SCIENCE CHINA
Information Sciences

June 2022, Vol. 65 160101:1–160101:12
https://doi.org/10.1007/s11432-020-3181-9

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

Onfocus detection: identifying individual-camera eye contact from unconstrained images

Dingwen ZHANG[1†], Bo WANG[2,4†], Gerong WANG[1†], Qiang ZHANG[1*], Jiajia ZHANG[1], Jungong HAN[3*] & Zheng YOU[2]

[1]*School of Mechano-Electronic Engineering, Xidian University, Xi'an* 710071*, China;*
[2]*State Key Laboratory of Precision Measurement Technology and Instruments, Tsinghua University, Beijing* 100084*, China;*
[3]*Computer Science Department, Aberystwyth University, Ceredigion* SY23 3FL*, UK;*
[4]*Beijing Jingzhen Medical Technology Ltd., Beijing* 100084*, China*

Received 16 September 2020/Revised 8 December 2020/Accepted 29 January 2021/Published online 20 April 2022

Abstract Onfocus detection aims at identifying whether the focus of the individual captured by a camera is on the camera or not. Based on the behavioral research, the focus of an individual during face-to-camera communication leads to a special type of eye contact, i.e., the individual-camera eye contact, which is a powerful signal in social communication and plays a crucial role in recognizing irregular individual status (e.g., lying or suffering mental disease) and special purposes (e.g., seeking help or attracting fans). Thus, developing effective onfocus detection algorithms is of significance for assisting the criminal investigation, disease discovery, and social behavior analysis. However, the review of the literature shows that very few efforts have been made toward the development of onfocus detector owing to the lack of large-scale public available datasets as well as the challenging nature of this task. To this end, this paper engages in the onfocus detection research by addressing the above two issues. Firstly, we build a large-scale onfocus detection dataset, named as the onfocus detection in the wild (OFDIW). It consists of 20623 images in unconstrained capture conditions (thus called "in the wild") and contains individuals with diverse emotions, ages, facial characteristics, and rich interactions with surrounding objects and background scenes. On top of that, we propose a novel end-to-end deep model, i.e., the eye-context interaction inferring network (ECIIN), for onfocus detection, which explores eye-context interaction via dynamic capsule routing. Finally, comprehensive experiments are conducted on the proposed OFDIW dataset to benchmark the existing learning models and demonstrate the effectiveness of the proposed ECIIN.

Keywords onfocus detection, deep neural network, capsule routing, computer vision, deep learning

Citation Zhang D W, Wang B, Wang G R, et al. Onfocus detection: identifying individual-camera eye contact from unconstrained images. Sci China Inf Sci, 2022, 65(6): 160101, https://doi.org/10.1007/s11432-020-3181-9

1 Introduction

In this work, we define onfocus detection as the task to identify whether the focus of the individual (human or animal) captured by a camera is on the camera or not, which is performed on images captured in unconstrained imaging conditions (so-called "in the wild") as shown in Figure 1. The focus of an individual during face-to-camera communication leads to individual-camera eye contact, which is a special type of eye contact and plays a crucial role in social communication. Such eye contact could reflect irregular individual status (e.g., lying [1] or suffering mental disease [2]) or special purposes (e.g., seeking help [3] or attracting fans [4]). To this end, onfocus detection tends to be instrumental to a wide range of real-world applications in the fields of criminal investigation, disease discovery, social behavior analysis and it also has value for facilitating robots for human-machine interaction and communication.

In spite of its academic value and practical significance, onfocus detection also presents great challenges in unconstrained capture conditions owing to the complex image scenes, unavoidable occlusion, diverse

* Corresponding author (email: qzhang@xidian.edu.cn, jungonghan77@gmail.com)
† Zhang D W, Wang B, and Wang G R have the same contribution to this work.

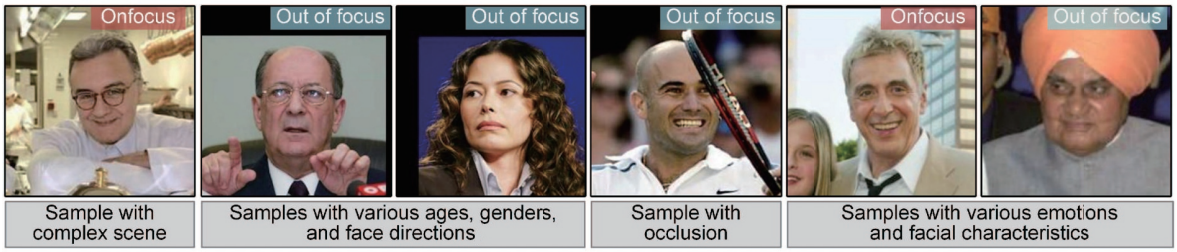

Figure 1 (Color online) Examples to illustrate the onfocus detection task and the challenge scenarios.

face directions and emotions, various facial features (especially for the features of the eyes), and imagery factors like blur, over-exposure, etc. Some examples to illustrate the challenges of onfocus detection in the wild (OFDIW) can also be referred to in Figure 1.

According to the existing literature, few efforts have been made toward the onfocus detection problem, leaving most of the aforementioned challenges under-studied[1)]. Besides, there is not a publicly available dataset for this task. To advance this research field, this paper establishes a large-scale onfocus detection dataset called OFDIW. OFDIW consists of 20623 unconstrained images which are collected from the LFW dataset [6] and the Oxford-IIIT Pet dataset [7], enabling us to study the individual-camera eye contact for both humans and pets. It is worth mentioning that human-pet communication becomes prevalent in our daily life. Many people raise pets and some companies raise pets and use them to help people relieve stress. However, owing to the barrier in language, human-pet communication heavily relies on eye contact. Thus, we think involving the onfocus study of pets might be beneficial and we thus choose two of the most common pets, i.e., dog and cat, for our study. Other pets will be studied in the future. Each image in this dataset is labeled with the corresponding ground-truth, i.e., either onfocus or out of focus, by human annotators. With the large-scale images and ground-truths, OFDIW can benefit methods based on the current deep learning models. Besides, the individuals contained in OFDIW have good diversity in emotions, ages, facial characteristics, and rich interactions with surrounding objects and background scenes. The samples also satisfy the imbalanced data distribution. All these factors make OFDIW closer to real-world scenarios. This dataset will be publicly accessible to other researchers to promote development in this area.

With OFDIW, we can accomplish the onfocus detection task by training the convolutional neural network (CNN) models, which is basically a classification problem. However, directly applying the conventional CNN model to classify the input image will not always work as it ignores an important factor—the eye-context interaction. As we know, eye regions provide critical cues for onfocus detection. However, by directly performing the conventional CNN model onto the input image, most of the extracted information actually comes from the context region as the eye region only occupies a very small fraction of the whole image. Although the context regions are also informative for onfocus detection, using them as the dominant cue will not obtain the correct onfocus detection results. Another strategy is to first localize the eye regions and then accomplish onfocus detection based on the features extracted on the eye regions. This strategy could obtain more accurate detection results. However, in cases when individual's gaze has the interaction with the context facial characteristics (e.g., face directions and locations) or context objects (e.g., balls or other people), the information extracted only from the eye regions is still insufficient for identifying the focus of the individual precisely (see examples in Figure 2). Thus, designing effective frameworks to explore eye-context interaction becomes a critical issue.

To address this issue, we propose a novel network model called eye-context interaction inferring network (ECIIN), in which context CNN extracts features with high activations on eye region, while eye CNN extracts features with high activations on some more detailed sub-regions in eye region. Then, capsules of context capsule network (CAP) and eye CAP are learned to reflect the context status and eye status, respectively. ECIIN is characterized by the following. (1) It is designed with a two-stream network architecture to extract the features of the eye region and the context region separately and then infer the eye-context interaction via an interaction inference stream. (2) It has an eye region mining module, which enables it to mine the important eye regions from each input image without any specific annotation on

1) It is worth mentioning that the method in [5] aims to estimate whether a certain image pixel (from the whole image scene) is at the camera's focus distance or not whereas our study tries to estimate whether the person within the image is staring at the camera or not.

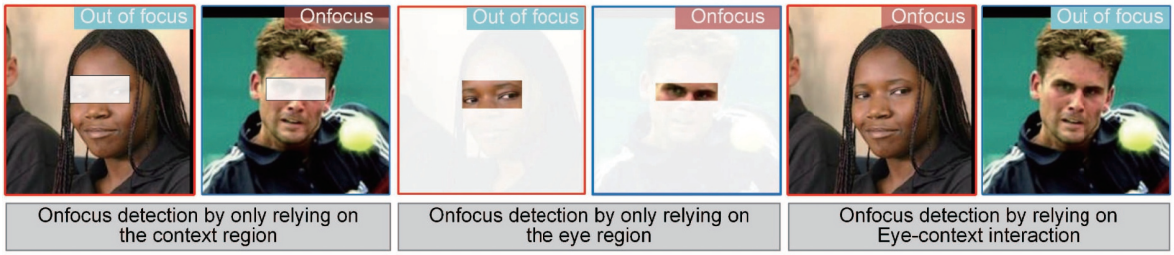

Figure 2 (Color online) Illustrate the importance of exploring eye-context interaction for onfocus detection. From the examples, we can observe that onfocus detection by only relying on either the context region (the first block) or the face region (the second block) may lead to incorrect results while considering the eye-context interaction (the third block) would obtain the correct results. The images framed with the same color are the same sample. The sample with the red frame shows the interaction of the eye region and the face direction, while the sample with the blue frame shows the interaction of the eye region and the nearby object, i.e., the ball.

them. (3) Considering the superior capacity of the capsule network [8] in modeling part-whole relationship and overcoming view changes, we introduce the dynamic capsule routing scheme into ECIIN to construct the eye capsules and context capsules, and meanwhile, model the interaction among the eye and context regions explicitly. To sum up, this work mainly contains three-fold contributions.

• We study the novel onfocus detection task and build a large-scale dataset, i.e., OFDIW. OFDIW will be publicly available to promote further development of this research field.

• We reveal the main challenges in this task and establish a novel end-to-end deep model, called ECIIN, to explore the informative eye-context interaction cues for onfocus detection.

• Comprehensive experiments on the OFDIW dataset have been conducted, which benchmarks the performance of the state-of-the-art eye contact detection, gaze estimation, image categorization networks, and fine-grained image classification models, and demonstrates the superior capacity of the proposed ECIIN. The project (containing both datasets and codes) is open at the website[2)].

2 Related work

Eye contact detection. While this work is closely related to the field of eye contact detection [9–11], the difference is that eye contact detection mainly studies the eye contact between two humans in face-to-face communication. Over there, the images or videos are usually captured by wearable imaging equipment, such as eye-tracking glasses [2,12,13], or in the controlled image capturing setup [14]. Under this circumstance, the focus of the individual captured by the camera would be dependent on his/her interaction with the observer wearing the imaging equipment. In contrast, in the investigated onfocus detection task, we study the individual-camera eye contact. Here, the images are captured by normal cameras instead of the wearable imaging equipment. Thus, the focuses of the individuals are mainly dependent on the mental activities or physical status of themselves rather than the interaction with the observer. Therefore, the image contents presented in the study of the individual-camera eye contact as well as the messages delivered through the individual-camera eye contact would be different from those in the common human-human eye contact. To this end, we treat the detection of individual-camera eye contact as a novel task and name it as onfocus detection. Furthermore, eye contact detection (ECD) is studied in indoor environment with carefully installed equipment. The images are captured with clear background and have equal distance to the camera, thus having the constrained imaging conditions. In contrast, our task is studied in open environments with arbitrary camera locations, thus being 'in the wild'. From this perspective, our task is beyond the exploration of ECD.

Gaze estimation. Gaze estimation [15–18] is another related topic to this work, which aims at inferring the continuous gaze directions from images or videos. Gaze estimation usually applies in determining the user's point of gaze on a display surface, such as the tablets and laptop screens, to facilitate human-machine interaction. Gaze estimation methods with this purpose can be used to provide helpful prior knowledge for eye contact detection and onfocus detection tasks and the methodologies proposed for gaze estimation can be used to identify potential eye contacts by moderate adjustments. But based on our investigation, in the studied individual-camera interaction scenario, whether the individual is onfocus or

2) https://github.com/wintercho/focus.

not plays a major role in the delivery of social communication signal, whereas the specific gaze direction in out of focus cases is usually without a clear subjective intention or caused by external disturbance. To this end, onfocus detection is a new research branch on gaze estimation. Notice that onfocus detection has its own value in practical applications. For example, when judging whether a prisoner is lying or not, one important sign is to see if the prisoner dares to look at the police in the eye. Under this circumstance, predicting eye gaze on other locations is not that important. What is more, the gaze estimation systems would spend more execution time as their algorithms are usually more complicated. Besides, as the gaze estimation methods are not designed specifically on judging onfocusness, they would have limited detection accuracy as demonstrated in Section 5. It is worth mentioning that using the annotation of this task to assist the learning of fine gaze estimation would be another potential research direction to benefit this community. A new trend in the field of gaze estimation is to predict gaze directions for multiple persons having group activities and interactions [19,20]. Gaze360 [19] uses the camera as a third perspective to observe individual interactions between groups. In contrast, we place cameras anywhere as a first perspective in the scene to study the individual-camera eye contact.

Relevant datasets. There are some existing datasets established for the study of gaze estimation or eye contact detection, which are relevant to our OFDIW dataset. Among the existing datasets, Columbia Gaze [14], Eyediap [21], and Gaze360 [19] are three widely used datasets for gaze estimation. However, these datasets are captured using controlled recording setups, not in real-world scenarios. The images in the MMDB dataset [22], MPIIGaze dataset [23], and GazeCapture dataset [24] are captured by eyeglasses, laptop cameras, and smartphones, respectively, in the context of face-to-face communication or human-machine(screen) interaction. Although these data are relevant to our task, they cannot be directly used to solve our problem. Taking Gaze360 [19] for example, it has 169935 images but only less than 200 images (0.1%) are onfocus while others are out of focus. Training on such data would be likely to bias the model towards a trivial solution for our task. This also explains why the state-of-the-art gaze models trained on the existing datasets cannot obtain good performance on our task. And that well justifies why spending efforts to establish a novel dataset for our task is necessary. Thus, these data are not suitable to be used in the study of onfocus detection. Compared to these datasets, OFDIW contains images that are captured in unconstrained real-world scenarios toward the study of individual-camera eye contact. To this end, our collected OFDIW cannot be replaced by any existing dataset to study the onfocus detection problem.

3 The OFDIW dataset

3.1 Image collection, split, and ground-truth annotation

Our OFDIW dataset consists of two subsets: unconstrained human face subset and pet face subset, respectively. The images in the human face subset (OFDIW-HF) are acquired from the famous LFW dataset [6], which is widely used for face recognition-related tasks. While the images in the pet face subset (OFDIW-PF) are acquired from the Oxford-IIIT Pet dataset [7], which mainly contain faces of dog and cat. OFDIW totally contains 20623 images, which are sufficient to implement the current deep learning-based frameworks for the onfocus detection task. Each of OFDIW-HF and OFDIW-PF is further split into the training set and the test set. The training sets of OFDIW-HF and OFDIW-PF contain 9924 and 5542 images, respectively. The test sets of OFDIW-HF and OFDIW-PF contain 3309 and 1848 images, respectively. Based on the data split, 75% of images in OFDIW are used for training while 25% for testing.

In order to complete the task of onfocus detection, camera can be placed anywhere in the scene to capture first-view images where the camera is on the normal line of the image plane that passes the image center. In other words, the camera is located anywhere on the subject's face, so the subject can be located anywhere in our task (see examples in Figure 3). Notice that the investigated problem mainly deals with one subject scenario, including one-to-one or one-to-multiple communications like staff interview, president speech, prisoner interrogation, etc. Thus, in our dataset, there is only one subject appearing in the image. During the annotation, we do not use the physical location to determine whether the object is onfocus or not. Instead, we asked the annotator to use a subjective way to feel whether the subject was looking at them. For each image, there are three human annotators to label whether the individual (either the human or the pet) in the image focuses on the camera or not and obtain the final

Figure 3 (Color online) Illustrate the location of the subject.

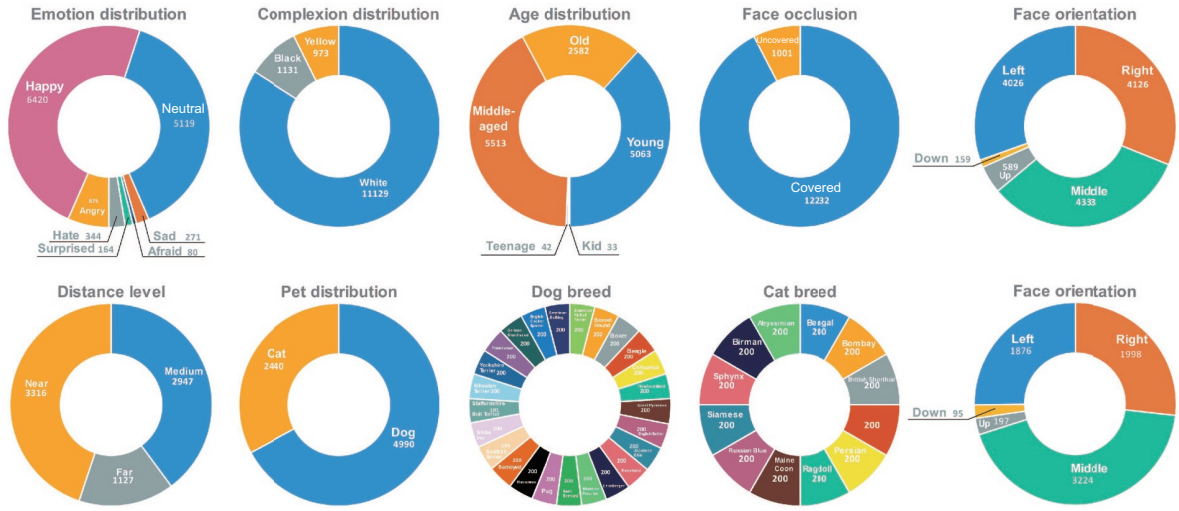

Figure 4 (Color online) Diversity of the collected OFDIW dataset. Please zoom in for details.

annotation by processing the majority voting on the three human annotators. The annotation processes are performed independently on different human annotators and different images.

3.2 Diversity and difficulty

To reveal the diversity of the OFDIW dataset, we also conduct statistics from different aspects. Specifically, for OFDIW-HF, we conduct statistics on different age levels of the individuals presented in the images, the complexity levels of the image scenes, the occlusion of the human faces, the emotion distribution of the individuals, the gender distribution of the individuals, and the face orientation distribution of the individual. For OFDIW-PF, we conduct statistics on the distance distribution between the face of the pet to the camera, the face directions of the pets, and the breed distributions of cat and dog, respectively.

From the statistics results shown in Figure 4, we can observe that OFDIW-HF exhibits a natural long-tailed distribution on emotion, age, and face orientation. In terms of emotion, most of the individuals present happy and neutral and there are also small groups of individuals who present angry, sad, surprise, etc. With respect to age, most of the individuals are middle-aged and young and there are also small groups of individuals who are kids and teenagers. In terms of face orientation, most of the individuals face left, right, or middle, while a few individuals face up and down. Among these images, most of the individuals (about 85%) are white while the rests are black and yellow. In addition, face occlusion occurs in more than 90% images. In OFDIW-PF, about 33% are cat images while the rests are dog images. It contains 25 dog breeds and 12 cat breeds in total, which are uniformly distributed. Among these images, most of the individuals (about 85%) are close to or have the medium distance to the camera, while a few individuals are far from the camera. The face orientation distribution of the pets has the similar long-tailed property as that of the humans.

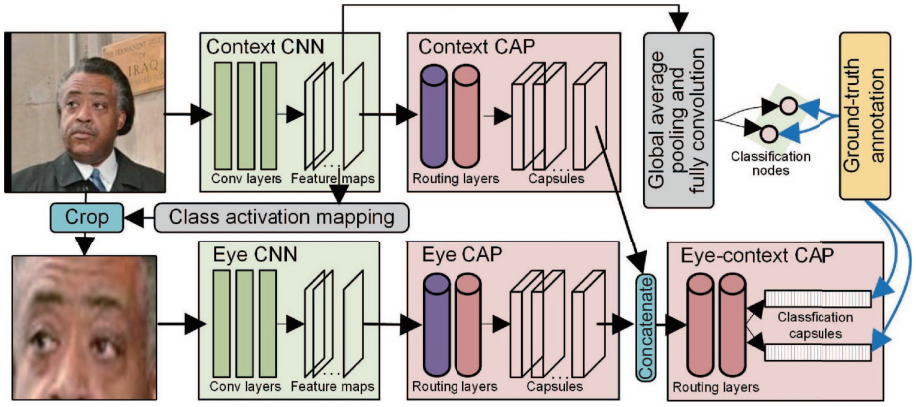

Figure 5 (Color online) Framework of the ECIIN for onfocus detection. The purple blocks indicate the primary capsule layers while the pink blocks indicate the conv-capsule layers. Blue arrows indicate the supervision information. The inputs of the two network streams are with the spatial size of 224 × 224. The whole network is trained in an end-to-end fashion.

4 Eye-context interaction inferring network

Given each input image, the proposed ECIIN model (see Figure 5) detects onfocus phenomenon by automatically mining the eye regions from the image and then modeling the interaction between the eye region and the context region. Different from the previous eye contact detection studies [10,25,26], ECIIN does not require face detection, facial landmarks tracking, or face pose estimation results to make the final decision. Instead, we introduce several carefully designed network modules into ECIIN to learn to explore such information cues implicitly by casting the whole learning problem as an image categorization task.

4.1 Network design

ECIIN first enables two network streams to separately extract information from the context region and the eye region, and then these two streams are fused into one stream to explore the eye-context interaction. For each of the eye region stream and the context region stream, we use two network blocks to extract useful information, i.e., the CNN block and the CAP block. Based on our intuition, the CNN block is first used to extract high-dimensional region feature to replace the original RGB pixel feature. CAP block is used to explore the spatial-hierarchical information upon the regions to obtain the final prediction. In the CNN block, conventional convolution layers are adopted to extract feature maps for the input image region. Here we follow the widely-used VGG network architecture [27] to implement our CNN block by using its conv1–conv5 layers. Then, considering that capsule network is good at exploring the part-whole relation, we formulate the eye-context relation (a critical issue for onfocus detection) as an abstract part-whole relation in our model, where a CAP block is used to construct primary capsules based on the features extracted by the CNN block and then conduct the conv-capsule mapping to model the part-whole relationship within the input image region (either the context region or the eye region).

4.2 CAP modeling

Given feature maps $\boldsymbol{F} \in \mathbb{R}^{(W\times H\times 512)}$ obtained from the CNN block, we construct the 4×4 pose matrices of the primary capsules by conducting n_p 1×1 convolutions on $\boldsymbol{F}$ and the 1-dimensional activations of the primary capsules by applying the sigmoid function on $\boldsymbol{F}$. This constructs $W\times H\times \frac{n_p}{4\times 4}$ primary capsules[3)] where each of them has a 4×4 pose matrix and a 1-dimensional activation. Then, in each of the convolutional capsule layers, we set the kernel size as k and the new capsule type as n_t, indicating that there will be n_t capsules in each location of the next layer and each of these capsules will be obtained by aggregating capsules from the current layer of its $k\times k$ surrounding regions. The aggregation is performed by a matrix transformation process and a dynamic routing process. Let $\{\boldsymbol{M}_i\}$ denote the pose matrices of the current capsule layer, while let $\{\boldsymbol{V}_{ij}\}$ denote the vote matrices of the next capsule layer. The matrix transformation process is performed by learning the transformation matrices $\{\boldsymbol{W}_{ij}\}$ such as $\boldsymbol{V}_{ij}=\boldsymbol{M}_i\boldsymbol{W}_{ij}$. The dynamic routing process is performed by the EM-based routing-by-agreement

3) Here we can interpret it as $\frac{n_p}{4\times 4}$ types of capsules while each type contains $W\times H$ entities at the corresponding positions.

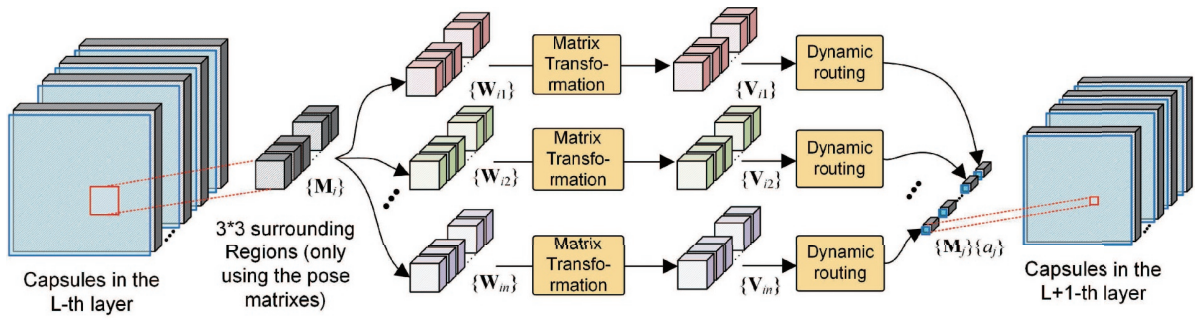

Figure 6 (Color online) Capsule convolution operation between two capsule convolutional layer, where $\{\boldsymbol{W}_{ij}\}$ and $\{\boldsymbol{V}_{ij}\}$ indicate the transformation matrices and the vote matrices, respectively. $\{\boldsymbol{M}_i\}$ and $\{\boldsymbol{M}_j\}$ indicate the capsules from the current layer and the next layer, respectively. The blue channel in the capsules indicates the activation while the gray channels indicate the pose matrix.

strategy [8] to obtain the capsules which consist of both the pose matrices $\{\boldsymbol{M}_j\}$ and the activations $\{a_j\}$ in the next layer. More specifically, for each dimension of the vectorized pose matrix $\boldsymbol{M}_j$, we formulate it as the expectation of a Gaussian distribution on the corresponding dimension of $\{\boldsymbol{V}_{ij}\}$. Let $p^h_{i|j}$ (h is the dimension index of the vectorized pose matrix) denote the probability of $\{v^h_{ij}\}$ belonging to the Gaussian distribution of the j-th capsule, and we calculate the activation of the j-th capsule using the minimum description length principle [8]:

$$a_j = \text{logistic}\left(\lambda\left(\beta_a - \beta_u \sum_i r_{ij} - \sum_h \text{cost}^h_j\right)\right), \quad \text{cost}^h_j = -\sum_i r_{ij} \ln p^h_{i|j}, \tag{1}$$

where β_a and β_u are two learnable parameters in the capsule routing process, cost^h_j indicates the cost for using the capsules in the current layer to active the j-th capsule in the next layer. r_{ij} is the weight for assigning the i-th capsule in the current layer to the j-th capsule in the next layer, which is initialized uniformly and then gradually updated in the routing process. λ is an inverse temperature parameter which increases at each iteration. An illustration of the above-mentioned convolutional capsule mapping operation is shown in Figure 6.

As for the network layers in the eye-context CAP block, similar operations are adopted. Specifically, it takes the concatenation of the context capsules and eye capsules as the input. Then, it has two routing layers. The first one implements the conv-capsule mapping to explore the eye-context relationship among the context capsules and eye capsules. While the second one infers the two classification capsules for final prediction.

4.3 Learning objective function

To train ECIIN, we follow [8] to adopt the spread loss by maximizing the gap between the activation of the target class and the activation of the other classes:

$$L = \sum_{p \neq t} \max(0, m - (a_t - a_p))^2, \tag{2}$$

where a_t and a_p indicate the activation of the target class and one another class, respectively. m is the measure of the margin between a certain wrong class and the target class, which is set to 0.2 at the beginning of the training process and then linearly increases to 0.9 during the subsequent learning process. To mine the eye region from each input image automatically, we connect the context CNN block with a global average pooling layer [28] followed by a fully connected layer to predict the two-dimensional classification vector (see Figure 5). Then, training this network branch through the cross-entropy loss would enable the network to mine discriminant image regions (through the class activation mapping technique [28]) which contribute a lot to the classification. From Figure 7 we can observe that the mined image regions are mainly the eye regions, which conforms to the intuition that the eye region would play a very important role in the onfocus detection task.

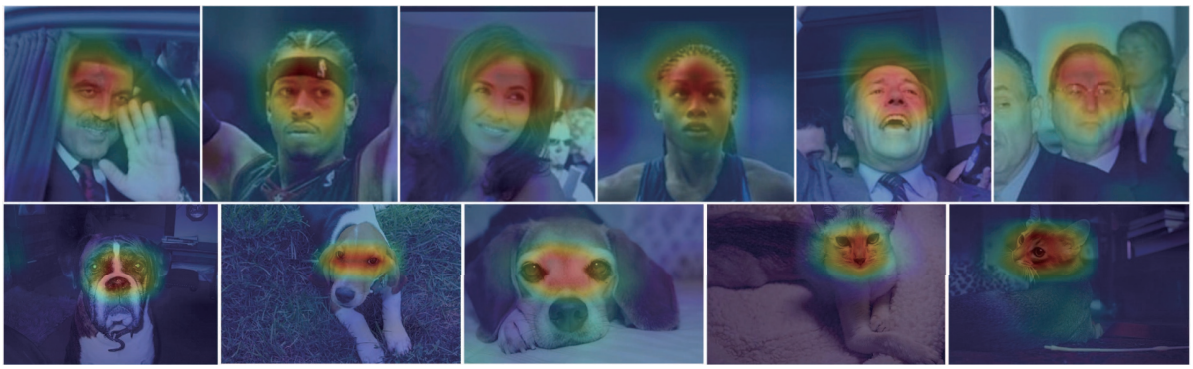

Figure 7 (Color online) Some examples of the eye regions mined by our approach, where regions with red masks are predicted with higher probability for being the eye regions.

5 Experiments

The experiments are implemented on both subsets of the built OFDIW dataset. On OFDIW-HF, we set $n_p = 512$, $n_t = 32$, kernel size k as 3. The strides of the context CAP block and eye CAP block are set to 2 while the stride of the eye-context CAP block is set to 1. On OFDIW-PF, considering that there are less data for training, we reduce the model parameter to avoid overfitting. Specifically, we add a max-pooling layer after the context CNN block and eye CNN block, and set $n_p = 256$, n_t as 8 for the context CAP block and eye CAP block while 4 for the eye-context CAP block. The stride is set to 1 for all the three CAP blocks. We adopt the widely used classification accuracy and F-measure in evaluation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \qquad \text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

where TP, TN, FP, FN indicate the numbers of true positive, true negative, false positive, and false negative samples, respectively. Precision and Recall are calculated by $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, and $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$.

5.1 Comparison to state-of-the-art methods

In this subsection, we first compare our approach with several existing eye contact detection and gaze estimation methods [2, 11, 15, 25, 29]. When compared with the gaze estimation methods, we simply modify the prediction head of their models for two-class classification. It is also worth mentioning that most eye contact detection and gaze estimation methods use the pre-trained facial landmark detector to locate the eye and other facial components from each input image firstly and then train the learning models. In contrast, our proposed approach directly performs on the raw input image without requiring the pre-trained facial landmark detector. The experimental results on two subsets of OFDIW are reported on the top blocks of Table 1 [2, 11, 15, 25, 27, 29–38]. From the comparison results on OFDIW-HF, it can be observed that our method obtains 4.21%–10.96% relative performance gains under the measurement of detection accuracy, and 2.11%–6.26% relative performance gains under the measurement of F-measure. While such improvements become stronger on OFDIW-PF. Based on our understanding, the reason for the failure of the eye contact detection and gaze estimation methods on OFDIW-PF might be that these methods cannot obtain precise locations of the eye and other facial components of pets. This demonstrates that our approach can work flexibly to detect individual-camera eye contact for both human and pets.

We also compared our approach with the state-of-the-art image categorization methods [27,35–38] and fine-grained image classification methods [30–34]. The corresponding comparison results are reported in the middle and bottom blocks of Table 1, respectively. Notice that both our model and the compared models use the parameters pre-trained on the ImageNet dataset [39]. From the comparison results, we can observe that directly using these state-of-the-art network architectures cannot obtain better performance than our proposed ECIIN model. Take the results on OFDIW-HF as an example. When comparing to the state-of-the-art image categorization networks and fine-grained image classification networks, our approach obtains 1.26%–4.55% relative performance gains under the detection accuracy. It worth mentioning that during the comparison of classification models, such improvements are not marginal according to the current literature, e.g., [40, 41], a new classification model usually exceeds the

Table 1 Comparison to the state-of-the-art methods, including the eye contact detection methods, gaze estimation methods, image categorization networks, and fine-grained image classification networks

Method	OFDIW-HF		OFDIW-PF	
	Accuracy	F-measure	Accuracy	F-measure
DEEPEC [11]	0.7939	0.8639	0.5411	0.6323
Gaze-lock detector [25]	0.8129	0.8821	0.5801	0.6450
PiCNN [2]	0.7996	0.8747	0.5584	0.5996
Multimodal CNN [15]	0.7634	0.8476	0.5487	0.6329
CA-Net [29]	0.8117	0.8809	0.6272	0.6518
NTSnet [30]	0.8190	0.8834	0.7030	0.7183
MPN-COV [31]	0.8362	0.8926	0.7192	0.7435
DFL [32]	0.8102	0.8814	0.7495	0.7504
BCNN [33]	0.8320	0.8919	0.7608	0.7600
DCL [34]	0.8214	0.8867	0.7576	0.7679
VGG16 [27]	0.8262	0.8897	0.7516	0.7536
Resnet50 [35]	0.8229	0.8837	0.7446	0.7495
Res2net50 [36]	0.8223	0.8861	0.7505	0.7572
Densenet121 [37]	0.8290	0.8904	0.7608	0.7676
Senet154 [38]	0.8193	0.8832	0.7673	0.7746
Ours	**0.8471**	**0.9007**	**0.7744**	**0.7747**

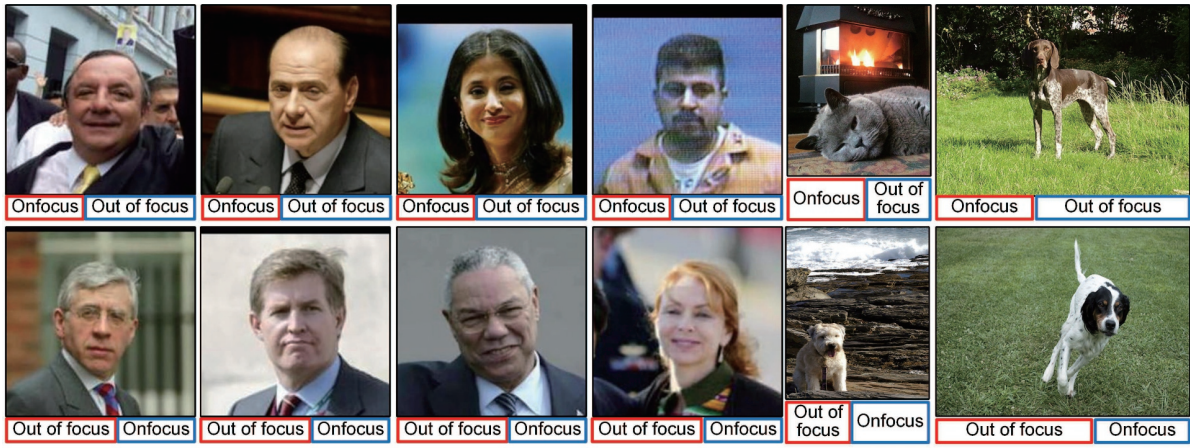

Figure 8 (Color online) Some examples to illustrate the failure cases of the proposed approach, where the ground-truth labels are shown in the red boxes while the prediction results are shown in the blue boxes.

existing models around 1% in accuracy. To our best knowledge, the advantage of our approach upon the existing image categorization models is that the proposed framework contains a two-stream architecture to model the eye-context interaction. While compared to the fine-grained classification models, the advantage of the proposed approach is the capacity to facilitate the part-whole relationship modeling for onfocus detection.

For model complexity, when compared to the gaze models, the complexity of our model is lower as gaze models usually need to integrate multiple deep models to accomplish the task. When compared to the compared classification networks, our model has higher complexity owing to the two-stream architecture and the capsule routing layers.

To analyze the failure cases of the proposed approach, we show examples in Figure 8. As can be seen, the failure cases mainly occur when one or two eyeballs of the subject (either human or pet) are hard to observe from the input image. Under this circumstance, the extracted feature cannot provide sufficient information to make precise predictions.

5.2 Ablation study

In the ablation study, we first study the key components of the proposed ECIIN model. Specifically, we compare the proposed approach with three baseline models. The first one is the 2S CNN model, which

Table 2 Ablation study of our ECIIN model on OFDIW-HF[a)]

Method	Accuracy	F-measure
2S CNN	0.8244	0.8890
2S CNN + Context CAP	0.8302	0.8881
2S CNN + 2S CAP	0.8401	0.8964
2S CNN + 2S CAP + EC CAP	0.8471	0.9007

a) 2S is short for two-stream; EC CAP is short for eye-context capsule block.

Table 3 Study on different strategies to generate the input eye regions of the eye CNN block

Strategy	Accuracy	F-measure
Weighted multiplying strategy (WM)	0.8395	0.8961
Concatenation strategy (CO)	0.8419	0.8970
Binary multiplying strategy (BM)	0.8465	0.8990
Cropping strategy (CR)	0.8471	0.9007

uses the two-stream CNN architecture with the decision-level fusion. In other words, the final decision of this baseline is obtained by averaging the predictions of the two CNN streams rather than applying a fusion block to integrate the features learned from the two CNN streams to generate the final prediction. The 2S CNN + Context CAP baseline adds the context CAP on top of 2S CNN, while the 2S CNN + 2S CAP baseline adds the eye CAP on top of 2S CNN + Context CAP. Similar to 2S CNN, these two baselines also adopt the decision-level fusion strategy to obtain the final prediction. Finally, our approach is implemented by further adding the eye-context CAP on top of 2S CNN + 2S CAP. From the experimental results reported in Table 2, we can observe that using the capsule blocks cannot only help better explore the information within the context image regions or the eye regions but also improve the final prediction by exploring the interaction between the eye and context regions.

In addition, we also study several strategies to generate the input eye regions of the eye CNN block. The studied strategies include the weighted multiplying strategy (WM), the concatenation strategy (CO), the binary multiplying strategy (BM), and the cropping strategy (CR). Specifically, the WM strategy directly implements the element-wise product between the original image and the class activation map (CAM) obtained from the context CNN to generate eye CNN's input. The CO strategy concatenates the original image and the class activation map as the input of the eye CNN. The BM strategy first binarizes the obtained CAM with the threshold of 0.8 and then implements the element-wise product between the original image and the binary CAM to generate eye CNN's input. The CR strategy uses the same threshold to binarize the CAM and then crops the rectangle image region from the original image according to the highlight locations on CAM. The experimental results reported in Table 3 demonstrate that the CR strategy works better than other strategies in terms of both accuracy and F-measure. To our knowledge, this might be due to the CR strategy can make the eye CNN and eye CAP better focus on the fine-level features around the eye regions.

6 Conclusion

This work studies a new task called onfocus detection, which aims at identifying whether the focus of the individual (human or animal) captured by a camera in unconstrained imaging conditions is on the camera or not. To solve this problem, we first build a large-scale onfocus detection dataset, named OFDIW, to facilitate the study in this field. The OFDIW dataset contains individuals with diverse emotions, ages, facial characteristics, and rich interactions with surrounding objects and background scenes. It consists of two subsets and totally 20623 unconstrained images. Then, we reveal the main challenges in the onfocus detection task and establish a novel end-to-end deep model, called ECIIN, to explore the informative eye-context interaction cues for onfocus detection. Comprehensive experiments on the OFDIW dataset demonstrate that the proposed model can obtain superior performance when compared to the state-of-the-art gaze estimation, eye contact detection, image categorization and fine-grained classification methods. Besides, comprehensive ablation studies are also conducted to verify the effectiveness of each component considered in our approach.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61876140,

61773301), Fundamental Research Funds for the Central Universities (Grant No. JBZ170401), and China Postdoctoral Support Scheme for Innovative Talents (Grant No. BX20180236).

References

1 Walczyk J J, Griffith D A, Yates R, et al. LIE detection by inducing cognitive load. Criminal Justice Behav, 2012, 39: 887–909
2 Chong E, Chanda K, Ye Z, et al. Detecting gaze towards eyes in natural social interactions and its use in child assessment. Proc ACM Interact Mob Wearable Ubiquitous Technol, 2017, 1: 1–20
3 Grossmann T. The eyes as windows into other minds. Perspect Psychol Sci, 2017, 12: 107–121
4 Prantl L, Heidekrueger P I, Broer P N, et al. Female eye attractiveness—Where beauty meets science. J Cranio-Maxillofacial Surgery, 2019, 47: 73–79
5 Zhao W, Zhao F, Wang D, et al. Defocus blur detection via multi-stream bottom-top-bottom network. IEEE Trans Pattern Anal Mach Intell, 2020, 42: 1884–1897
6 Huang G B, Learned-Miller E. Labeled Faces in the Wild: Updates and New Reporting Procedures. Technical Report, 14-003. 2014
7 Parkhi O M, Vedaldi A, Zisserman A, et al. Cats and dogs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 3498–3505
8 Hinton G E, Sabour S, Frosst N. Matrix capsules with EM routing. In: Proceedings of International Conference on Learning Representations, Vancouver, 2018
9 Qodseya M, Panta F, Sedes F. Visual-based eye contact detection in multi-person interactions. In: Proceedings of 2019 International Conference on Content-Based Multimedia Indexing, Dublin, 2019. 1–6
10 Li B, Zhang H, He L, et al. Eye contact detection via CNN-based gaze direction regression. In: Proceedings of 2019 WRC Symposium on Advanced Robotics and Automation, Beijing, 2019. 425–431
11 Mitsuzumi Y, Nakazawa A, Nishida T. DEEP eye contact detector: robust eye contact bid detection using convolutional neural network. In: Proceedings of British Machine Vision Conference, London, 2017
12 Ye Z, Li Y, Fathi A, et al. Detecting eye contact using wearable eye-tracking glasses. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, 2012. 699–704
13 Zhang X, Sugano Y, Bulling A. Everyday eye contact detection using unsupervised gaze target discovery. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, Québec City, 2017. 193–203
14 Smith B A, Yin Q, Feiner S K, et al. Gaze locking: passive eye contact detection for human-object interaction. In: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, St. Andrews Scotland, 2013. 271–280
15 Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 4511–4520
16 Huang Q, Veeraraghavan A, Sabharwal A. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision Appl, 2017, 28: 445–461
17 Park S, Spurr A, Hilliges O. Deep pictorial gaze estimation. In: Proceedings of the European Conference on Computer Vision, Munich, 2018. 721–738
18 Lian D, Hu L, Luo W, et al. Multiview multitask gaze estimation with deep convolutional neural networks. IEEE Trans Neural Netw Learn Syst, 2019, 30: 3010–3023
19 Kellnhofer P, Recasens A, Stent S, et al. Gaze360: physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, Long Beach, 2019. 6912–6921
20 Müller P, Huang M X, Zhang X, et al. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw Poland, 2018. 1–10
21 Mora K A F, Monay F, Odobez J M. Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras. In: Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, 2014. 255–258
22 Rehg J, Abowd G, Rozga A, et al. Decoding children's social behavior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland Oregon, 2013. 3414–3421
23 Zhang X, Sugano Y, Fritz M, et al. It's written all over your face: full-face appearance-based gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu Hawaii, 2017. 51–60
24 Krafka K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas Nevada, 2016. 2176–2184
25 Parekh V, Subramanian R, Jawahar C V. Eye contact detection via deep neural networks. In: Proceedings of International Conference on Human-Computer Interaction, Vancouver, 2017. 366–374
26 Ye Z, Li Y, Liu Y, et al. Detecting bids for eye contact using a wearable camera. In: Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Ljubljana, 2015. 1–8
27 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
28 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas Nevada, 2016. 2921–2929
29 Cheng Y, Huang S, Wang F, et al. A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, New York, 2020. 10623–10630
30 Yang Z, Luo T, Wang D, et al. Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision, Munich, 2018. 420–435
31 Li P, Xie J, Wang Q, et al. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 947–955

32 Wang Y, Morariu V I, Davis L S. Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 4148–4157
33 Lin T Y, RoyChowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 1309–1322
34 Chen Y, Bai Y, Zhang W, et al. Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 5157–5166
35 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
36 Gao S H, Cheng M M, Zhao K, et al. Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 652–662
37 Huang G, Liu Z, Pleiss G, et al. Convolutional networks with dense connectivity. IEEE Trans Pattern Anal Mach Intell, 2019. doi: 10.1109/TPAMI.2019.2918284
38 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 7132–7141
39 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115: 211–252
40 Wang G, Wang K, Lin L. Adaptively connected neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 1781–1790
41 Tokozume Y, Ushiku Y, Harada T. Between-class learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 5486–5494