ArbGaze: Gaze Estimation from Arbitrary-Sized Low-Resolution Images

Hee Gyon Kim and Ju Yong Chang

Abstract: The goal of gaze estimation is to estimate a gaze vector from an image containing a face or eye(s). Most existing studies use pre-defined fixed-resolution images to estimate the gaze vector. However, images captured from in-the-wild environments may have various resolutions, and variation in resolution can degrade gaze estimation performance. To address this problem, a gaze estimation method from arbitrary-sized low-resolution images is proposed. The basic idea of the proposed method is to combine knowledge distillation and feature adaptation. Knowledge distillation helps the gaze estimator for arbitrary-sized images generate a feature map similar to that from a high-resolution image. Feature adaptation makes creating a feature map adaptive to various resolutions of an input image possible by using a low-resolution image and its scale information together. It is shown that combining these two ideas improves gaze estimation performance substantially in the ablation study. It is also demonstrated that the proposed method can be generalized to other popularly used gaze estimation models through experiments using various backbones.

Keywords: gaze estimation; knowledge distillation; feature adaptation; deep neural network

1. Introduction

The human eye gaze tells what a person’s interests are, and it can be used as a medium for non-verbal communication. As a means of communication, the gaze can be used in many areas, such as human–computer interaction [1], human–robot interaction [2], virtual reality [3], augmented reality [4], and autonomous driving [5]. The goal of the gaze estimation task in computer vision is to estimate the gaze information of a subject from an input image that includes the subject’s face or eye(s). The gaze information can be expressed as a target display’s two-dimensional (2D) pixel coordinates to which the subject’s gaze is fixated or as a three-dimensional (3D) gaze vector in the camera coordinate system. The target display varies depending on the environment in which the image is acquired, such as a mobile phone [6], a laptop [7], and a tablet [8].

Existing gaze estimation methods estimate gaze information from images acquired in a controlled environment or from synthetic images [9–16]. Many studies assume that the resolution of the face or eye(s) extracted from the input image is similar. In addition, high-resolution facial images obtained in environments where the distance between the camera and the subject is relatively close (e.g., mobile phones, tablets, and laptops) are available for gaze estimation [6,7,17,18]. However, in an in-the-wild environment where the distance between the camera and the subject is not constant or in an environment where the distance between the subject and the target display is relatively far (e.g., TV), low-resolution images of various sizes can be acquired. These low-resolution input images degrade the performance of the gaze estimation method. Our study addresses this problem in gaze estimation from arbitrary-sized low-resolution images, as shown in Figure 1a. It should be noted that the aspect ratio of the input image (i.e., the width ratio to the height of the image) is assumed to be constant in our study. Addressing the variation of the input image’s aspect ratio is beyond our study’s scope.
Figure 1. The problem to be addressed in this study (i.e., gaze estimation from arbitrary-sized low-resolution images) is shown in (a). Low-resolution images of various sizes can degrade the gaze estimation performance of the existing gaze estimation model. The following approaches can be used to solve this problem: (b) a method based on an arbitrary SR model, (c) a method based on several different gaze estimation models corresponding to various resolutions, and (d) the proposed method that is based on a single model.

To address this problem, super-resolution (SR) methods with an arbitrary scale factor [19–21] can be utilized. The basic idea is to reconstruct a high-resolution image from a low-resolution image and feed the resulting image into the gaze estimation model, as shown in Figure 1b. However, the SR method requires much memory and time for training, and noise or artifacts in the reconstructed image can disturb correct gaze estimation. Figure 1c shows another approach in which different gaze estimation models are trained according to each resolution to cover input images of various resolutions. However, this method requires a large number of models corresponding to multiple resolutions.

To efficiently perform gaze estimation from arbitrary-sized low-resolution images, we propose a method (i.e., ArbGaze) that utilizes knowledge distillation [22] and feature adaptation [20] together. In knowledge distillation, the knowledge of a larger network (i.e., teacher) is generally transferred to a small network (i.e., student). On the contrary, the teacher and student have a network structure of almost the same complexity in our study. The difference between the two networks is in their input images: the teacher and student use high-resolution images and arbitrary-sized low-resolution images for gaze estimation, respectively. Therefore, the teacher can learn privileged knowledge that the student cannot obtain from low-resolution images. It makes the intermediate features of the teacher contain more abundant information for gaze estimation than the student. Our basic idea is to distill the teacher’s knowledge into the student, allowing the student to generate features similar to the rich features of the teacher.

However, models with fixed weights do not have sufficient complexity to adapt to input images of arbitrary resolution. Therefore, the proposed student is constructed using the feature adaptation module based on conditional convolution (i.e., CondConv) [23] that has a larger capacity than the general convolution operation. A scale factor can be defined between high-resolution images for the teacher and arbitrary-sized low-resolution images for the student. This scale factor indicates the ratio of the resolution of the high-resolution image and that of the low-resolution image. The larger the scale factor, the greater the difference in size between the two images. The feature adaptation module outputs a new scale-adapted feature map from a feature map obtained from an arbitrary-sized low-resolution image using this scale information. Intuitively, knowledge distillation supervises the student on what features to create, and feature adaptation supervises the student on how to construct those features.
In summary, the research question to be answered in this study is whether efficient gaze estimation is possible from an arbitrary-sized low-resolution image. Our hypothesis is that knowledge distillation and feature adaptation can be utilized together to construct an efficient and effective gaze estimation model. The contributions of this study are as follows.

- A novel method is proposed to solve the problem of gaze estimation from low-resolution input images of arbitrary size, which has been rarely addressed previously.
- The proposed method significantly improves the gaze estimation performance for low-resolution images of various sizes by combining knowledge distillation and feature adaptation.

2. Related Works

This section introduces existing gaze estimation studies and problems that may arise in gaze estimation. Knowledge distillation and feature adaptation are utilized in the proposed method for gaze estimation from arbitrary-sized low-resolution input images. Therefore, knowledge distillation and feature adaptation for various scales are briefly reviewed.

2.1. Gaze Estimation

Gaze estimation methods can be divided into model-based and appearance-based approaches [6,24–27]. In the model-based approach, the physical structure of the eye, such as corneal reflex, iris, or pupil movement, is geometrically modeled for gaze estimation [28–30]. These methods typically rely on special equipment, such as infrared or RGB-D cameras. On the contrary, the appearance-based approach directly uses input images. It can estimate the gaze information from data acquired through a general camera, such as a webcam [6–8,10–18,31,32]. Owing to this advantage, the appearance-based approach has dominated gaze estimation studies recently. Such appearance-based methods are now briefly reviewed.

Early gaze estimation methods estimated gaze through various machine learning techniques, such as random forest [9,10], adaptive linear regression [33], and support vector regression [6]. Recently, gaze estimation studies using large-scale datasets [6–12,14,17,34,35] and deep learning have been actively conducted. Deep learning-based gaze estimation methods can be mainly divided into 2D gaze estimation studies [6,8,36–38] and 3D gaze estimation studies [7,10–15,18,31]. The goal of 2D gaze estimation is to estimate the 2D coordinates of a pixel on which the gaze is directed on a specific display, such as a cell phone or laptop. In the 3D gaze estimation method, the gaze vector in the camera coordinate system is usually estimated. The gaze vector can be a 3D direction vector [10,11,31] or an angle vector composed of pitch and yaw [12–18].

In gaze estimation, various issues may arise in addition to the problem we focus on in this paper. Anatomical differences between people can degrade gaze estimation performance. To solve this inter-person problem, ref. [39] proposed a person-specific few-shot learning method using a few calibration samples. In this method, a personalized gaze estimator based on a rotation-aware latent representation of gaze is trained using meta-learning. Ref. [40] addressed the issue of gaze estimation datasets. Building a new gaze estimation dataset is expensive, and the existing datasets are annotated under different environments. Therefore, in [40], an unsupervised representation learning method that does not use ground-truth information was presented. It uses gaze redirection, which generates an output image with the same eye as the input image but with redirected gaze, as an auxiliary task for unsupervised learning. Ref. [41] addressed the difficulty of acquiring 3D gaze information in an in-the-wild environment. One interesting idea to solve this is to utilize the geometric constraint of people looking at each other in the video. In [41], looking-at-each-other labels for in-the-wild images were obtained based on such an idea, and a weakly-supervised training method for gaze estimation based on the labels was proposed. Ref. [42] addressed the domain adaptation problem in gaze estimation. To generalize the gaze estimator to a new domain, ref. [42] proposed a plug-and-play gaze adaptation framework, where outliers of the outputs of pre-trained models in the target domain are considered noisy labels, helping the collaborative learning process.
The issues above are important but not related to the motivation of our study, which is presented as follows. Most appearance-based methods, including the works of literature above, assume that an input image of fixed size is used. Therefore, to use the existing methods, it is necessary to resample the low-resolution images of various sizes into images of a fixed size. Unfortunately, loss of information in the general resampling process can cause deterioration of gaze estimation performance. Our paper addresses this issue and proposes an effective gaze estimation method from an arbitrary-sized low-resolution image.

2.2. Knowledge Distillation

A straightforward way to increase the performance of deep learning models is to utilize large-scale datasets with complex networks. However, this requires substantial computation and memory usage. Knowledge distillation is proposed to allow relatively simple models (i.e., student) to achieve similar performance to complex models (i.e., teacher) [22]. As the complex teacher model has high capacity, it can learn richer knowledge from data than the simple student model. This knowledge of the teacher model is transferred to a lightweight student model via a loss function between the two models, which is trained to output intermediate features similar to the teacher model. The teacher–student framework of knowledge distillation is adopted from this perspective for the proposed gaze estimation method. A loss function is applied between the teacher model trained on high-resolution images and the student model trained on arbitrary-sized low-resolution images, causing the student model to output intermediate features similar to the teacher model. These intermediate features contain information that is difficult to learn from low-resolution images, which helps improve the gaze estimation performance of the student model.

2.3. Scale- or Resolution-Dependent Feature Adaptation

Consistently addressing low-resolution images of different sizes is a critical issue in various areas of computer vision, including gaze estimation. Ref. [19] proposed a meta upscale module that reconstructs high-resolution images from low-resolution feature maps over multiple scale factors. This module is trained to output weights for image reconstruction using the positional relationship of the corresponding pixels between the low-resolution feature map and the high-resolution image together with the scale factor. For human shape and pose estimation, ref. [43] proposed a resolution-aware network (i.e., RA-net) to handle various resolutions of input images. RA-net adapts the feature map by multiplying the feature map extracted from the input image by a resolution-aware matrix composed of parameters including resolution information. Similar to [19], ref. [20] proposed a feature adaptation block to handle low-resolution images of various sizes in SR. Based on CondConv [23], the feature adaptation block adapts the output feature map of the baseline network to the scale factor by increasing the similarity between intermediate feature maps for low-resolution input images of various sizes. Unlike previous studies addressing SR [19,20] and human shape estimation [43] tasks, feature adaptation is performed for gaze estimation from arbitrary-sized low-resolution images in our study, where the feature adaptation block of [20] is adopted.

3. Proposed Method

Our framework consists of a teacher network $N^T$ and a student network $N^S$, as shown in Figure 2. $N^T$ and $N^S$ output a 3D gaze vector from fixed-sized ($H_{HR} \times W_{HR}$) high-resolution image $I_{HR}$ and arbitrary-sized ($H_{LR} \times W_{LR}$) low-resolution image $I_{LR}$, respectively. Our framework adopts offline distillation, so $N^T$ is trained before $N^S$. $I^S_{LR}$, an input image of $N^S$, may have various resolutions. $I^S_{LR}$ is converted into an image, $I^{BC}_{LR}$, with the same resolution as the high-resolution image through bicubic interpolation before being processed in the student. Feature adaptation modules ($A_1, \ldots, A_N$) in $N^T$ produce intermediate features ($F^A_1, \ldots, F^A_N$) adaptive to arbitrary resolution using the input scale.
factor $s$. The scale factor $s$ is defined as follows under the assumption that $I^{HR}$ and $I^{LR}$ have the same aspect ratio:

$$s := \frac{H^{HR}}{H^{LR}} = \frac{W^{HR}}{W^{LR}}.$$  

(1)

In the training of the student $N^S$, the teacher $N^T$ transfers information that is difficult to learn from low-resolution images to $N^S$ through the distillation loss $L^S_{KD}$ between the intermediate features of the two networks. Using the feature adaptation module and distillation loss improves the gaze estimation performance of the student $N^S$ for arbitrary-sized low-resolution images.

Figure 2. Overview of the proposed method.

3.1. Teacher Network

The teacher network $N^T$ outputs a gaze vector $\hat{g}^T \in \mathbb{R}^3$ from a high-resolution input image $I^{HR}$. ResNet-18 [44] is used to construct $N^T$. As shown in Figure 2, $N^T$ consists of one convolutional layer, $N$ residual modules, one global average pooling layer, and two fully connected (FC) layers. As a residual module, the building block of [44] is used. The output feature map $F^T_n$ of the $n$-th residual module $R^T_n$ is used for the next layer’s input and the distillation loss. The output $F^T_N$ of the last residual module is passed through the global average pooling layer and two FC layers to produce the estimated gaze vector $\hat{g}^T$.

The loss function $L^T_{gaze}$ used to learn $N^T$ is defined as the angular error between two gaze vectors, which is as follows:

$$L^T_{gaze} = \arccos\left(\frac{g \cdot \hat{g}^T}{\|g\|_2 \|\hat{g}^T\|_2}\right),$$  

(2)

where $g \in \mathbb{R}^3$ and $\|\cdot\|_2$ denote the ground-truth gaze vector corresponding to the input image and 2-norm, respectively.

3.2. Student Network

The student network $N^S$ outputs a gaze vector $\hat{g}^S$ from an arbitrary-sized low-resolution image $I^{LR}$ and its scale factor $s$. As shown in Figure 2, $N^S$ has a structure in which a bicubic interpolation module and $N$ feature adaptation modules $(A_1, \ldots, A_N)$ are added to $N^T$, where each feature adaptation module is placed after each residual module. First, the bicubic interpolation is applied to an arbitrary-sized low-resolution image $I^{LR}$ to output an image $I^{BC}$ having the same resolution as $I^{HR}$. Then, $I^{BC}$ is fed into the convolutional layer and passed through the residual module and feature adaptation
module alternately. $F^A_n$ computed by the last feature adaptation module is fed into the global average pooling layer, and the gaze vector estimation $\hat{g}^S$ is generated through two FC layers.

The feature adaptation module produces the feature map adaptive to the scale factor of the input low-resolution image from the feature map computed by the residual module. The feature adaptation module increases the similarity between the intermediate feature maps of the teacher and the student over multiple scale factors, and, as a result, allows the student to output teacher-like feature maps over multiple scale factors [20]. The feature adaptation module is composed of a masking block and a scale-conditioned layer, as shown in Figure 3a. The scale-conditioned layer is a 2D convolutional layer based on CondConv [23]. This layer consists of the two FC layers and $K$ experts (i.e., $W_1, \ldots, W_K$), as shown in Figure 3b. The result calculated through the linear combination of $K$ routing weights (i.e., $\alpha_1, \ldots, \alpha_K$) produced by two FC layers and $K$ experts is used for 2D convolution as kernel weights $W$. The masking block consists of four convolutional layers followed by a sigmoid function. The feature map computed by the sigmoid function, composed of values between 0 and 1, is a kind of mask that determines the pixel position of the feature map to which the scale information generated by the scale-conditioned layer is added [20]. From the masking block’s output and the scale-conditioned layer’s output, a feature map including scale information is calculated through element-wise multiplication. Finally, it is added to the input feature map $F^S_n$ to output a scale-adapted feature $F^A_n$.

![Figure 3](image)

The loss function $L^S$ applied to the student network consists of $L^S_{\text{gaze}}$, which measures the angular error and distillation loss $L^S_{\text{KD}}$ as follows:

$$L^S = L^S_{\text{gaze}} + \lambda L^S_{\text{KD}},$$

where $\lambda$ is a hyperparameter that controls the relative strength between the two loss functions. $L^S_{\text{gaze}}$ is defined similarly to $L^T_{\text{gaze}}$ in Equation (2), and the distillation loss $L^S_{\text{KD}}$ to make the student network generate features similar to the teacher network is defined as follows:

$$L^S_{\text{KD}} = \sum_{n=1}^{N} \text{MSE}(F^T_n, F^A_n),$$

where $\text{MSE}(\cdot, \cdot)$ denotes the mean squared error between two feature maps, and $N$ represents the number of feature adaptation modules.

3.3. Model Training and Testing

Training the proposed ArbGaze model consists of two steps: (i) training the teacher network using high-resolution images and gaze loss in Equation (2), and (ii) training the student network using arbitrary-sized low-resolution images and gaze/distillation losses in Equation (3). When training the student network, the teacher network is used to calculate the distillation loss, but the weights of the teacher network are frozen and not updated. During the test, the teacher network is discarded, and the student network is applied to
arbitrary-sized low-resolution test images to output estimated gaze vectors, which are used to evaluate the proposed method.

4. Experimental Results

4.1. Datasets

The publicly available gaze estimation datasets UTMultiview (UTMV) [10] and MPI-IGaze (MPII) [17] are used to evaluate the proposed method. The UTMV dataset provides real and synthetic data. Among them, only real data consisting of 8 head poses and 160 gaze directions for each subject are used. As a result, the dataset contains 64,000 (50 × 8 × 160) images obtained from 50 subjects. The MPII dataset is acquired via a laptop in an unconstrained environment and contains 45,000 images from 15 subjects. Both datasets provide images that have undergone the same preprocessing process, and these images are used as high-resolution images (i.e., $I_{HR}$) for training the teacher network. The preprocessing process first rotates a virtual camera to remove the roll rotation from the head pose and to point it toward the center of the eye, then sets the virtual camera at a fixed distance from the subject. A grayscale image of $H_{HR} \times W_{HR} = 36 \times 60$ resolution for one eye is generated from the virtual camera, and additional histogram equalization is applied to the generated image. Low-resolution images (i.e., $I_{LR}$) for learning the student network are acquired from high-resolution images through bicubic interpolation. Note that these artificially synthesized low-resolution images are also used for testing of the proposed method. The values obtained by sampling from 1.1 to 4.0 at intervals of 0.1 are used as the scale factor $s$, and in this case, the smallest resolution is 9 × 15. Three-fold cross-validation and leave-one-out cross-validation are performed for UTMV and MPII, respectively. 50 UTMV subjects are divided into three groups, {0–16}, {17–33}, and {34–49}, based on the subject ID to conduct the experiment.

4.2. Implementation Details

The teacher and student networks are constructed by modifying the FC layer of ResNet-18 [44]. Their hyperparameters, such as kernel size and stride of the convolutional layer, are all the same as in ResNet-18. As ResNet is a model for classifying input images into 1000 classes, the existing FC layer is removed and then two FC layers and a ReLU activation function are added for gaze estimation. The size of the two added FC layers is 128 and 3. Our framework uses four residual modules and four feature adaptation modules (i.e., $N = 4$).

The values of {kernel size, stride, padding} of the four convolutional layers in the feature adaptation module are {3, 1, 1}. Except in the last layer, the convolutional layer is followed by batch normalization and ReLU activation functions. The activation function following the last convolutional layer is a sigmoid function, as shown in Figure 3a. The number of output channels depends on the number of input channels. Assuming that the number of input channels is $C$, the number of output channels is reduced to $C/4$, $C/16$, $C/16$, 1. The size of the FC layers in the scale-conditioned layer is 64 and 3, and the size of the expert is $C \times C \times 3 \times 3$.

In the student network, the parameters of all layers, except the feature adaptation module, are initialized to be the same as the parameters of the pre-trained teacher network. Adam [45] is used with a momentum of 0.9 for training the teacher and student networks. In all experiments, the learning rate is kept constant at $5 \times 10^{-3}$, and the mini-batch size is 256. $\lambda$ in Equation (3) is experimentally set to 10.

4.3. Ablation Study

Our framework uses knowledge distillation and feature adaptation together. For the ablation study, the feature adaptation module is removed from the student network, and the network is learned using only $L_{gaze}$ without distillation loss. The trained model is used as a baseline for comparison with the proposed method. That is, Baseline represents a general backbone-based model trained using only the gaze loss $L_{gaze}^{S}$ from input images of various
sizes. Figure 4 shows the results of ablation experiments, which proves that all components used in the proposed method (i.e., feature adaptation (FA) and knowledge distillation (KD)) contribute to the performance improvement. Baseline + FA and Baseline + KD represent models in which the feature adaptation module and distillation loss $L_{KD}$ are added to the baseline, respectively. Baseline + FA + KD is the proposed method, which uses knowledge distillation and feature adaptation at the same time. HR image represents the gaze estimation result of the teacher network on the high-resolution image, which can be considered the most positive performance expectation that the proposed method can perform.

According to the experimental results, Baseline + FA achieves slightly improved results compared to Baseline, which shows that the capacity increase in the model due to the addition of the feature adaptation module helps the model adapt to input images of various sizes. However, this result corresponds to lower performance compared with Baseline + KD. Such an outcome reveals that the teacher’s information transferred through knowledge distillation is more effective for gaze estimation than the increase in the model’s capacity. Finally, Baseline + FA + KD, which is the proposed method, shows the best performance for all scale factors by using information extracted from high-resolution images and increasing the model’s capacity. This result justifies that the proposed components (FA and KD) improve the performance of our method.

The UTMV dataset is acquired in a laboratory environment with eight fixed head poses. However, the MPII dataset is acquired in an unconstrained environment, so the head pose variation is more diverse. The roll rotation of the head pose is removed through preprocessing in both datasets. Still, the processed images in the MPII dataset contain more pitch and yaw rotation variations than those in the UTMV dataset. Therefore, the MPII dataset is more complex and challenging than the UTMV dataset. Nevertheless, the performance improvement of gaze estimation by the proposed method is higher in the MPII dataset than in the UTMV dataset. In particular, when the scale factor is smaller than 2.5, the proposed method shows comparable performance with the HR image result. Therefore, performance improvement by the proposed method (Baseline + FA + KD) over Baseline is valid for input images containing various head poses.

For statistical tests, the paired t-test is performed using the mean gaze errors of the proposed method (Baseline + FA + KD) and the Baseline over 15 leave-one-person-out cross-validation folds of the MPII dataset. Figure 5 visualizes the calculated $p$-values. For all scale factors except $s = 1.1$, the $p$-value is less than 0.05. As the scale factor increases, the $p$-value decreases. This result shows that the performance gain of the proposed method is statistically significant, especially for high scale factors.
1. Comparison to Baselines

Gaze estimation for an arbitrary-sized low-resolution input image can be performed through two other methods, as shown in Figure 1b,c. The first method (i.e., Arbitrary SR baseline) applies SR to a low-resolution image of any size and uses the resultant image for gaze estimation. Among the SR methods for low-resolution images of arbitrary size, Meta-RDN [19] is used in our experiments. In the second method (i.e., Multiple gaze baseline), different gaze estimation models are trained according to various resolutions of input images. To this end, various teacher networks corresponding to various resolutions are trained. For comparison with the above two baselines, 30 scale factors (1.1, 1.2, ..., 4.0) are used; thus, a total of 30 networks are trained for the Multiple gaze baseline. In the case of the Arbitrary SR baseline, the FLOPs in Meta-RDN vary according to the scale factor, so the average value for 30 scale factors is calculated.

Table 1 shows the results of comparing our method with the above two baselines. Params and FLOPs represent the numbers of learnable parameters (in mega) and floating point operations (in giga) of the model, respectively. The last six columns represent gaze estimation errors according to integer scale factors (i.e., 2, 3, and 4) on UTMV and MPII datasets. According to Table 1, the Arbitrary SR baseline shows a gaze estimation error of approximately 0.2 degrees lower than that of our framework. However, this baseline has about three times as many parameters and about 100 times more FLOPs than our framework. The Multiple gaze baseline has the highest number of parameters but the smallest number of FLOPs. It is because there are 30 different models for 30 scale factors, and the gaze vector is estimated through just one model in the inference time. The proposed method has almost the same FLOPs as the Multiple gaze baseline but shows fewer parameters and gaze estimation errors. Through this experiment, it is demonstrated that the proposed method can perform efficient gaze estimation in terms of memory and inference time for arbitrary-sized low-resolution images.

Table 1. Quantitative comparison to other baselines for gaze estimation from arbitrary-sized low-resolution images. The best results are shown in bold, and the second best results are underlined.

| Method               | Params (M) ↓ | FLOPs (G) ↓ | UTMV (Degree) ↓ | MPII (Degree) ↓ |
|----------------------|-------------|------------|-----------------|-----------------|
|                      |             |            | s = 2 | s = 3 | s = 4 | s = 2 | s = 3 | s = 4 |
| Arbitrary SR baseline| 33.36       | 23.51      | 5.60  | 5.73  | 5.97  | 5.21  | 5.26  | 5.37  |
| Multiple gaze baseline| 337.09    | 0.20       | 5.47  | 5.90  | 6.45  | 5.32  | 5.60  | 5.82  |
| ArbGaze (ours)       | 12.08       | 0.22       | 5.61  | 5.83  | 6.28  | 5.23  | 5.32  | 5.58  |

4.5. Generalization of the Proposed Method

ResNet [44] and VGG [46] are used as feature extractors (i.e., backbones) in various gaze estimation studies [14,16,25,34,35,47,48]. In the proposed method, ResNet-18 is used as the backbone. It is important to determine whether the ideas of feature adaptation and knowledge distillation proposed in this study can be applied to other backbones.
with various complexity. Therefore, experiments are performed by changing the existing ResNet-18-based teacher and student structure to investigate the generalization performance of the proposed method to other models. The teacher network based on ResNet-18 includes four residual modules. The number of building blocks \([4,4,4,4]\) that compose each residual module is \([2,2,2,2]\). The number of these building blocks is changed to create simple and complex models compared with our original teacher network. The models used in the experiment are ResNet-10 and ResNet-34, which are composed of building blocks of \([1,1,1,1]\) and \([3,4,6,3]\), respectively. In the case of VGG, VGG-13, VGG-16, and VGG-19 with batch normalization are used. Experiments are conducted using the MPII dataset acquired in an unconstrained environment.

Figure 6 shows the experimental results for the two backbones. The results indicate that the proposed method improves the gaze estimation performance for arbitrary-sized low-resolution images regardless of the structure or complexity of the model. In the case of the ResNet backbone, increasing the model complexity improves the gaze estimation performance of the baseline and the proposed method. In the case of the VGG backbone, increasing model complexity does not result in performance improvement. VGG has relatively many parameters compared with ResNet, which causes overfitting. As a result, the simplest VGG-13-based model \((VGG13 + FA + KD)\) shows the best performance for most scales.

Figure 6. Mean gaze error on MPII dataset with different backbones: (a) ResNet and (b) VGG.

4.6. Model Compression

Knowledge distillation was originally proposed to make simple models perform similarly to complex models with high performance \([22]\). Therefore, most studies using knowledge distillation focus on reducing the model size. In our study, the distillation loss only transfers knowledge from the teacher network to the student network. In this section, the model compression aspect of knowledge distillation is further investigated. In Section 4.5, the performance of the ResNet backbone increases as the model becomes more complex. However, the performance of the VGG backbone does not increase due to overfitting, so an experiment on model compression is conducted using the ResNet backbone.

Figure 7 shows the experimental results of model compression using ResNet. ResNet-34 and ResNet-18 are used for the teacher network in Figure 7a,b, respectively. In both cases, the student network is based on ResNet-10. According to Figure 7a, ResNet34→ResNet34 outperforms ResNet10→ResNet10. This finding reveals that using a complex backbone improves gaze estimation performance. Using a lighter student network (ResNet34→ResNet10) slightly reduces the performance compared with ResNet34→ResNet34, but it still achieves better performance than ResNet10→ResNet10. Figure 7b also shows similar experimental results. Model compression is evidently effective in the proposed method.
4.7. Qualitative Results

Figure 8 shows the visualization of the gaze vector estimated through the proposed method. The top three and bottom three rows show the results for the UTMV and MPII datasets, respectively. Each column shows the results for a scale factor sampled at intervals of 0.5 from 1.0 to 4.0. A scale factor of 1.0 means that a high-resolution (36 × 60) image is used. The blue arrow indicates the ground-truth gaze vector projected on the image. The green and red arrows correspond to the results estimated through the proposed method and baseline, respectively. As shown in Figure 8, the gaze vector inferred by the proposed method points in a direction more similar to the ground-truth vector than the gaze vector from the baseline. Hence, the proposed method can perform a more accurate gaze estimation than the baseline from input images of various resolutions.

Figure 8. Gaze vector visualization. The top and bottom three rows correspond to the UTMV and MPII datasets, respectively. $s$ denotes the scale factor. The blue arrow indicates the ground-truth gaze vector. The green and red arrows indicate estimated gaze vectors by the proposed and baseline methods, respectively.
4.8. Quantitative Comparison with Gaze Estimation Methods from Fixed-Sized Images

In most existing appearance-based gaze estimation methods, it is assumed that an input eye or face image of a fixed size is given. Although the goal of our study is different from theirs, the performance of the proposed method on fixed-size input images is presented for completeness. Among the various backbones in Section 4.5, a VGG13-based model (VGG13 + FA + KD) is selected and used for gaze estimation from fixed-size input images of the MPII dataset. A quantitative comparison with the existing methods is given in Table 2. The proposed method achieves the best performance among the compared eye-image-based methods. These results show that the proposed method achieves competitive performance with other recent methods for fixed-size input images, even though our approach was learned for multiple scale factors for gaze estimation from arbitrary-sized low-resolution images.

Table 2. Quantitative comparison to other existing gaze estimation methods from fixed-sized images on MPII dataset. The best results are shown in bold, and the second best results are underlined.

| Method                  | Input     | Gaze Error (Degree) ↓ |
|-------------------------|-----------|-----------------------|
| RF [10]                 | eye image | 7.99                  |
| Mnist [17]              | eye image | 6.30                  |
| GazeNet [48]            | eye image | 5.83                  |
| ARE-Net [49]            | eye image | 5.02                  |
| Proposed (VGG13 + FA + KD) | eye image | 4.88                  |
| Full Face [7]           | face image| 4.90                  |
| RT-Gaze [14]            | face image| 4.30                  |
| FAR-Net [50]            | face image| 4.30                  |

4.9. Discussion

Our thorough experiments show that the two contributions of our study presented in Section 1 are achieved. First, as reviewed in Section 2, existing gaze estimation studies rarely address arbitrary-sized low-resolution images [6,24–27]. In our study, two baselines visualized in Figure 1 are implemented to address the problem and quantitatively compared with the proposed method in Section 4.4. Compared to the Arbitrary SR Baseline, which depends on an external SR network, and the Multiple gaze baseline, which depends on multiple gaze models, the proposed method consists of a single novel model, so it can efficiently perform gaze estimation. Second, the proposed method is based on a novel combination of knowledge distillation [22] and feature adaptation [19,20,23,43]. According to the experiments in Section 4.3, both components contribute substantially to the performance improvement of the proposed method for gaze estimation compared to the naive Baseline.

However, the method proposed in this study has the following limitations. First, as mentioned in Section 1, in the proposed method, the aspect ratio of the input image is assumed to be constant, which means that the size of the input image is not genuinely arbitrary. Since human eyes are considered to have a similar aspect ratio, the aspect ratio of the input images will not be much different. However, the proposed method requires input images of a fixed aspect ratio, which is one of the limitations of the proposed method. Next, as mentioned in Section 4.1, the MPII and UTMV datasets used for training and evaluating the proposed method consist of high-resolution images. Therefore, the low-resolution images for learning the proposed method are artificially synthesized through downsampling from the corresponding high-resolution images. These synthetic images are not guaranteed to have the same distribution as real low-resolution images. Therefore, another limitation of our study is that the proposed method is trained and evaluated on synthetic images rather than real ones.
5. Conclusions

In this study, a gaze estimation method (i.e., ArbGaze) from arbitrary-sized low-resolution images is proposed. The proposed method adopts the teacher–student framework and transfers the teacher model information obtained from high-resolution images to the student, a gaze estimation model for arbitrary-sized images, through knowledge distillation. The feature adaptation module helps the student network effectively learn the information transferred from the teacher network through CondConv parameterized with scale information. It is empirically proved that the proposed method is more efficient and effective than baseline models that can perform gaze estimation from arbitrary-sized low-resolution images. In addition, it is demonstrated that the proposed method can be easily generalized to other gaze estimation models through experiments on various backbones. Our future work is to extend the proposed approach to other computer vision tasks and apply ArbGaze to various application areas, such as human–computer interaction, human–robot interaction, virtual reality, augmented reality, and autonomous driving.

Author Contributions: Conceptualization, H.G.K. and J.Y.C.; methodology, H.G.K.; software, H.G.K.; validation, H.G.K.; formal analysis, H.G.K.; investigation, H.G.K.; resources, J.Y.C.; data curation, H.G.K.; writing—original draft preparation, H.G.K.; writing—review and editing, J.Y.C.; visualization, H.G.K.; supervision, J.Y.C.; project administration, J.Y.C.; funding acquisition, J.Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Culture Technology R&D Program by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Name: Development of a functional content platform for stress relief, Project Number: R2020060003, Contribution Rate: 95%). The present research has been conducted by the Research Grant of Kwangwoon University in 2022 (Contribution Rate: 5%).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Chen, W.X.; Cui, X.; Zheng, J.; Zhang, J.M.; Chen, S.; Yao, Y.D. Gaze Gestures and Their Applications in human-computer interaction with a head-mounted display. arXiv 2019, arXiv:1910.07428.
2. Chakraborty, P.; Ahmed, S.; Yousuf, M.A.; Azad, A.; Alyami, S.A.; Moni, M.A. A Human-Robot Interaction System Calculating Visual Focus of Human’s Attention Level. IEEE Access 2021, 9, 93409–93421. [CrossRef]
3. Liu, H.; Qin, H. Perceptual Self-Position Estimation Based on Gaze Tracking in Virtual Reality. Virtual Real. 2022, 26, 1359–4338. [CrossRef]
4. Wang, Z.; Zhao, Y.; Lu, F. Gaze-Vergence-Controlled See-Through Vision in Augmented Reality. arXiv 2022, arXiv:2207.02645.
5. Yuan, G.; Wang, Y.; Yan, H.; Fu, X. Self-calibrated driver gaze estimation via gaze pattern learning. Knowl.-Based Syst. 2022, 235, 107630. [CrossRef]
6. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; Torralba, A. Eye tracking for everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2184.
7. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. 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, HI, USA, 21–26 July 2017; pp. 51–60.
8. Huang, Q.; Veeraraghavan, A.; Sabharwal, A. TabletGaze: Datasheet and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Mach. Vis. Appl. 2017, 28, 445–461. [CrossRef]
9. Smith, B.A.; Yin, Q.; Feiner, S.K.; Nayar, S.K. 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, UK, 8–11 October 2013; pp. 271–280.
10. Sugano, Y.; Matsushita, Y.; Sato, Y. Learning-by-synthesis for appearance-based 3d gaze estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1821–1828.
11. Wood, E.; Baltrušaitis, T.; Morency, L.P.; Robinson, P.; Bulling, A. Learning an appearance-based gaze estimator from one million synthesised images. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, Charleston, SC, USA, 14–17 March 2016; pp. 131–138.
12. Zhu, W.; Deng, H. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy; 22–29 October 2017; pp. 3143–3152.
13. Wang, K.; Zhao, R.; Ji, Q. A hierarchical generative model for eye image synthesis and eye gaze estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA; 18–22 June 2018; pp. 440–448.
14. Fischer, T.; Chang, H.J.; Demiris, Y. Rt-gene: Real-time eye gaze estimation in natural environments. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany; 8–14 September 2018; pp. 334–352.
15. He, Z.; Spurr, A.; Zhang, X.; Hilliges, O. Photo-realistic monocular gaze redirection using generative adversarial networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea; 27 October–2 November 2019; pp. 6932–6941.
16. Yu, Y.; Liu, G.; Odobej, J.M. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA; 14–19 June 2020; pp. 11937–11946.
17. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA; 7–12 June 2015; pp. 4511–4520.
18. Ranjan, R.; De Mello, S.; Kautz, J. Light-weight head pose invariant gaze tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA; 18–22 June 2018; pp. 2156–2164.
19. Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; Sun, J. Meta-SR: A magnification-arbitrary network for super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA; 14–19 June 2020; pp. 1575–1584.
20. Wang, L.; Wang, Y.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning a single network for scale-arbitrary super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada; 10–17 October 2021; pp. 4801–4810.
21. Behjati, P.; Rodriguez, P.; Mehr, A.; Hupont, I.; Tena, C.F.; Gonzalez, J. Overnet: Lightweight multi-scale super-resolution with overscaling network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA; 3–8 January 2021; pp. 2694–2703.
22. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. arXiv 2015, arXiv:1503.02531.
23. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. Adv. Neural Inf. Process. Syst. 2019, 32.
24. Hansen, D.W.; Ji, Q. In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. 2009, 32, 478–500. [CrossRef] [PubMed]
25. Pathirana, P.; Senarath, S.; Meedeniya, D.; Jayarathna, S. Eye gaze estimation: A survey on deep learning-based approaches. Expert Syst. Appl. 2022, 199, 116894. [CrossRef]
26. Cheng, Y.; Wang, H.; Bao, Y.; Lu, F. Appearance-based Gaze Estimation with Deep Learning: A Review and Benchmark. arXiv 2021, arXiv:2104.12668.
27. Cazzato, D.; Leo, M.; Distante, C.; Voos, H. When I Look into Your Eyes: A Survey on Computer Vision Contributions for Human Gaze Estimation and Tracking. Sensors 2020, 20, 3739. [CrossRef] [PubMed]
28. Yoo, D.H.; Chung, M.J. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. Comput. Vis. Image Underst. 2005, 98, 25–51. [CrossRef]
29. Chen, J.; Ji, Q. 3D gaze estimation with a single camera without IR illumination. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA; 8–11 December 2008; IEEE: Piscataway, NJ, USA; 2008; pp. 1–4.
30. Hansen, D.W.; Pecce, A.E. Eye tracking in the wild. Comput. Vis. Image Underst. 2005, 98, 155–181. [CrossRef]
31. Palmero, C.; Selva, J.; Bagheri, M.A.; Escalera, S. Recurrent cnn for 3d gaze estimation using appearance and shape cues. arXiv 2018, arXiv:1805.03064.
32. Park, S.; Zhang, X.; Bulling, A.; Hilliges, O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland; 14–17 June 2018; pp. 1–10.
33. Lu, F.; Sugano, Y.; Okabe, T.; Sato, Y. Adaptive linear regression for appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. 2014, 36, 2033–2046. [CrossRef]
34. Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea; 27 October–2 November 2019; pp. 6912–6921.
35. Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; Hilliges, O. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK; 23–28 August 2020; Springer: Berlin/Heidelberg, Germany; 2020; pp. 365–381.
36. Chang, Z.; Matias Di Martino, J.; Qiu, Q.; Espinosa, S.; Sapiro, G. Salgaze: Personalizing gaze estimation using visual saliency. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea; 27–28 October 2019.
37. Guo, T.; Liu, Y.; Zhang, H.; Liu, X.; Kwak, Y.; In Yoo, B.; Han, J.J.; Choi, C. A generalized and robust method towards practical gaze estimation on smart phone. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.

38. He, J.; Pham, K.; Valliappan, N.; Xu, P.; Roberts, C.; Lagun, D.; Navalpakkam, V. On-device few-shot personalization for real-time gaze estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.

39. Park, S.; Mello, S.D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; Kautz, J. Few-shot adaptive gaze estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9368–9377.

40. Yu, Y.; Odobez, J.M. Unsupervised representation learning for gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7314–7324.

41. Kothari, R.; De Mello, S.; Iqbal, U.; Byeon, W.; Park, S.; Kautz, J. Weakly-supervised physically unconstrained gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9980–9989.

42. Liu, Y.; Liu, R.; Wang, H.; Lu, F. Generalizing gaze estimation with outlier-guided collaborative adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 3835–3844.

43. Xu, X.; Chen, H.; Moreno-Noguer, F.; Jeni, L.A.; Torre, F.D.I. 3d human shape and pose from a single low-resolution image with self-supervised learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 284–300.

44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv 2014, arXiv:1412.6980.

46. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

47. Chen, Z.; Shi, E. Appearance-based gaze estimation using dilated-convolutions. In Proceedings of the Asian Conference on Computer Vision, Perth, WA, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 309–324.

48. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 41, 162–175. [CrossRef] [PubMed]

49. Cheng, Y.; Lu, F.; Zhang, X. Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 105–121.

50. Cheng, Y.; Zhang, X.; Lu, F.; Sato, Y. Gaze Estimation by Exploring Two-Eye Asymmetry. IEEE Trans. Image Process. 2020, 29, 5259–5272. [CrossRef] [PubMed]