.943	0.894	0.899	0.808
$\text{Unet}_{\text{small}}$	4.321M	0.964	0.932	0.933	0.947	0.863	4.32M	0.963	0.931	0.937	0.877
Unet	17.268M	0.967	0.937	0.938	0.943	0.878	17.266M	0.966	0.936	0.942	0.886
E_k(lmk)	3.125M	0.968	0.938	0.941	0.948	0.878	3.122M	0.964	0.932	0.937	0.878
LS^{2}E-Seg+E_k(lmk)	3.125M	0.971	0.944	0.944	0.953	0.889	3.122M	0.971	0.945	0.952	0.897
LS^{2}E-Seg*+E_k(lmk)	3.125M	**0.973**	**0.948**	**0.948**	**0.956**	**0.894**	3.122M	**0.972**	**0.947**	**0.953**	**0.901**

of $\{\lambda_{rec}, \lambda_{cons}, \lambda_{reg}\}$ are $\{1, 2, 0.5\}$ respectively. The λ_{adv} was initialized as 10 and was divided by a factor of 10 every 5000 iterations during the training. We resize the images in TEyeD to 192×144 and images in UnitySeg to 200 ×120 as input resolution.

3) Measurement: We evaluate the eye semantic segmentation performance of different methods and settings via IoU (Intersection-Over-Union) scores and F1 scores. The IoU of class $i \in$\{background, iris, pupil, sclera\} is define as:

$$\text{IoU}_\text{i} = \frac{|\text{P}_\text{i} \cap \text{G}_\text{i}|}{|\text{P}_\text{i} \cup \text{G}_\text{i}|} \tag{10}$$

where P_i, G_i are respectively the region of class i from the ground truth and predicted mask. We report the mean IoU and mean F1 scores of four classes (iris, pupil, sclera, background) and three single classes (iris, pupil, sclera) IoU evaluated on the TEyeD and UnitySeg (without pupil IoU) in our experiments.

B. Experimental results of eye segmentation

1) Under unsupervised protocol: In unsupervised protocol, we trained the proposed models with unlabelled data and get the predicted landmarks and segmentation maps directly. We compared with a CycleGAN [33] which is trained using unpaired training images and segmentation maps not in the training set. We did not compare with other unsupervised segmentation methods because it is difficult to ensure that the unsupervised segmented regions are the expected pupil, iris, and sclera. Table IV reports the mean IoUs and F1 on TEyeD and UnitySeg datasets using **CycleGAN**, **LS^{2}E-Seg** and **LS^{2}E-Seg***. LS^{2}E-Seg was trained with unlabelled training images while LS^{2}E-Seg* was trained with unlabelled training and test images.

It is shown that Our LS^{2}E-Seg and LS^{2}E-Seg* outperform the CycleGAN on segmentation metrics. In addition, training our model using test images in a self-training manner can improve the segmentation results on the test dataset.

We present the results of several supervised segmentation methods in the following Sec IV-B.2. The results show that our self-supervised methods LS^{2}E-Seg and LS^{2}E-Seg* are comparable to other supervised methods.

We also evaluated our method qualitatively. We adopted the LS^{2}E-Seg method on TEyeD and UnitySeg datasets and illustrated the learned landmarks and segmentations in Fig. 5. Our method can obtain accurate eye landmarks and eye segmentation maps in different conditions.

2) Under supervised protocol: To investigate whether the proposed LS^{2}E-Seg can further improve the supervised eye segmentation, we evaluated the methods under supervised protocol, where the annotations of landmarks or segmentations for the training images were used. We compared our self-supervised methods finetuned by training set with RITnet, Resnet50, Unet, $\text{Unet}_{\text{small}}$, E_k(Seg) and E_k(Lmk).

RITnet [5] is the champion method of the OpenEDS [12] 2019 eye semantic segmentation challenge. We reproduce their method using the open source code.

Resnet50 is a keypoint detector based on Resnet50 [13] trained by the MSE Loss. We convert the predicted keypoints into eye segmentation maps as the segmentation results.

Unet is widely used in the medical image segmentation area and has been proven effective sufficiently. We trained

the Unet model using the standard Cross Entropy loss.

$\text{Unet}_{\text{small}}$ is created by reducing the channels of Unet to a quarter of the original model, so that the number of $\text{Unet}_{\text{small}}$ model's parameters is closed to the encoder E_k.

E_k(Seg) is created by changing the last convolution layer of E_k from 1 channel to m channels for m part segmentation.We trained E_k(Seg) using cross entropy loss and take it as the supervised eye semantic segmentation baseline.

E_k(Lmk) is a supervised model to generate a keypoint pictorial representation for eye images. We used the ground truth of keypoint pictorial representation to train E_k by the Mean Square Loss and obtain E_k(Lmk). We obtain eye semantic segmentation maps by the method mentioned in Sec. III-B as the segmentation results.

We pre-trained our model LS^{2}E-Seg and LS^{2}E-Seg* without labels and finetuned the self-supervised models using the same way with E_k(Lmk). We named the two models as **LS^{2}E-Seg+E_k(lmk)** and **LS^{2}E-Seg*+E_k(lmk)** respectively. Table IV reports the mean IoUs and F1 on TEyeD and UnitySeg datasets using supervised methods above.

Our results show LS^{2}E-Seg-pretrained model provides a good initialization for eye segmentation. When we fine-tuned the pre-trained model LS^{2}E-Seg on labelled training set, we achieve higher mF1 and IoUs than E_k(Lmk) trained from scratch. In addition, the results of LS^{2}E-Seg* + E_k(Lmk) is the highest in both two datasets.

The results also show that E_k(Lmk) can achieve comparable or better results than other segmentation methods with fewer parameters. It indicates that converting the learned keypoint pictorial representation to the segmentation map is an effective method for eye semantic segmentation.

C. Discussion of the size of the prior set

In Sec IV-B, we use 500k ground truth landmarks in TEyeD and 20k landmarks in UnitySeg as the landmarks' prior set. To study the importance of the prior set size in the proposed method, we use the varying proportion of the original prior set to train LS^{2}E-Seg. Table II show that our method retains most of the performance when decreasing the size of the prior set.

TABLE II
VARYING PROPORTION OF THE ORIGINAL PRIOR SET.
THE RESULTS ARE REPORTED BY MIOU.

Data proportion	1%	5%	10%	25%	50%	100%
TEyeD	0.843	0.848	0.856	0.867	0.877	0.884
UnitySeg	0.852	0.857	0.865	0.870	0.881	0.886

D. Discussion of the proportion of labels

To further show self-supervised pre-trained models are better than random initialization for supervised training, we finetuned our self supervised models with different proportions of training data and compared the results with other methods trained with the **same** labelled data from scratch. We present the mIoU of different methods finetuned on different proportion of labels in Fig. 6. **LS^{2}E-Seg$^-$** in Fig. 6 stands for the LS^{2}E-Seg trained with 5% of our prior set. The results show that our LS^{2}E-Seg-pretrained models can achieve better

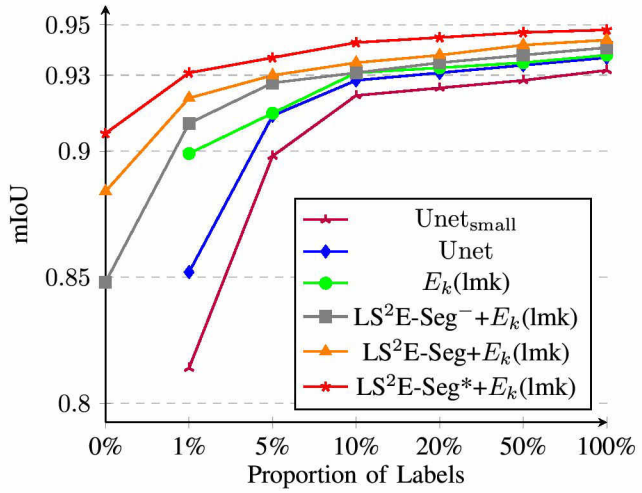

Fig. 6. Comparison of different methods finetuned on different proportion of labels, evaluated on TEyeD.

results with same labelled data or comparable results with less labelled data than supervised methods. For example, we use 5% data to finetune LS^{2}E-Seg$^-$ initialization (0.927) and obtain higher mIoU than supervised method E_k(Lmk) using 5% data (0.915). And we achieve the performance of E_k(Lmk) (0.938) trained with 100% data using only 10% data with LS^{2}E-Seg* (0.937).

E. Ablation Study

To investigate the effect of three key components in our method: keypoints' representation prior (P), keypoint representation bottleneck (B), a symmetric architecture (S), we ablate one of these components at once. We set $\lambda_{adv} = 0$ to remove keypoints' representation prior and remove keypoint representation bottleneck by dropping functions ϕ and ψ. In addition, we change the symmetric architecture by not reconstruct I_2 in (5).

Table III shows the effect of ablating one of these components on our experimental datasets. Our results show that the basic conditional auto-encoder method cannot finish the self-supervised eye segmentation task without the keypoint representation prior. In addition, the keypoint representation bottleneck can significantly import the segmentation performance (mIoU: $0.764 \rightarrow 0.884$ on TEyeD and $0.831 \rightarrow 0.886$ on UnitySeg). At last, the symmetric architecture is also useful for our segmentation model (mIoU: $0.878 \rightarrow 0.884$ on TEyeD and $0.882 \rightarrow 0.886$ on UnitySeg).

F. Comparison with direct segmentation map translation

We analyze the reasons for not using eye segmentation maps as our prior to force E_k to translate input images to segmentation map directly in Sec III-C. Table IV shows the results of direct segmentation map translation methods with and without keypoint bottleneck and the results demonstrate their disadvantages quantitatively. We find that the keypoint representation bottleneck does not improve the results of directly segmentation map translation methods, because we only make use of a neural network for mapping the keypoint coordinates to segmentation maps and the mapping is not as accurate as the fitting method in Sec. III-B.

TABLE III

ABLATION STUDY ON TEYED (THE TOP) AND UNITYSEG (THE BOTTOM). P: KEYPOINTS' REPRESENTATION PRIOR. B: KEYPOINT BOTTLENECK. S: SYMMETRIC ARCHITECTURE.

S	B	P	mF1	mIoU	IoU(iris)	IoU(pupil)	IoU(sclera)
✓	✓	✓	0.937	0.884	0.890	0.923	0.789
	✓	✓	0.933	0.878	0.881	0.917	0.782
✓		✓	0.861	0.764	0.751	0.671	0.669
✓			0.480	0.338	0.329	0.186	0.197

S	B	P	mF1	mIoU	IoU(iris)	IoU(pupil)	IoU(sclera)
✓	✓	✓	0.938	0.886	0.887	—	0.797
	✓	✓	0.934	0.882	0.884	—	0.793
✓		✓	0.905	0.831	0.783	—	0.737
✓			0.639	0.537	0.542	—	0.144

TABLE IV

COMPARISON WITH DIRECT SEGMENTATION MAP TRANSLATION.

Datasets	TEyeD		UnitySeg	
Methods	mF1	mIoU	mF1	mIoU
Ours	0.937	0.884	0.938	0.886
Segmentation Map Translation	0.894	0.813	0.909	0.838
Segmentation Map Translation w/o keypoint bottleneck	0.901	0.824	0.912	0.842

G. Appearance and Shape Disentanglement

Given two arbitrary eye images I_1 and I_2, our model can generate a novel image with appearance from I_1 and eye shape from I_2. The new image is created by $G(E_a(I_1), E_k(I_2))$. In the same way, we can create the new image $G(E_a(I_2), E_k(I_1))$. Fig. 7 shows some examples combining one eye appearance with another eye shape.

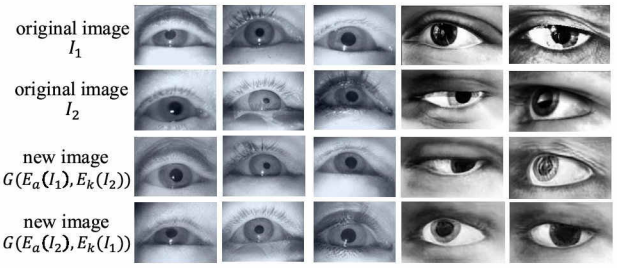

Fig. 7. **Disentanglement between Appearance and Shape**. The new images in bottle two rows are generated by exchanging the appearance and shape of two different images in top rows.

V. CONCLUSIONS AND FUTURE WORKS

We presented a self-supervised eye semantic segmentation method which has two procedures: training a self-supervised landmark detector using a symmetrical auto-encoder architecture and an eye keypoint prior, converting the detected landmark to the corresponding segmentation map. We have shown the effectiveness of our method on TEyeD and UnitySeg. This paper substitutes the unpaired keypoint prior for paired data and labels in the eye-parsing area. In the future, we would like to investigate more efficient methods to utilize the unpaired prior.

REFERENCES

[1] B. Adegoke, E. Omidiora, S. Falohun, and J. Ojo. Iris segmentation: a survey. *IJMER*, 2013.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

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

[4] F. L. Bookstein. Fitting conic sections to scattered data. *Computer graphics and image processing*, 9(1):56–71, 1979.

[5] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz. Ritnet: Real-time semantic segmentation of the eye for gaze tracking. In *ICCV Workshop*, 2019.

[6] J. H. Cho, U. Mall, K. Bala, and B. Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, 2021.

[7] A. Das, U. Pal, M. Blumenstein, and M. A. F. Ballester. Sclera recognition-a survey. In *IAPR*, 2013.

[8] J. Daugman. High confidence visual recognition of persons by a test of statistical independence. *TPAMI*, 15(11), 1993.

[9] A. Fitzgibbon, M. Pilu, and R. B. Fisher. Direct least square fitting of ellipses. *TPAMI*, 21(5), 1999.

[10] W. Fuhl, G. Kasneci, and E. Kasneci. Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. *CoRR*, abs/2102.02115, 2021.

[11] W. Fuhl, W. Rosenstiel, and E. Kasneci. 500,000 images closer to eyelid and pupil segmentation. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 2019.

[12] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. Openeds: Open eye dataset. *CoRR*, abs/1905.03702, 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[16] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NIPS*, 2018.

[17] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *CVPR*, 2020.

[18] X. Ji, J. F. Henriques, and A. Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019.

[19] P. Kansal and S. Devanathan. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. In *ICCV Workshop*, pages 3688–3693. IEEE, 2019.

[20] S.-H. Kim, G.-S. Lee, H.-J. Yang, et al. Eye semantic segmentation with a lightweight model. In *ICCV Workshop*. IEEE, 2019.

[21] R. S. Kothari, A. K. Chaudhary, R. J. Bailey, J. B. Pelz, and G. J. Diaz. Ellseg: An ellipse segmentation framework for robust gaze tracking. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2757–2767, 2021.

[22] S. Lian, Z. Luo, Z. Zhong, X. Lin, S. Su, and S. Li. Attention guided u-net for accurate iris segmentation. *Journal of Visual Communication and Image Representation*, 56:296–304, 2018.

[23] B. R. Mitchell et al. *The Spatial Inductive Bias of Deep Learning*. PhD thesis, Johns Hopkins University, 2017.

[24] R. A. Naqvi and W.-K. Loh. Sclera-net: Accurate sclera segmentation in various sensor images based on residual encoder and decoder network. *IEEE Access*, 7:98208–98227, 2019.

[25] Y. Ouali, C. Hudelot, and M. Tami. Autoregressive unsupervised image segmentation. In *ECCV*. Springer, 2020.

[26] J. Perry and A. Fernandez. Minenet: A dilated cnn for semantic segmentation of eye features. In *ICCV Workshop*, pages 0–0, 2019.

[27] 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*, pages 234–241. Springer, 2015.

[28] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. 2017.

[29] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017.

[30] R. P. Wildes. Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348–1363, 1997.

[31] E. Wood, T. Baltrusaitis, L. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In P. Qvarfordt and D. W. Hansen, editors, *ETRA*, pages 131–138. ACM, 2016.

[32] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018.

[33] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.