360-Degree Gaze Estimation in the Wild Using Multiple Zoom Scales

Ashesh
ashesh276@gmail.com
Chu-Song Chen
chusong@csie.ntu.edu.tw
Hsuan-Tien Lin
htlin@csie.ntu.edu.tw

National Taiwan University
Taipei, Taiwan

Abstract

Gaze estimation involves predicting where the person is looking at within an image or video. Technically, the gaze information can be inferred from two different magnification levels: face orientation and eye orientation. The inference is not always feasible for gaze estimation in the wild, given the lack of clear eye patches in conditions like extreme left/right gazes or occlusions. In this work, we design a model that mimics humans' ability to estimate the gaze by aggregating from focused looks, each at a different magnification level of the face area. The model avoids the need to extract clear eye patches and at the same time addresses another important issue of face-scale variation for gaze estimation in the wild. We further extend the model to handle the challenging task of 360-degree gaze estimation by encoding the backward gazes in the polar representation along with a robust averaging scheme. Experiment results on the ETH-XGaze dataset, which does not contain scale-varying faces, demonstrate the model's effectiveness to assimilate information from multiple scales. For other benchmark datasets with many scale-varying faces (Gaze360 and RT-GENE), the proposed model achieves state-of-the-art performance for gaze estimation when using either images or videos. Our code and pretrained models can be accessed at https://github.com/ashesh-0/MultiZoomGaze.

1 Introduction

Gaze estimation is a critical task in computer vision for understanding human intentions and has significant potential to be used in unconstrained environments—i.e., in the wild. Saliency detection [26, 28], human-robot interaction [21, 24], virtual reality [23, 27] are some of its main applications. The goal of gaze estimation is to predict the gaze direction of a given person accurately. The predicted direction can be 2D [11, 16], like the (x,y) coordinate of a mobile or laptop screen, or 3D [31, 35], reflecting the real-world reference system.

This work focuses on 3D gaze estimation. As shown in Figure 1(Top), the task of 3D gaze estimation in the wild has two unique properties that make it challenging. Firstly, there can be occlusions, extremely left/right face orientations, and even backward gazes. For those subjects, details of the face and eye regions are often incomplete or totally absent. These two regions, each at a different magnification level, arguably carry most information

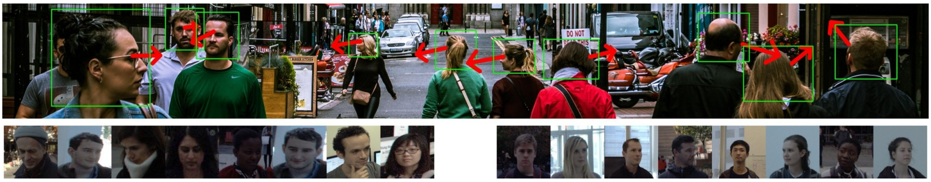


Figure 1: (Top) Our model MSA produces qualitatively good gaze prediction (red arrow) in the wild with backward gazes and varying head sizes. Image taken from www.pexels.com/photo/photo-of-people-walking-on-street-2416653/. (Bottom) Head crops from Gaze360 dataset. The head area in the image is referred to as the "scale" of the input image in this paper. Compared to right image set, left set has larger scale.

for the gaze. Even when eye patches are available, they are of varying resolution because of the varying camera-person distance. Recent approaches that leverage the face and eye information for gaze estimation [2, 4] generally rely on separate and therefore *complete* and *high-resolution* patches of the face and eyes. Therefore, those approaches are not suitable for our task of gaze estimation in the wild. Secondly, images in the wild setting have varying scales, where "scale" informally means the size of the head region. Varying scales come from varying camera-person distances and also result in varying resolutions of the subjects. Such variations cannot be easily handled by usual convolutional neural networks (CNNs) [13, 17]. In particular, we observe that the test performance varies significantly with varying scale.

One simple approach to tackle the scale variation issue is to use a bounding box to crop and resize the head region, where the bounding box can be generated from a head detector, a face detector or a facial keypoint detector. For example, Kellnhofer et al. [15] work with the head crops of the raw images. However, producing a tight head-bounding box in the wild is challenging because of significant variations in the scale, background, head orientations and lighting conditions. Therefore, scale variation still persists even after re-scaling the head-crops, as shown in Figure 1(Bottom). The aforementioned issue of incomplete facial information also makes it hard to leverage face or facial keypoint detectors for obtaining the bounding box. That is, approaches that re-scale with bounding boxes do not fully solve the scale variation issue in this setting.

Most approaches have addressed the scale variation by leveraging image warping [31] to normalize the input image [3, 6, 25, 37, 38]. Those image-warping approaches re-orient the camera virtually such that the head orientation and the camera-person distance to the virtual camera can be fixed [25, 37, 38]. But image warping typically requires locating the midpoint of the eyes accurately, and the midpoint is not always available under the aforementioned issue of incomplete facial information. That is, image-warping approaches are not satisfactory for gaze estimation in the wild, either.

In this paper, we tackle the two challenges of gaze estimation in the wild, namely extracting information from multiple magnification levels with incomplete facial information and significant scale variations, from a different perspective. Instead of bounding and normalizing the head region to *one* scale in advance, we seek to mimic what humans do when trying to accurately estimate a gaze: take successively focused looks at the head (face) region. In particular, we expand an input image to siblings scaled with different zooms, and aggregate the 2D feature maps from all siblings to make the final estimation. During feature extraction, every feature-map has an expression in all scales. Then, spatial max-pooling is applied on multiple 2D feature maps to get one 2D feature map of the same dimension, which reflects the intent to pick the best expression for each feature. The aggregation with multiple scales costs a mild increase of model complexity, but significantly improves the model by augmenting and aggregating with domain-justified transforms (zoom-in). Those transforms are simpler than the ones used in image-warping approaches, thus avoiding the complication of locating the midpoint between the eyes. In addition, aggregating multiple scales *within* the model instead of resizing to one scale *in advance* arguably makes the proposed model more robust than bounding-box-based approaches.

Another key challenge that we identify for estimation in the wild is the ambiguous representation of $-\pi$ and π for the yaw angle that causes discontinuity in the parametric space. For example, the Gaze360 dataset [15] contains backward gazes, for which the magnitude of the yaw angle is greater than $\pi/2$. To resolve the ambiguity, we propose to encode the yaw angle using the complex exponential (polar) representation. This allows estimating the yaw angle from two different trigonometric functions. We analyze the two estimates and propose a weighted averaging scheme that takes the best of them to further improve the prediction.

The contributions of the paper can be summarized as follows: 1) To the best of our knowledge, we are the first to extract information from multiple scales for solving gaze estimation in the wild. Our strength is in the simplicity of our approach—not requiring sophisticated external modules for extracting eye patches, head pose nor facial keypoints. Instead, our approach works with a simple head detector that does not have to be perfectly tight. The simplicity makes it possible to easily couple the approach across multiple backbones and different input types (images or videos). 2) To improve backward gaze prediction, which is critical for estimation in the wild, we propose to use the polar representation along with a robust averaging scheme to estimate the yaw angle. 3) Our proposed approach achieves state-of-the-art performance on Gaze360 [15], a benchmark dataset in the wild and RT-GENE [6], another dataset in the wild with a natural environment. We also demonstrate the effectiveness of the approach on ETH-Xgaze [39], a dataset collected with a controlled environment.

2 Proposed Model

We first present our model for single-image input gaze estimation with full 360° variations in the yaw angle. We then extend the model to a more complicated task of video (sequential) gaze estimation and then to a simpler task where backward gazes are absent.

Problem Formulation. The state-of-the-art work for gaze estimation in the wild [15] converts the raw images (or video frames) in the wild, like Figure 1(Top), to head-crop images using an off-the-shelf head detector to construct the training and test data. We follow the same construction. The construction makes each (head-crop) image roughly face-centered.

Given an input head-crop image I, the 3D gaze estimation task for images aims to predict the ground-truth yaw angle θ_g and pitch angle ϕ_g of the gaze direction of the subject within I. We denote the predicted yaw and pitch as θ_p and ϕ_p respectively. The corresponding task for videos, identical to [15] takes a sequence of video frames $V_{0:2T} = \{I_0, I_1, ..., I_{T-1}, I_T, .., I_{2T}\}$ as the input and aims to predict the yaw and pitch angles for the frame I_T.

For evaluation, we convert the target gaze and its prediction from polar co-ordinate representation (θ, ϕ) to a unit vector in 3D cartesian co-ordinates. We then evaluate on angular error, which is the angle between the target and predicted unit vectors.

Target Domain Transformation. A gaze is said to be a backward gaze when $\theta_g \in \{[-\pi, -\pi/2] \cup [\pi/2, \pi]\}$. The model proposed in [15] directly predicts θ and ϕ, which we discovered to

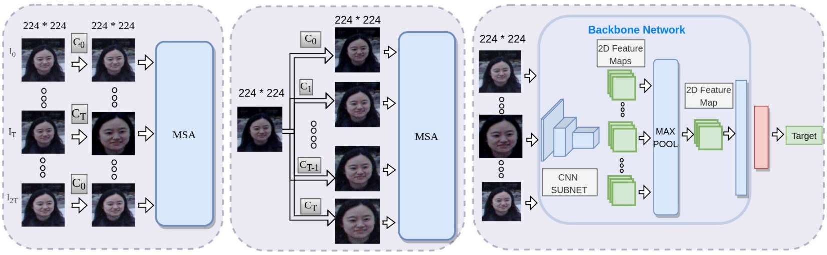


Figure 2: MSA (right) and data pre-processing for static (middle) and sequence (left) model.

be causing considerable losses on backward gazes due to a discontinuity in the yaw domain: the yaw value jumps discontinuously from π to $-\pi$. Refer supplemental for details. We tackle this problem via target space transformation: we predict $\sin(\theta)$, $\cos(\theta)$, and $\sin(\phi)$. We use tanh activation to ensure apppropriate prediction range.

To estimate θ from the predicted $\sin(\theta)$ and $\cos(\theta)$, we use a two-step method. First, we estimate θ in two ways. In the sine-based way, we estimate yaw θ_S from $\sin(\theta)$ and $\text{sign}(\cos(\theta))$;[1] in the cosine-based way, we estimate yaw θ_C using $\cos(\theta)$ and $\text{sign}(\sin(\theta))$.

In the second step, we calculate our final estimate of yaw, θ_p. Initially we used $\theta_p = \theta_{SC}$ where $\theta_{SC} = (\theta_S + \theta_C)/2$. However, θ_S is much more accurate than θ_C around 0° which can be observed in the distribution of θ_g, θ_S, and θ_C shown in supplemental. From the dip in the distribution of θ_C around 0° primarily and that of θ_S slightly around $\pm 90°$, one can say that model is having difficulty in predicting those regions. We argue that the low derivative of the tanh activation function near ± 1 makes it difficult for the model to predict values very close to ± 1, and that the high derivative of the $\sin^{-1}$ and $\cos^{-1}$ functions around 1 discourages the prediction of angles close to $\pm 90°$ and 0° respectively. We address this using a weighted averaging scheme. Specifically, our yaw prediction θ_{WSC} is defined as $\theta_{WSC} = w * \theta_S + (1-w) * \theta_C$. Here, w is defined as $w = |\cos((\theta_S + \theta_C)/2)|$. Defined this way, in practice, θ_S is assigned greater weight when θ_g is near 0° and θ_C is assigned greater weight when θ_g is near $\pm\pi/2$.

Loss Function. Following the Gaze360 paper [15], we use Pinball loss for all experiments on the dataset. In generic terms, if y_g is the target and y_p is prediction, then the quantile loss L_τ for quantile τ is defined as

$$L_\tau(y_p, \sigma, y_g) = \max(\tau y_\tau, -(1-\tau) y_\tau) \qquad y_\tau = \begin{cases} y_g - (y_p - \sigma) & \text{for } \tau \le 0.5 \\ y_g - (y_p + \sigma) & \text{otherwise} \end{cases} \tag{1}$$

Here, σ can be understood as the uncertainty in prediction, which is yet another output from our network. This formulation is used on all three gaze targets, namely $\sin(\theta)$, $\cos(\theta)$, and $\sin(\phi)$. The final loss is the average of these losses over quantiles $\tau = 0.1$ and $\tau = 0.9$. See [15] for more details.

Architecture for Static Model. In Gaze360 [15], as shown in Figure 1 (Bottom), head crops have varying scales. Additionally, the pinball static model [15] together with the proposed sine-cosine transformation does not handle images of varying scales any better (refer to the Experiments section for details). We use center-cropping along with spatial max-pooling to handle scale robustly and extract multi-scale features.

[1] $\text{sign}(\cdot)$ returns $+1$ on a non-negative number, and -1 otherwise.

The architecture of the proposed MSA (Multiple Scale Aggregation) model and data preprocessing for single image input case is shown in Figure 2 (Middle and Right). The input image is center-cropped with multiple sizes $\text{CCropL} = \{C_i, i \in [0,T]\}$ and subsequently scaled back to the original size, yielding a set of images with varying scales, which are then fed into the proposed MSA module. It comprises of a pre-trained backbone network and a final dense layer. Images are fed into the CNN portion of the backbone network to produce 2D feature maps for differently scaled images. We then use max-pooling on these stacked 2D feature maps along the scale dimension, and pass the output thereof through the backbone head, which is comprised of dense layers. Finally, the output from the backbone head is passed through a dense layer and an appropriate activation to yield the predictions. To predict $\sin(\theta)$, $\cos(\theta)$, and $\sin(\phi)$, we use tanh activation. To predict σ, we use sigmoid activation. We then use the θ_{WSC} formulation to get θ.

Architecture of Sequence Model. With sequence model, we use MSA with a different data preprocessing as shown in Figure 2 (Left). Given input image sequence $I_0, I_1..I_{2T}$, we center-crop the images with sizes $C_0, C_1..C_{T-1}, C_T, C_{T-1}..C_1, C_0$ respectively and rescale them back to original size. We also introduce the constraint that $C_i > C_{i+1} \forall i$. The specific crop-size ordering and constraint is placed to implicitly preserve information about frame ordering. Notably, C_T, the crop size for the target frame I_T, is the smallest which gives the greatest zoom-in effect. This should encourage the model to extract more micro-level details from eye region in I_T. These rescaled images are then fed to MSA to get the prediction.

Changes for Non-Backward Gaze Estimation. When solving the simpler non-backward gaze estimation, the absence of yaw discontinuity means that the proposed sine-cosine transformation is not needed. Therefore, in this setting, we directly predict θ and ϕ which essentially reduces the dimension of last dense layer from 4 to 3. We test this setting on the RT-GENE dataset [6] and ETH-Xgaze [39] dataset.

3 Experiments

3.1 Implementation Details and Data

Implementation Details Besides using pretrained Resnet18 [10] for the backbone as done in [15], MobileNet [29], SqueezeNet [12], ShuffleNet [20], and Hardnet [1] were also used. All weights get updated during training. Like [15], $T = 3$ was used for sequence model. For all experiments on all datasets and both model types, unless specified otherwise, we used [224,200,175,150] for CCropL, obtained empirically. More details given in supplemental.

Data We conducted experiments on Gaze360 [15], RT-GENE [6] and ETH-XGaze [39]. Gaze360 includes images of both indoor and outdoor scenes, with a full 360° range of yaw angles. The camera-to-person distance varies considerably (1 m to 3 m) leading to variable head sizes and image resolutions. Most existing datasets [11, 16, 22, 30, 31, 38] don't contain this much variation in camera-person distance and yaw angle. Using 238 subjects, this dataset comprises 129K training images, 17K validation images, and 26K test images. Similar to [15], we only work with head crops provided in the dataset.

RT-GENE [6] also has varying camera-person distances (80–280 cm). RT-GENE includes four sets of images: two original sets and two inpainted sets. There are two versions of both sets: the raw version and a MTCNN [36] based normalized version. As in [5], we find the inpainted sets to be noisy and therefore we only use the Raw-Original and Normalized-

Model + Backbone	All 360	Front 180	Back
Pinball Static [15]	15.6	13.4	23.5
Pinball Static [15] Re-Implemented	15.6	12.8	26.0
Spatial Weights CNN [38]	20.7	16.8	34.9
Spatial Weights CNN [38] + Resnet	15.5	12.8	25.4
CA-Net [4]	18.2	15.3	28.6
CA-Net [4] + Resnet	15.2	12.8	23.6
Static+avg	14.4	12.8	20.4
Static+wavg	14.4	12.7	20.4
MSA+raw	15.8	12.4	28.1
MSA+avg	14.0	12.3	19.9
MSA	**13.9**	**12.2**	**19.9**

Model	All 360	Front 180	Back
Pinball LSTM [15]	13.5	11.4	21.1
SSA+avg+Seq	13.1	11.5	18.9
SSA+avg+Seq + LSTM	13.2	11.5	19.4
SSA+wavg+Seq + LSTM	13.1	11.4	19.3
SSA+wavg+Seq	13.0	11.4	**18.8**
MSA+avg+Seq + LSTM	12.7	10.9	19.2
MSA+avg+Seq	**12.5**	10.8	19.0
MSA+Seq + LSTM	12.7	10.9	19.2
MSA+Seq	**12.5**	**10.7**	19.0

Table 1: Performance comparison on Gaze360 dataset [15]. For each configuration, three models were trained. Average angular error is reported in this table. Resnet18 is used as backbone in our model and its variants. (Left) Comparison on Static models. (Right) Comparison on Sequential (Temporal) models.

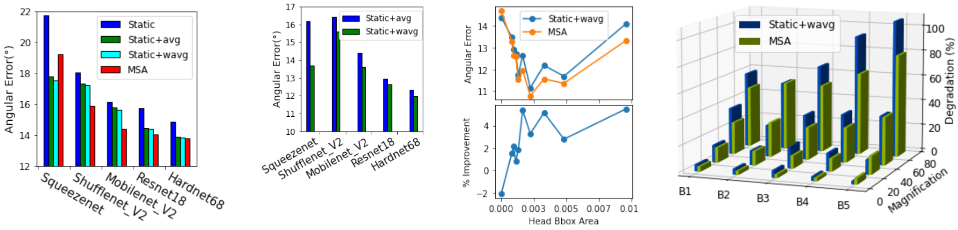
Original set. However, since the normalized version used facial keypoints for rescaling the image, we note that those image sets are therefore not relevant for the unconstrained setting.

ETH-XGaze data covers a wide range of head poses, gaze angles and lighting conditions and has high resolution images. It has 1M images comprising of 80 subjects in train set and 15 in test set. While the pre-processed images, which we use, have fixed camera-person distance, the wide range of head poses and gaze angles makes it a useful dataset for evaluation of in the wild performance.

3.2 Model Performance Comparison

Benchmarks. Due to the challenging nature of the dataset, we could not find any other work on Gaze360 dataset apart from the original work [15] of which we directly report the performance from their paper. On implementing their Static model [15], we found different performance in few yaw ranges and so we report our implementation as 'Pinball Static [15] Re-Implemented'. We adapt work of Cheng et al. [4] which also makes use of features from multiple scales by extracting features from eye region and face. We use the eye bounding boxes provided with the dataset for this work. We also adapt work of Zhang et al [38] where, similar to us, they work with just face images. For both of above mentioned works, we also report results by replacing their face feature extractor (backbone network) with Resnet18 and the loss with pinball loss so as to have a more authentic comparison.

Ablation Study. We compare our MSA model with multiple variants of itself. Here, we mention differing points for each variant against our model. MSA+avg and MSA uses the θ_{SC} and θ_{WSC} formulation for yaw respectively. MSA+raw directly predicts θ, ϕ and uncertainty σ, and does not use the sine-cosine transformation. Static+avg and Static+wavg do not use the multi-crop based idea and can be understood as being MSA+avg and MSA



(a) (b) (c) (d)

Figure 3: All results are computed on Gaze360 dataset. For (a), (b) and (d), B1, B2, B3, B4, and B5 denote Squeezenet, Shufflenet_V2, Mobilenet_V2, Resnet18, and Hardnet68 backbones respectively.(a) Performance of Static model variants. (b) Angular error on samples having front ±20° as yaw. (c) Top: Angular error variation with variation in Head bounding bbox area (in fractions) on front 180° images. Head bbox area is binned into 10 equal sized bins and average angular error is reported for each bin. Bottom: Percentage benefit of our MSA model over Static+wavg model with Head bbox area variation.(d) Percentage increase in average angular error on front 180° gazes with varying amount of magnification.

respectively with CCropL $= [224], T = 1$. SSA, used in sequential model is MSA with no multicrop, i.e, CCropL $= [224, 224, 224, 224], T = 3$.

Performance. In Table 1 (left) we compare Static model performance. Firstly, by comparing Static+avg model's performance (20.4°) on Back gazes with [15] (23.5°), benefit of sine-cosine transformation can be seen. Next, benefit of multi-scale aggregation can be seen on Front 180 gazes by observing MSA+raw's performance (12.4°) with [15](13.4°). Finally, MSA model outperforms on all three gaze categories namely front 180°— $|\theta_g| \in [0°, 90°]$ (12.2°), back — $|\theta_g| \in [90°, 180°]$ (19.9°) and overall (13.9°). It is worth noting the inferior performance of [4] (overall 15.2°) which used both eye and face patches. It shows that absence of clear eye patches in such conditions significantly hampers the performance.

In Table 1 (right), we show that in Sequence model type, our model MSA+Seq outperforms the benchmarks on all three gaze categories. We also show the performance on using the LSTM module as an aggregation module instead of our max-pool operator for a closer comparison to [15]. More details are given in the supplementary material, where we do an experiment of varying the aggregation modules.

Individual Effect of Different Components. Figure 3 (a) shows the performance gains achieved by different components. We observe the benefit of the sine-cosine transformation by comparing Static with Static+avg, and observe the benefit of multiple zoom scales by comparing MSA with Static+wavg. We argue that reason for inferior performance of MSA with Squeezenet was that it being the lightest backbone (1.2M parameters), could not manage multiple scales. Finally, as θ_{WSC} was introduced to fix issues with θ_{SC} around primarily 0°, its benefit can be observed on frontal gazes in Figure 3 (b).

Performance Variation With Head Crop Area. In Gaze360 dataset, the authors provide the head bounding box dimensions as a fraction of the original fixed size full body image. Firstly, as seen in Figure 3 (c), for front 180° gazes, we find considerable variation in performance with change in head bounding box area. Secondly, MSA gets better performance consistently on most bounding box area bins.

Performance Variation with Camera-Person Distance We look at how much MSA outperforms the Static+wavg model on front 180 gazes as we vary the distance. We binned

Method	Bkb	Image	Input	Err
Spatial weights CNN [38]	-	NOrig	Face	10
RT-Gaze [6]	-	NOrig	Eye	8.6
RT-Gaze [6]	-	NOrig + NIn	Eye	7.7
FAR-Net [5]	-	NOrig	Face + Eye	8.4
Static	B4	Orig	Face	7.9
MSA+raw	B4	Orig	Face	7.1
Static	B5	Orig	Face	7.0
MSA+raw	B5	Orig	Face	**6.7**
Static	B4	NOrig	Face	7.2
MSA+raw	B4	NOrig	Face	7.3
MSA+raw+175	B4	NOrig	Face	6.9

Table 2: Performance comparison on RT-GENE dataset. NOrig, Orig and NIn stand for Normalized-Original, Raw-Original and Normalized-Inpainted image type respectively. B4 and B5 stands for Resnet18 and Hardnet68 respectively.

Method	Bkb	Err
ETH-XGaze [39]	–	4.5
Pinball Static [15]	B4	4.4
MSA+raw	B4	4.1
MSA+raw	B5	4.0

Table 3: Performance comparison on ETH-XGaze dataset. B4 and B5 stands for Resnet18 and Hardnet68 respectively

images on camera-person distance into three equal bins. We show percentage improvement achieved by MSA over Static+wavg in Table 5. MSA gives more benefit for larger distances.

Performance Degradation with Scale Perturbations. Here we quantify performance degradation if a zoomed-in/zoomed-out image is given instead as input at evaluation. For every magnification level, we created one zoomed-in image and one zoomed-out image. For a magnification of c, we center-cropped the original image with size $224-c$ and rescaled it back to size 224 to create a zoomed-in image. To produce a zoomed-out image, we padded the original image with $c/2$ pixels on the boundary and rescaled it down to 224. Thus, the greater the magnification c, the greater the zoom-in and zoom-out effect. In Figure 3 (d), we plot the percentage increase in angular error (averaged over zoom-in and zoom-out) with the amount of magnification on front 180° gazes. In the majority of cases across backbones and magnification levels, MSA has lower percentage increment in angular error as compared to Static+wavg thereby showing robustness of multi-zoom approach on scale variations.

On Time Complexity and GPU RAM MSA+Seq takes same time and RAM as Pinball LSTM [15]. MSA takes L times more time w.r.t Pinball Static [15], L being the number of crop sizes. However, unlike LSTM, MSA can be parallelized since the 2D feature map computation for each magnification level is independent and so MSA+Seq can run faster than Pinball LSTM [15] and MSA can achieve comparable speed to Pinball Static [15]. If we parallelize MSA, then it however would take L times more GPU RAM over Pinball Static [15]. Otherwise, 2 times GPU RAM is needed for which one would need to compute feature map for each cropsize one by one while simultaneously updating the max feature map stored in a buffer. But since all our experiments needed just 4 crop-sizes, we argue that even for Static model type, this is not a big limitation for MSA. MSA with Resnet18 evaluates 279 images in 1s wall time and takes 2.5GB GPU RAM on a single 2080 Ti GPU with 64 batch size and 4 workers. With 1 batch size and 1 worker, it is 80 images/sec.

Comparison with feature-map upscaling One way to get a faster model could be to rescale feature map instead of rescaling input image. Inspired from [9], here, we passed the image through the initial portion of the backbone network (Resnet18) to get a 2D feature map. We

Res	All 360	Front 180
$7*7$	14.5	12.8
$14*14$	14.5	12.8
$28*28$	14.5	12.6

Table 4: Performance of upscaling based idea on Gaze360 dataset. First column, Res stands for feature map resolution

Distance(m)	% Improvement
0.9-1.8	3.9%
1.8-2.7	4.5%
2.7-3.5	5.2%

Table 5: Variation of % improvement of MSA over Static+wavg model w.r.t camera-person distance on Gaze360.

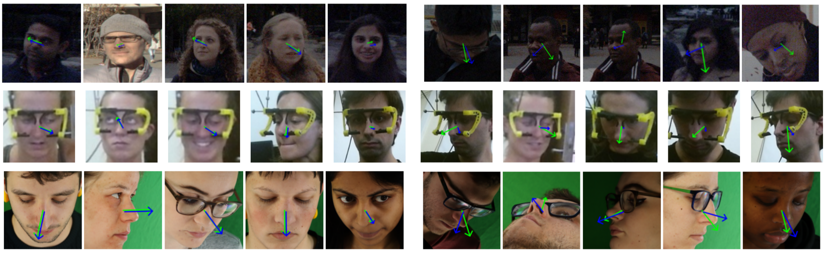


Figure 4: Showing some of the better (Left) and worse (Right) performing images for MSA model on Gaze360 (Row 1), RT-GENE (Row 2) and ETH-Xgaze (Row 3) dataset. Green and purple arrow denote the ground truth and predicted gaze directions respectively.

then center-cropped the feature map with different sizes and rescaled them back. We obtain these cropsizes by transforming CCropL. For 14*14 feature map, C_i crop-size will transform to $\frac{C_i*14}{224}$. These rescaled feature-maps are then averaged and passed through the remaining network to get the prediction. We used θ_{WSC} formulation. From Table 4, we find feature-map upscaling to be worse than MSA for all feature-map resolutions.

Qualitative Analysis In Figure 4, we look at few of the better and worse performing images from all three datasets. One could see that in cases where (1) the head orientation is highly oblique and when (2) eyeball is deviated from center position but is less visible, the model performs poorly. Performance on some backward gazes were also worse but we didn't include them here since there is little visual cue to glean from them. Model naturally performs best on relatively more frequent head poses.

Model performance on RT-GENE and ETH-XGaze In Table 2, we see that MSA+raw gets state-of-the-art results on RT-GENE. We directly report the performance of Spatial weights CNN and RT-Gaze on RT-GENE from [6]. Likewise for FAR-Net from [5]. However, MSA+raw(7.3°) with default CCropL does not outperform Static model(7.2°) with Normalized-Original (NOrig) images. This makes sense because unlike Raw-Original, NOrig images do not have significant variations in scale and also are a bit zoomed in. Consequently, the proposed minimum cropsize of 150 in CCropL proves too low and results in inferior performance. However, when CCropL is set to $[224, 208, 191, 175]$ (MSA+raw+175 model), which has the higher min-cropsize 175, it outperforms Static model. As mentioned in Section 3.1, comparing on Normalized-Original is unfair to the in the wild setting on which we focus on. Next, Table 3 compares the performance on ETH-XGaze dataset. It is important to note that here, we worked with normalized images— the camera-head distance is fixed. In absence of scale variation, the outperformance of MSA+raw model over Static pinball model [15] points to its ability to extract information from multiple scales.

4 Related Work

Initial work on gaze estimation was model-based [8, 14, 32, 33], involving modeling the geometry of the eye and then using this for gaze estimation. With the rise of computational resources, the increase in dataset sizes, and the emergence of deep learning, appearance-based approaches [3, 6, 25, 34, 37] came to being. As the name suggests, this approach involves estimating the gaze directly from the image appearance. A typical appearance-based model is a neural network which takes as input an image containing a human face or an eye patch and then predicts the gaze from this. Unlike the model-based approach, here the setup is quite practical and the environment more relaxed. Most appearance-based models require a single camera with lower requirements for high-resolution images. However, since this does not capture eye geometry, it is highly susceptible to changes in head pose. The fact that several appearance-based approaches take the head pose as input along with the image [7, 18, 19, 31] is a testament to the sensitivity of this approach to head pose changes.

Initial appearance-based models worked with eye patches [37]. Subsequently, face images began to be used as input [38]. More recently, both face and eye patch have been used together as input, demonstrating the presence of useful information at multiple scales [2, 4]. As discussed before, one problem with this is the need for bounding boxes for the eyes, especially when dealing with large gaze angles, low-resolution images, or occlusions.

As with the development in models, there has been a gradual evolution of datasets. With recent datasets we see a larger number of participants [16], greater variation in head pose and gaze angle [6, 15], and greater variation in camera-person distance, background variations, and so on. Notably, the Gaze360 dataset [15] has several interesting properties. With full 360° variation in yaw, significant camera-person distance variations, and both indoor and outdoor background settings, Gaze360 is a promising dataset for 3D gaze estimation in the wild. In the context of scale, we also find the RT-GENE dataset [6] to be quite useful, due to its significant variation in camera-person distance (80–280 cm).

5 Conclusion

We present a novel approach for gaze estimation in the wild where we extract information from different magnification levels by aggregating features from images belonging to these levels. Our work is simple and can be easily coupled with different input types (images or videos). Furthermore, the approach reaches state-of-the-art performance for gaze estimation in the wild without relying on more complicated components like eye patch detector, face pose detector, or image warping. The multi-zoom approach has potentials for other vision problems in the wild as well, where the scale variation is a crucial issue.

Acknowledgement

We thank the anonymous reviewers for valuable suggestions. This work is partially supported by the Ministry of Science and Technology of Taiwan via the grants MOST 107-2628-E-002-008-MY3, 110-2628-E-002-013, and 110-2218-E-002-033-MBK. We also thank the National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational resources.

References

[1] Ping Chao, Chao-Yang Kao, Yushan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. HarDNet: A Low Memory Traffic Network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3551–3560, Seoul, Korea (South), October 2019. IEEE. ISBN 9781728148038. doi: 10.1109/ICCV.2019.00365. URL https://ieeexplore.ieee.org/document/9010717/.

[2] Zhaokang Chen and Bertram E. Shi. Appearance-Based Gaze Estimation Using Dilated-Convolutions. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, volume 11366, pages 309–324. Springer International Publishing, Cham, 2019. ISBN 9783030208752 9783030208769. doi: 10.1007/978-3-030-20876-9_20. URL http://link.springer.com/10.1007/978-3-030-20876-9_20.

[3] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 105–121, Cham, 2018. Springer International Publishing.

[4] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10623–10630, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i07.6636. URL https://aaai.org/ojs/index.php/AAAI/article/view/6636.

[5] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze Estimation by Exploring Two-Eye Asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[6] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 339–357, Cham, 2018. Springer International Publishing.

[7] Kenneth Alberto Funes Mora and Jean-Marc Odobez. Gaze estimation from multimodal Kinect data. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–30, June 2012.

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

[9] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patchnetvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14141–14152, June 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

ISBN 9781467388511. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

[11] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5-6):445–461, August 2017. ISSN 0932-8092, 1432-1769. doi: 10.1007/s00138-017-0852-4. URL http://link.springer.com/10.1007/s00138-017-0852-4.

[12] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv:1602.07360 [cs]*, November 2016.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial Transformer Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[14] Li Jianfeng and Li Shigang. Eye-Model-Based Gaze Estimation by RGB-D Camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 606–610, June 2014.

[15] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6911–6920, October 2019.

[16] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184, June 2016.

[17] Karel Lenc and Andrea Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. *International Journal of Computer Vision*, 127 (5):456–476, May 2019.

[18] Feng Lu, Takahiro Okabe, Yusuke Sugano, and Yoichi Sato. A Head Pose-free Approach for Appearance-based Gaze Estimation. In *Procedings of the British Machine Vision Conference 2011*, pages 126.1–126.11, Dundee, 2011. British Machine Vision Association.

[19] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Head pose-free appearance-based gaze sensing via eye image synthesis. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1008–1011, November 2012.

[20] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11218, pages 122–138. Springer International Publishing, Cham, 2018. ISBN 9783030012632 9783030012649. doi: 10.1007/978-3-030-01264-9_8. URL http://link.springer.com/10.1007/978-3-030-01264-9_8.

[21] A. Jung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A.t Crof. Meet Me where I'm Gazing: How Shared Attention Gaze Affects Human-Robot Handover Timing. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 334–341, March 2014.

[22] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 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 - ETRA '14*, pages 255–258, Safety Harbor, Florida, 2014. ACM Press.

[23] Nitish Padmanaban, Robert Konrad, Tal Stramer, Emily A. Cooper, and Gordon Wetzstein. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences*, 114(9):2183–2188, February 2017.

[24] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054, October 2016.

[25] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-Shot Adaptive Gaze Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9367–9376, Seoul, Korea (South), October 2019. IEEE.

[26] Daniel Parks, Ali Borji, and Laurent Itti. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research*, 116:113–126, November 2015.

[27] PatneyAnjul, SalviMarco, KimJoohwan, KaplanyanAnton, WymanChris, BentyNir, LuebkeDavid, and LefohnAaron. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, November 2016.

[28] Dmitry Rudoy, Dan B. Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning Video Saliency from Human Gaze Using Candidate Selection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1147–1154, June 2013.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, June 2018. doi: 10.1109/CVPR.2018.00474.

[30] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *UIST '13*, 2013.

[31] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, June 2014.

[32] Kang Wang and Qiang Ji. Hybrid model and appearance based eye tracking with kinect. In *ETRA*, 2016.

[33] Kang Wang, Shen Wang, and Qiang Ji. Deep eye fixation map learning for calibration-free eye gaze tracking. In *ETRA*, 2016.

[34] Kang Wang, Rui Zhao, and Qiang Ji. A Hierarchical Generative Model for Eye Image Synthesis and Eye Gaze Estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 440–448, Salt Lake City, UT, June 2018. IEEE.

[35] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing Eye Tracking With Bayesian Adversarial Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11899–11908, June 2019.

[36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. ISSN 1070-9908, 1558-2361. doi: 10.1109/LSP.2016.2603342. URL http://ieeexplore.ieee.org/document/7553523/.

[37] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, June 2015.

[38] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, July 2017.

[39] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 365–381. Springer International Publishing, 2020.