A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation

Yihua Cheng,[1] **Shiyao Huang,**[2] **Fei Wang,**[2] **Chen Qian,**[2] **Feng Lu**[1,3,4*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, China
[2]SenseTime Co., Ltd. [3]Peng Cheng Laboratory, Shenzhen, China
[4]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, China
{yihua_c, lufeng}@buaa.edu.cn, {huangshiyao, wangfei, qianchen}@sensetime.com

Abstract

Human gaze is essential for various appealing applications. Aiming at more accurate gaze estimation, a series of recent works propose to utilize face and eye images simultaneously. Nevertheless, face and eye images only serve as independent or parallel feature sources in those works, the intrinsic correlation between their features is overlooked. In this paper we make the following contributions: 1) We propose a coarse-to-fine strategy which estimates a basic gaze direction from face image and refines it with corresponding residual predicted from eye images. 2) Guided by the proposed strategy, we design a framework which introduces a bi-gram model to bridge gaze residual and basic gaze direction, and an attention component to adaptively acquire suitable fine-grained feature. 3) Integrating the above innovations, we construct a coarse-to-fine adaptive network named CA-Net and achieve state-of-the-art performances on MPIIGaze and EyeDiap.

Introduction

Human gaze implicates important cues for applications such as saliency detection (Alshawi, Long, and AlRegib 2018), human-computer interaction (Zhang, Sugano, and Bulling 2017) and virtual reality (Patney et al. 2016).

Gaze estimation methods can be divided into model-based methods and appearance-based methods. Model-based methods generally achieve accurate gaze estimation with dedicated devices, but are mostly limited to laboratory environment due to short working distance (typically within 60cm) and high failure rate in the wild. Appearance-based methods attract much attention recently, they require only a webcam to capture images and directly learn the mapping from images to gaze directions. As human eye appearance can be influenced by various factors in the wild such as head pose, CNN-based methods are proposed and significantly outperform classical methods thanks to CNN's superior ability in learning very complex mapping functions.

CNN-based methods estimate gaze directions from face or eye images. Zhang *et al.* (Zhang et al. 2015) first propose
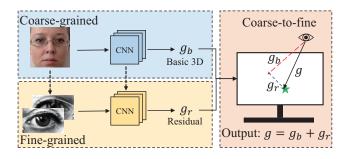
Figure 1: The process of coarse-to-fine strategy. We extract coarse-grained feature from face image to estimate basic gaze direction g_b and extract fine-grained feature from eye images to estimate gaze residual g_r. We use g_r to refine g_b and acquire the outputted gaze direction g.

a network to estimate gaze directions from eye images. Afterwards face images are put forward (Zhang et al. 2017a). Recently, methods simultaneously utilizing face and eye images achieve even better results (Chen and Shi 2018). Nevertheless, previous methods treat face and eye images only as independent or parallel feature sources, thus neglecting their intrinsic relationship at the level of feature granularity. In fact, eye image provide fine-grained feature focusing on gaze, while face image supplies coarse-grained feature with richer information.

To make full use of the feature relationship between face and eyes, we propose a coarse-to-fine strategy in this paper, which achieve state-of-the-art performances on the most common benchmarks. The core idea of proposed coarse-to-fine strategy is to **estimate a basic gaze direction from face image and refine it with corresponding residual predicted from eye images.** Specifically, since face image carries richer information than eye images, we utilize it to estimate an approximation of gaze direction. Then we extract fine-grained feature from eye images to refine the basic gaze direction and finally acquire coarse-to-fine gaze direction.

As shown in Figure 1, we first design a CNN to extract coarse-grained feature from face image and predict the basic gaze direction. Next, another CNN is set up to extract the fine-grained feature from eye images and generate

gaze residual. At last, we acquire the final gaze direction by adding the basic gaze direction and gaze residual vectorially.

However, there still remain two key problems to be considered. The first problem is how to ensure the estimated gaze residual is effective for refining its corresponding base gaze direction. The second problem is how to enforce the fine-grained features extracted from eye images to be suitable for estimating the gaze residual. Inspired from NLP algorithms, we generalize the coarse-to-fine process as a bi-gram model to solve the first problem. The bi-gram model bridges gaze residual and basic gaze direction, and produces gaze residual coupling with basic gaze direction. For the second problem, an attention component is proposed to adaptively acquire suitable fine-grained features.

Integrating above algorithms, we finally propose the coarse-to-fine adaptive network (CA-Net) for gaze estimation, which can adaptively acquire suitable fine-grained feature and estimates 3D gaze directions in a coarse-to-fine way. To the best of our knowledge, we are the first to consider the intrinsic correlation between face and eye images and propose a framework for coarse-to-fine gaze estimation.

In summary, the contributions of this work are threefold:

1. We propose a novel coarse-to-fine strategy for gaze estimation.

2. We propose an ingenious framework for coarse-to-fine gaze estimation. The framework introduces a bi-gram model, which bridges coarse-grained gaze estimation and fine-grained gaze estimation, and an attention component, which adaptively acquires suitable fine-grained feature.

3. Based on the proposed framework, we design a network named CA-Net and achieve state-of-the-art performances on MPIIGaze and EyeDiap.

Related work

Gaze estimation methods can be simply divided into model-based and appearance-based (Hansen and Ji 2010).

Model-based methods

Model-based methods can estimate gaze with good accuracy by building geometric eye models (Guestrin and Eizenman 2006). They typically fit the model by detecting eye features such as near infrared corneal reflections (Nakazawa and Nitschke 2012), pupil center (Valenti, Sebe, and Gevers 2012), and iris contours (Funes Mora and Odobez 2014; Xiong et al. 2014). However, the detection of eye features may require dedicated devices such as infrared lights, stereo/high-definition cameras, and RBG-D cameras (Funes Mora and Odobez 2014; Xiong et al. 2014). Meanwhile, model-based methods usually have limited working distances between the user and the camera. These limitations show that model-based methods are more suitable for controlled environment, e.g., the laboratory setting, rather than outdoor settings.

Appearance-based methods

Most of appearance-based methods only require a webcam to capture images and learn the mapping function from im-

ages to the corresponding gaze (Tan, Kriegman, and Ahuja 2002). The loose requirement attracts much attention for appearance-based methods. Up to now, many methods such as Neural networks (Baluja and Pomerleau 1994; Xu, Machin, and Sheppard 1998), adaptive linear regression (Lu et al. 2014b), Gaussian process regression (Williams, Blake, and Cipolla 2006) and dimension reduction (Lu et al. 2017) have been proposed to learn the mapping function. In order to handle arbitrary head motion, images can be used to learn more complex mapping functions (Lu et al. 2014a; 2015). However, learning a generic mapping function is still challenging because of the highly non-liner of mapping function.

Recently, CNNs-based methods show better accuracy than conventional appearance methods. Zhang et al. (Zhang et al. 2015) first proposed a CNNs-based method to estimate gaze, the method was designed based on LeNet (Lecun et al. 1998) and estimates gaze from eye images. Yu et al. (Yu, Liu, and Odobez 2018) proposed a multitask gaze estimation model with landmark constrain, they estimate gaze from eye images. Fischer et al. (Fischer, Chang, and Demiris 2018) extracted feature from two-eye images with VGG-16 (Karen and Andrew 2014) to estimate gaze, they use an ensemble scheme to increase robustness of proposed method. Cheng et al. (Cheng, Lu, and Zhang 2018) proposed a CNNs-based network which uses two-eye images as inputs and utilizes the two-eye asymmetry to optimize whole network.

Meanwhile, recent studies prove face images is effective in CNNs-based methods. Krafka et al. (Krafka et al. 2016) implemented the CNNs-based gaze tracker in the mobile devices, it estimates gaze from face and eye images. Zhang et al. (Zhang et al. 2017a) proposed a spatial weights CNN to estimate gaze from face images. Deng et al. (Zhu and Deng 2017) proposed a CNNs-based method with geometry constraints, it uses face and eye images as inputs and can estimate gaze in free-head setting. Zhao et al. (Chen and Shi 2018) proposed a CNNs-based method using dilated convolution to estimate gaze from face and eye images. Xiong et al. (Xiong and Kim 2019) combines the mixed effects model with CNN and estimates gaze from face images.

Method

In this section, we introduce the architecture of our CA-Net, which can adaptively acquire suitable fine-grained feature and estimate the gaze directions in a coarse-to-fine way.

Overview

The core idea of the coarse-to-fine strategy is to estimate a basic gaze direction from face image and refine it with corresponding residual predicted from eye images. We propose the CA-Net based on the coarse-to-fine strategy.

The CA-Net contains two subnets: Face-Net and Eye-Net. Face-Net extracts coarse-grained feature from face image and estimates the basic gaze direction. Eye-Net estimates gaze residual from two eye images to refine the basic gaze direction. Next, We first propose an attention component to adaptively assign weights for two-eye features. A suitable eye feature is acquired by adding the weighted two-eye features together. In addition, since the gaze residual
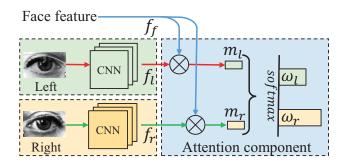
Figure 2: The architecture of proposed attention component. It adaptively assigns weights for left and right eyes.
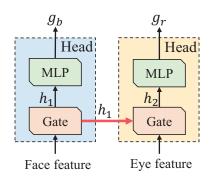

Figure 3: The coarse-to-fine process is generalized as a bi-gram model, which bridges gaze residual g_r and corresponding basic gaze direction g_b.

is associated with the basic gaze estimation, we generalize the coarse-to-fine process as a bi-gram model to bridge the Face-net and Eye-net, and produces gaze residual coupling with basic gaze direction. Finally, CA-Net output gaze direction by adding the basic gaze direction and gaze residual together.

The rests of this section are organized as follows. We first introduce the process of feature generation, in which we propose an attention component to adaptively assign weights for two-eye features. We then introduce the coarse-to-fine strategy, which can be generalized as a bi-gram model. Next, we detail the architecture of proposed CA-Net and define the loss function of CA-Net. At last, we present the implementation details at the end of this section.

Feature generation

A key point of coarse-to-fine strategy is to acquire suitable feature, especially for estimating gaze residual. Therefore, we first describe the process of feature generation.

The face feature is used to estimate basic gaze direction. Therefore, a common CNN is used to extract the coarse-grained face feature from face images. As for eye feature, we also respectively use a CNN to extract features from two eye images. However, after acquiring the left and right eye features, a key problem is how to obtain suitable eye feature from two-eye features to accurately estimate gaze residual.

There are at least two factors we need to consider. First, as for different basic gaze directions, the suitable eye features can be different. Second, two eye appearances have different reliabilities for gaze estimation (Cheng, Lu, and Zhang 2018) because of the in-the-wild setting such as free-head. Those two factors both influence the acquirement of suitable eye feature. In order to tackle above factors, we propose an attention component which can adaptively assign weights for two eyes. The suitable eye feature is produced by summing the weighted left and right eye features.

The attention component is inspired by attention mechanisms, which are widely used in NLP (Vaswani et al. 2017). The architecture of proposed component is shown in Fig. 2. In particular, as for left eye image, a score m_l is produced by left eye feature f_l and face feature f_f. As for right eye image, a score m_r is produced by right eye feature f_r and face feature f_f. Then, a softmax layer is used to balance the scores

m_l and m_r and the weights w_l and w_r are outputted for left eye and right eye. The method which produces scores from feature is various, in the paper, we directly use the method proposed in (Dzmitry Bahdanau 2015).

The proposed component has following properties:

1. Score m_l is related with face feature, which is used to predict basic gaze directions. It means that basic gaze directions can decide the size of m_l. This corresponds with the first factor we describe above.

2. Score m_l is related with left eye feature. On the other words, m_l is related with the left eye appearance. It corresponds with the second factor we describe above.

3. Score m_l is irrelevant to right eye feature, it is reasonable that the scores of left eyes are irrelevant to right eyes.

4. w_l is generated by comparing m_l and m_r. Although the scores of left eyes are irrelevant to right eyes, the final weight should be generated with considering both the scores of two eyes.

5. Above properties also suit for m_r and w_r.

We also formulate the process of our implementation as follows. we acquire the score of left eyes by

$$m_l = v^T tanh(W_1^T f_f + W_2^T f_l), \quad (1)$$

and acquire the score of right eyes by

$$m_r = v^T tanh(W_1^T f_f + W_2^T f_r), \quad (2)$$

where v, W_1 and W_2 are learned parameters and are implemented by fully connected layers.

Meanwhile, a softmax layer is used to balance the scores of left and right and outputs the weights.

$$[w_l, w_r] = softmax([m_l, m_r]). \quad (3)$$

The final eye feature f_e can be acquired by

$$f_e = w_l * f_l + w_r * f_r. \quad (4)$$

Coarse-to-fine gaze estimation

After acquiring the features, it is still unknown how to perform gaze estimation in a coarse-to-fine way. A straightforward solution is that learn a mapping function to estimate the basic gaze direction from coarse-grained face feature and then learn another mapping function to estimate the gaze residual from fine-grained eye feature. However, this solution has two problems. First, it do not consider the relation between basic gaze directions and gaze residuals. Second, for the estimation of gaze residual, although eye images are finer than face images, it is bad to directly discard the face feature. Therefore, we generalize the coarse-to-fine process as a bi-gram model. The architecture of the bi-gram model is shown in Fig. 3, we omit the process of feature generation which can be various.

Specifically, as show in Fig. 3, the face feature is processed by a gate function to produce state h_1. Then, on one hand, the state h_1 is used to estimate the basic gaze direction. On the other hand, the state h_1 is delivered into the next gate and produces the state h_2 with eye feature. The state h_2 is used to estimate gaze residuals. The gate function can be various. The main task of designed gate is to filter the previous states and reduce the influence of previous task on current task. We use GRU (Cho et al. 2014) in this work.

The coarse-to-fine process can be understood as follows. The basic gaze directions are directly estimated from state h_1, which is produced by face feature. The gaze residuals are estimated from state h_2, which is generated from h_1 and eye feature. This means the process of estimating gaze residuals is related with basic gaze directions. Meanwhile, with delivering h_1, the face feature is also implicitly used to estimate gaze residuals rather than discarding. Moreover, since the face feature includes much coarse-grained information, a learned gate is used to adaptively filter the state h_1.

The process of learned gate is shown as follow:

$$z_i = \sigma(W_z \cdot [h_i, f]). \tag{5}$$

$$r_i = \sigma(W_r \cdot [h_i, f]). \tag{6}$$

$$\tilde{h_{i+1}} = ReLU(W_h \cdot [r_i * h_i, f]). \tag{7}$$

$$h_{i+1} = (1 - z_i) * h_i + z_i * \tilde{h_{i+1}}, \tag{8}$$

where f represents the corresponding feature. W_z, W_r and W_h are learned parameters, which can be implemented with fully connected layers. The h_0 is set as a zero matrix.

CA-Net

By integrating above algorithms, we propose CA-Net which can adaptively acquire suitable eye features and estimates 3D gaze directions in a coarse-to-fine way. The architecture of the proposed CA-Net is shown in Fig. 4. It contains two subnets, which are Face-Net and Eye-Net.

Face-Net uses face images as input to estimate the basic gaze directions. We first design a CNN to extract face feature from face images. After acquiring the face feature, we deliver the face feature into the head component (detail in
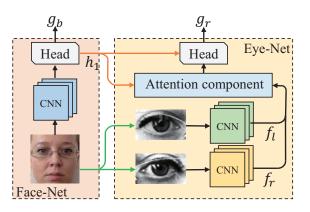

Figure 4: The architecture of CA-Net, which estimates gaze in coarse-to-fine way. The Face-Net estimates basic gaze directions from face images. The Eye-Net estimates gaze residuals from eye images.

Fig. 3). The basic gaze direction g_b and state h_1 are produced by the head component.

Eye-Net uses two eye images as inputs. Two CNNs are designed to extract left eye feature f_l and right eye feature f_r. Then, a attention component is used to fusion f_l and f_r (detail in Fig. 2). We input the state h_1 rather than face feature into the attention component to guiding the generation of eye feature. After acquiring eye feature, we send the eye feature with h_1 into a head component to estimate gaze residuals g_r.

The final output of our CA-Net is

$$g = g_b + g_r \tag{9}$$

Given the ground truth g^*, the loss function of CA-Net is defined as

$$Loss = \alpha * \mathcal{L}(g_b, g^*) + \beta * \mathcal{L}(g, g^*), \tag{10}$$

where $\mathcal{L}$ is defined as

$$\mathcal{L}(a, b) = \arccos\left(\frac{a \cdot b}{\|a\|\|b\|}\right). \tag{11}$$

We empirically set $\alpha = 1$ and $\beta = 2$. On the one hand, this loss function encourages CA-Net to estimate an accurate basic gaze direction. On the other hand, we assign a larger weight for g than g_b to ensure CA-Net can get a more accurate gaze direction than the basic gaze direction.

Implementation detail

The inputs of CA-Net are 224*224*3 face images, 36*60 gray-scale left and right eye images.

The CNN in Face-Net consists of thirteen convolutional blocks. Each block contains one convolutional layer, one ReLU and one Batch Normalization (Ioffe and Szegedy 2015). The sizes and strides of all convolutional kernels are set as 3*3 and 1. The numbers of convolutional kernel are (64, 64, 128, 128, 256, 256, 256, 256, 256, 256, 512, 512, 1024). We also insert one max pool layer after the second, fourth, seventh and tenth convolutional blocks. The sizes of max-pooling layers are 2*2 and strides are 2*2. A global average pooling layer (GAP) is used after the thirteenth block
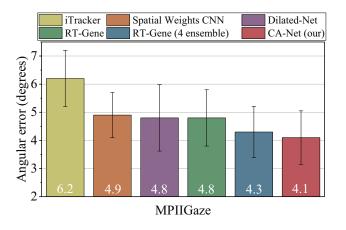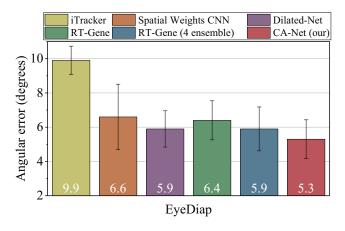
Figure 5: Performance in MPIIGaze dataset

Figure 6: Performance in EyeDiap dataset

and output 1024D feature. Final, the 1024D feature is sent to a fully connected layer (FC) to output 256D face features.

The CNN for the left eye in Eye-Net consists of ten convolutional blocks. The numbers of convolutional kernel are (64, 64 ,128, 128, 128, 256, 256, 256, 512, 1024). The strides of the second, fifth, eighth convolutional kernels are set as 1. The sizes of all convolutional kernels are set as 3*3. Same as Face-Net, a GAP and FC are final used to output the 256D left eye feature. Meanwhile, for the right eye, the same CNN is designed and output 256D right eye feature.

We implement CA-Net by using Pytorch. We train the whole network in 200 epochs with 32 batch size. The Learning rate is set as 0.001. We initialize the weights of all layers with MSRA initialization (He et al. 2015).

Experiment

Dataset

The experiments are conducted in two popular gaze estimation datasets: MPIIGaze (Zhang et al. 2017b) and Eye-Diap (Mora, Monay, and Odobez 2014).

MPIIGaze is the largest dataset for appearance-based gaze estimation which provides 3D gaze directions. It is common used in the evaluation of appearance-based methods (Zhang et al. 2017a; Ranjan, De Mello, and Kautz 2018; Liu et al. 2018; Cheng, Lu, and Zhang 2018; Xiong and Kim 2019). MPIIGaze dataset contains 213,659 images which are captured from 15 subjects. Note that, MPIIGaze provides a standard evaluation protocol, which selects 3000 images for each subject to compose the evaluation set. We conduct experiments in the evaluation set rather than the full set.

EyeDiap dataset contains a set of video clips of 16 participants. The videos are collected under two visual target sessions, which are screen target and 3D floating ball. We use the videos collected under screen target sessions and sample one image per fifteen frames to construct the evaluation set. Note that, since two subjects lack the videos in the screen target session, we obtain the images of 14 subjects finally.

Data preprocessing

We follow the process proposed in (Zhang et al. 2017b) to normalize the two datasets. Specifically, the goal of appearance-based gaze estimation is to estimate gaze directions from eye appearances. However, since head pose has six freedoms, eye appearances are various in the real world. This complicates the gaze estimation task. Therefore, we eliminate the translation in head pose by rotating the virtual camera and the roll in head pose by wrapping images. In addition, we crop eye images from normalized face images with provided landmarks by the dataset. Note that the landmark can be also automatically detected by various face detection algorithms (Brandon, Ludwiczuk, and Mahadev 2016). The eye images are histogram-equalized and converted into gray scale to eliminate the influence of illumination. Note that, the images provided by MPIIGaze has been normalized, we only apply the normalization into EyeDiap.

Comparison with appearance based methods

We first conduct an experiment to compare the performance of the proposed method with other appearance-based methods. The experiment is conducted in both MPIIGaze and EyeDiap. Note that, for the two datasets, we both apply the leave-one-person-out strategy to obtain robust results.

We choose four methods as compared methods, which are *iTracker* (Krafka et al. 2016), *Spatial weights CNN* (Zhang et al. 2017a), *Dilated-Net* (Chen and Shi 2018) and *RT-Gene* (Fischer, Chang, and Demiris 2018). Since the accuracy of *RT-Gene* can be improved by four models ensemble, we also show the result of the model ensemble and call it as *RT-Gene (4 ensemble)* to distinguish from *RT-Gene*. Note that, currently, the best reported performance in MPIIGaze is achieved by *RT-Gene (4 ensemble)*.

Fig. 5 shows the result in MPIIGaze dataset. The performances of *Spatial weights CNN*, *Dilated-Net* and *RT-Gene* are all around $4.8°$. The *RT-Gene (4 ensemble)* can improve the performance by a large margin using ensemble scheme, which is $4.3°$. Our CA-Net achieves the state-of-the-art performance as $4.1°$ in the MPIIGaze dataset. The CA-Net has $0.7°$ improvement compared with *RT-Gene* and also has

Table 1: Comparison between appearance-based methods.

Methods	MPIIGaze	EyeDiap
iTracker	6.2°	9.9°
Spatial Weights CNN	4.9°	6.6°
RT-Gene	4.8°	6.4°
Dilated-Net	4.8°	5.9○
RT-Gene(4 ensemble)	4.3°	5.9°
Faze (Park et al. 2019)	5.2°	–
MeNet (Xiong and Kim 2019)	4.9°	–
Our method	**4.1°**	**5.3°**

Table 2: Ablation study.

Methods	Performance
Gate ablation	4.46°
Attention ablation	4.50°
Face-Net	4.58°
Eye-Net	5.01°
Joint-Net	5.00°
Face-Net (CA)	4.65°
CA-Net	**4.14°**

$0.2°$ improvement compared with *RT-Gene (4 ensemble)*. Note that, our CA-Net achieves the state-of-the-art performance without ensemble scheme. The accuracy also can be further improved using ensemble.

we re-implement the Dilated-Net according to the original paper and use the author provided source codes for the rest methods. Fig. 6 shows all the results. *iTracker* has the worst performance because of the shallow network. *Spatial weights CNN* and *RT-Gene* have $6.6°$ and $6.4°$ performance in EyeDiap. However, *Dilated-Net* significantly outperforms *Spatial weights CNN* and *RT-Gene*. It shows the a performance as *RT-Gene (4 ensemble)* which is $5.9°$. Our CA-Net achieves the best performance as $5.3°$ in EyeDiap and has $0.6°$ improvement compared with *Dilated-Net*.

The good performance in two datasets demonstrates the advantage of the proposed CA-Net. In addition, since some recent appearance-based methods do not provide source code and the methods also are difficult to re-implement , we carefully show the reported accuracy in Table 1 for reference. In order to get a fair comparison, we only show the accuracy in MPIIGaze because the MPIIGaze dataset provides a standard evaluation set.

Ablation study

In order to demonstrate the effectiveness of each component in the CA-Net, we perform ablation study in MPIIGaze.

Ablation study about components. We first perform ablation study to demonstrate the effect of the attention component and gate component.Specifically, we evaluate two extra methods which are *Gate ablation* and *Attention ablation*.

Gate ablation ablates the learned gate from CA-Net, it directly concatenates face feature with eye feature to estimate gaze residuals. Note that, we do not modify the attention component, where the face feature is also inputted into the attention component to guide the generation of eye feature.

Attention ablation ablates the attention component from CA-Net, it assigns fixed weights as 0.5 for both left eye and right eye to generate the fine-grained eye feature.

The result is shown in the second row of Table 2. The performance of *Gate ablation* shows $0.32°$ decrease compared with CA-Net. Meanwhile, the performances of *Attention ablation* have $0.46°$ decrease than CA-Net. It demonstrates the advantages of attention component and learned gate.

Ablation study about network. The proposed CA-Net shows the best performance in two datasets. However, it is still uncertain whether the coarse-to-fine strategy can im-

prove the performance. In order to prove the advantages of coarse-to-fine, we perform ablation study on the network.

We respectively evaluate each subnet in CA-Net. Totally three methods are evaluated. *Face-Net.* We directly use the Face-Net to estimate gaze. *Eye-Net.* We directly use the Eye-Net to estimate gaze from two eye images. Note that, the attention component is not used in this method. We generate eye feature by directly concatenating the left eye feature and right eye feature. *Joint-Net.*We use the same architecture as CA-Net to extract face feature, left eye feature and right eye feature. The gaze directions are estimated by the joint feature which is generated by concatenating the three features. We also provide the performance of basic gaze directions in CA-Net and call it as *Face-Net (CA)*.

The result is shown in Table 2. *Face-Net* shows the best performance between compared methods, which is $4.58°$. *Face-Net (CA)* achieves $4.65°$ performance which has $0.7°$ decrease compared with *Face-Net*. However, with the coarse-to-fine strategy, CA-Net achieves $0.51°$ improvement than *Face-Net (CA)* and significantly outperforms other methods with $4.14°$ performance. This demonstrates the advantages of the proposed coarse-to-fine strategy.

Moreover, although the backbone of *Joint-Net* is the same as CA-Net, CA-Net achieves $0.86°$ improvement than *Joint-Net*. It is benefited from the proposed coarse-to-fine strategy.

Additional analysis

In order to show the advantages of the algorithms proposed in CA-Net, We perform some additional analysis in MPIIGaze and summarize the results into Table 3. The performance of each method is shown in the column of "Refine". In addition, we also show the performance of basic gaze directions in Table 3 and list the results in the column of "Basic". We call it as basic performance in the rest parts.

Coarse-to-fine v.s. Fine-to-coarse. The core of our paper is the coarse-to-fine strategy. In order to further validate the correctness of the coarse-to-fine strategy, we evaluate the performance of *Fine-to-coarse* . As for *Fine-to-coarse*, it means to estimate a basic gaze direction from eye images and refine it with residual predicted from face image.

As shown in Table 3, it is obvious that our CA-Net, i.e. coarse-to-fine strategy, achieves better performance than *Fine-to-coarse*. With only changing the strategy, *Fine-to-coarse* has $0.49°$ decrease compared with CA-Net. It demonstrates the advantages of our coarse-to-fine strategy. In addition, an interesting observation is that the performance of *Fine-to-coarse* is similar with the performance of

Table 3: Additional analysis about different algorithms.

Methods	Basic	Refined
Fine-to-coarse	5.14°	5.00°
One gram	4.42°	4.43°
Face attention	4.47°	4.30°
Eye attention	4.43°	4.34°
CA-Net	**4.65°**	**4.14°**

Eye-Net (show in Table 2) while our CA-Net can improve the performance by a large margin compared with *Face-Net* (show in Table 2). It proves the correctness of the proposed coarse-to-fine strategy.

bi-gram v.s. One gram. In order to estimate the gaze direction in a coarse-to-fine way, one key idea is that gaze residuals are associated with basic gaze directions. Based on the idea, we generalize the coarse-to-fine way as a bi-gram model. However, it is uncertain whether the bi-gram model is useful. In this part, we provide a comparison between bi-gram model and one gram model to show the advantages of bi-gram model. In particular, we simply use a zero matrix to replace the delivered face feature, where the gaze residuals are only estimated from eye feature. Note that, we do not modify the attention component. The fine-grained eye feature is also generated with the guiding of face feature.

As shown in Table 3, *One gram* shows a better basic performance than CA-Net. However, without the information about basic gaze directions, the fine-grained eye feature can not further refine the basic gaze direction. Final, the *One gram* has 0.29° decrease than CA-Net. The result demonstrates the usefulness of bi-gram model.

Attention component v.s. other weight generations In order to acquire suitable fine-grained feature to estimate gaze residuals, we propose an attention component to adaptively assign weights for left and right eyes. Specifically, the attention component learns the eye weights from face feature and corresponding eye feature. In order to show the advantages of the proposed attention component, in this part, we conduct comparison by replacing the weight component.

There are two weight generations chosen for comparing. *Face attention* generates the weights of two eyes from face feature. *Eye attention* generates the weights of two eyes from corresponding eye features. The results are shown in Table 3. A suitable baseline is *Attention ablation* (show in Table 2), which achieve 4.5° performance. As shown in Table 3, *Face attention* and *Eye attention* show the better performance compared with *Abalte Attention*. It demonstrates that the face feature and corresponding eye feature both are useful for the coarse-to-fine gaze estimation. Meanwhile, they both show worse performance than CA-Net. This demonstrates the advantages of the proposed attention component.

Visual results. We also show some visual results in Fig. 7. It is obvious that our method can perform well in different cases. In addition, as shown in the sixth and seventh subfigures in Fig. 7, our CA-Net can also produce accurate gaze directions when the gaze direction deviates from the face direction. This demonstrates that our method not only focuses on face images but also is sensitive to the eye region.

Figure 7: Some visual results of estimated 3D gaze.

Conclusion

In this paper, we propose a coarse-to-fine strategy to estimate gaze directions. The process of the coarse-to-fine strategy is to estimate a basic gaze direction from face image and refine it with residual predicted from eye images. A key point of the coarse-to-fine strategy is the estimation of gaze residuals. In order to accurately estimate gaze residuals, we propose an attention component to adaptively assign weights for eye images and to obtain suitable eye feature. In addition, we also generalize the coarse-to-fine process as a bi-gram model to bridge the basic gaze directions and gaze residuals. Based on above algorithms, we propose CA-Net, which can adaptively acquire suitable fine-grained feature and estimates 3D gaze directions in a coarse-to-fine way. Experiments show the CA-Net achieves state-of-the-art performance in MPIIGaze and EyeDiap.

References

Alshawi, T.; Long, Z.; and AlRegib, G. 2018. Unsupervised uncertainty estimation using spatiotemporal cues in video saliency detection. *IEEE Transactions on Image Processing* 27(6):2818–2827.

Baluja, S., and Pomerleau, D. 1994. Non-intrusive gaze tracking using artificial neural networks. In *NIPS*, 753–760.

Brandon, A.; Ludwiczuk, B.; and Mahadev, S. 2016. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Chen, Z., and Shi, B. E. 2018. Appearance-based gaze estimation using dilated-convolutions. In *ACCV*.

Cheng, Y.; Lu, F.; and Zhang, X. 2018. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *ECCV*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dzmitry Bahdanau, Kyunghyun Cho, Y. B. 2015. Neural machine translation by jointly learning to align and translate. In *international conference on learning representations*.

Fischer, T.; Chang, H.; and Demiris, Y. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*.

Funes Mora, K. A., and Odobez, J.-M. 2014. Geometric

generative gaze estimation (g3e) for remote rgb-d cameras. In *CVPR*, 1773–1780.

Guestrin, E., and Eizenman, M. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. on Biomedical Engineering* 53(6):1124–1133.

Hansen, D., and Ji, Q. 2010. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. on PAMI* 32(3):478–500.

He, K.; Zhang, X.; Ren, S.; and Jian, S. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift.

Karen, S., and Andrew, Z. 2014. Very deep convolutional networks for large-scale image recognition. In *CVPR*.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *CVPR*, 2176–2184.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Liu, G.; Yu, Y.; Funes Mora, K. A.; and Odobez, J.-M. 2018. A differential approach for gaze estimation with calibration. In *29TH BRITISH MACHINE VISION CONFERENCE*.

Lu, F.; Okabe, T.; Sugano, Y.; and Sato, Y. 2014a. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing (IVC)* 32(3):169–179.

Lu, F.; Sugano, Y.; Okabe, T.; and Sato, Y. 2014b. Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36(10):2033–2046.

Lu, F.; Okabe, T.; Sugano, Y.; and Sato, Y. 2015. Gaze estimation from eye appearance: A head pose-free method via eye image synthesis. *IEEE Transactions on Image Processing (TIP)* 24(11):3680–3693.

Lu, F.; Chen, X.; ; and Sato, Y. 2017. Appearance-based gaze estimation via uncalibrated gaze pattern recovery. *IEEE Transactions on Image Processing (TIP)* 26(4):1543–1553.

Mora, K. A. F.; Monay, F.; and Odobez, J. M. 2014. Eyediap:a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Symposium on Eye Tracking Research and Applications*, 255–258.

Nakazawa, A., and Nitschke, C. 2012. Point of gaze estimation through corneal surface reflection in an active illumination environment. In *ECCV*, 159–172.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-shot adaptive gaze estimation. In *ICCV*.

Patney, A.; Kim, J.; Salvi, M.; Kaplanyan, A.; and Luebke, D. 2016. Perceptually-based foveated virtual reality. In *Acm Siggraph Emerging Technologies*.

Ranjan, R.; De Mello, S.; and Kautz, J. 2018. Light-weight head pose invariant gaze tracking. In *CVPRW*.

Tan, K.; Kriegman, D.; and Ahuja, N. 2002. Appearance-based eye gaze estimation. In *WACV*, 191–195.

Valenti, R.; Sebe, N.; and Gevers, T. 2012. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 21(2):802–815.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

Williams, O.; Blake, A.; and Cipolla, R. 2006. Sparse and semi-supervised visual mapping with the S^{3}GP. In *CVPR*, 230–237.

Xiong, Y., and Kim, H. J. 2019. Mixed effects neural networks (menets) with applications to gaze estimation. In *CVPR*.

Xiong, X.; Liu, Z.; Cai, Q.; and Zhang, Z. 2014. Eye gaze tracking using an rgbd camera: a comparison with a rgb solution. *PETMEI Workshop* 1113–1121.

Xu, L. Q.; Machin, D.; and Sheppard, P. 1998. A novel approach to real-time non-intrusive gaze finding. In *BMVC*, 428–437.

Yu, Y.; Liu, G.; and Odobez, J.-M. 2018. Deep multitask gaze estimation with a constrained landmark-gaze model. In *ECCV workshop*.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *CVPR*, 4511–4520.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017a. It's written all over your face: Full-face appearance-based gaze estimation. In *CVPR Workshop)*.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017b. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *TPAMI* PP.

Zhang, X.; Sugano, Y.; and Bulling, A. 2017. Everyday eye contact detection using unsupervised gaze target discovery. In *UIST*.

Zhu, W., and Deng, H. 2017. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *ICCV*.