Human Attention in Fine-grained Classification

Yao Rong[1]
yao.rong@uni-tuebingen.de

Wenjia Xu[2]
xuwenjia16@mails.ucas.ac.cn

Zeynep Akata[1,3]
zeynep.akata@uni-tuebingen.de

Enkelejda Kasneci[1]
enkelejda.kasneci@uni-tuebingen.de

[1] University of Tübingen, Germany

[2] University of Chinese Academy of Sciences, Beijing, China

[3] Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

The way humans attend to, process and classify a given image has the potential to vastly benefit the performance of deep learning models. Exploiting where humans are focusing can rectify models when they are deviating from essential features for correct decisions. To validate that human attention contains valuable information for decision-making processes such as fine-grained classification, we compare human attention and model explanations in discovering important features. Towards this goal, we collect human gaze data for the fine-grained classification dataset CUB and build a dataset named CUB-GHA (Gaze-based Human Attention). Furthermore, we propose the Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN) to integrate human gaze knowledge into classification models. We implement our proposals in CUB-GHA and the recently released medical dataset CXR-Eye of chest X-ray images, which includes gaze data collected from a radiologist. Our result reveals that integrating human attention knowledge benefits classification effectively, e.g. improving the baseline by 4.38% on CXR. Hence, our work provides not only valuable insights into understanding human attention in fine-grained classification, but also contributes to future research in integrating human gaze with computer vision tasks. CUB-GHA and code are available at https://github.com/yaorong0921/CUB-GHA.

1 Introduction

Through a lifelong learning process, humans have developed a selective attentional mechanism, which has received attention in many areas of artificial intelligence [56]. As human attention can be revealed from gaze data, it bears the potential to explain our behavior and decisions [32]. Many computer vision applications embrace human gaze information to detect salient objects for solving tasks [20, 35, 40]. To visually illustrate human attention in these tasks, it is common to add a Gaussian filter on fixation points to form a feature map [16], which is also called *saliency* map [22] (see Figure 1). Similar to how gaze explains

Figure 1: Overview of our proposed methodology. HA saliency map is used to obtain attention area which is used to enhance the training dataset in Gaze Augmentation Training (Left), while it is used as extra knowledge and fused together with the image knowledge in the Knowledge Fusion Network (Right).

human decisions, the post-hoc attention of a network, i.e. model explanation, tries to reveal important regions for neural network decision-making [13, 51, 56, 41, 43, 59]. Both can be visualized by means of saliency maps, thus allowing the study of similarities and differences between them. In this context, several previous works show that humans and models are looking at different regions when performing the same task [8, 57]. However, it is not clear whether a feature discovered by a human is more efficient for solving a given task or not. Our work addresses this research gap and the hypotheses that (1) human attention focuses on essential features for solving the task (e.g. fine-grained classification); (2) using human attention also allows improving model performance in accomplishing the task. To validate the first hypothesis, we first capture and present human attention in the style of a saliency map. We compare the regions that human attention covers with the ones that are discovered by the model (model explanation), and show that human attention hints on the regions that are more discriminative in the classification. We propose two modules which make use of the essential features revealed by human gaze to validate the second hypothesis: we use Gaze Augmentation Training (GAT) to train a better classifier and a Knowledge Fusion Network (KFN) to integrate the human attention knowledge into models.

Our contributions are as follows: (1) We collect human gaze data for the fine-grained data set CUB, enhance it by incorporating human attention and coin this new dataset as **CUB-GHA** (Gazed-based Human Attention). For this novel dataset, we also validate the efficiency of human gaze data in discovering discriminative features. (2) We propose two novel modules to incorporate human attention knowledge in classification tasks: Gaze Augmentation Training (GAT) and Knowledge Fusion Network (KFN). (3) To showcase the relevance of our work for highly relevant applications, we evaluate our methods not only on our novel CUB-GHA dataset, but also on chest radiograph images from a recently released dataset CXR-Eye (which contains also gaze data). Our work shows that human attention knowledge can be successfully integrated in classification models and help improve the model performance with regard to the state-of-the-art in different classification tasks.

2   Related Work

Human Gaze in Machine Learning. Recent developments in hardware devices allow for the precise recording of eye movements in different activities, ranging from human-computer interaction [27, 28] to complex and dynamic real-world tasks, such as driving [4, 49] and robotics [3, 39, 47]. Furthermore, the way that visual information is processed can reveal

information about a person's strategy or level of expertise [6]. In the medical domain, researchers have validated that gaze data reveals patterns which can benefit AI models, as for disease (Pneumonia and Congestive Heart Failure) classification [19]. In computer vision, gaze data has proven its usefulness in various applications [20, 33, 35, 40]. E.g., [20] collects gaze (coordinates, duration, etc.) vectors for 60 bird classes in dataset [43] to form embeddings for zero-shot learning. [33] compares the attention map generated by an attention module (two convolutional layers) with human attention maps generated by the data from [20] and shows that human attention surpasses the attention module. [35] proposes a photograph cropping system using the collected fixation data to identify important content and compute the best crop. Eye tracking data is also used to extract dominant objects in videos [40]. Different from previous works which use gaze for specific tasks [20, 35, 40], our proposal GAT leverages human attention to train a better backbone which can be used in many different tasks and frameworks. Moreover, we evaluate GAT and KFN for two different classification tasks and thus show the general validity of our methods.

Attention Module in Fine-grained Classification. Many previous works [11, 15, 23, 24, 25, 38, 42, 53, 57, 58, 60] integrate attention modules in networks to localize the parts which are important for fine-grained classifications and make use of the information of the discriminative parts to improve the models' performance. [11, 23, 24, 25, 58] adopt the Recurrent Attention Model (RAM) [29], where an attention agent is deployed to predict locations of the discriminative regions, and train the classifier based on these cropped regions. The attention agent is trained with a reinforcement learning algorithm to address the non-differentiability due to the cropping operation. However, the architecture of this attention model is cumbersome with high computational cost. [15, 42, 53, 57, 58, 60], on the other hand, design attention modules using the output from intermediate layers in networks and enforce it to capture discriminative features. Compared to previous works, we do not use the intermediate outputs from networks to generate model attention but use human attention maps. Our method augments the training set with regions cropped according to human attention and thus accomplishes training a better classifier. We compare our method with previous works and demonstrate the profit of exploiting human attention in Section 5.

3 CUB-GHA Dataset

In this section, we first provide the details of our gaze data collection paradigm and then analyze the effect of machine explanation and human attention to the fine-grained classification model. To collect gaze data, we employ the CUB-200-2011 (CUB) [46] dataset with 11,788 images from 200 bird classes incorporating various annotations: image-level attributes, body part locations, and text descriptions of the bird. Our annotation leads to a human-gaze enhanced version, i.e. CUB-GHA.

We choose the fine-grained CUB dataset for two reasons: 1) The difference between two similar classes lies in local and compositional attributes, which can be precisely captured by human gaze. For instance, it is challenging to achieve a measure for unified human attention when comparing a bear and a horse as there are many differences between them. In contrast, distinguishing between two similar birds with different throat colors presents a more unified problem (as shown in Figure 2). 2) The CUB dataset is widely used for various computer vision tasks, such as fine-grained classification [10, 11, 58], zero-shot learning [1, 50, 51, 55], explainable artificial intelligence [2, 7, 17], etc. Thus, our CUB-GHA may serve as a valuable foundation for exploring the effect of human attention on those

Figure 2: **(a)** Eye tracker set-up: We use a Tobii Spectrum eye-tracker to capture gaze information at a high frequency of 1200 Hz. **(b)** Data collection: Step 1 represents a schematic overview of the image comparison task where two images of different species are freely viewed. In Step 2, a randomly selected example of one of the species is shown to the user for which gaze data is then collected. To gamify this setting, the user is asked to choose the correct class in Step 3. **(c)** Preparing human attention data: we visualize human attention in Gaussian-based saliency maps.

tasks.

3.1   Gaze Data Collection

Collection Framework. As illustrated in [20], humans fixate on class-discriminative features when they observe two very similar classes. In this paper, we adopt an image comparison game [20], where we encourage participants to look at the discriminative features when comparing two similar images from different categories. The comparison task is designed to be challenging to provide more powerful insights, i.e. two classes in one comparison pair are chosen to be very similar.

A schematic overview of our data collection is presented in Figure 2. Figure 2 (a) shows the experimental setup including a picture of the eye-tracker (Tobii Spectrum Eye Tracker, sampling at 1200 Hz) and the chin rest as well as the display (1920 × 1080 resolution). The chin rest is used to ensure precise recordings of the eye movements. Each image is re-scaled to fit to the screen and placed at the center. The average distance between the participant's nose and the screen is approximately 60 *cm*. The comparison task consists of three steps shown in Figure 2 (b). In step 1, we present two representative images at the same time, each from one bird class of the CUB dataset, e.g. representative images of Barn Swallow and Tree Swallow. We choose the comparison pairs under the same sub-classes, and then different persons manually check the visual similarity to make sure that the comparison is not too simple. The participants are allowed to observe the images for as long as they want. When the participant is ready for the classification task, in step 2, an image from one of the two classes of the CUB dataset is shown. The participant has to choose which category the image belongs to by viewing the image. Note that the image shown for classification is displayed for only 3 seconds to avoid explorative gaze behavior unrelated to the task. One collection session includes one image from each class, meaning that there are 200 images reviewed per session. Every image in CUB is reviewed by five different participants. 25 subjects (19 males and 6 females with mean age 27.64 ± 4.15) participate in the experiment. Although the participants do not take part in the same number of sessions and instances, we make sure that every participant views all classes in every session. It is worth noting that all participants are domain novices with no specific knowledge about birds.

Gaze Data Preparation. The raw gaze data is preprocessed to extract fixation locations

using the Velocity-Threshold Identification (I-VT) algorithm [50]. The resulting fixation points offered in the dataset include coordinates and duration information. Based on this information, we generate saliency maps for human gaze as shown in Figure 2 (c). Every fixation location is modelled as a Gaussian distribution $G(\mu, \sigma^2)$, where σ is 75 pixels (in the displace resolution), according to the ratio of the distance to the screen and the approximate foveal area of $2°$. The duration of the fixation is then used as a weight for its Gaussian distribution. Finally, the saliency map is presented in grayscale image form. From here on, we note the human attention saliency map generated from gaze data as HA.

3.2 Gaze Data Analysis

In this section, we validate the hypothesis that *HA covers discriminative regions for the fine-grained classification*. Given the same image and the same (visual) task, HA and model explanation (ME) reveal regions which are important in making decisions for humans and models, respectively. Thus, we compare HA with four MEs provided by a trained classifier (vanilla ResNet-50 [12]) with a classification score of 85.58% on CUB , and validate that HA is able to discover features that better differentiate the bird from other bird classes. The four ME used are Class Activations Maps (CAM) [59], Gradient-based CAM (Grad-CAM) [36], InputXGradient (IxG) [41], and IntegratedGradients (IG) [43].

For quantitative comparison, we compare HA and ME using the keep and retrain (KAR) procedure (proposed in the appendix to [13]) to validate if the important regions highlighted by HA and ME help the model to make decisions. Concretely, we gradually insert important pixels to a blank image according to their values in HA or ME saliency maps. The modified percentage of pixels is [5,10,15,20,25,30,50,70,90]. After a certain amount of pixels are inserted, we retrain a new model using the modified train images and report the accuracy on modified test sets. Modified images at 5%, 20% and 70% of pixels inserted using Grad-CAM are shown in Figure. 3 (middle). The intuition behind this is that the class-discriminative information should be included in the pixels that are evaluated as very important; with more pixels inserted which are relatively less important, the model performance will not improve much. If a saliency map selects the informative features as being the impor-

Figure 3: Comparison of HA and ME in discriminative feature discovery. **Top:** Test accuracy on modified datasets using different saliency maps. The x-axis is the insertion percentage and the y-axis is the accuracy on test set. The AUC of each curve is reported in zoom-in image. **Middle:** modified images (using Grad-CAM as an example). **Bottom:** Illustration of HA and four MEs.

tant ones for classification, the increase of accuracy at the beginning of insertion is rapid, i.e. the resulting higher Area Under the Curve (AUC) indicates a better feature importance estimate.

The keep and retrain curves and the AUC scores for each method are shown in Figure. 3 (top), and the qualitative saliency maps for HA and four MEs for one image are shown in the

bottom. We see that HA and MEs do not focus on the same image regions: humans consider the white feathers on the black wing as a more important feature, while the model uses the yellow head as the most important feature (see the original image in Figure 1). HA discovers more informative and important features for the fine-grained classification model than the MEs do, e.g. HA obtains an AUC score of 0.716 compared to Grad-CAM (0.706) and IG (0.702). With only 5% important pixels revealed, the model trained with HA modified images can reach an accuracy of 81% while the model trained with ME modified images only reaches an accuracy of around 70%. More details of the analyses can be found in the supplementary material.

4   Methodology

In this section, we introduce how we incorporate the gaze information to improve the classification performance, i.e. using gaze to augment training data (GAT) or as an extra information source (KFN). The illustration of the architecture is shown in Figure 1.

4.1   Gaze Augmentation Training

Motivated by the assumption that the model should pay attention to the discriminative image regions (highlighted by HA), we enhance our model's reaction to those regions by adding them as augmentation in training as illustrated in Figure 1 (left).

To get the k augmentation images for the input image $I \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 3}$ (where $\mathcal{H}$ and $\mathcal{W}$ represent the width and height of the input image), we implement a sliding window algorithm to find areas which contain human attention. A window with the size of (w, h) slides on the HA map $A \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 1}$ from the upper left to the right bottom corner (with stride size s in both dimensions). We rank all the window areas according to the averaged pixel values inside windows and get k cropped images according to top-k highest scores. We resize the cropped images to the half of the width and height of the I, i.e. $I' \in \mathcal{R}^{\frac{\mathcal{H}}{2} \times \frac{\mathcal{W}}{2} \times 3}$, as suggested in [11, 38, 57] where the attended regions are resized into smaller sizes. I' has the same label y as I does. To get various regions, we use various window sizes and the non-maximum suppression. The training set is extended to $I \cup I'$. We train the model on the enlarged dataset with cross-entropy loss. Note that GAT just needs human gaze information in training and the model takes only original images as inputs in the test phase.

4.2   Knowledge Fusion Network

As shown in Fig. 1 (right), our KFN is a two-branch network that fuses the knowledge from HA and the original image features together. The first branch is the image knowledge branch. This branch takes the original images $I_o \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 3}$ as the input, where $\mathcal{H}$ and $\mathcal{W}$ represent the width and height of the input image, respectively. We use a CNN backbone $f_o(\cdot)$ to extract image feature $f_o(I_o) \in \mathcal{R}^{D_o}$ from I_o, where D_o denotes the dimension of the feature channel. Another branch, the HA knowledge branch, incorporates the gaze features of this image. We multiply the gaze information (HA) with the input image by $I_g = I_o \odot A$, where $A \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 1}$ is the HA saliency map. Through this operation, pixels in the image get different weights from the gaze: the area where humans pay attention to is brighter than the rest. I_g contains visual features which are important for the classification. Another CNN backbone $f_g(\cdot)$ is utilized to extract the gaze feature as $f_g(I_g) \in \mathcal{R}^{D_g}$. Then the gaze feature

and original image feature are concatenated together to form the fused feature $f(I_o, I_g) \in \mathcal{R}^{(D_o+D_g)}$. It this way, we integrate HA into a multiclass classification task to study the potential of HA to improve the performance of the image classifier. The whole network is trained with cross-entropy loss.

5 Experiment

In this section, we first introduce datasets and implementation details. Then we show the results of our proposed GAT and KFN. To show the general validity of our methods, We test on two datasets: CUB-GHA and Eye Gaze Data for Chest X-rays (CXR-Eye) [18].

5.1 Datasets and implementation details

CUB-GHA includes 11788 images in total, with 5994 images for training and 5794 for validation [46]. Each image contains eye gaze data from 5 participants. CXR-Eye includes 1083 chest X-ray images with gaze data from a radiologist while performing routine radiology readings [18]. The goal of this dataset is to make a prediction based on the chest X-ray image, whether the subject has one of two clinically prevalent diseases (pneumonia or congestive heart failure (CHF)), or the subject is healthy (normal). The human gaze data is also visualized in the saliency map style. Each image is annotated with one label out of three classes. We choose this dataset because it is a unique human gaze dataset in the medical domain. For such safety-critical applications (e.g. computer-aided diagnosis), we believe the integration of human attention can increase the acceptance and trust of these applications among users.

In our experiments on the CUB dataset, the input images are resized to 448×448 (the images are cropped to this size with the smaller edge first resized to 448) and then randomly flipped horizontally in training. We use the SGD optimizer [34] with an initial learning rate of 0.001. In the experiments on the CXR dataset, the input images are resized to 224×224 and a random horizontal flip is used in training. We use the Adam optimizer [21] with an initial learning rate of 0.0005. Since the CXR-Eye dataset is relatively small, we run 5-fold cross validation and report the average accuracy of the five validation sets as the final score. All experiments are run for totally 100 epochs training on a single NVIDIA GeForce RTX 3090 and the learning rate decreases after every 50 epochs by a factor of 0.1.

For GAT and KFN, we use ResNet-50 [12] and EffiecientNet-b5 [44] pretrained on ImageNet as backbones on CUB and CXR, respectively. In GAT, we crop the original image using three sets (large, medium and small) of window sizes (more details can be found in the supplementary material). Inside each set of window sizes, we run a sliding window algorithm and get k augmentation images for each image in the training set. Concretely, k is set to 2 for large, 3 for medium and 4 for small scale, which results in 9 augmentation images in total. When combining GAT and KFN, we use the GAT trained classifier as backbone in our KFN and fine-tune the KFN for only 20 epochs.

5.2 Evaluation on CUB-GHA

Ablation study. To measure the influence of GAT and KFN on the fine-grained classification, we design an ablation study on the CUB dataset where we train a ResNet-50 with cross-entropy loss as the baseline, and several variants by adding GAT and KFN training

modules to the baseline. From the results shown in Table 1, we observe that both GAT and KFN can improve the fine-grained classification accuracy by a large margin. GAT (with HA) improves the baseline model by 2.42% to 88%, which indicates that human gaze falls on areas containing discriminative features for classification. When using HA in KFN, the accuracy score is increased from 85.58% to 86.99%, which demonstrates that KFN integrates the knowledge of human attention successfully. To show the effectiveness and uniqueness of HA knowledge, we use two machine explanation methods Grad-CAM [36] and IG [43] as the saliency maps, replacing HA in GAT and KFN. HA surpasses both methods in the GAT and KFN modules, e.g. KFN (HA) gains 86.99% while KFN (IG) gains 85.66%. It indicates that human gaze contains unique knowledge that can not be acquired by the model itself. From the result of GAT+KFN, we observe that the combination of both exceeds using any of them alone.

Method	Acc.
MixUp [54]	86.23
CutMix [52]	86.15
SnapMix [14]	87.75
Ours (GAT)	**88.00**
OSME+MAMC[42]	86.30
TASN [58]	87.90
API [60]	87.70
ACNet [15]	88.10
Ours (KFN+GAT)	**88.66**

Table 2: Comparison with the state-of-the-art methods on CUB. **Top:** Comparison of GAT with data augmentation methods. **Bottom:** Comparison of GAT+KFN with attention-based models.

Method		Acc.
ResNet-50 [12]		85.58
GAT	Grad-CAM [36]	87.68
	IG [43]	87.73
	HA	88.00
KFN	Grad-CAM [36]	85.04
	IG [43]	85.66
	HA	86.99
GAT+KFN	HA	**88.66**

Table 1: Ablations study of GAT and KFN on CUB. "Acc." denotes the accuracy in %.

Comparison with state-of-the-art. We compare our proposed modules with several state-of-the-art methods. Note that for a fair comparison, we compare with the results of using ResNet-50 as the backbone and the input resolution of 448×448. First, we compare our GAT with other data augmentation methods, i.e., MixUp [54], CutMix [52] and SnapMix [14] in Table 2 (top). The difference between our GAT and other data augmentation methods is that we do not generate synthetic images. MixUp combines two images and their labels linearly, while the rest replace one part of the image with one part from other images. Our GAT simply extends the dataset with the cropped images, which introduces very low computation cost to train the classifier. Among all these works, training a ResNet-50 with GAT outperforms with other state-of-the-art augmentation methods and achieves an accuracy of 88%. Moreover, this better trained backbone can be combined easily with other framework to further improve the performance, for instance we combine it with our KFN and thus get better results.

We compare our full network with the attention-based methods on CUB in Table 2 (bottom). We choose these methods (OSME+MAMC [42], TASN [58], API [60] and ACNet [15]) due to their high performance and relevance in simulating human attention by attention modules. They apply attention modules to capture discriminative features from the intermediate output in the network, while we use and integrate the HA directly. For instance, [15, 42] applies several layers on the top of the output of the residual block to obtain the re-

Method	S3N [9]	S3N + GAT (Ours)	CrossX [26]	CrossX + GAT (Ours)	MMAL [53]	MMAL + GAT (Ours)
Accuracy	87.95%	88.91%	87.70%	88.51%	89.25%	**89.53%**

Table 3: Combining our GAT model with the state-of-the-art methods on CUB.

Figure 4: Illustration of model explanations using HA. Two improved examples and one failure example of our model are shown. For each example, we show the input and misclassification classes; HA saliency map, model explanation of our model, and the baseline model.

gion features; API [60] simulates the comparison behavior of humans as our participants do in the data collection in order to learn discriminative representations. Our full network outperforms all state-of-the-art models, achieving 88.66% compared to the attention networks API (87.70%) and ACNet (88.10%). The high performance of our KFN and GAT validates that human gaze can benefit a model's performance in the task.

We combine our module with other state-of-the-art models flexibly and thus improve the performance. In Table 3, we show our re-implementations with official code and our improvement by combining our GAT in S3N [9], CrossX [26] and MMAL [53] models. Please note that no HA information is needed in the inference phase. Our combination of MMAL and GAT improves MMAL from 89.25% to 89.53%. We improve CrossX from 87.70% to 88.51% and S3N from 87.95% to 88.91%, which also surpass the best results given in [9, 26].

Qualitative results. We show two examples from two classes whose accuracy is improved the most compared to the baseline model (vanilla ResNet-50), and one example of a class where our model fails to classify correctly in Figure 4. In the first example, the baseline model looks at the belly of an Orange Crowned Warbler and misclassifies it as a Nashville Warbler who also has a yellow fluffy belly. Our model instead focuses on the throat, which is discriminative between the two classes: an Orange Crowned Warbler has a yellow throat, while a Nashville Warbler has a clear mixture of gray and yellow colors on its throat. In the second example, the discriminative feature is the tail. The baseline model mistakes the background as the tail, while our model localizes the tail successfully. Moreover, our model explanation is also more compact and similar to the human saliency map. In the third example, we show a failure of our model: Our model attends to the feet instead of beak which causes the misclassification of a Caspian Tern as an Elegant Tern. Although our model aligns with the human attention, it puts more weight on the feet of birds, since the color of feet is an important feature for distinguishing between a Caspian Tern and a Common Tern (or an Artic Tern).

5.3 Evaluation on CXR-Eye

Comparison with state-of-the-art. The state-of-the-art work on CXR-Eye [19] uses the Efficient-b5 [44] as the classifier, however, it deploys random splits to create training, validation and test sets. For a fair comparison, we re-run its network using our 5-fold cross validation setting and report the average of five validation accuracies as the score for this

method. The result of this baseline is 70.97%. When implementing GAT, the result is improved to 71.86%; when implementing KFN, the accuracy is improved by 3.45% to 74.42%. The full model (GAT+KFN) achieves 75.35% exceeding Efficient-b5 [19] by 4.38%. When comparing the performance boost from GAT and KFN, the KFN improves the model on CUB more than GAT. The reason for the difference is how the gaze data is collected.

In CXR-Eye, the gaze data of the radiologist is collected in an interpretation routine. From the examples shown in Figure 5 (sec. column), we see that fixations spread over many locations (light blue area). These locations may play an important role in diagnoses, but GAT localizes the area that the radiologist fixates for relatively longer time. KFN can integrate the knowledge of all potential locations therefore improves the performance by a larger margin.

Qualitative results. To study the influence of integrating HA into the network, we compare the model explanation (Grad-CAM [36]) of each branch in KFN and the qualitative results are shown in Figure 5. From the figure, we see that the HA branch follows more the human attention while the image branch is focusing different areas.

In the first example (top), human attention focuses more on the left side than the right and the HA branch also does, while the image branch looks more on the right side. The image branch in the second example concentrates on a wrong area, but the HA branch corrects the attentive area to the right. Therefore, KFN improves the performance compared to a model only using images. Most importantly, incorporating gaze knowledge helps to increase the trust and acceptance of the model-based decision in applications such as medical diagnostics, since the model aligns with human behavior.

Figure 5: Illustration of the influence of using HA in model explanation. **Left to Right:** the original Chest X-ray image; HA saliency map; Model explanation of the Image Branch (w/o HA knowledge) and Model explanation of the HA Branch.

6 Conclusion

In this work, we investigate human attention in classification tasks on the CUB and CXR datasets. In particular, we collect a new gaze dataset, CUB-GHA, and show that human attention focuses on the discriminative regions for a fine-grained classification task. To study the hypothesis that human attention helps a model in the decision-making, we propose the Gaze Augmentation Training and Knowledge Fusion Network which integrate human attention knowledge into the network. Our proposed method improves the accuracy in classification by a large margin on both datasets, showing the general validity of our methods. Thus, our work indicates that human attention provides hints on distinct features in different classification tasks.

The aim of our work is to demonstrate the potential benefit of human gaze data in classification. As a by-product of this work, we provide the research community with a gaze-enriched dataset CUB-GHA, which can be incorporated with other existing comprehensive annotations (textual explanations, attributes and bounding boxes, etc.). Researchers can therefore validate multiple applications, where human gaze is required in the interaction with a machine.

7 Acknowledgement

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2015.

[2] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *ECCV*, 2018.

[3] Reuben M Aronson, Thiago Santini, Thomas C Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot shared manipulation. In *HRI*, 2018.

[4] Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *ITSM*, 2017.

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 2018.

[6] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ETRA*, 2020.

[7] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. In *NeuIPs*, 2019.

[8] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 2017.

[9] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019.

[10] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *NeuIPs*, 2018.

[11] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *NeuIPs*, 2019.

[14] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *AAAI*, 2021.

[15] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *CVPR*, 2020.

[16] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[17] Atsushi Kanehira and Tatsuya Harada. Learning to explain with complemental examples. In *CVPR*, 2019.

[18] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy Wu, Matthew Tong, Arjun Sharma, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth Krupinski, and Mehdi Moradi. Eye Gaze Data for Chest X-rays (version 1.0.0), 2020. URL https://physionet.org/content/egd-cxr/1.0.0/.

[19] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 2021.

[20] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze embeddings for zero-shot image classification. In *CVPR*, July 2017.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[22] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *Journal of Vision*, 2016.

[23] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *ICCV*, 2017.

[24] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[25] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding, and Yuanqing Lin. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *AAAI*, 2017.

[26] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S. Davis, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *ICCV*, 2019.

[27] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 2014.

[28] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *TPAMI*, 2014.

[29] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *NeuIPs*, 2014.

[30] Anneli Olsen. The tobii i-vt fixation filter. *Tobii Technology*, 2012.

[31] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018.

[32] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual review of neuroscience*, 1990.

[33] LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 2020.

[34] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[35] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, 2006.

[36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[37] Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *ACL*, 2020.

[38] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. In *ICLRW*, 2015.

[39] Ali Shafti, Pavel Orlov, and A Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *ICRA*, 2019.

[40] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *CVPR*, 2015.

[41] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*. PMLR, 2017.

[42] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018.

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*. PMLR, 2019.

[45] Joan N Vickers. *Perception, cognition, and decision training: The quiet eye in action*. Human Kinetics, 2007.

[46] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[47] Daniel Weber, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling location proposals of unknown objects through gaze information for human-robot interaction. In *IROS*, 2020.

[48] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[49] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *WACV*, 2020.

[50] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

[51] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeuIPs*. Curran Associates, Inc., 2020.

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[53] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MMM*. Springer, 2021.

[54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[55] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017.

[56] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI*, volume 2020, 2020.

[57] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017.

[58] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, 2019.

[59] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[60] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *AAAI*, 2020.

Supplemental Materials

In this document, we provide technical details about our data collection and experiments. First, we explain how we set the standard deviation of the Gaussian distribution in the Human Attention (HA) saliency map generation and show more analyses on gaze data including the relationship between human fixation points and the discriminative attributes of birds. In addition, quantitative and qualitative comparisons between the model explanations (MEs) and HA are demonstrated. In the second section, we introduce implementation details (e.g. sliding window sizes) in the Gaze Augmentation Training (GAT).

1 CUB-GHA

1.1 HA Saliency Map Generation

Figure S1 illustrates a human observing an image on the eye-tracker display. As mentioned in the paper, we post process every fixation location as a Gaussian distribution $N(\mu, \sigma^2)$ on the HA saliency map, where σ is 75 pixels (in the display's resolution). We calculate the standard deviation σ as follows. In our experiment setup, the distance d between the human eye and the eye-tracker display is 60 cm, and the visual angle θ is set to $2°$ following [45]. In this case, $l = \tan 2° \cdot d = 21$ mm. According to the settings of display, in the horizontal direction the length of the display is 530 mm and the resolution is 1920 pixels. Therefore, we can get that $l = 21$ mm covers approximately 75 pixels on the display. We set 75 pixels as the standard deviation with the image rescaled to the display resolution (1920 × 1080). The saliency map is rescaled to its original size afterwards.

Figure S1: Illustration of a human observing an image on the eye-tracker display.

1.2 Gaze Data Analysis

In this section, we validate that the attributes discovered by our collected human gaze data are discriminative for the fine-grained classification. CUB includes ground-truth attributes for each image and they are 312-dimension binary vectors. We use them to conduct the ground-truth discriminative attributes of each bird class in the dataset. There are 100 comparison pairs in the data collection experiments, and we compare each image from the first class with every image in the second class. For instance, if there are M images in the first class and N in the second one, there are in total $M \cdot N$ combinations between the two classes. For each combination, we conduct a comparison attribute vector where 1 is set if that attribute entry is the same for both images, or 0 if not the same, i.e. the comparison attribute vector is also

Figure S2: Histogram of the number of focused bird body parts in CUB-GHA. **Y-axis** refers to the amount of images with the certain number of parts (**X-axis**).

a 312-dimension binary vector. We sum $M \cdot N$ comparison attribute vectors together to have one 312-dimension vector representing the ground-truth discriminative attribute for these two classes. For instance if the attribute `has-wing-color::brown` in the comparison vector is 354, it means that the attribute `has-wing-color::brown` differs in the 354 image pairs. In the end, we group the attributes into seven body parts (head, beak, breast, belly, back, wing, leg). For example, we sum up all the attribute values in the comparison vector that are related to the wing, and the sum represents the difference of the wing between the given two classes. The body part with the highest sum is the most discriminative body part between the two classes.

When our participants look at the image, they always focus on the discriminative body parts of the bird. The body part which human gaze falls in should contain the largest number of different attributes between the compared two classes. With the help of body part center coordinates in each image, we can assign every fixation (collected for this image from five participants) to its nearest body part according to the distance between the center coordinate and the fixation coordinate. In Figure S2, we show the histogram of the number of focused bird body parts on the whole CUB-GHA dataset. We see that there are three body parts focused by humans in 3855 images. Most of the images (92.52%) include less than five parts focused in the dataset. In very few images, our participants view all seven parts of the bird. In each image, we sum up the duration of fixations belonging to one body part and use it to represent the amount of human attention on that part. A longer duration sum indicates more attention participants have paid. We rank the seven body parts for each image according to the duration sums and calculate the rate that the top-k focused body parts hit the most discriminative one (which is conducted from the ground-truth attributes). The hit rate is shown in Table S1. From the results, we see that our participants discover the most discriminative body part in 84.4% of the images correctly. Within four parts that participants consider to be important for the classification, the ground-truth distinct body part is found in 98.3% of the images. This result shows that human gaze data in CUB-GHA hints on discriminative body parts/attributes in the classification.

Top-k	1	2	3	4
Hit rate (%)	84.40	93.60	97.18	98.31

Table S1: Hit rate of the most discriminative body part. Top-k refers to the k longest focused body parts by humans in CUB-GHA.

Figure S3: Modified images in the Keep and Retrain procedure. The pixels are inserted according to the importance in the estimation maps. **Top to bottom**: importance estimation maps (saliency maps), modified images using top 5%, 10% and 20% important pixels in saliency maps. **Left to right**: HA, ME-Gradient-based CAM (Grad-CAM) [36], Class Activations Maps (CAM) [59], InputXGradient (IxG) [41] and Integrated Gradients (IG) [43].

1.3 Comparison between ME and HA

In this section, we provide more details and results of comparing MEs and HA. We use the KAR (keep and retrain) procedure [13] and the concrete procedure works as follows: given an input image $I \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 3}$ and the importance estimation map $A \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 1}$, where H and W represent the width and height of input image, respectively; A can be the HA or ME saliency map. We construct a mask $M \in \mathcal{R}^{\mathcal{H} \times \mathcal{W} \times 1}$ to filter the pixels in I. First, we sort A in a descending order to A^R according to the attention values. Then we binarize A by taking the top p percent of pixels in A^R as one and others as zero:

$$M(x,y) = \begin{cases} 1.0, & \text{if } (x,y) \in P \\ 0.0, & \text{otherwise} \end{cases},$$

where P are the indices of top ranked p percent pixels. We apply the mask M to filter the corresponding image I in the training and testing set: $I' = M \odot I$, so that only the top p percent of the most important features are observed by the network. After such a modification of the dataset, we train a new model and compare the test accuracy. This procedure aims at evaluating whether the important feature estimated by A (i.e. model or human attention) is critical to the classification or not. A good estimation A encodes important features in a small amount of pixels. In other words, a higher accuracy with such small amount of pixels indicates that the given features are more important. We generate the new dataset using an insertion percentage $p = [5,10,15,20,15,30,50,70,90]$ and train the vanille ResNet-50 [12] using the same hyper-parameters as in the baseline training. We run this procedure three times independently from random initialization for each estimation map and report the average accuracy on the test set.

Figure S3 illustrates the qualitative results of the modified images using HA and MEs. These differences can be observed when using 5% and 10% as insertion percentages. If we compare HA and MEs, they focus on a similar area after 20% pixels are inserted: the wing and head parts. When comparing among MEs, CAM and IxG are similar to Grad-CAM and

IG, respectively. In this example, Grad-CAM/CAM pays more attention to the head, while IG/IxG focuses more on the body. From the qualitative comparison results, we see that the HA and MEs estimate different parts of the bird as being the most important ones for the classification task, especially regarding the first 10% important regions.

We also conduct a quantitative similarity comparison between HA and MEs. We evaluate on different metrics: Kullback-Leibler divergence (KL-D), correlation coefficient (CC) and similarity (SIM), which are often used in comparisons of how similar two images are [6]; rank-correlation (Rank-Co) as introduced in [8]; shuffled AUC metric (sAUC) evaluating every pixel in saliency maps as a classification task; information gain (IG) measuring the performance over a baseline [6, 33]. CAM is very similar to Grad-CAM, e.g. Grad-CAM achieving 0.565 on CC and 1.242 on KL-D, while CAM achieving 0.563 and 1.248, respectively. Additionally, we observe IG and IxG achieving similar performances on these metrics, i.e. 0.699 for IG v.s. 0.694 for IxG on CC, and 1.318 for IG v.s. 1.310 for IxG on KL-D. These similarities can be seen from the qualitative results as well. From all different metrics, we see that the Grad-CAM tends to be the most similar to HA, as Grad-CAM achieves the highest scores in all six metrics. This is consistent with the results from the KAR that Grad-CAM achieves the best performance among all MEs.

	KL-D ↓	CC ↑	SIM ↑	Rank-Co ↑	sAUC ↑	IG ↑
CAM	1.248	0.563	0.399	**0.761**	0.460	0.938
Grad-CAM	**1.242**	**0.565**	**0.415**	**0.761**	**0.508**	**1.376**
IG	1.318	0.546	0.361	0.699	0.436	0.921
IxG	1.310	0.543	0.375	0.694	0.461	1.001

Table S2: Similarity comparison between MEs and HA saliency map. (↓: the lower the better; ↑: the higher the better.)

2 GAT Experiments

	Small	Medium	Large
CUB-GHA	(123,134) (134,123) (123,123) (134,134)	(174,190) (190,174) (174,174) (190,190)	(246,264) (269,246)
CXR-Eye	(87,95) (95,87) (95,95) (87,87)	(123,135) (135,123) (123,123) (135,135)	(180,190) (190,180)

Table S3: Sliding window size used in GAT.

Concrete sliding windows sizes (w, h) used for each dataset in GAT experiments are listed in Table S3. For the CUB-GHA dataset, we choose the sliding window sizes based on the averaged size of bird bounding boxes: the width is 246 and the height is 269 if images are resized to 448×448. Therefore, we use 246 and 269 as sizes for the large scale. The medium window size is conducted using the factor of $\frac{\sqrt{2}}{2}$ to have the half of the bounding box area, i.e. we use 174 and 190 as window size options. The factor used in the small scale is 0.5. For the CXR-Eye dataset, we choose 0.8 and 0.85 as factors with respect to the resized image size 224×224 for the large window size, i.e. two options are 180 and 190. Similarly, factors for the medium window size are 0.55 and 0.6. The small window sizes are scaled based on the medium window sizes by the factor of $\frac{\sqrt{2}}{2}$. The motivation of using different sliding window sizes is to get different parts which are discriminative for the classification. To avoid very similar cropped areas, we choose 0.25 as the iou threshold in the non-maximum

suppression. Table S4 lists the ablation study of using different numbers of cropped areas (k) in the augmentation training on two datasets. (2,2,2) denotes that two cropped areas are picked up from each window scale to form the augmentation training set. We choose (2,3,4) as the final setting since it gives relatively better results on both datasets. Figure S4 illustrates the augmentation images using the setting (2,3,4) in three sets of window scales on both datasets.

(L,M,S)	CUB (%)	CXR (%)
(2,2,2)	87.50	71.03
(2,3,2)	88.06	71.58
(2,3,3)	88.00	71.86
(2,3,4)	88.00	72.21

Table S4: Results of using different window size settings on CUB-GHA and CXR-Eye. The number of windows used in large, medium and small size is shown on the left. The accuracy is in %.

Figure S4: Illustration of cropped images used in the Gaze Augmentation Training. **Left and Right:** HA saliency maps used for augmentation on CUB-GHA and CXR-Eye. **Middle:** cropped images in three scales (large, medium and small).