Mining Gaze for Contrastive Learning toward Computer-Assisted Diagnosis

Zihao Zhao1*, Sheng Wang1,2,3*, Qian Wang1,4, Dinggang Shen1,3,4†

1School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
2School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
3Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
4Shanghai Clinical Research and Trial Center, Shanghai, China
{zhaozh2022, qianwang, dgshen}@shanghaitech.edu.cn, wsheng@sjtu.edu.cn

Abstract

Obtaining large-scale radiology reports can be difficult for medical images due to various reasons, limiting the effectiveness of contrastive pre-training in the medical image domain and underscoring the need for alternative methods. In this paper, we propose eye-tracking as an alternative to text reports, as it allows for the passive collection of gaze signals without disturbing radiologist’s routine diagnosis process. By tracking the gaze of radiologists as they read and diagnose medical images, we can understand their visual attention and clinical reasoning. When a radiologist has similar gazes for two medical images, it may indicate semantic similarity for diagnosis, and these images should be treated as positive pairs when pre-training a computer-assisted diagnosis (CAD) network through contrastive learning. Accordingly, we introduce the Medical contrastive Gaze Image Pre-training (McGIP) as a plug-and-play module for contrastive learning frameworks. McGIP uses radiologist’s gaze to guide contrastive pre-training. We evaluate our method using two representative types of medical images and two common types of gaze data. The experimental results demonstrate the practicality of McGIP, indicating its high potential for various clinical scenarios and applications.

Introduction

Gaze is a rich bio-signal that provides information on where an individual’s eyes are directed. The collection of gaze data has significantly advanced in recent years in terms of ease, cost (typically under 300 USD), and speed (Elfares et al. 2023; Uppal, Kim, and Singh 2023; Valliappan et al. 2020; Wan et al. 2021). Gaze data has been extensively researched and utilized across multiple fields, such as marketing (Wedel and Pieters 2017), robotics (Aronson and Admoni 2022; Aronson, Almutlak, and Admoni 2021; Palinko et al. 2016; Biswas et al. 2022; Manzi et al. 2020), virtual reality (Hu et al. 2019, 2021a; Matthews, Uribe-Quevedo, and Theodorou 2020; Hu et al. 2021b; Hu 2020). In addition to these fields, eye tracking has also gained attention in the medical imaging domain as a low-cost and convenient tool (Wang et al. 2022; Ma et al. 2023; Karargyris et al. 2021). For example, the Gaze Meets ML Workshop, endorsed by MICCAI, was held in 2022 to explore the application of gaze data in medical image analysis (Organizers 2022). One advantage of eye tracker in clinical medical imaging settings is that they do not require radiologists to open additional software programs or tools. Instead, the eye tracker can be easily attached beneath the monitor, allowing the radiologist to continue working with their existing tools and software. This can save time and effort compared to traditional annotation tools like drawing masks, circles, or boxes, which require radiologists to switch between different programs and interfaces. Moreover, eye tracking can provide additional data that cannot be captured by traditional annotation tools, such as insight into the radiologist’s attentional processes and decision-making strategies.

In order to utilize these information in gaze, contrastive learning is a natural framework to choose since it has already successfully mined information in many cross modality data (Radford et al. 2021). In contrastive pre-training, images sharing similar semantics should be considered positive pairs, and vice versa. Conventional approaches (Chen et al. 2020) create positive pair by randomly augmenting an image twice as illustrated in Figure 1. 1) One straightforward improvement to generate better positive pairs is to use radi-
Figure 2: We show examples, with images of similar semantics corresponding to similar gaze patterns. On the left, there are four breast mammographic images, among which two are benign masses (green boxes) and two are malignant masses (blue boxes). The distributions of gaze points are similar across two benign masses, and also across two malign masses. On the right, there are two dental X-ray images of different patients. The yellow and red boxes indicate wisdom teeth on the upper and lower jaws, respectively. Across two images, the teeth of the same location have similar gaze heatmaps, corresponding to shared anatomical roles and the underlying image semantics.

In this work, we propose Medical contrastive Gaze Image Pre-training (McGIP) to use radiologist’s gaze to generate additional positive pairs for medical images. As a substitute for radiological reports or diagnosis labels, gaze data are 1) easy to access, 2) highly variant and 3) directly related to visual pattern of each lesion. Originating from observations during gaze collection, we found that medical images corresponding to similar gaze patterns, when read by a radiologist, are often positively paired and thus should be drawn closer in the latent space, as illustrated by Figure 1. Specifically, a radiologist delivered similar gaze patterns when presented with medical images of the same semantical type. Two examples in Figure 2 illustrate our observation.

On the left of Figure 2, we show four examples of breast mammography corresponding to benign (in green boxes) and malignant (in blue boxes) masses, respectively. When looking at the exemplar benign masses in green boxes (BI-RADS 3, with cancer probability less than 2%), the radiologist often shows a more “scattered” pattern for the distribution of the gaze points. In contrast, when looking at malignant masses (BI-RADS 4C, with cancer probability 50-95%, and BI-RADS 5, with cancer probability at least 95%), radiologist tends to have a much more “centered” gaze pattern.

On the right of Figure 2, we show two dental panoramic X-ray images. When zooming into tooth #1 (denoted with yellow box), one can notice that the gaze heatmaps are always similar when radiologist views images from two different patients. Meanwhile, the gaze heatmaps of tooth #32 (denoted with red boxes) are also similarly low in magnitude across patients, implying little attention devoted to the molars from the radiologist. Other clinical researches also report that different types of abnormality can lead to different gaze patterns (Kundel et al. 2008; Voisin et al. 2013).

We design different schemes to properly measure gaze similarity under different conditions, so that our proposed method can be generalized to various types of medical images. In summary, our main contributions are as follows.

- To the best of our knowledge, McGIP is the first work to utilize human gaze as an alternative role of medical reports to guide contrastive pre-training.
- We investigate three schemes of gaze similarity evaluation, to serve different types of medical images and also representations of gaze data.
- We validate McGIP on two very different medical image diagnosis tasks of breast mammography and dental panoramic X-rays. The performance shows its effectiveness and generalizability for potential applications.
The code implementation of our method is released at https://github.com/zhaozh10/McGIP.

**Related Works**

In this section we first introduce gaze and its application in radiology. Then we briefly cover the topic of contrastive learning and corresponding positive pair selection.

**Gaze in Radiology**

Visual attention has proven a useful tool to understand and interpret radiologist’s reasoning and clinical decision. Carmody, Nodine, and Kundel (1981) published one of the first eye-tracking studies in the field of radiology, where they studied the detection of lung nodules in chest X-ray films. In mammography, a strong correlation is found between gaze patterns and lesion detection performance (Kundel et al. 2008; Voisin et al. 2013).

Recently, studies start to investigate the potential of gaze in medical image analysis from the deep learning perspective. Mall, Brennan, and Mello-Thoms (2018) and Mall, Krupinski, and Mello-Thoms (2019) investigated the relationship between human visual attention and CNN in finding missing cancer in mammography. Karargyris et al. (2021) developed a dataset of chest X-ray images and gaze coordinates. They used a multi-task framework to perform classification for diagnosis and prediction of the gaze heatmap from radiologists at the same time. Wang et al. (2022) proposed Gaze-Attention Net, which used gaze as extra supervision other than only ground-truth labels.

**Contrastive Learning**

Large-scale contrastive pre-training has become popular due to its generalizability to many scenarios and robustness against overfitting (Radford et al. 2021). There are several attempts to utilize contrastive pre-training in medical image analysis. Sowrirajan et al. (2021) proposed MoCo-CXR to produce the models with better representations and initialization for detection of abnormalities in chest X-rays. Azizi et al. (2021) utilized multiple images of the underlying pathology per patient to construct more informative positive pairs for multi-instance contrastive learning. These works have adopted image-augmentation-based semantic-unaware strategies to generate positive pairs.

In the early days of contrastive learning, good representation requires a large number of negative pairs in a batch (Chen et al. 2020; He et al. 2020). However, more recently, negative pairs are shown to be less necessary for learning a good representation. That is, the number of negative pairs may have a limited influence on the representation quality when the framework is designed properly (Caron et al. 2021). In this paper, the roles of positive pairs and their impact on the learned representations will also be our focus.

While it is critical to design positive pairs in contrastive learning, most existing frameworks apply semantic-unaware data augmentation that is adapted from the conventional supervised learning (He et al. 2020; Chen et al. 2020; Grill et al. 2020). Recent studies have found that a semantic-aware contrastive learning process can perform better. Selvaraju et al. (2021) proposed CAST to use unsupervised saliency maps to sample the crops. Peng et al. (2022) proposed ContrastiveCrop for augmentation of semantic-aware cropping.

**Method**

This section first discusses the collection and processing of gaze signals. Then, we propose different gaze similarity evaluation schemes for structured and unstructured medical images. Finally, we use gaze similarity to appropriately generate positive pairs and integrate them with multiple contrastive pre-training frameworks.

**Gaze Collection and Processing**

Gaze collection can be a seamless process when radiologists conduct routine diagnoses. Specifically, we used a Tobii pro nano remote eye-tracker to collect binocular gaze data at 60Hz. A radiologist with ten years of experience was invited to read and diagnose images on a computer, e.g., from a breast mammography dataset. A graphic user interface was developed to adapt to the typical clinical workflow, offering functions of multi-window viewing and interactive operations such as zooming, contrasting, etc. In this way, the gaze data can be collected from a nearly real diagnosis environment for the radiologist, which reduces interference during collection (Ma et al. 2023).

The recorded eye-tracking data consists of a temporal sequence of points, each of which comes with the location and the timestamp. The gaze points need to be further categorized into saccade points (fast eye movement) and fixation points (yellow dots in Figure 3, denoted as \( f_{ij} \) for multi-match algorithm (Dewhurst et al. 2012)). Specifically, the centroid of a small cluster of the raw gaze points is considered as a fixation point, with the time-lapse of the cluster as its duration. Note that the saccade points indicate rapid eye movement, which corresponds to object searching in a large field-of-view for human vision system. Thus, the saccade points carry global features of the images. In contrast, the locations and duration of fixation points, corresponding to radiologist’s visual focus on specific regions in the image, can reveal local features. A patch containing normal tissue only may have fewer than ten fixation points, while a lesion patch can have much more. The gaze points can be expressed by either gaze sequence (preserving both temporal and spatial information of the gaze) or gaze heatmap (providing spatial distribution only).

**Gaze Similarity Evaluation for Different Scenarios**

It is pivotal to evaluate gaze similarity in our work, as images with similar gaze patterns are presumably close to each other in semantics. We have divided medical images into two categories roughly, namely **structured** and **unstructured** images. In structured images, the patients are typically well-positioned and imaged following strict clinical protocols, and radiologist’s gaze tends to be similar in the same regions across different images. The reason is that those regions typically correspond to the same anatomic structure as illustrated by the example in the right of Figure 2. The unstructured images, in comparison, are very different – the patients...
Unstructured image + gaze sequence

Unstructured image +

Structured image +

Figure 3: Three schemes to compute gaze similarity for different image types and gaze representations. On the left, for unstructured images, we extract five features for each gaze sequence, and calculate the inter-sequence similarity by the multi-match algorithm (Dewhurst et al. 2012). In the middle, also for unstructured images, we use heatmap as gaze representation, and calculate the similarity by gaze moment. On the right, for structured images, we dHash each heatmap into an $8 \times 8$ code.

are not imaged with predefined position, and the images, e.g. mammography and pathological image, are usually interpreted at patch-level. In this case, the global anatomical structure is not a major clue to support the diagnosis, as shown in the left of Figure 2.

For a batch of images $[x_1, x_2, ..., x_n]$, we assume an affinity matrix for gaze similarity. In this work, we investigate three different ways to evaluate gaze similarity for different categories of medical images. For the unstructured images, we utilize both gaze sequence (in the form of $[time, location]$ for each fixation point) and gaze heatmap, as the two data formats are both commonly adopted. And for the structured images, we propose to evaluate the gaze similarity by referring to the gaze heatmap.

**Unstructured Image + Gaze Sequence.** Holmqvist et al. (2011) described gaze sequences from five different features, i.e. shape, length, direction, position and duration. We further calculate the inter-sequence gaze similarity from these five features based on multi-match (Dewhurst et al. 2012). That is, the comparison between two gaze sequences of varying lengths is considered as a string-editing problem, where the minimum editing cost serves the dissimilarity.

Assuming two gaze sequences $G_1$ and $G_2$ contain 3 and 4 points as shown in Figure 3, $f_{11}, f_{21}$ are two fixation points in $G_1$ and $G_2$, $l_{11}, l_{21}$ are offsets between two consecutive fixations, and $d_{11}, d_{21}$ indicate durations of $f_{11}$ and $f_{21}$. The inter-point duration editing cost is the relative duration difference, i.e. $|d_{11} - d_{21}|/\text{max}(d_{11}, d_{22})$. We thus construct $S_{\text{dur}} \in \mathbb{R}^{4 \times 3}$ based on pairwise duration editing cost. The minimum $\text{duration}$ editing cost is the minimum travel cost from the top left of $S_{\text{dur}}$ to the bottom right, which can be formulated as a classic dynamic programming problem. Analogously, the position, shape, length and direction costs can also be computed through corresponding editing costs as shown in Figure 3. The multi-match shall first measure the similarity between two gaze sequences from the five above-mentioned aspects. The overall similarity, denoted by $A_{ij}$, is then calculated by weighted summation. Similarly, the $A_{ij}$ of gaze sequences $G_i$ and $G_j$ is also computed using the same approach.

**Unstructured Image + Gaze Heatmap.** An even more common way to represent gaze data is to use the heatmap, which is generated by convoluting the raw gaze points with Gaussian filters (Le Meur and Baccino 2013). Based on the heatmaps, we adopt image moment to measure their similarity. In general, the family of moments is defined as

$$
\mu_{pq} = \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \, dx \, dy,
$$

(1)

where $f(x, y)$ is gaze heatmap, $(\bar{x}, \bar{y})$ is the centroid of $f(x, y)$ and $p, q \in \mathbb{N}$ define the order of moment. Then, the first invariant of Hu-moment (Hu 1962) is adopted here to measure the dispersion of the heatmap in both row and column directions:

$$
\phi_1 = \frac{\mu_{20} + \mu_{02}}{\mu_{00}},
$$

(2)

where $\mu_{20}$ and $\mu_{02}$ is the moment of inertia. We also use $\mu_{00}$, which is the weight moment and duration in our case. Thus, the gaze is described by its moment vector $m_{\text{gaze}} = [\mu_{00}, \phi_1]$. Beware that gaze moment is introduced here to measure the difference. In this case, the affinity between two moment vectors $m^i_{\text{gaze}}$ and $m^j_{\text{gaze}}$ is defined as

$$
A_{ij} = A_{ij} = \alpha(1 - \delta(\mu_{00}, \mu_{00}^j)) + (1 - \alpha)(1 - \delta(\phi_1^i, \phi_1^j)),
$$

(3)

where the affinity is defined as one minus the normalized $L_1$ distance between two moment vectors: $\delta(x, y) = \frac{L_1(x, y)}{\text{max}(x, y)}$, and $\alpha$ is a manually selected hyper-parameter.

**Structured Image + Gaze Heatmap.** In this circumstance, we adopt dHash (Maharana 2016), a widely-used
image matching algorithm, to embed a heatmap into 64-bit hash code and then measure the similarity. As in Figure 3, dHash first resizes the gaze heatmap $G$ into $\tilde{G} \in \mathbb{R}^{8 \times 9}$, thus filtering out much high-frequency components yet preserving lasting fixations. In each row, we compute the difference between adjacent pixels and get $D \in \mathbb{R}^{8 \times 8}$. Here, the computation happens in the row direction, because the teeth in our exemplar dataset are aligned in this way. The binary mask $H \in \mathbb{B}^{8 \times 8}$ is then encoded by thresholding $D > 0$.

The similarity $A_{ij}$ between two heatmaps can thus be measured as cosine similarity by flattening $H$'s.

Contrastive Pre-training with Gaze

Typical contrastive learning methods usually construct one positive pair for each sample, while the proposed McGIP can construct several positive pairs for each sample in a batch. For a batch of images $\{x_1, x_2, ..., x_n\}$, $A$ is the affinity matrix for gaze similarity. Denote the constraint function in contrastive learning as $CST(\cdot)$, which represents InfoNCE in MoCo and $L_2$ distance in BYOL for example. The overall loss function for a batch is

$$L = \frac{\sum_{i,j=1}^{n} 1_{A_{ij} \geq t}(A_{ij}) \cdot CST(Net(x_i), Net(\tilde{x}_j))}{\sum_{i,j=1}^{n} 1_{A_{ij} \geq t}(A_{ij})},$$

(4)

where $1$ refers to indicator function, $\tilde{x}_j$ denotes the transformed view of $x_j$, and “Net” indicates the encoder of contrastive learning. The term $t$ here is a threshold to determine whether the two images are similar enough to be a positive pair. The above Eq. (4) will be the optimization objective during each iteration. Moreover, gaze data inherently contains noise and uncertainty, which is especially significant on unstructured images. We correspondingly introduce $p$, denoting confidence score, for unstructured images. When selecting positive unstructured image pairs based on gaze heatmaps, we only consider it a true positive with a possibility of $p$ if the gaze similarity is higher than $t$.

Experimental Results

Datasets and Metrics

We conduct experiments on two datasets: INbreast (Moreira et al. 2012) and Tufts dental dataset (Panetta et al. 2021).

The INbreast dataset (Moreira et al. 2012) includes 410 full-field digital mammography images collected from 115 patients. We invited a radiologist with 10 years of experience to diagnose the images in this dataset, and collected the eye-movement data. The gaze data was collected using a Tobii pro nano eye-tracker, and pre-processed with the toolbox proposed in Ma et al. (2023). We randomly split the INbreast dataset for five-fold cross-validation, with 80% images for training and 20% for test. To inspect the performance, we report the accuracy (ACC), area-under-curve for malignant masses (M-AUC) and area-under-curve for all three classes (AUC) on the testing data. Here M-AUC is specially calculated since malignant masses are critical to diagnose, whose risks to cancer can be more than 10 times higher than benign masses (Liberman and Menell 2002). We use multi-match and gaze moment to calculate the gaze similarity on this dataset, respectively. The threshold of gaze similarity is set to 0.7. For gaze moment, $p$ is set to 0.5 in all experiments.

The Tufts dataset (Panetta et al. 2021) is composed of 1000 panoramic dental X-ray images, together with processed gaze heatmaps. We choose 70% and 10% of images for training and validation, while the remaining 20% of images constitute the testing set. Accuracy (ACC) and area-under-curve (AUC) are reported as the performance indicators. We use dHash to calculate the gaze similarity and set the threshold to 0.7.

All the experiments are implemented with PyTorch 1.13.0 on a single NVIDIA RTX3060. Unless otherwise specified, all networks are trained for 200 epochs using Adam optimizer with the learning rate ($lr$) set to $2e^{-5}$ in the pre-training. Fine-tuning and linear probing for final classification are trained for 10 epochs (INbreast) and 20 epochs (Tufts) with Adam optimizer ($lr$: $2e^{-5}$). All pre-training methods are initialized from ImageNet pre-trained weights.

Performance on Diagnosis Tasks

In order to demonstrate the generalizability and practicality of McGIP, we test different diagnosis tasks on the two datasets. In particular, we evaluate the fine-tuning performance, which is popularly adopted in medical image studies (Azizi et al. 2021).

The results of INbreast dataset are reported in Table 1, where three different backbone models ResNet18 (RN18), ResNet50 (RN50), and ResNet101 (RN101) are adopted to evaluate the effectiveness of contrastive pre-training. Compared to the conventional way to supervise the pre-training with ImageNet-1K, the performance of McGIP improves constantly. Moreover, compared to existing contrastive learning such as MoCo v2 (Sowrirajan et al. 2021) (denoted as MoCo), McGIP constantly improves with different backbones in ACC (from 83.42% to 85.41% averaged over three backbones). The same trend is also observed for other evaluation metrics, such as AUC and M-AUC, in most cases. Notably, although the AUC of McGIP is slightly lower when using the RN101 backbone, it demonstrates a significant advantage in terms of M-AUC, which is a more critical metric for accurate breast cancer diagnosis. Similarly, for the classification task related to the panoramic X-ray image dataset, we report the results for only BYOL with three backbones in Table 2 due to page limit. Still, McGIP constantly offers the best fine-tuning performance among all compared settings.

In summary, McGIP effectively improves contrastive learning in the final diagnosis performance, while our method is notably agnostic to different network backbones.

Representation Quality Analysis

While previous results confirm that McGIP can effectively boost classification performance with various contrastive learning frameworks, it is more interesting to inspect representation quality after contrastive pre-training. We visualize the point-to-point affinity in Figure 4, which highlights the quality of the learned representation of McGIP. Specifically, several image patches are randomly selected from the testing set of INbreast and then resized to $224 \times 224$. Based on the pre-trained backbone of RN50, we use linear-probing.
weights and derive the high-resolution feature map for each patch. Then, we randomly select a point inside the malignant mass (e.g., marked by a positive dot in Figure 4), and calculate its affinity with all other points in the patch by cosine similarity of their corresponding feature vectors. The resulting affinity maps are shown in individual rows, while the columns are for different pre-training schemes. SimSiam+McGIP delivers pulse-like patterns. Although BYOL shows slight improvements over ImageNet, it still fails to group some cases, as indicated by the last row in Figure 4. In contrast, our method exhibits significant improvements. In the last column of the figure, the boundaries between high- and low-affinity regions are clearly defined and mostly aligned with the mass contours. We attribute such superiority to the spatial semantics introduced via gaze signals, which contribute to better classification performance as evidenced by Table 1.

### Gaze Sequence vs. Gaze Heatmap

In the method section, we introduce two approaches for measuring gaze similarity in unstructured images: multi-match for gaze sequence and gaze moment for gaze heatmap. The fine-tuning (FT) and linear-probing (LP) performances of these two approaches are compared in Table 3, based on INbreast with RN50. The two approaches perform similarly while offering various benefits. Gaze moment, being a renaissance variant of Hu-moment algorithm (Hu 1962), has a straightforward and elegant form. Gaze sequence preserves more information because there are timestamps for individual spatial locations and the multi-match algorithm describes gaze similarity from multiple perspectives. In conclusion, we recommend that users choose between these two approaches based on their specific requirements. By offering multiple options for measuring gaze similarity, we hope to enable researchers to choose the most appropriate approach for their scenarios.

### Gaze vs. Ground-Truth

We conducted an empirical study to evaluate the effectiveness of using gaze data as a form of weak supervision compared to supervised contrastive learning using different backbone models (Khosla et al. 2020), which is based on ground-truth labels. To assess the performance of gaze data,
BYOL + McGIP

Abnormality
Positive
ImageNet
BYOL
BYOL + McGIP
Negative
ImageNet
BYOL
BYOL + McGIP

Figure 4: Image affinity analysis. We selected a point from feature map and calculated similarity with all other points. Natural image supervised pretrained, semantic-unaware pretrained, and McGIP pretrained weights, are used to illustrate with both positive points (abnormality) and negative points (non-abnormality). Brighter color denotes more similar points.

![Image Affinity Analysis](image)

Table 4: Comparison between gaze-guided and ground-truth labels-guided contrastive learning using different backbone models (default contrastive learning method: BYOL).

<table>
<thead>
<tr>
<th>Method</th>
<th>INbreast M-AUC</th>
<th>INbreast AUC</th>
<th>INbreast ACC</th>
<th>Tufts AUC</th>
<th>Tufts ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RN18+GT</td>
<td>93.01±1.78</td>
<td>90.86±1.32</td>
<td>84.32±0.66</td>
<td>60.53</td>
<td>59.00</td>
</tr>
<tr>
<td>RN18+Gaze</td>
<td>95.83±0.63</td>
<td>94.96±1.13</td>
<td>85.14±0.85</td>
<td>62.91</td>
<td>65.00</td>
</tr>
<tr>
<td>RN50+GT</td>
<td>96.02±0.88</td>
<td>88.96±1.61</td>
<td>85.14±0.85</td>
<td>59.29</td>
<td>58.00</td>
</tr>
<tr>
<td>RN50+Gaze</td>
<td>97.07±0.75</td>
<td>93.80±0.79</td>
<td>87.57±1.01</td>
<td>61.35</td>
<td>67.50</td>
</tr>
<tr>
<td>RN101+GT</td>
<td>94.81±2.48</td>
<td>90.60±2.97</td>
<td>85.95±1.08</td>
<td>59.17</td>
<td>63.50</td>
</tr>
<tr>
<td>RN101+Gaze</td>
<td>95.46±2.67</td>
<td>90.09±3.08</td>
<td>86.76±0.54</td>
<td>61.14</td>
<td>64.50</td>
</tr>
</tbody>
</table>

Figure 5: Linear-probing performance (backbone: RN50) on the INbreast dataset with (a) different pre-training epochs and (b) different similarity thresholds.

![Linear-Probing Performance](image)

Conclusion

In this paper, we start with an observation that images sharing similar semantics usually have similar gaze patterns. Therefore we explore radiologists’ gazes for contrastive pretraining. We propose McGIP, a simple module to the existing contrastive learning, to shift more attention on the images with similar gaze patterns. Three schemes to evaluate gaze similarity are investigated given different medical scenarios. The superior performance on two different clinical tasks show the practicability and generalizability of McGIP.
Acknowledgments
This work is supported in part by The Key R&D Program of Guangdong Province, China (grant number 2021B0101420006).

References
Aronson, R. M.; and Admoni, H. 2022. Gaze complements control input for goal prediction during assisted teleoperation. In Robotics science and systems.
volume 10952, 1095216. International Society for Optics and Photonics.


