Teacher-generated spatial-attention labels boost robustness and accuracy of contrastive models

Yushi Yao*
Waymo
yushiy@waymo.com

Chang Ye*
Google
yechang@google.com

Junfeng He†
Google
junfenghe@google.com

Gamaleldin F. Elsayed†
Google
gamaleldin@google.com

Abstract

Human spatial attention conveys information about the regions of visual scenes that are important for performing visual tasks. Prior work has shown that the information about human attention can be leveraged to benefit various supervised vision tasks. Might providing this weak form of supervision be useful for self-supervised representation learning? Addressing this question requires collecting large datasets with human attention labels. Yet, collecting such large scale data is very expensive. To address this challenge, we construct an auxiliary teacher model to predict human attention, trained on a relatively small labeled dataset. This teacher model allows us to generate image (pseudo) attention labels for ImageNet. We then train a model with a primary contrastive objective; to this standard configuration, we add a simple output head trained to predict the attention map for each image, guided by the pseudo labels from teacher model. We measure the quality of learned representations by evaluating classification performance from the frozen learned embeddings as well as performance on image retrieval tasks (see supplementary material). We find that the spatial-attention maps predicted from the contrastive model trained with teacher guidance aligns better with human attention compared to vanilla contrastive models. Moreover, we find that our approach improves classification accuracy and robustness of the contrastive models on ImageNet and ImageNet-C. Further, we find that model representations become more useful for image retrieval task as measured by precision-recall performance on ImageNet, ImageNet-C, CIFAR10, and CIFAR10-C datasets.

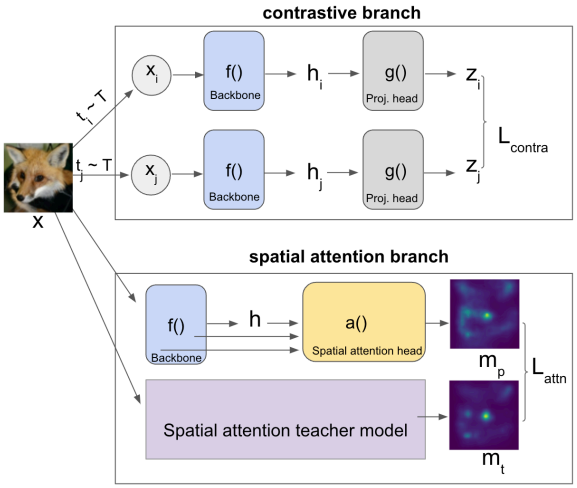

Figure 1. Illustration. A teacher model is trained to predict human spatial-attention from a small dataset. Then the model is used to provide attention labels for larger dataset, which are used as additional targets for contrastive models.

1. Introduction

Deep learning models have made significant progress and obtained notable success on various vision tasks. Despite these promising results, humans continue to perform better than deep learning models in many applications. A notable reason is that deep learning models have a tendency to learn "short-cuts", i.e., giving significance to physically meaningless patterns or exploiting features which are predictive in some settings, but not causal [20]. Examples include focusing on less significant features such as background and textures [13]. These models yield representations that are less generalizable and lead to models that are highly sensitive to small pixel modulations [42].

Human vision on the other hand is known to be much more robust and generalizable. One major difference between human and machine vision is that humans tend to

*Equal technical contribution.
†Equal leadership and advising contribution
Correspondence to:
junfenghe@google.com & gamaleldin@google.com

focus on specific regions in visual scene [45]. These locations often reflect regions salient or useful to perform a specific vision task. Machines, instead, initially place equal significance to all regions. A natural question is: will it be beneficial if machine vision models is guided by human spatial attention?

Human spatial attention has been shown to benefit computer vision models in supervised tasks, such as classification [32]. Yet, it is still a question whether adding a form of weak supervision in the form of human spatial attention could similarly benefit self-supervised models that are trained end-to-end. Self-supervised models typically need a large amount of data to yield good representations. To test if training weakly supervised models with human spatial attention cues, we will need to collect a large volume of human spatial attention labels, which is a very expensive process that requires either using trackers to record eye movements [5,43,52] or asking humans to highlight regions that they attend to [25, 27]. This process is prohibitively tedious and costly for datasets with millions of examples.

In this work, we test the hypothesis that a weak supervision in the form of human spatial attention is beneficial for representation learning for models trained with a contrastive objective. Inspired by knowledge distillation and self-training using teacher models [47, 49], we address the challenge of obtaining spatial attention labels on large scale image datasets by using machine pseudo-labeling. We train a teacher model on a set of limited ground truth human spatial attention labels, and use this teacher model to generate spatial attention pseudo-labels for the large ImageNet benchmark. We are then able to utilize the generated spatial attention maps in the contrastive models, and discover that this approach yields representations that are highly predictive of human spatial attention. Further, we find that the learned representations are better as measured by higher accuracy and robustness on classification downstream tasks, and higher precision and recall on image retrieval tasks. Interestingly, we find that the gains from using teacher models to provide pseudo labels are larger than using the limited ground truth human labels directly when training contrastive models, and the gains are larger for contrastive models than when applying same method to supervised models.

In summary, our contributions are as follows:

- We create a dataset with spatial attention maps for the ImageNet [37] benchmark by first training a teacher model to predict human spatial attention labels from Salicon dataset [25] and then use the model to label ImageNet examples
- We use spatial-attention labels from the teacher model as an additional prediction target to models trained with contrastive objective.
- We find that the proposed method can learn better representation, leading to better accuracy and robustness for downstream classification tasks (on ImageNet and ImageNet-C), and better performance on retrieval tasks (on ImageNet, ImageNet-C, CIFAR-10, and CIFAR10-C).

Trained teacher model is available at:
https://github.com/google-research/google-research/tree/master/human_attention/

2. Related work

Contrastive learning: Contrastive learning has gained popularity in the past few years for self-supervised and semi-supervised representation learning. In general, contrastive learning aims to learn similar representations for similar data pairs and different representations for different pairs. SimCLR [6] utilized MLP projection heads and strong data augmentation for constructing similar pairs and have demonstrated great gains in image classification downstream tasks. To form the contrastive loss for each mini-batch with N examples, the similar data pairs are constructed from two augmentations of the same image and different pairs from the other images within a batch, and then computing the NT-Xent loss. A different formulation is used in [50] by encouraging the empirical cross correlation of the representations of two versions of augmented mini-batch to be close to identity. In [17], it is further proposed to build large dictionaries for self-supervised learning (MOCO), and moreover in [7], better results are achieved on image classification and object detection tasks when combining advances from SimCLR and MOCO. Follow up work by [14] managed to obtain good performance without the need for dissimilar pairs by encouraging the representation of similar pairs across two versions of the network (trained network and exponential moving average version) to be similar. Further, in [8] it is shown that simple Siamese networks can still learn good representations without the need for dissimilar pairs, large batches or momentum encoders, etc .

Human spatial attention: Human visual system has developed an attention mechanism that focuses on regions in the visual space that are of interest or highly informative to the vision task [12, 48]. Eye trackers are often used to collect human spatial attention [5, 43, 52]. Many gaze data sets [2] have been collected with these eye trackers. Besides eye trackers, human spatial attention data can also be collected via mouse tracking [25, 27], e.g., users see a blurry version of an image, then click on regions they want to see more clearly, mimicking human's peripheral vision based on neurophysiological studies [19, 27]. Both eye tracking and mouse tracking methods are very expensive, which limit the number of examples in those datasets. Due to the relatively lower cost of mouse tracking, one can often generate relatively larger attention data from this method than

eye tracking. For example, Salicon [25] dataset is one of the largest spatial attention datasets, contains around 20K images, each with attention labels from 50-60 participants, via a mouse tracking system, under free viewing setting . Yet, this data is still orders of magnitude smaller than those needed to train contrastive representation learning models.

Spatial attention/saliency prediction models aim at predicting which areas in an image will be salient to human attention and attract eye fixation, usually with collected gaze/attention data as ground-truth. Early works in saliency prediction usually define saliency through a set of hand crafted features such as color difference, contrast, intensities, etc. [21, 28]. Recent works [1, 19, 24, 30, 31, 36] leverage the power of deep neural networks and are often trained/fine-tuned on large scale gaze data sets like Salicon [2, 25].

Spatial attention of computer vision models: Spatial attention in neural networks can be mainly categorized into post-hoc attention like class activation map (CAM) ([51]), and trainable attention (e.g., [15, 23, 46]). Post-hoc spatial attention methods have been proposed to estimate regions in the image that are important or give rise to model decisions, often for model interpretation. In supervised settings where classification labels are known, the simplest and most direct method is class activation map (CAM) [51]. CAM uses class labels to extract the feature map that is most informative about the true class of an image. Grad-CAM [39] generalizes the CAM to apply to any model with any downstream task. [38] proposed to use Grad-CAM to design augmentation policies in self-supervised learning to tackle weak performance in complex scene images with many objects. ContraCAM [35] applies Grad-CAM assuming downstream task of contrastive learning, thus allowing computing spatial attention maps with no class label supervision. In [35], it is proposed to utilize the spatial attention information learned from ContraCAM to design data augmentation strategies to discourage contextual and background biases in a scene. Unlike [38] and [35] that uses spatial attention to design augmentation policies, here we focus on an end-to-end framework to predict spatial attention targets instead.

Recently, there are also several papers [10, 11, 16, 33, 41, 44]) exploring the similarity/difference of model spatial attention vs human spatial attention, for the task of VQA (visual question answering) [10], object detection [11], reinforcement learning [16], etc. Among existing works, [32] is the most relevant to ours, since it conducted experiments to use human spatial attention to supervise model spatial attention, for three tasks (salient object segmentation, video action recognition and fine-grained image classification) and demonstrated that human spatial attention is beneficial. However, it still remains a question whether such benefits could be extended to contrastive representation learning. In terms of strategy to utilize spatial attention labels, in [32], attention labels are used as spatial weighting, while we designed auxiliary task to predict spatial attention labels.

Teacher model pseudo-labeling: Previous work on knowledge distillation and machine self training has demonstrated that machine teaching machines approaches may address the challenge of labeling large datasets. In image classification, [49] demonstrated that training a model to classify images then use that model to provide pseudo-labels improved classification performance. Related idea is applied in [47] for language models, which showed that the knowledge learned by language models pretrained on text corpus could be distilled to generate new datasets for common reasoning, and that training common reasoning models on this new data largely improve performance. Inspired by these successes, we train a teacher model on smaller human attention data and use this model to generate new spatial attention pseudo labels for ImageNet benchmark (see Figure 1 and Figure 2).

3. Methods

3.1. The Contrastive Learning Framework

Contrastive learning is one of the most popular self-supervised representation learning methods. It learns an embedding space so that similar data (positive) pairs are mapped to be close in the embedding space and different (negative) pairs are mapped to be far away. In practice, positive pairs are often generated by applying data augmentation to one image like adding noises, cropping, etc and negative pairs are different examples in the mini-batch. Among contrastive learning methods, SimCLR framework [6] has shown solid performance, which we choose here as our main method. For each batch, the images are augmented in two different ways. Then we feed them into ResNet feature extractor backbone and compute NT-Xent loss to minimize the difference between augmentations of the same images and maximize the differences between different images. NT-Xent loss can be computed as follows [6]:

$$l_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{k \neq i} \exp(sim(z_i, z_k)/\tau)} \quad (1)$$

where z_i is the embedding of ith example, and $sim(\cdot, \cdot)$ is the cosine similarity between two embeddings.

3.2. ImageNet-Attn: spatial attention maps for ImageNet generated by a teacher model

To train contrastive learning models, we need large datasets with millions of examples. Thus, to address the question whether human attention is beneficial for contrastive models, we will need large human spatial attention labeled dataset. However, there is no spatial attention dataset available with that size on typical large image

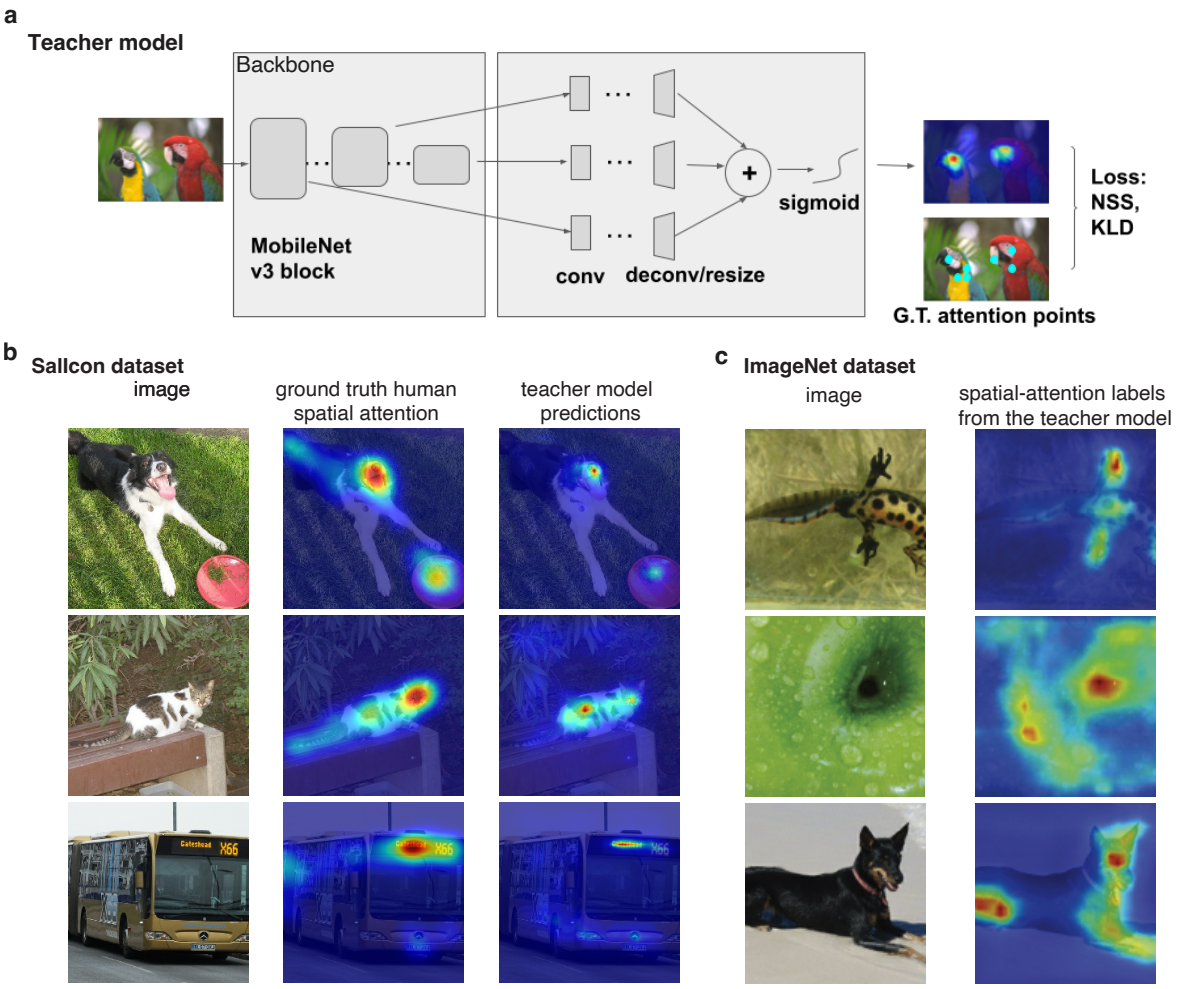

Figure 2. (a) The architecture of teacher model to predict attention. The teacher model was trained from random initialization to predict ground truth (G.T.) human spatial attention labels. (b) Examples from Salicon dataset, with G.T. human attention heatmaps, and teacher model predictions. (c) Examples of spatial-attention pseudo-labels generated by the teacher model for ImageNet dataset (ImageNet-Attn).

benchmark such as ImageNet. The largest attention data set available is Salicon [25], consisting of twenty thousand examples only, orders of magnitude smaller than ImageNet, which has around 14 million examples.

To generate attention labels for ImageNet images, we train a teacher model on Salicon attention data set to predict human attention ground truth labels (see Figure 2b). Then we use this teacher model to create a new dataset (ImageNet-Attn) with spatial attention maps for ImageNet (see Figure 2c for some examples). The teacher model architecture is illustrated in Figure 2a. Specifically, we used MobileNet-V3-small as the backbone, and embeddings from 4 layers (conv 2,4,6,8) are extracted. On each embedding, we applied two conv layers (the first conv layer with 3×3 kernel, number of channels matching the input embedding, max pooling of 3 and relu; the second conv layer with 1×1 kernel , 1 channel and relu). We bilinearly resized the output to the same resolution of input. The result of the 4 branches are then summed to yield 1 feature, followed up by a sigmoid function to obtain spatial-attention map. Note that using multiple intermediate (both early and late) layers is one of the key ingredient to make the spatial attention teacher model successful. This strategy mimics human visual attention, which is known to be affected by both low-level characteristics such as color, intensity, texture, as well as high level characteristics like shape and object, etc. Essentially, our teacher model has a similar architecture as in [36], but is further simplified with more efficient backbone, less channels/layers, etc, so that it can be trained from scratch on Salicon data with randomly initialized backbone to avoid any leak of class label information, while most existing attention prediction models [1,36] needs to finetune with a pretrained classification backbone network.

3.3. Training contrastive models with spatial attention maps

Spatial Attention Branch The overall training framework is shown in Figure 1. This spatial attention prediction framework consists of two branches: the contrastive branch and the spatial attention branch. The contrastive branch is the same as the original SimCLR method [6], which applies augmentations to image x to get different variants x_i and x_j, and learns the representation h_i and h_j via a feature extractor backbone network (e.g., ResNet), then use a projection head to map h_i/h_j to z_i/z_j, where the contrastive loss is applied.

For the spatial-attention branch, it takes as an input not only the final embedding h, but also early intermediate layer embeddings (following the same reason of including both high level and low level visual cues as in teacher model), and predicts an spatial attention heatmap m_p. More specifically, we apply a global average pooling on the output of the last three blocks of the ResNet backbone. Then, we select the max channel for each of three block output (after average pooling), and resize with bilinear interpolation to the image resolution, which has some resemblance to CAM approaches [51]. Finally we stack the representations together, pass them into a linear readout layer (with a bias term), and use the output as our final spatial attention prediction m_p. We keep the spatial attention head as simple as possible so that the guide on attention head output can be back propagated to representation more directly.

We use the ImageNet-Attn data generated by the teacher model as target for each ImageNet example, denoted as m_t, we can then train the network spatial attention output m_p to be close to m_t. We hypothesize that this method regularizes the training of the feature extractor backbone rather than explicitly enforce the network to generate masked representations that match the spatial attention maps. Note that for attention branch, there is no augmentation applied to each image x, since human attention is not invariant to transformation (e.g., a human looking at a cropped image may attend to different region compared to a consistent crop of human attention map of the original image).

Loss function The loss function L consists of two terms:

$$L = L_{contra} + L_{attn} \tag{2}$$

L_{contra} is the contrastive loss, or more specifically $L_{contra} = \sum_{i,j} l_{i,j}$ and $l_{i,j}$ is defined in Equation (1), the same as in [6]. L_{attn} is the attention regularization loss, and more specifically

$$L_{attn} = \sum_i (\lambda KLD(m_i^p, m_i^t) - \beta NSS(m_i^p, p_i^t)) \tag{3}$$

where λ and β are two weighting parameters ($\lambda = 1.0$ and $\beta = 0.1$ in our work). m_i^p is the predicted 2D spatial-attention map from the attention head of our proposed model for i-th example (original image, not augmented ones), m_i^t is the pseudo spatial-attention target map predicted by the teacher model for i-th example, and $KLD()$ is the KL divergence [1]. Besides KLD loss, we also use $NSS()$, the Normalized Scanpath Saliency loss [4]. Those two losses are typically used for human attention predictions [3,9]. KLD is often used to match the spatial-attention target distribution, and NSS is typically added on top as it is generally observed to help generate attention maps that are in perceptual agreement with human judgements (see [3,34]). For NSS loss, the larger the better, so there is a negative sign before it. NSS loss needs gaze/attention points instead of heatmap as ground-truth, so we extract pseudo gaze/attention points p_i^t from the attention heatmap m_i^t. To obtain p_i^t, we first extract the point with highest value in current spatial-attention map m_i^t, then generate a new spatial-attention map by subtracting a Gaussian blur around the extracted point from the current attention map. The process is repeated with the new attention map until the maximal value of the attention map is smaller than a threshold (more specifically, $0.2 * max\_sal$, where max_sal is the maximal value in m_i^t). [2]

Table 1. ImageNet Top-1 classification accuracy for different models (mean ± SE for 3 seeds, except for * which means best result from all runs).

Model	Accuracy (%)
Contrastive	67.61 ± 0.04
Contrastive attn. teacher	**68.23 ± 0.08**
Contrastive attn. co-train	66.35 ± 0.12
Contrastive attn. with explicit attention block [32]	61.15*
Contrastive attn. with self attention mask [32]	49.65*
Contrastive attn. with ContraCAM [35] as attention	67.71*
Supervised	75.91 ± 0.10
Supervised attn. teacher	76.02 ± 0.04
Supervised (ResNet-18)	69.17 ± 0.07
Supervised (ResNet-18) attn. teacher	69.30 ± 0.04

4. Experiments

4.1. Implementation details

In our experiments,in Eq.1, we choose $\tau = 0.1$ and $N = 2048$. In Eq.2, we used equal weighting of 1.0 for the SimCLR and attention losses.

The weight for KLD and NSS losses are 1 and 0.1, re-

[1] m_i^p is normalized when computing KLD, i.e., divided by its pixel value sum, so that it becomes a distribution with all pixel value sum equal to 1. Similarly for m_i^t.

[2] Our process of extracting gaze points from heatmaps is to inverse the process of generating heatmaps from gaze points. Due to the uncertainty about the gaze locations, researchers typically apply Gaussian blur for each gaze point, then sum all Gaussians to generate the heatmap (see [3, 26]). We follow the inverse of this process to obtain gaze points from heatmap, similarly as done in [22].

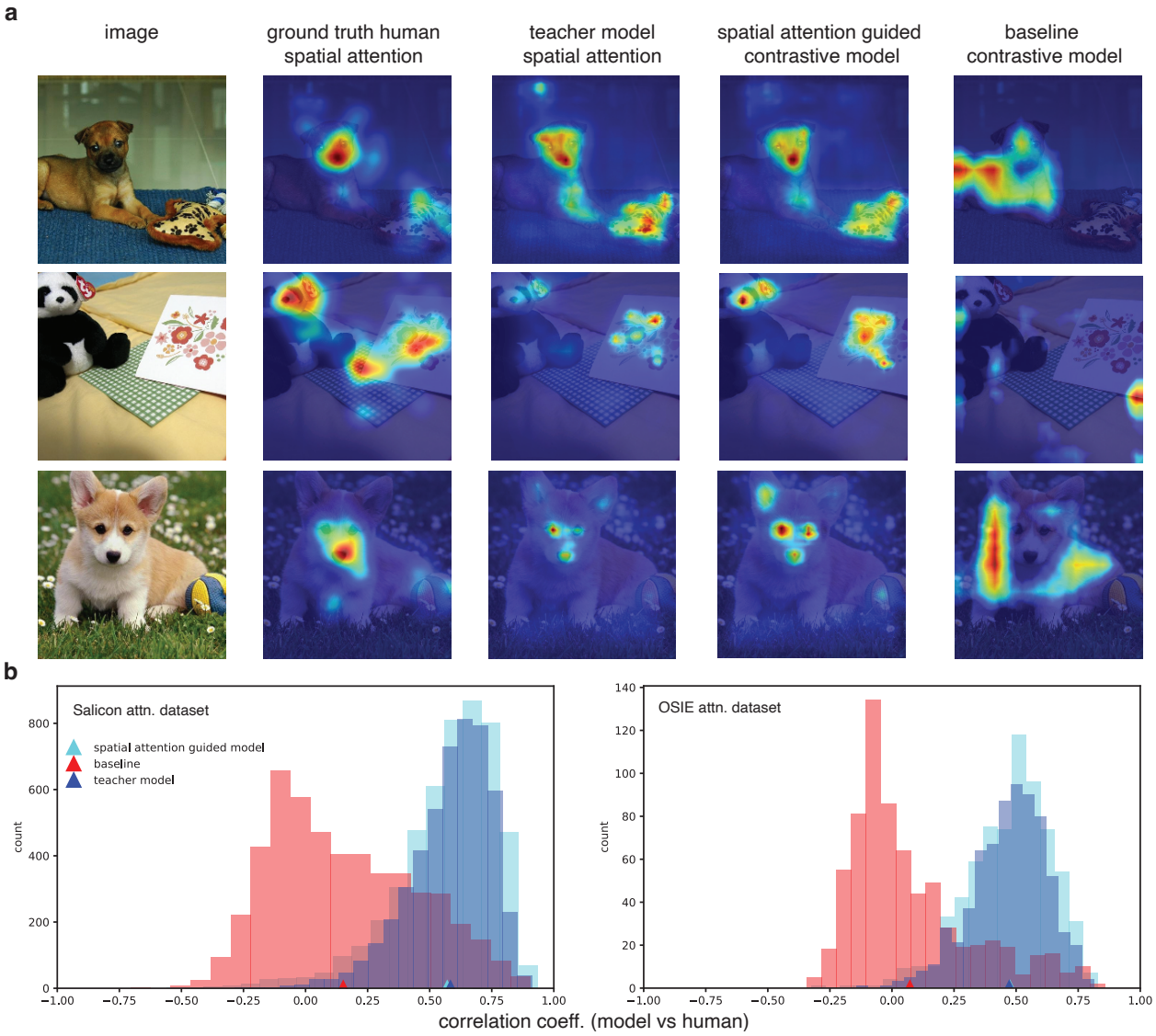

Figure 3. (a) Examples comparing spatial attention maps predicted by different models vs ground truth human attention data on OSIE dataset [43]. (b) Distribution of correlation coefficients between model predicted attention maps vs ground truth human attention maps on Salicon [25] validation set (left) and OSIE dataset (right) [43]. Three models are shown: (1) Teacher model trained on Salicon [25] training set. (2) Baseline SimCLR model (3) The proposed spatial attention guided SimCLR model trained with pseudo-attention labels provided by the teacher model. Please note that both baseline and our model are trained to predict target attention labels, with main difference that the baseline model has a stop gradient placed between the prediction layer and network features to prevent spatial-attention targets to inform network features. Also note that our teacher model is trained on Salicon where the attention is collected by mouse tracking, while the results in (a) and (c) here are on OSIE data, where the attention are collected by eye trackers. The results demonstrated both the teacher model and the attention head in proposed model has a good generalization capacity for new kinds of attention labels.

spectively. We obtained those weights by hyperparameter search from 0.1–5 for KLD and 0.01–1 for NSS, based on ImageNet accuracy of a separate dev set (80%-20% split on training set for parameter search with 20% as dev set).

4.2. Spatial attention guided models are highly predictive of human attention

In this section, we explore whether the use of auxiliary teacher model to provide spatial attention pseudo labels on ImageNet better aligns contrastive model's attention with human attention. We define aligning model spatial attention here as the ability to predict spatial attention mask from

the model backbone features by a simple readout layer discussed in Section 3. We trained two ResNet-50 backbones using the SimCLR objective from [6]. We added additional spatial-attention losses as discussed in Section 3. For the baseline model, we placed a stop gradient operation between the backbone features and the attention projection head to prevent attention information from leaking to the backbone features, whereas for the attention guided model, we allowed the learned attention gradients to flow back to the backbone.

We evaluated the degree the predicted attention maps is aligned with human attention by performing correlation analysis, which is typically used to compare human and model spatial attention. We measured Pearson's correlation coefficient between the model predicted attention and ground truth human spatial-attention maps from Salicon [25] validation set (Note that the teacher model is trained with Salicon training set). We find that both teacher model and spatial attention guided contrastive model are highly correlated with human attention (Figure 3b), while the baseline contrastive model is much less correlated.

To test whether this alignment with human attention is general, beyond just Salicon data, we obtained another attention dataset on OSIE images [43] (Figure 3a). This data more faithfully reflects human attention as it collected from (mobile) eye tracker [43]. We find that the baseline model is weakly positively correlated with human attention (ttest: $\rho = 0.07$ $p < 0.001$) suggesting that the contrastive loss produces features that are predictive of human attention to some extent. Yet, the correlation was generally close to 0 and explains only 0.5% of data variance. The spatial attention guided model has a much stronger correlation to human attention (ttest: $\rho = 0.48$ $p < 0.001$) than the baseline model (Two samples ttest: $p < 0.001$), and thus more faithfully reflecting human visual attention (See Fig 3a for qualitative examples and Fig 3b bottom for quantitative analysis). Further, the correlation with human attention for the teacher model and the attention guided contrastive model were quite similar (Two samples ttest: $p = 0.7$).

4.3. Spatial attention guided models are more accurate than baselines

We evaluate the quality of the representations learned by spatial attention guidance framework using the typical contrastive learning evaluation criteria: fitting an ImageNet [37] linear classifier on top of the frozen representation (in practice we place stop gradient at the end of the backbone and train the classifier concurrently while training the backbone). We compute Top 1 accuracy on ImageNet validation set and compare the results with baselines. As shown in Table 1, we observe around 0.6% accuracy gain on ImageNet compared to vanilla SimCLR. We further explore an alternative way of incorporating human attention data. Rather than using pseudo attention labels on ImageNet from the teacher model, we add Salicon data to the training data, and directly predict attention ground truth labels for Salicon data with spatial attention head (Contrastive attn. co-train in Table 1) . More specifically, we pass 2048 ImageNet images and N=2048 Salicon images into our backbone (N is obtained by hyperparameter search 512–2048 based on evaluation on the separate dev set) to compute backbone features. Then, we compute the SimCLR loss from ImageNet features and use Salicon images features to predict spatial attention maps, and we use Salicon attention ground truth labels to compute the KLD and NSS losses. Interestingly, we find this method to lead to worse performance compared to using the teacher model generated spatial-attention labels, which gives evidence that the teacher model is generalizing its knowledge about human attention data beyond the limited Salicon training data.

One important question is whether the classic trainable attention method applies to contrastive learning. For example, in [32], an attention block/map is used to adjust the weights of the embedding, where the attention block can be self learned implicitly, or guided by human attention ground-truth. It shows both implicit self-learned attention block and attention block guided by human attention can improve the accuracy of fine-grained classification (Table VII in [32]). We adopted the approach described in their paper: add an attention block after the representation embedding, and use the attention block's output to mask the representation embedding. We apply it to the contrastive models, and the results are shown in Table 1 as "explicit attention mask" (the attention block is guided by human attention) and "self attention mask" (self attention). As observed, both yield poor performance, since the contrastive attention with self-attention mask method learns an attention mask from the contrastive loss rather than to predict the teacher targets, the model may be getting weaker supervision.

Another important question is whether our current spatial attention targets can be replaced by other spatial attention approach like ContraCAM [35]. In our method, we predicted the spatial attention with a linear layer. Whereas in ContraCAM a contrastive loss is used to construct the attention prediction. In both cases, we trained the predictions to match the target attention maps obtained from the teacher model. The result is shown in Table 1 as "Contrastive with ContraCAM as attention". However, we don't see much accuracy improvements with ContraCAM as the model attention, showing that the linear simple prediction leads to better performance.

To investigate whether the spatial attention guidance framework benefits supervised models in the same way as contrastive models, we applied the same approach for supervised models. Supervised models similarly benefit from

this framework, yet the gain is limited compared to the contrastive models perhaps due to the higher accuracy the supervised models achieves compared to the contrastive models. To control for accuracy, we used a supervised model trained with smaller backbone (ResNet-18), which gives comparable accuracy to the bigger ResNet-50 backbone that is trained with contrastive objective. Even when controlling the mismatch in accuracy, the supervised model gain from providing spatial-attention supervision is limited compared to contrastive models (Table 1).

4.4. Spatial attention guided models are more robust than baselines

Human vision is very robust to noise or small pixel modulations, compared to computer vision models. Here, we hypothesize that models that aligns better with human attention may learn more robust representations. We test this hypothesis using ImageNet-C dataset [18]. We take the representations learned from the proposed model and train a linear classifier on ImageNet training data. We then evaluate classification performance on ImageNet-C at various corruption/noise types. Table Supp.1 in the supplementary compares the classification performance on ImageNet-C for baseline contrastive model and the proposed models trained with spatial-attention teacher guidance, average over the 5 different corruption/noise magnitudes in ImageNet-C. We find that contrastive models trained with teacher guidance outperforms baseline consistently, suggesting that the representation learned by spatial-attention guided model is indeed more robust.

4.5. Spatial attention guided models generate better representation for retrieval

Besides classification, we tested the quality of the representations for another downstream task: image retrieval. We use the model to extract representation for ImageNet validation set, and use the representations to run image retrieval. 5000 randomly chosen validation images are used as query, while the rest validation images are database images. For each query image, top k retrieval results are returned by sorting the cosine distance between representations of database images and the query. The retrieval accuracy is measured with the standard precision recall curve [40], where precision and recall are computed by checking whether results have the same class label as the query. Mean precision and recall is obtained as the average across queries. We choose different k values, and draw the mean precision-recall curve. Besides ImageNet (”clean” in Fig Supp.1a in the supplementary), we also use ImageNet-C (”noise level 1-5”). Results from ”fog” corruption is shown in In Figure Supp.1a in the supplementary,, while results for other corruptions can be found in the supplementary too. The representation extracted with the Contrastive attn teacher model outperforms the baseline Constrastive model in this retrieval task, on both ImageNet images, and ImageNet-C images for most corruption types and levels.

Moreover, we evaluate transfer to other dataset for the retrieval task. We take the model trained on ImageNet to extract representations for CIFAR-10 [29] test set and CIFAR-10-C [18] (”clean” and ”noise level 1-5” in Fig Supp.1b in the supplementary respectively), without retraining/finetuning. 1000 images are used as query and the rest are used as database. Results for ”fog” noise are shown in Fig Supp.1b in the supplementary, while results for other noises are shown in supplementary too. As shown, the proposed attention guided model outperforms the baseline model on both clean images and almost all noise/corruption types and levels.

5. Conclusion

In this work, we tested the hypothesis that utilizing human spatial attention can be beneficial for obtaining better representation for contrastive models. We overcome the challenge of obtaining human spatial attention labels for large dataset by utilizing a teacher model trained on limited human attention labels to provide pseudo-attention labels for ImageNet. We augmented training of a commonly-used network (ResNet50) trained with SimCLR contrastive objective with pseudo-spatial attention labels from the teacher model. Our results demonstrate that contrastive models trained with those pseudo-attention labels are more predictive of human attention and we obtain better representations.

Despite the gains observed on downstream tasks are not large, the gains were consistent across 3 tasks (classification, robustness, and image retrieval). Thus, taken together, our findings support the above hypothesis. The limited gains in classification may suggest that spatial attention alone may not be enough to achieve SOTA, but perhaps in conjunction with other architectural and methodological improvements. Another limitation is that our exploration was conducted on natural images only, but other interesting applications may include other domains like medical images or autonomous driving. When domains becomes very different, it may be required to retrain teacher models to capture those new domains (e.g., train on attention data that comes from medical experts for medical images). These are interesting and important questions that may be addressed in future work.

6. Acknowledgement

We are grateful to Yifan Tian, Josh McAdams, Zhuode Liu, and Ethan Steinberg for assistance with analyses. We also thank Mike Mozer, Kai Kohlhoff, Simon Kornblith, and Ting Chen for useful discussions.

References

[1] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3, 4

[2] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. *http://saliency.mit.edu*, 2012. 2, 3

[3] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE TPAMI*, 41(3):740–757, 2018. 5

[4] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019. 5

[5] Jorge Paolo Casas and Chandramouli Chandrasekaran. openeyetrack-a high speed multi-threaded eye tracker for head-fixed applications. *Journal of open source software*, 4(42), 2019. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 5, 7

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[9] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE TIP*, 27(10):5142–5154, 2018. 5

[10] 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*, 163:90–100, 2017. 3

[11] Mohammad K Ebrahimpour, J Ben Falandays, Samuel Spevack, and David C Noelle. Do humans look where deep convolutional neural networks "attend"? In *International Symposium on Visual Computing*, pages 53–65. Springer, 2019. 3

[12] Simone Frintrop, Erich Rome, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):1–39, 2010. 2

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2

[15] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022. 3

[16] Suna Sihang Guo, Ruohan Zhang, Bo Liu, Yifeng Zhu, Dana Ballard, Mary Hayhoe, and Peter Stone. Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 34:25370–25385, 2021. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 8

[19] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015. 2, 3

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 1

[21] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 3

[22] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 5

[23] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018. 3

[24] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020. 3

[25] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 2, 3, 4, 6, 7

[26] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 5

[27] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):1–40, 2017. 2

[28] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. 3

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 8

[30] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014. 3

[31] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017. 3

[32] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2020. 2, 3, 5, 7

[33] Thomas Langlois, Haicheng Zhao, Erin Grant, Ishita Dasgupta, Tom Griffiths, and Nori Jacoby. Passive attention in artificial neural networks predicts human visual selectivity. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[34] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 2022. 5

[35] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 5, 7

[36] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. Tidying deep saliency prediction architectures. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10241–10247. IEEE, 2020. 3, 4

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2, 7

[38] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11058–11067, 2021. 3

[39] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600, 2019. 3

[40] Nikhil V Shirahatti and Kobus Barnard. Evaluating image retrieval. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 955–961. IEEE, 2005. 8

[41] Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. Vqa-mhug: A gaze dataset to study multimodal neural attention in visual question answering. *arXiv preprint arXiv:2109.13116*, 2021. 3

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[43] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature communications*, 11(1):1–12, 2020. 2, 6, 7

[44] Leonard Elia Van Dyck, Roland Kwitt, Sebastian Jochen Denzler, and Walter Roland Gruber. Comparing object recognition in humans and deep convolutional neural networks—an eye tracking study. *Frontiers in Neuroscience*, page 1326, 2021. 3

[45] Brian A Wandell. *Foundations of vision*. Sinauer Associates, 1995. 2

[46] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 3

[47] Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021. 2, 3

[48] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, 5(6):495–501, 2004. 2

[49] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2, 3

[50] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3, 5

[52] Jan Zimmermann, Yuriria Vazquez, Paul W Glimcher, Bijan Pesaran, and Kenway Louie. Oculomatic: high speed, reliable, and accurate open-source eye tracking for humans and non-human primates. *Journal of neuroscience methods*, 270:138–146, 2016. 2