Principled Analysis of DeepRARE for Unsupervised Saliency Prediction

Johannes Bertram
University of Tübingen, Tübingen AI Center, Germany
johannes.bertram@student.uni-tuebingen.de

Abstract

DeepRARE (Mancas, Kong, and Gosselin 2020) is a compelling approach to visual saliency prediction. It combines the traditional idea of low-level pop-out with modern high-level deep features. DeepRARE assumes high saliency for rare VGG16 (Simonyan and Zisserman 2015) features within an image without needing saliency supervision. This paper evaluates DeepRARE on the MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.) using the principled information gain metric and conducts a detailed error case inspection. Error case analysis reveals that human faces and text are insufficiently attended to. Adding face and text detectors improves model performance suggesting potential improvements of DeepRARE by using stronger pre-trained features. DeepRARE performs at least 15% better than low-level saliency models, but still 60% short of state-of-the-art models.

1 Introduction

The amount of visual information available to humans exceeds the brain's capacity to process it (Broadbent 1958). This creates a bottleneck that necessitates filtering. This filtering is achieved through visual attention. Modeling visual attention can help understand perception and has several direct applications, such as robotics, visual marketing, object detection, and more (Mancas and Le Meur 2016).

What attracts visual attention is often called salient and is modeled in co-called saliency models. Saliency is largely task-independent: Regardless of the possible tasks performed while viewing a scene, specific objects catch attention (Kümmerer and Bethge 2023). For example, a red blinking light will grab our attention and is thus considered salient. Saliency was first introduced as a low-level property. Thus, the first saliency map models, which predict saliency across entire images, are based on low-level features such as those realized by V1 (Itti, Koch, and Niebur 1998).

The first key ingredient of any saliency model is the features it uses. In the field of visual saliency modeling, the question of whether low-level or high-level features attract attention has been debated for a long time. However, the best saliency models of the last few years all use high-level features, suggesting that they are important for predicting fixations (Kümmerer and Bethge 2023; Kümmerer, Wallis, and Bethge 2016; Linardos et al. 2021). The rise of deep learning enabled researchers to build models that use the high-level features of deep learning models trained on large image classification datasets.

The second key ingredient is determining the importance of the detected features for visual saliency. This problem is difficult and highly context-dependent: While a red light will be salient in most cases, an exception is a scene containing many red lights. Here, a single green light will stand out among the red lights. Thus, just assigning a high importance to the feature "red" does not suffice. Instead, for example, this effect can be modeled by RARE (Riche, Mancas, Gosselin, et al. 2012): RARE uses simple features and assumes that humans attend to features that are rare within an image. Nowadays, state-of-the-art is achieved by using strong pre-trained feature extractors and then training a model to

predict human fixations using these deep features (Kümmerer and Bethge 2023; Kümmerer, Wallis, and Bethge 2016; Linardos et al. 2021). This approach of determining the feature importance is called transfer learning and comes with the need for supervised data and the computational cost of training the models on visual saliency.

Here, DeepRARE (Mancas, Kong, and Gosselin 2020; Kong et al. 2021) comes in. It combines the promising idea of RARE to determine feature importance with the use of deep features: Both low-level and high-level features are extracted from VGG16 (Simonyan and Zisserman 2015). Then, those features that are rare within an image are predicted to capture visual attention. To date, DeepRARE has not been evaluated on the well-known MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.). Further, there has not been a detailed analysis of DeepRARE's strengths and weaknesses, specifically regarding its rarity-based feature importance determination and the quality of its deep features. Here, I aim to address these open questions. The specific contributions of this paper are:

- Analysis of the benefits of incorporating deep, high-level features by comparing DeepRARE with RARE.
- Detailed error case analysis and visualization in comparison to state-of-the-art model DeepGaze IIE (Linardos et al. 2021) revealing DeepRARE's insufficient detection of faces and text as salient objects.
- Improved performance using additional face and text detectors; this showcases the potential of stronger feature extractors while also highlighting the limitations of the rarity approach when compared to transfer learning.
- Principled quantitative evaluation of DeepRARE with the MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.) pipeline.
- Identifying DeepRARE as the top-performing unsupervised visual saliency model.

2 Related work

DeepFeat (Mahdi and Qin 2017) is a comparable architecture to DeepRARE. Instead of VGG16 (Simonyan and Zisserman 2015), it uses ResNet (He et al. 2015) trained on ImageNet (Russakovsky et al. 2015) for feature extraction. DeepFeat also works without additional supervision on eye movements. Instead, the saliency prediction is obtained by combining bottom-up and top-down maps based on the convolution features. The bottom-up component is obtained by taking the difference between fine- and coarse-grained versions of the same feature maps. Intuitively, if they agree, then there are no edges or gratings in these regions and thus nothing to attend to from a bottom-up perspective. If edges are present, the fine and coarse-grained feature maps will disagree. This disagreement indicates visually salient regions. The top-down component uses the class-activation-mapping (CAM) (Zhou et al. 2015) where weights are derived from the predicted class probabilities. This assigns higher saliency to those regions that lead to predicting the objects present in the image. Finally, a center bias is added.

The reported scores of DeepFeat on MIT1003 (Judd, Ehinger, et al. 2009) are very similar to those of DeepRARE. However, DeepRARE is more flexible than the CAM component (Zhou et al. 2015) of DeepFeat which cannot be applied to every CNN architecture.

3 Evaluation

The quantitative evaluation of saliency maps is by no means trivial. Evaluating with different metrics leads to highly inconsistent results. Also, a saliency map is a loosely defined term. The only agreed-upon characteristic is that regions with higher saliency values attract more attention compared to regions with lower saliency. However, it is unclear how much more attention these regions attract and how to interpret the saliency values. To solve this, Kümmerer and Bethge 2023 propose to reinterpret saliency in a Bayesian framework where a posterior probability of fixations given the image is calculated. The usage of posterior probabilities solves the previous problem, as, for example, a doubled probability indicates that humans look at this region twice as likely.

The Bayesian approach gives rise to a new metric that also allows to judge how much better a model is compared to another one based on the metric. This is possible using the models' log-likelihoods

where a log-likelihood of one bit more indicates that, on average, a model finds the data twice as likely. If choosing a sensible baseline, the log-likelihood difference to the baseline is called the information gain (IG) (Kümmerer and Bethge 2023). Information gain is measured in bit/fix and is inherently interpretable: A difference in information gains of 0.1 indicates that the better models outperforms the worse model by 10%. This information gain is the primary metric used for evaluation in this paper. As a baseline, a simple center bias is used.

As DeepRARE is not a probabilistic model, the saliency map it produces has to be transformed into a posterior distribution. To do so, the pysaliency package (Kümmerer 2025) from the MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.) is used. Specifically, to obtain the posterior, a center bias, a blur, and a monotonic non-linearity transforming the saliency map (Kümmerer, Wallis, and Bethge 2015; Kümmerer 2025) are optimized on the MIT1003 dataset (Judd, Ehinger, et al. 2009). Most evaluations and visualizations in this paper are performed on the publicly available MIT1003 dataset as is also done for the original DeepRARE (Mancas, Kong, and Gosselin 2020; Institute 2025). One has to note that the center bias and non-linearity are thus optimized and evaluated on the same dataset, while the model itself is fixed. For the final evaluation, the test set MIT300 (Judd, Durand, and Torralba 2012) is used.

4 Models

4.1 RARE

The original RARE model (Riche, Mancas, Gosselin, et al. 2012) uses simple bottom-up features. The feature importance for saliency is determined using the rarity approach: High saliency is predicted for rare feature activation regions in an image. The features are luminance and chrominance values as well as Gabor filter activations applied to each color channel. Therefore, RARE serves as a representative of low-level feature models. The comparison with RARE provides an argument in the low-level versus high-level debate in saliency modeling.

4.2 DeepRARE

DeepRARE (Mancas, Kong, and Gosselin 2020) uses a VGG16 (Simonyan and Zisserman 2015) network trained on ImageNet (Russakovsky et al. 2015). DeepRARE uses a similar rarity approach as RARE for predicting feature importance without supervision. The idea of DeepRARE is not restricted to VGG16 but can be adapted to different deep learning feature extractors.

In detail, DeepRARE uses all convolutional layers of VGG16 and groups them into five groups:

- Low-level: Layers 1-2
- Low-level: Layers 4-5
- Middle-level: Layers 7-9
- Middle-level: Layers 11-13
- High-level: Layers 15-17

The input image is resized to a shape of 224×224 pixels and given to the network. The activations of each layer are extracted. Then, using these layer activations, the rarity idea of (Riche, Mancas, Duvinage, et al. 2013) is implemented:

$$R(i) = -log(p(i))$$

This rarity function is based on the histograms of feature activations from every feature map using eleven bins. Then, $p(i)$ is the occurrence probability of pixels in bin i. This rarity is applied on every feature map of every convolutional layer. The resulting saliency map of the image is obtained by fusing and back projecting this rarity to the pixels of the input image. The fusing is done in three steps:

- In every convolutional layer, the feature map rarities are fused according to Itti 2004.
- The layers are fused in the same way into the five levels from above.

- The five resulting rarity maps are summed.

As humans tend to look at human faces, feature map 105 from layer 15, which detects large faces, is added on top. The result is a saliency map produced by DeepRARE. For evaluation, a center bias, blur, and monotonic non-linear transform (Kümmerer, Wallis, and Bethge 2015; Kümmerer 2025) are added (Institute 2025).

Critically, DeepRARE uses features from every layer of the deep neural network. As feature visualizations indicate, the earliest layers often learn low-level filters comparable to Gabor filters. On the other hand, the model also uses high-level features from the later layers. These high-level features are not used in RARE. Therefore, a detailed comparison of DeepRARE and RARE sheds light on the importance of high-level features.

4.3 DeepGaze IIE

Performance and error case comparisons are performed with respect to the state-of-the-art saliency model DeepGaze IIE (Linardos et al. 2021). The DeepGaze model family (Kümmerer, Wallis, and Bethge 2016; Linardos et al. 2021) uses different feature extractors and then trains an additional MLP taking the extracted features as input and eye tracking data as ground truth for training. State-of-the-art is achieved with DeepGaze IIE by combining several different deep feature extractors. The DeepGaze models are part of the dominant approach of transfer learning for determining feature importance for visual saliency. The readout MLP that predicts saliency uses features from several layers of the feature extractor, similar to DeepRARE. DeepGaze II (Kümmerer, Wallis, and Bethge 2016) even uses VGG19 (Simonyan and Zisserman 2015) features which are related to the VGG16 features that DeepRARE uses. The key difference between the DeepGaze models and DeepRARE thus is the transfer learning approach with eye tracking data.

5 Experiments

In this section, several models are compared on the MIT1003 (Judd, Ehinger, et al. 2009) dataset. All models' saliency maps are turned into posteriors using the pysaliency package (Kümmerer 2025). This includes the optimization of smoothing, center bias, and monotonic non-linearity on MIT1003. Only in Section 5.5, the performance is measured on the test set MIT300 (Judd, Durand, and Torralba 2012) after optimizing on MIT1003.

5.1 Evaluation of DeepRARE against RARE

The first comparison is performed between DeepRARE and its predecessor RARE. While both models use the rarity approach to determine feature importance, RARE only uses low-level features compared to DeepRARE low- and high-level deep features.

Table 1: Information Gains (IG), optimized AUCs (after optimizing the center bias and non-linearity on MIT1003) and AUC (before optimizing) on MIT1003 (Judd, Ehinger, et al. 2009)

Model	IG [bit/fixs]	optimized AUC	AUC
DeepGaze IIE (Linardos et al. 2021)	1.0984	0.8892	
DeepRARE with Face and Text Detector	0.7971	0.8583	0.8205
DeepRARE with Text Detector	0.7383	0.8550	0.8172
DeepRARE with Face Detector	0.7282	0.8519	0.8149
DeepRARE (Mancas, Kong, and Gosselin 2020)	0.6733	0.8488	0.8117
DeepRARE without Face Feature	0.6413	0.8471	0.8089
RARE (Riche, Mancas, Gosselin, et al. 2012)	0.5047	0.8363	0.7819
Only Face and Text Detector	0.4414	0.8243	0.5643

5.1.1 Results

As Table 1 shows, DeepRARE outperforms RARE on MIT1003. While the AUC score differences are often small, there is a clear improvement in information gain compared to the center bias baseline.

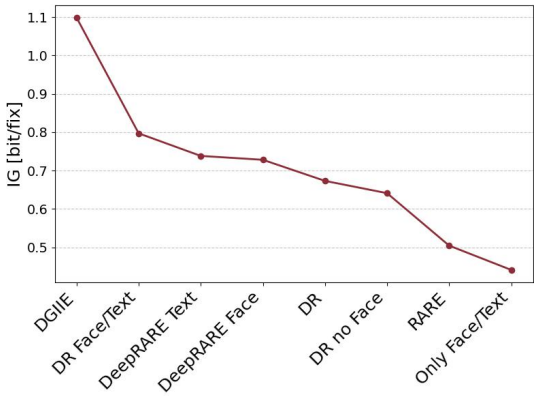
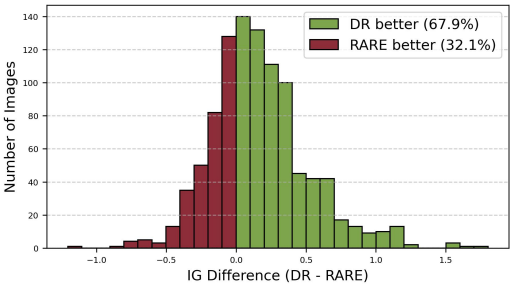
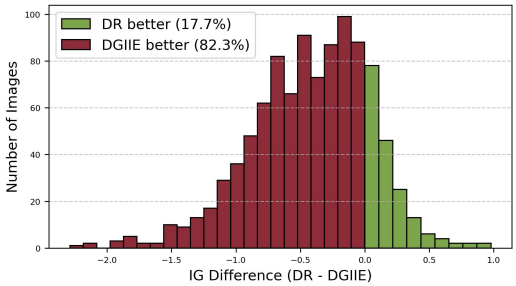


Figure 1: Information Gains (IG) in bit/fix on MIT1003 (Judd, Ehinger, et al. 2009). Models: DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE with added face and text detector (DR Face/Text) (Section 5.3), DeepRARE with added text detector (DR Text) (Section 5.3), DeepRARE with added face detector (DR Face) (Section 5.3), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with deactivated built-in face feature (Section 5.4), RARE (Riche, Mancas, Gosselin, et al. 2012), face and text detector with center bias but without any saliency model (Face/Text) (Section 5.4). Exact values and AUC scores in Table 1



(a) Histogram of IG differences between DeepRARE and RARE



(b) Histogram of IG differences between DeepRARE and DeepGaze IIE

Figure 2: Histograms comparing the performance of DeepRARE (DR) to a) RARE and b) DeepGaze IIE on individual images. The information gains (IG) are calculated for each image individually and then subtracted to obtain the histograms

Figure 2a highlights the superiority of DeepRARE: RARE is outperformed by DeepRARE on about a third of the images. The information gain difference shows that DeepRARE is 17% better. While images where RARE is more than 0.5 $bits/fix$ better than DeepRARE are outliers, there are many images where DeepRARE outperforms RARE by more than 0.5.

Inspecting the images where DeepRARE achieves the highest IG compared to RARE (Figure 3) shows that RARE fails to predict high enough saliency for faces and text. More generally, the saliency values that RARE predicts for different regions in an image often do not vary much. There is no object pop-out effect in saliency scores. This pop-out can be observed in DeepRARE for high-level objects like faces. The reason for the better results of RARE in Figure 4 is less apparent. It seems as though DeepRARE focuses too much on regions that, while somewhat salient, are not the most salient part of the image. For example, the center in Figure 4a, the middle window in Figure 4b, and the mountain outline in Figure 4c too salient in DeepRARE's prediction.

5.1.2 Discussion

DeepRARE can be viewed as a strictly more powerful extension of RARE: It contains low-level features comparable to RARE's features in the early network layers and additionally uses high-level features. The results show that the high-level features are useful for images where some object is more salient than can be predicted by just detecting it without having access to its meaning. This

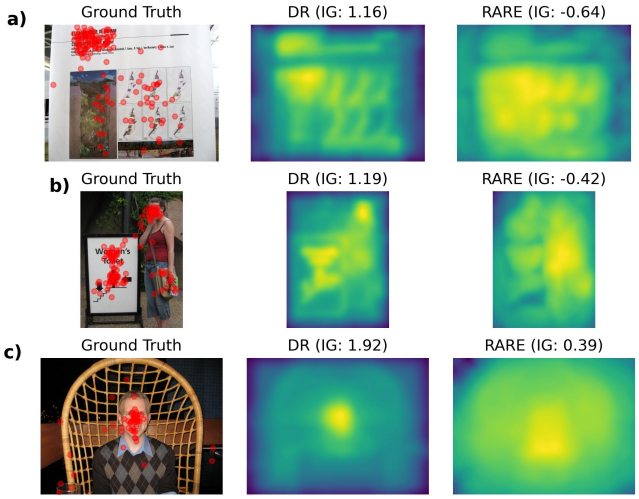


Figure 3: Image with ground truth fixations as red dots, predicted posteriors of DeepRARE (DR) (Mancas, Kong, and Gosselin 2020) and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with highest IG scores of DeepRARE compared to RARE. More examples in Figure 19

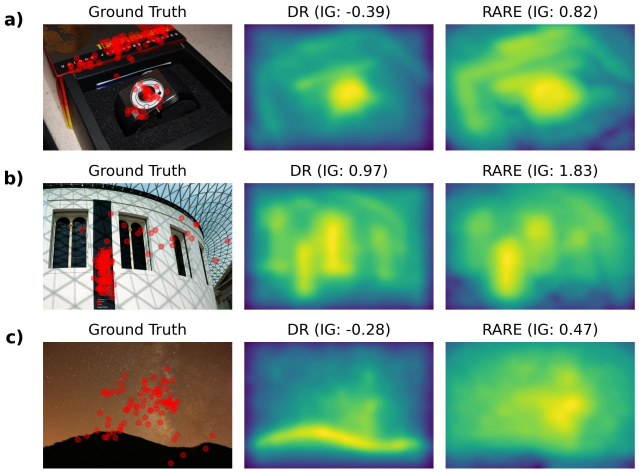


Figure 4: Image with ground truth fixations as red dots, predicted posteriors of DeepRARE (DR) (Mancas, Kong, and Gosselin 2020) and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with highest IG scores of RARE compared to DeepRARE. More examples in Figure 20

is what RARE does: Lines of text and images are detected by their edges but are not recognized as an object that is highly significant for gaze predictions. The cases in which RARE is better than DeepRARE seem to originate in a suboptimal weighting of different features detected in the image by DeepRARE. The weighting of features is not learned but is arbitrary and handcrafted. This is an inherent problem of the unsupervised saliency prediction approach used in RARE and DeepRARE. Still, as DeepRARE achieves higher scores, the weighting seems to work decently well in most cases.

5.2 Evaluation of DeepRARE against DeepGaze IIE

An approach that does not rely on this handcrafted combination of features is transfer learning. Here, the importance of features is learned using eye fixations as supervision. Therefore, DeepRARE is now compared to the state-of-the-art model DeepGaze IIE which uses transfer learning.

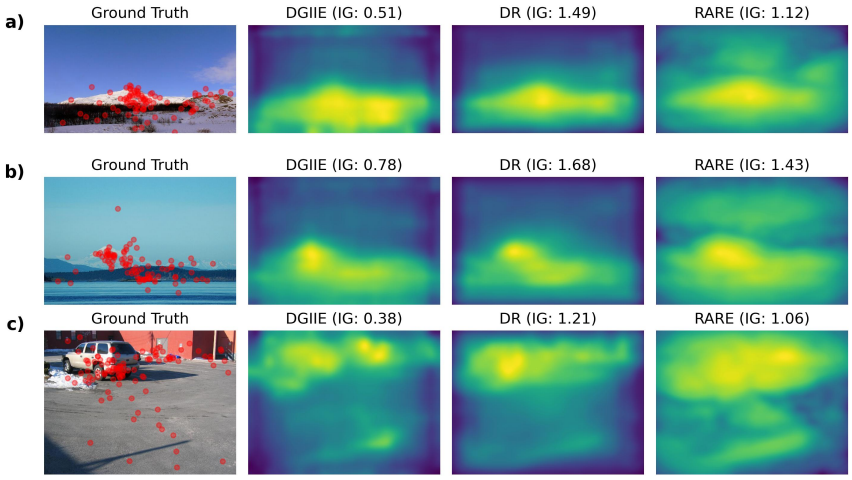


Figure 5: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020) and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with highest IG scores of DeepRARE compared to DeepGaze IIE. More examples in Figure 21

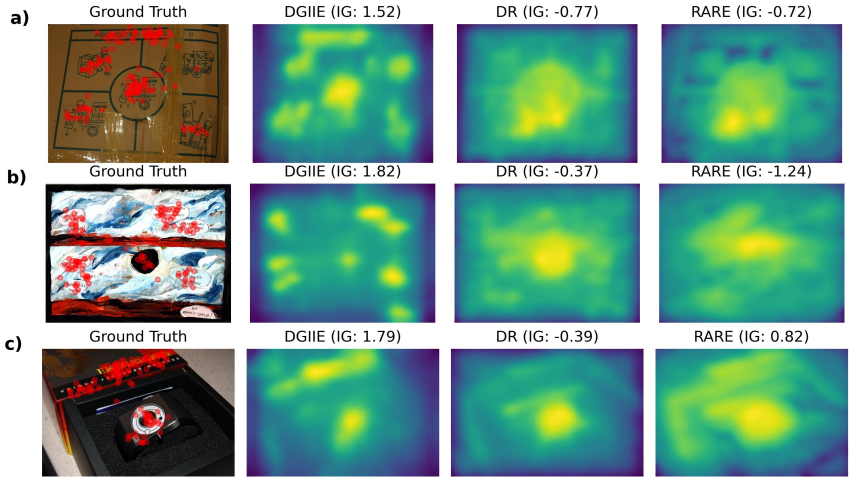


Figure 6: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020) and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with lowest IG scores of DeepRARE compared to DeepGaze IIE. More examples in Figure 22

5.2.1 Results

From Table 1 and Figure 2b one can see that DeepGaze IIE outperforms DeepRARE by 40% and on more than 80% of the images. DeepRARE does outperform on a few images that contain large regions with almost constant color, such as sky and road Figure 5. In general, DeepRARE performs well on images where a clear pop-out of a few objects is present (see also Figure 15). The benefits of DeegGazeIIE can be observed on images with text (Figure 6) and, if looking at further examples in the appendix (Figure 22), also on images containing faces. Especially when the faces and text segments are not centered in the image, DeepRARE fails to adequately attend to them.

5.2.2 Discussion

It is difficult to pinpoint an exact reason why DeepRARE outperforms DeepGaze IIE on these simple images. It might be the case that in simple images with few features, the rarity approach of DeepRARE is especially effective. The benefits of DeepGaze IIE can be identified, on the other hand. As eye tracking data confirms, humans pay extraordinary attention to text and faces in images. This is learned by DeepGaze IIE with eye tracking supervision. DeepRARE attends less to faces and text and thus significantly lacks in performance on these images. This is the case even though the VGG16 face feature is manually added to the DeepRARE output. This indicates that one main shortcoming of DeepRARE is the insufficient saliency prediction for faces and text.

5.3 Improving Face and Text Detection

As an attempt to investigate the effect of face and text detection in DeepRARE, this experiment uses external face and text detectors on top of DeepRARE.

5.3.1 Method

As face detector, MTCNN (Paz Centeno 2019) is used. MTCNN is added to the saliency prediction in the same way as the original face detector of DeepRARE, which is disabled. The outputs of MTCNN are rectangles containing faces in the image. The rectangles are constant at a value between 0 and 1 indicating the face detection probability of the face detector. A threshold is applied to only show faces with probability ≥ 0.5. This threshold is set by visual inspection. These rectangles are multiplied with a Gaussian-shaped curve centered in the middle of the rectangle. As the rectangles are always aligned with the image axes, the Gaussian has a diagonal covariance matrix with the diagonal entries depending on the width and height of the rectangle. The Gaussian is scaled to have a value of 1 at its mode. This ensures smoother face detection regions.

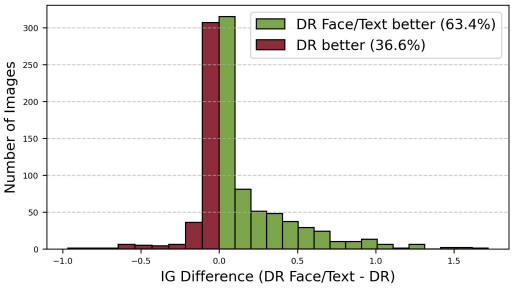
Similarly, the text detection is performed with EasyOCR (JaidedAI 2020). The outputs are again rectangles multiplied with Gaussians and added to the output in the same way. The threshold for text detection probabilities is set to 0.1 based on visual inspection.

The VGG16 face feature does not output face detection probabilities but feature activations that can reach values well above 1000. Therefore, a scaling factor for the face and text detectors is needed. These scaling factors are set using a grid search with the unoptimized AUC on MIT1003 as metric. As Table 1 shows, the IG, optimized AUC and unoptimized AUC are consistent across models. Therefore, using the unoptimized AUC serves as a compute-efficient proxy for information gain where the optimization over center bias, smoothing, and non-linearity does not have to be performed for every model used for the grid search.

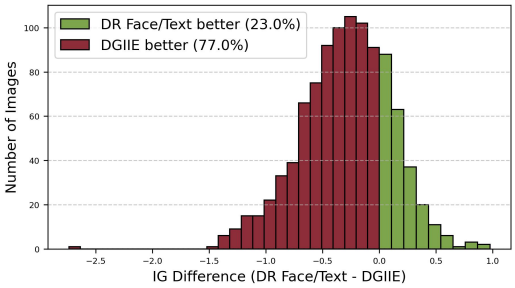
5.3.2 Results

First, additional face and text detection does not change model scores on most images (Figure 7a. Minor differences are expected because of different optimization results, but most images do not contain faces or text. Therefore, just using the face and text detection with a center bias but without any saliency model does not achieve high scores (Table 1.

Apart from these minor score changes on images without faces or text, there are more score improvements observable with the face and text detectors. Figure 8 displays the images where the face and text detector have the highest benefit. In these, DeepRARE puts high saliency to many regions while the ground-truth fixations focus on face and text only. The face and text detectors enable DeepRARE



(a) Histogram of IG differences between DeepRARE with face and text detection and DeepRARE



(b) Histogram of IG differences between DeepRARE with face and text detection and DeepGaze IIE

Figure 7: Histograms comparing the performance of DeepRARE with face and text detection (DR Face/Text) to a) DeepRARE and b) DeepGaze IIE on individual images. The information gains (IG) are calculated for each image individually and then subtracted to obtain the histograms

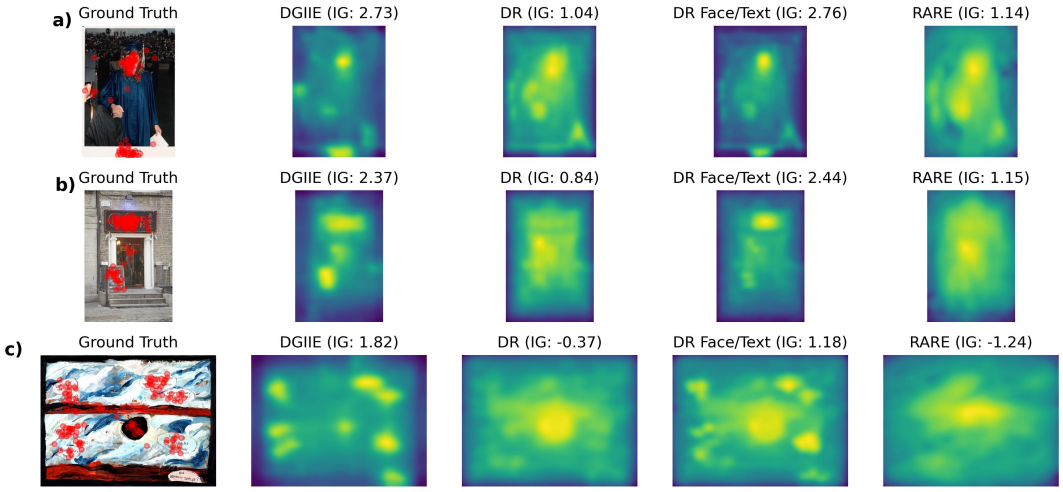


Figure 8: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with highest IG scores of DeepRARE with face and text detector compared to DeepRARE. More examples in Figure 25

to adequately focus on these regions. However, Figure 9 shows that the text detection can also be misleading and humans do not always focus solely on the text. On average, face and text detection is still beneficial, leading to performance gains of 12% (Table 1).

Comparing Figure 7b to Figure 2b shows that the scores of almost all images where DeepGaze IIE outperformed DeepRARE by more than 1.5 bits are significantly improved. In the remaining poor-performance images, text detection overfocuses or does not detect important text regions (Figure 10). The information gains on the simple images where DeepRARE already outperformed DeepGaze IIE Figure 5 are unchanged, as expected from the absence of faces and text in these simple images.

5.3.3 Discussion

These results show that, although overall beneficial, the text and face detection are not on par with DeepGaze IIE's capabilities. Minor improvements could be made with the more rigorous tuning of the hyperparameters mentioned in Section 5.3.1. However, the biggest improvements would likely require supervision on eye tracking data to learn the correct importance of faces and text depending on the context. Moreover, DeepGaze IIE still clearly outperforms DeepRARE with face and text detection on most images. The main improvement of DeepGaze IIE over DeepRARE thus seems

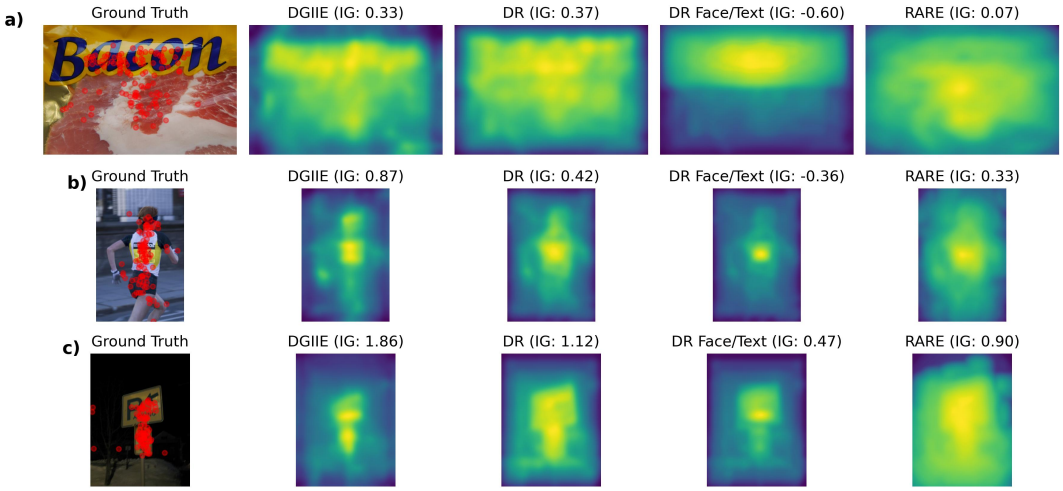


Figure 9: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with lowest IG scores of DeepRARE with face and text detector compared to DeepRARE. More examples in Figure 26

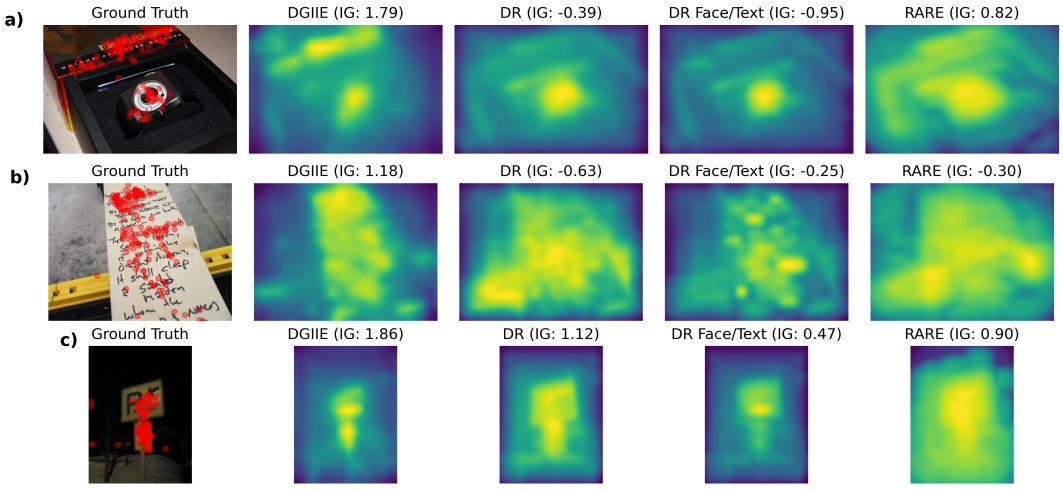


Figure 10: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 3 images with highest IG scores of DeepRARE with face and text detector compared to DeepGaze IIE. More examples in Figure 27

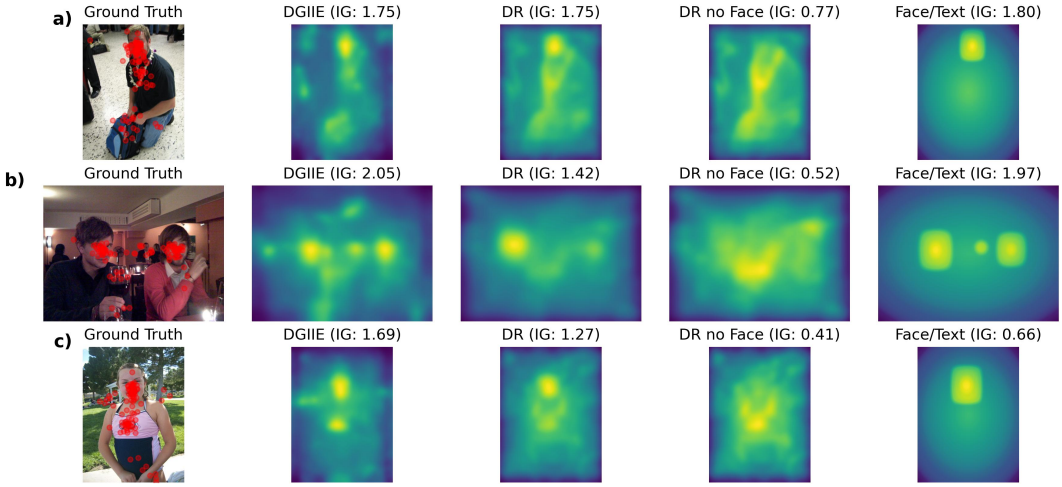


Figure 11: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE without built-in face feature (DR no Face) Section 5.3 and face and text detector with center bias instead of saliency model (Face/Text). Selection: Top 3 images with highest IG scores of DeepRARE compared to DeepRARE without built-in face feature. More examples in Figure 29

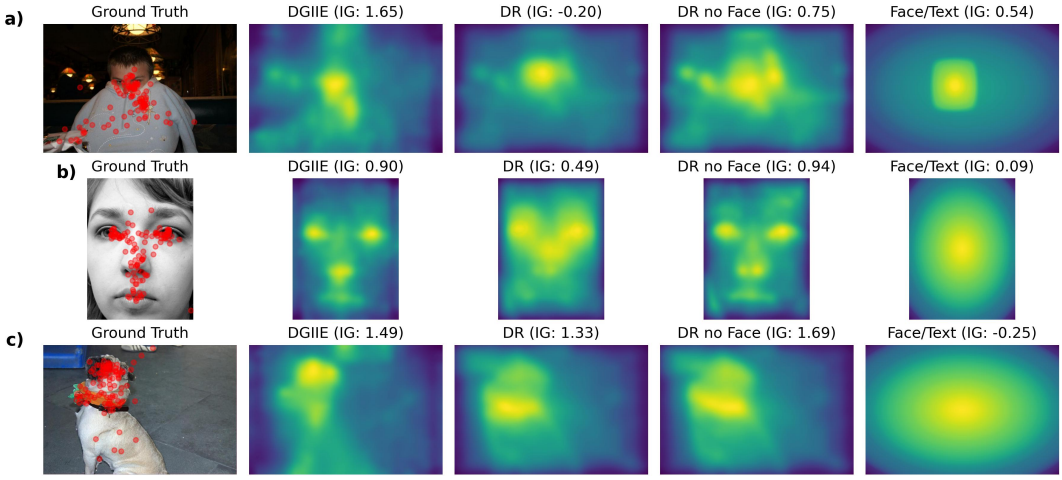


Figure 12: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE without built-in face feature (DR no Face) Section 5.3 and face and text detector with center bias instead of saliency model (Face/Text). Selection: Top 3 images with lowest IG scores of DeepRARE compared to DeepRARE without built-in face feature. More examples in Figure 30

to stem from the better weighting of features enabled by transfer learning and not only from better features.

5.4 Investigation of DeepRARE's built-in Face Feature

As DeepRARE includes a face feature, but Section 5.2 shows struggles with attending to faces, this section is devoted to investigating this face feature and its shortcomings compared to the additional face detector from Section 5.3.

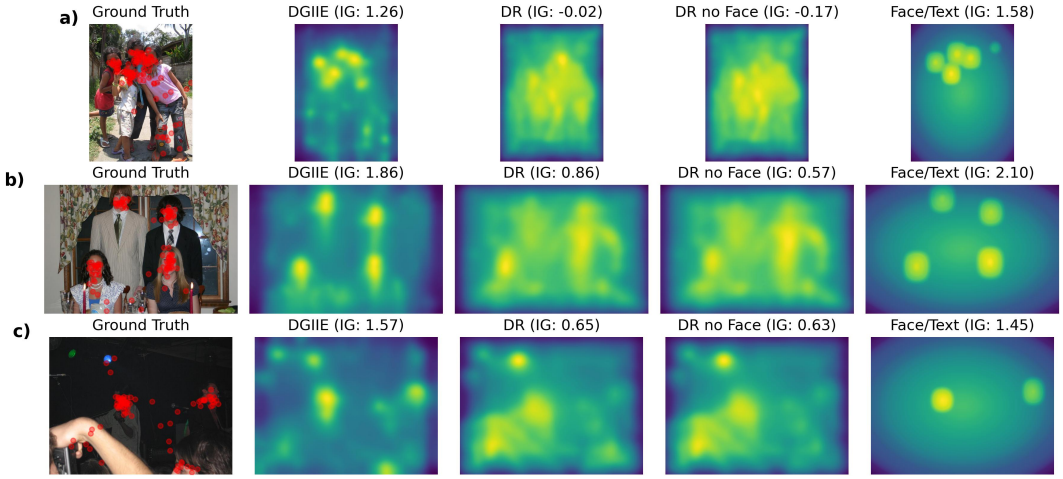


Figure 13: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE without built-in face feature (DR no Face) Section 5.3 and face and text detector with center bias instead of saliency model (Face/Text). Selection: Top 3 images with highest IG scores of DeepRARE with additional face detector compared to DeepRARE. More examples in Figure 31

5.4.1 Results

Overall, the face feature is a useful addition to DeepRARE as can be seen from the improved scores Table 1. Figure 11 underlines this by showing that the built-in face feature is working in some images. There, the face feature correctly shifts attention to the faces when added and also improves the information gain. These images are selected based on the highest performance gains stemming from the addition of this face feature to DeepRARE and only show faces. Thus the selected feature in VGG16 is connected with face detection.

However, the shortcomings of the face feature are apparent in Figure 12 and Figure 13. The addition of the face feature can also worsen DeepRARE's performance by wrongly attributing too much attention to faces. Additionally, the face feature also detects faces of dogs and cats and, as the drop in information gain suggests, attributes too much saliency to them (Figure 12). Additionally, not all human faces are detected by the built-in face feature (Figure 13). This can lead to significant drops in performance on these images. The MTCNN face detection does detect these faces, however, leading to high scores even without a saliency model.

5.4.2 Discussion

The built-in face feature detects faces and thus improves model performance. It also detects non-human faces which can also be beneficial. The performance drops due to using the face feature can likely be explained by wrong predicted feature importance with the rarity approach rather than being a problem of the features. However, the critical shortcoming of the face feature seems to be false negatives. Faces are not always detected, better face features lead to improved scores. This is additional evidence for the importance of high-level features for saliency prediction. Moreover, this showcases the possibility of improving DeepRARE's overall performance by using stronger features in general.

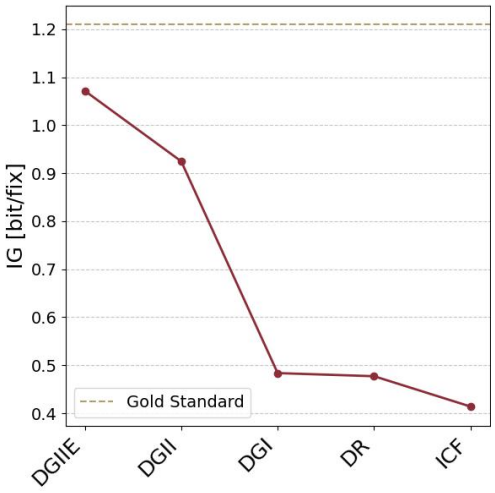
5.5 Evaluation on MIT300

Until now, all evaluations have been performed on the publicly available training dataset MIT1003. For a final quantitative evaluation of DeepRARE, the scores on the test set MIT300 (Judd, Durand, and Torralba 2012) from the MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.) are calculated. The models are still optimized for information gain in terms of the center bias and non-linearity on MIT1003, but critically they are not optimized on MIT300. The scores of

other models from the literature are directly taken from the MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.).

5.5.1 Results

The scores on MIT300 (Judd, Durand, and Torralba 2012) of DeepRARE along with several other models for comparison can be found in Figure 14. First, DeepRARE outperforms its predecessor RARE as well as the strongest low-level feature model CovSal by solid margins. ICF and eDN are the first, and, to date, best performing, deep learning saliency models not using transfer learning. DeepRARE also beats these two models by at least 6% DeepRARE's performance is comparable to that of DeepGaze I, an early transfer learning model. However, the later versions of DeepGaze with better features outperform DeepRARE by more than 44%. Most notably, DeepGaze II does so using the related VGG19 features only.



(a) Information Gains (IG) in bit/fix on MIT300. Models: DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepGaze II (DGII) (Kümmerer, Wallis, and Bethge 2016), DeepGaze I (DGI) (Kümmerer, Theis, and Bethge 2015), DeepRARE (DR) (Riche, Mancas, Gosselin, et al. 2012), and Intensity Contrast Features (ICF) (Kummerer et al. 2017). Exact values and AUC scores in Figure 14b.

Model	IG [bit/fix]	AUC
Gold Standard	1.7366	0.9341
DeepGaze IIE	1.0715	0.8829
DeepGaze II	0.9247	0.8733
DeepGaze I	0.4836	0.8427
DeepRARE	0.4771	0.8382
ICF	0.4140	0.8330
eDN		0.8171
CovSal		0.8116
RARE		0.7700

(b) MIT300 IG and AUC scores and Gold Standard from Kümmerer, Bylinskii, et al. n.d. Models: DeepGaze IIE (Linardos et al. 2021), DeepGaze II (Kümmerer, Wallis, and Bethge 2016), DeepGaze I (Kümmerer, Theis, and Bethge 2015), DeepRARE (Riche, Mancas, Gosselin, et al. 2012), and Intensity Contrast Features (ICF) (Kummerer et al. 2017), Ensembles of Deep Networks (eDN) (Vig, Dorr, and Cox 2014), CovSal (E. Erdem and A. Erdem 2013), RARE (Riche, Mancas, Gosselin, et al. 2012)

Figure 14: Comparison of models based on Information Gains (IG) and AUC on MIT300.

5.5.2 Discussion

The results show that DeepRARE is the most capable model in the MIT/Tuebingen Saliency Benchmark (Kümmerer, Bylinskii, et al. n.d.) that does not use transfer learning. All low-level models are significantly outperformed. Therefore, high-level features are necessary for good saliency prediction.

However, DeepRARE is outperformed by transfer learning models, even when they use comparable features. Hence, the rarity approach lacks behind transfer learning when it comes to determining feature importance for saliency prediction.

6 Conclusion

In this paper, DeepRARE is evaluated both by inspecting example performances and error cases compared to other saliency prediction models and by principled quantitative evaluation. The quality of DeepRARE's features and the ability of the rarity approach to predict saliency based on these features are investigated.

DeepRARE outperforms RARE and other low-level feature approaches. Therefore, high-level features are critical for saliency prediction. DeepRARE's high-level features work decently well at detecting relevant objects, but as the face and text detectors show can be improved further. This shows that using stronger features for DeepRARE will yield further performance gains.

Compared to state-of-the-art models, the main reason for lower performance stems from the advantages of transfer learning compared to the rarity approach. In many cases, rarity fails to predict accurate importance scores for the detected features. This comparison is not fair as transfer learning uses additional resources in the form of supervision on eye tracking data. When comparing DeepRARE to other unsupervised approaches, it achieves the highest performance showcasing the strength of the rarity approach in this unsupervised setting.

7 Future Research Directions

The performance gains from the face and the text detectors highlight that high-level features are essential for DeepRARE's performance. Also, DeepGaze shows that, although a different framework is used, using better features and combining feature extractors enabled higher performances. Therefore, future research could explore the performance ceiling of DeepRARE by using several stronger feature extractors.

Additionally, one could explore the performance of DeepRARE on the P3 and O3 datasets that due to their pop-out effects are an ideal fit for the rarity approach. For this, one would need a proper ground truth by performing eye-tracking experiments on P3 and O3. As DeepGaze struggles with these datasets, one could attempt a fusion of the two approaches.

References

Mancas, Matei, Phutphalla Kong, and Bernard Gosselin (May 2020). *Visual Attention: Deep Rare Features*. arXiv:2005.12073 [cs]. DOI: 10.48550/arXiv.2005.12073. URL: http://arxiv.org/abs/2005.12073 (visited on 12/05/2024).

Simonyan, Karen and Andrew Zisserman (Apr. 2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556 [cs]. DOI: 10.48550/arXiv.1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 12/30/2024).

Kümmerer, Matthias, Zoya Bylinskii, et al. (n.d.). *MIT/Tübingen Saliency Benchmark*. https://saliency.tuebingen.ai/.

Broadbent, D.E. (1958). "CHAPTER 9 - IMMEDIATE MEMORY AND THE SHIFTING OF ATTENTION". In: *Perception and Communication*. Ed. by D.E. Broadbent. Pergamon, pp. 210–243. ISBN: 978-1-4832-0079-8. DOI: https://doi.org/10.1016/B978-1-4832-0079-8.50011-6. URL: https://www.sciencedirect.com/science/article/pii/B9781483200798500116.

Mancas, Matei and Olivier Le Meur (Sept. 2016). "Applications of Saliency Models". In: *From Human Attention to Computational Attention. A Multidisciplinary Approach*. Ed. by Matei Mancas et al. Vol. 10. Springer Series in Cognitive and Neural Systems. Springer, pp. 331–377. DOI: 10.1007/978-1-4939-3435-5_18. URL: https://inria.hal.science/hal-01393254.

Kümmerer, Matthias and Matthias Bethge (Sept. 2023). "Predicting Visual Fixations". en. In: *Annual Review of Vision Science* 9.1, pp. 269–291. ISSN: 2374-4642, 2374-4650. DOI: 10.1146/annurev-vision-120822-072528. URL: https://www.annualreviews.org/doi/10.1146/annurev-vision-120822-072528 (visited on 12/05/2024).

Itti, L., C. Koch, and E. Niebur (Nov. 1998). "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11, pp. 1254–1259. ISSN: 01628828. DOI: 10.1109/34.730558. URL: http://ieeexplore.ieee.org/document/730558/ (visited on 12/05/2024).

Kümmerer, Matthias, Thomas S. A. Wallis, and Matthias Bethge (Oct. 2016). *DeepGaze II: Reading fixations from deep features trained on object recognition*. arXiv:1610.01563 [cs]. DOI: 10.48550/arXiv.1610.01563. URL: http://arxiv.org/abs/1610.01563 (visited on 12/05/2024).

Linardos, Akis et al. (Sept. 2021). *DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling*. arXiv:2105.12441 [cs]. DOI: 10.48550/arXiv.2105.12441. URL: http://arxiv.org/abs/2105.12441 (visited on 12/05/2024).

Riche, Nicolas, Matei Mancas, Bernard Gosselin, et al. (Sept. 2012). "Rare: A new bottom-up saliency model". In: *2012 19th IEEE International Conference on Image Processing*. Orlando, FL, USA: IEEE, pp. 641–644. ISBN: 978-1-4673-2533-2 978-1-4673-2534-9 978-1-4673-2532-5. DOI: 10.1109/ICIP.2012.6466941. URL: http://ieeexplore.ieee.org/document/6466941/ (visited on 12/05/2024).

Kong, Phutphalla et al. (Sept. 2021). *DeepRare: Generic Unsupervised Visual Attention Models*. arXiv:2109.11439 [cs]. DOI: 10.48550/arXiv.2109.11439. URL: http://arxiv.org/abs/2109.11439 (visited on 12/05/2024).

Mahdi, Ali and Jun Qin (Sept. 2017). *DeepFeat: A Bottom Up and Top Down Saliency Model Based on Deep Features of Convolutional Neural Nets*. arXiv:1709.02495 [cs]. DOI: 10.48550/arXiv.1709.02495. URL: http://arxiv.org/abs/1709.02495 (visited on 12/05/2024).

He, Kaiming et al. (2015). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

Russakovsky, Olga et al. (2015). *ImageNet Large Scale Visual Recognition Challenge*. arXiv: 1409.0575 [cs.CV]. URL: https://arxiv.org/abs/1409.0575.

Zhou, Bolei et al. (2015). *Learning Deep Features for Discriminative Localization*. arXiv: 1512.04150 [cs.CV]. URL: https://arxiv.org/abs/1512.04150.

Judd, Tilke, Krista Ehinger, et al. (2009). "Learning to predict where humans look". In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 2106–2113. DOI: 10.1109/ICCV.2009.5459462.

Kümmerer, Matthias (2025). *pysaliency: Python Framework for Saliency Modeling and Evaluation*. Accessed: 2025-01-08. URL: https://github.com/matthias-k/pysaliency.

Kümmerer, Matthias, Thomas S. A. Wallis, and Matthias Bethge (2015). "Information-theoretic model comparison unifies saliency metrics". In: *Proceedings of the National Academy of Sciences* 112.52, pp. 16054–16059. DOI: 10.1073/pnas.1510393112. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1510393112. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1510393112.

Institute, NUMEDIART (2025). *VisualAttention-DeepRare2019: Paper Results Reproduction*. Accessed: 2025-01-08. URL: https://github.com/numediart/VisualAttention-DeepRare2019/tree/master?tab=readme-ov-file#paper-results-reproduction.

Judd, Tilke, Frédo Durand, and Antonio Torralba (2012). "A Benchmark of Computational Models of Saliency to Predict Human Fixations". In: URL: https://api.semanticscholar.org/CorpusID:16750736.

Riche, Nicolas, Matei Mancas, Matthieu Duvinage, et al. (2013). "RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis". In: *Signal Process. Image Commun.* 28, pp. 642–658. URL: https://api.semanticscholar.org/CorpusID:44184156.

Itti, L. (2004). "Automatic foveation for video compression using a neurobiological model of visual attention". In: *IEEE Transactions on Image Processing* 13.10, pp. 1304–1318. DOI: 10.1109/TIP.2004.834657.

Paz Centeno, Iván de (2019). *MTCNN: Implementation of MTCNN (Multi-task Cascaded Convolutional Neural Networks) for face detection in TensorFlow and Keras*. https://github.com/ipazc/mtcnn. Accessed: 2025-01-12.

JaidedAI (2020). *EasyOCR: Ready-to-use OCR with 80+ supported languages in Python*. https://github.com/JaidedAI/EasyOCR/tree/master. Accessed: 2025-01-12.

Kümmerer, Matthias, Lucas Theis, and Matthias Bethge (Apr. 2015). *Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet*. arXiv:1411.1045 [cs]. DOI: 10.48550/arXiv.1411.1045. URL: http://arxiv.org/abs/1411.1045 (visited on 12/05/2024).

Kummerer, Matthias et al. (Oct. 2017). "Understanding Low- and High-Level Contributions to Fixation Prediction". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Vig, Eleonora, Michael Dorr, and David Cox (2014). "Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798–2805. DOI: 10.1109/CVPR.2014.358.

Erdem, Erkut and Aykut Erdem (Mar. 2013). "Visual saliency estimation by nonlinearly integrating features using region covariances". In: *Journal of Vision* 13.4, pp. 11–11. ISSN: 1534-7362. DOI: 10.1167/13.4.11. eprint: https://arvojournals.org/arvo/content_public/journal/jov/932809/i1534-7362-13-4-11.pdf. URL: https://doi.org/10.1167/13.4.11.

8 Appendix: Additional Visualizations

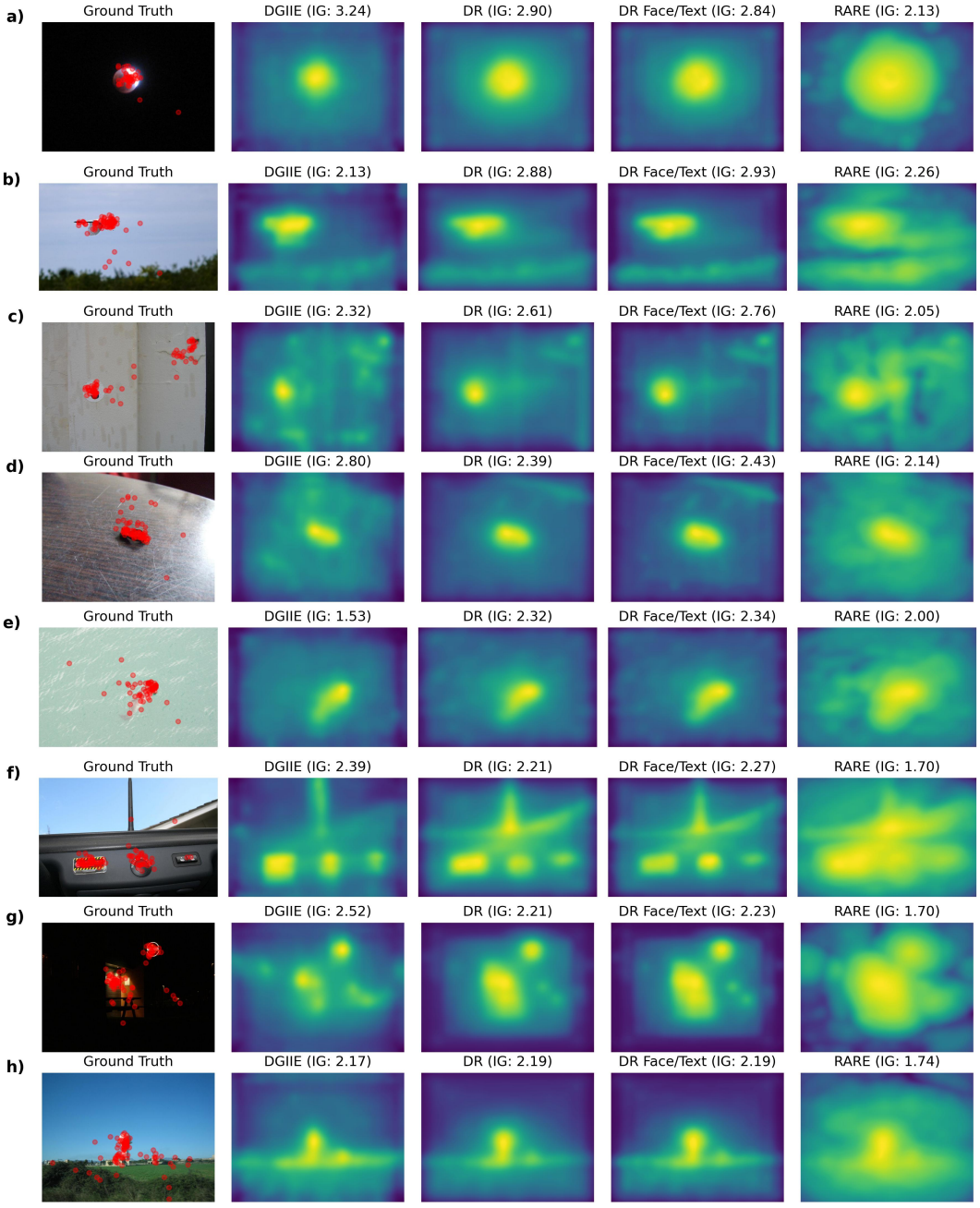


Figure 15: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of DeepRARE

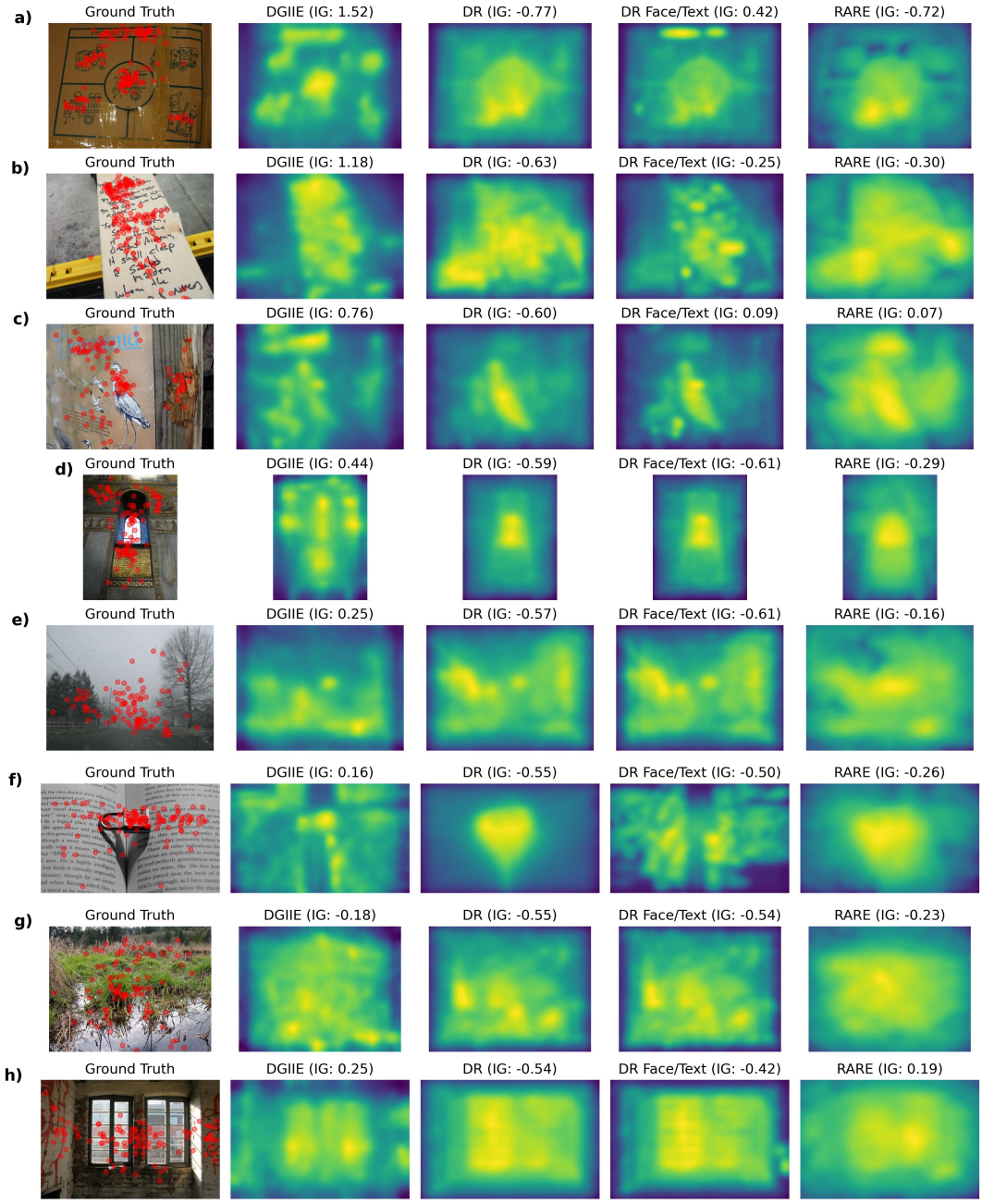


Figure 16: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with lowest IG scores of DeepRARE

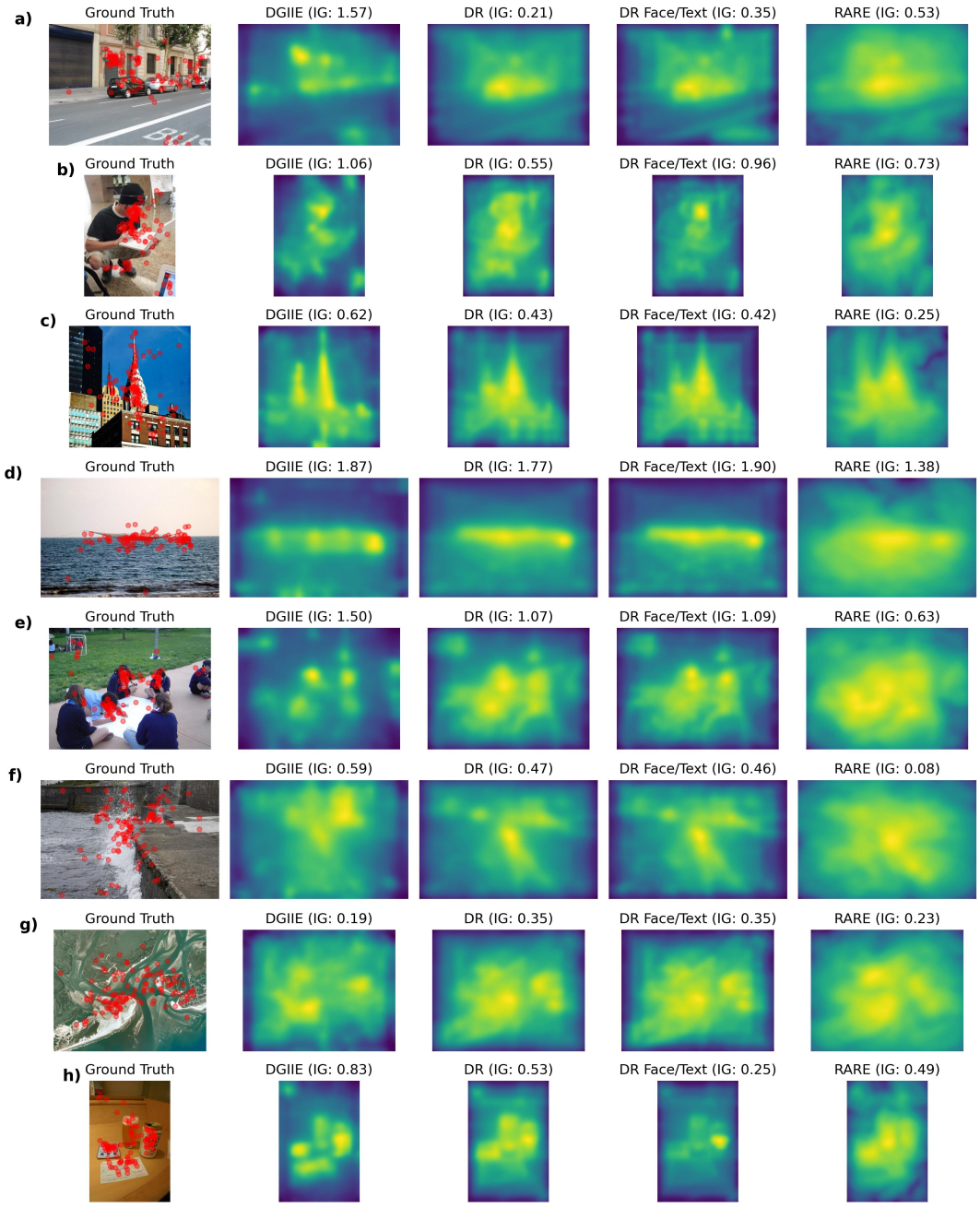


Figure 17: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: 8 random images

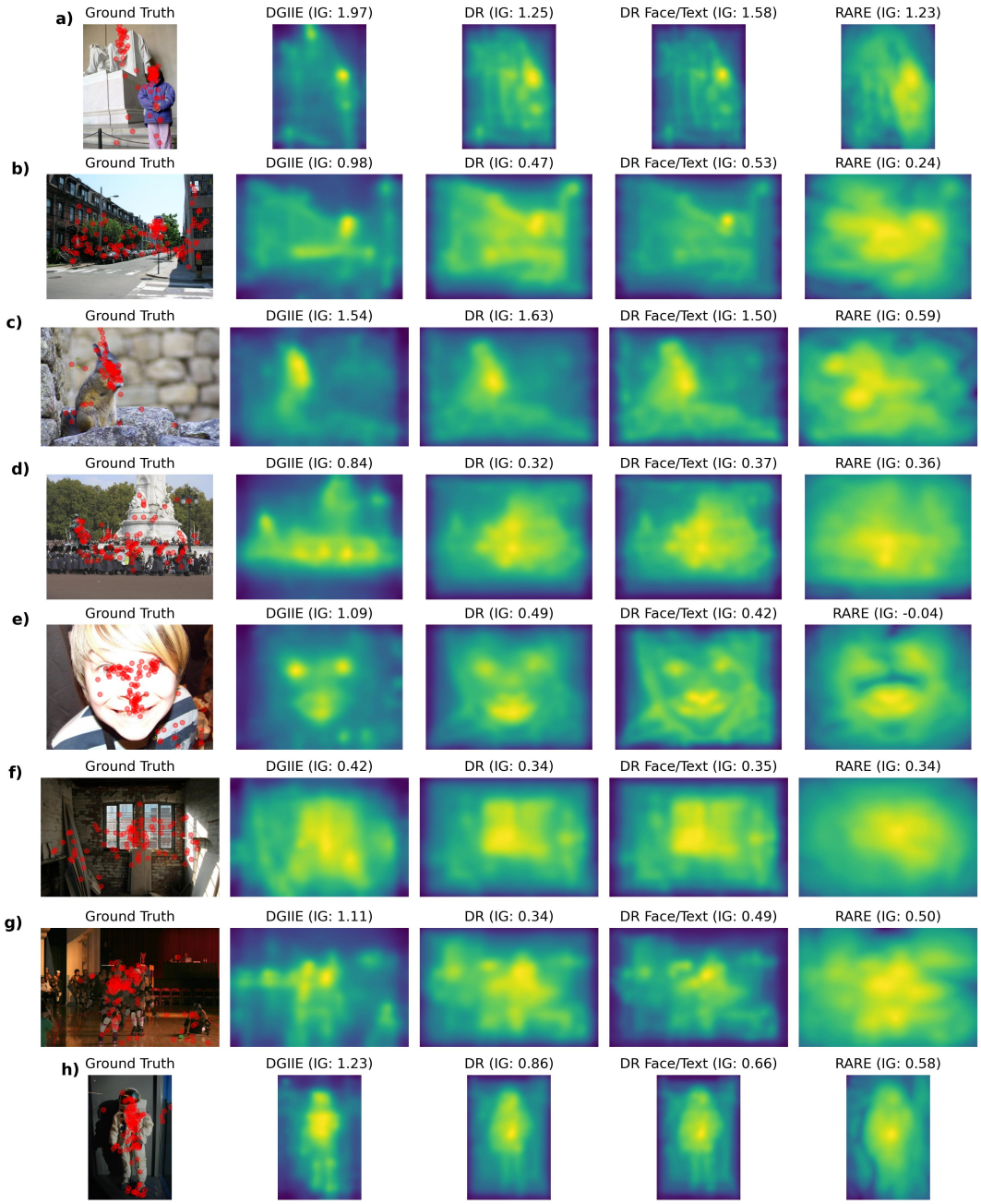


Figure 18: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: 8 random images

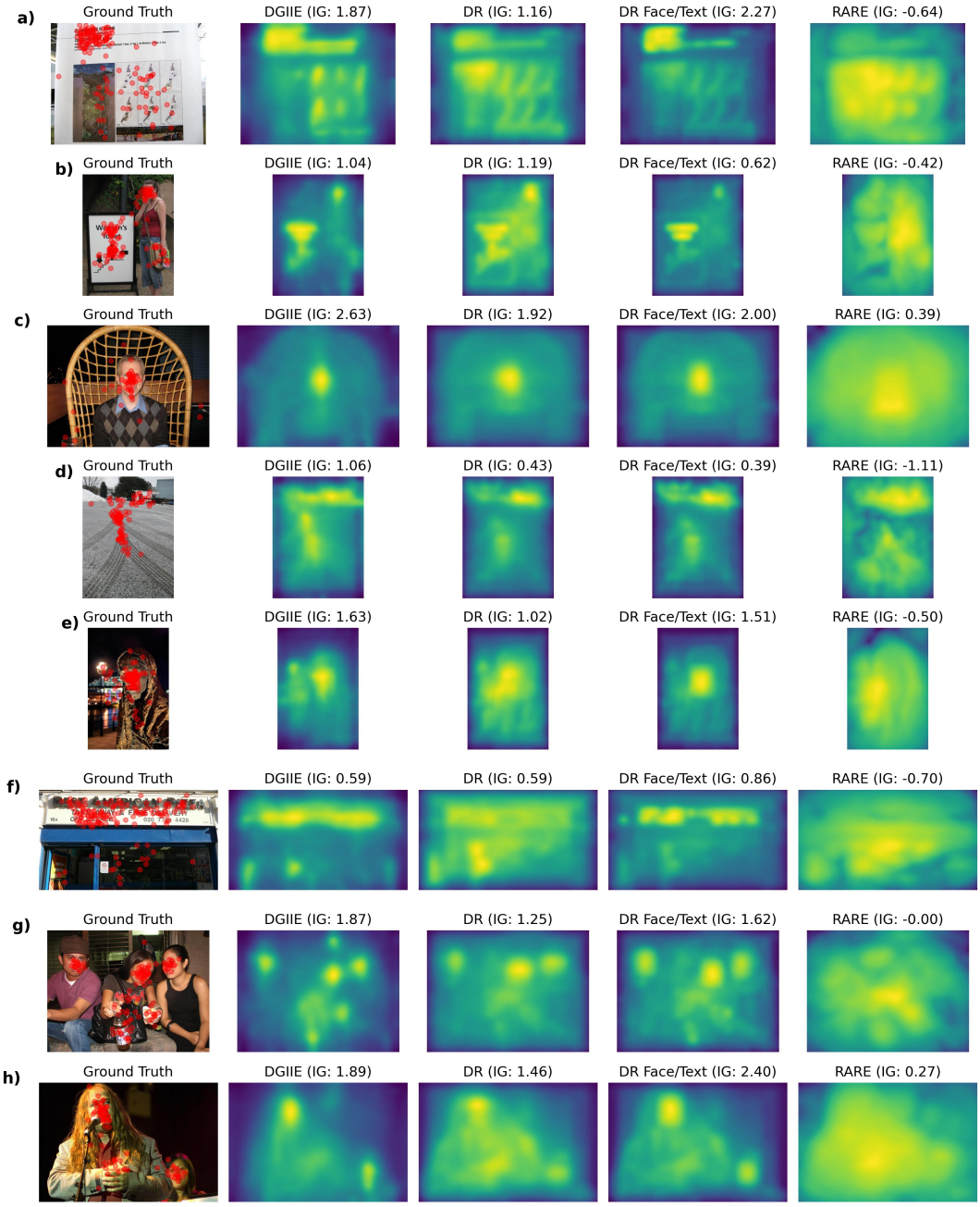


Figure 19: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of DeepRARE compared to RARE. More examples of Figure 3

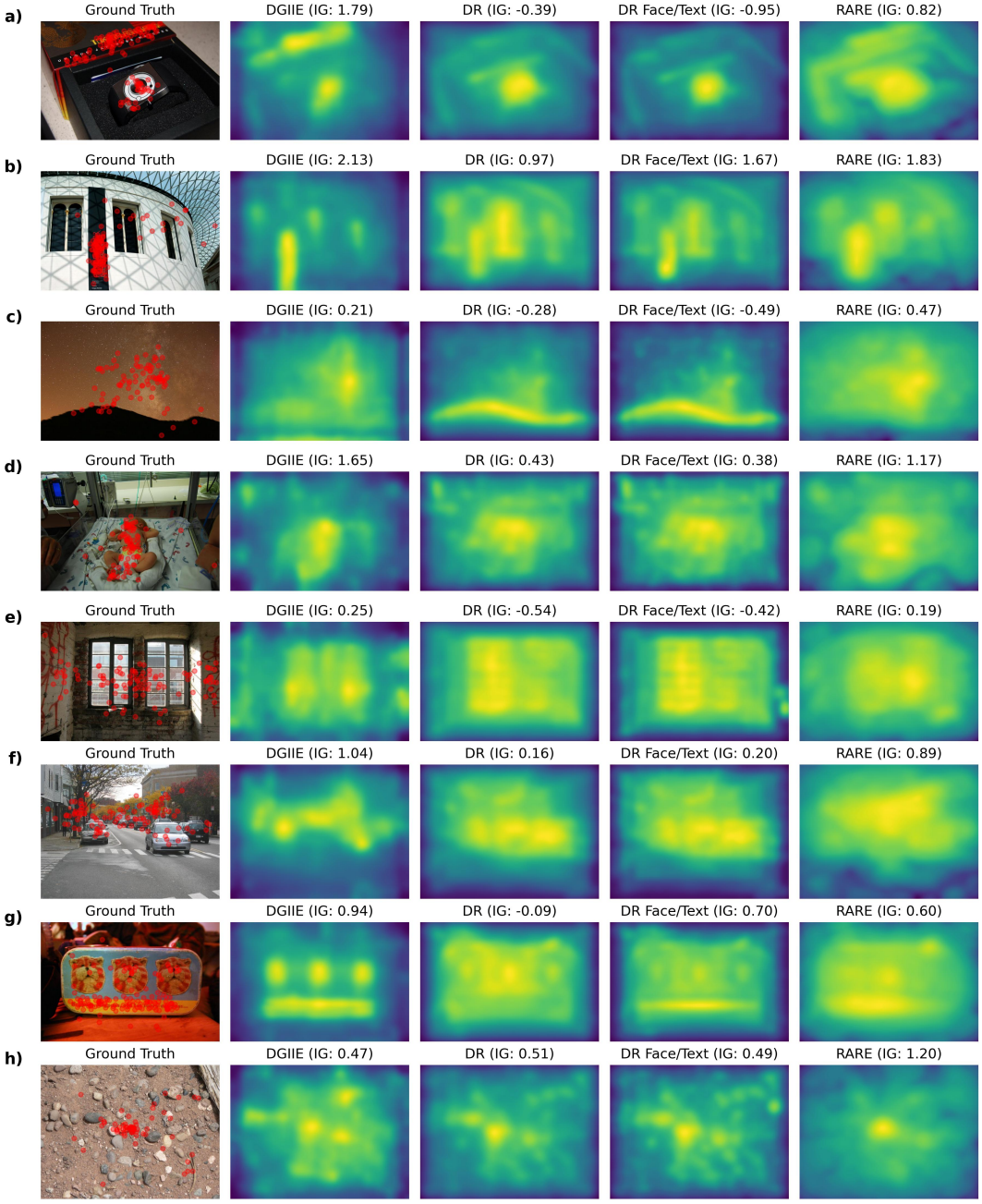


Figure 20: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of RARE compared to DeepRARE. More examples of Figure 4

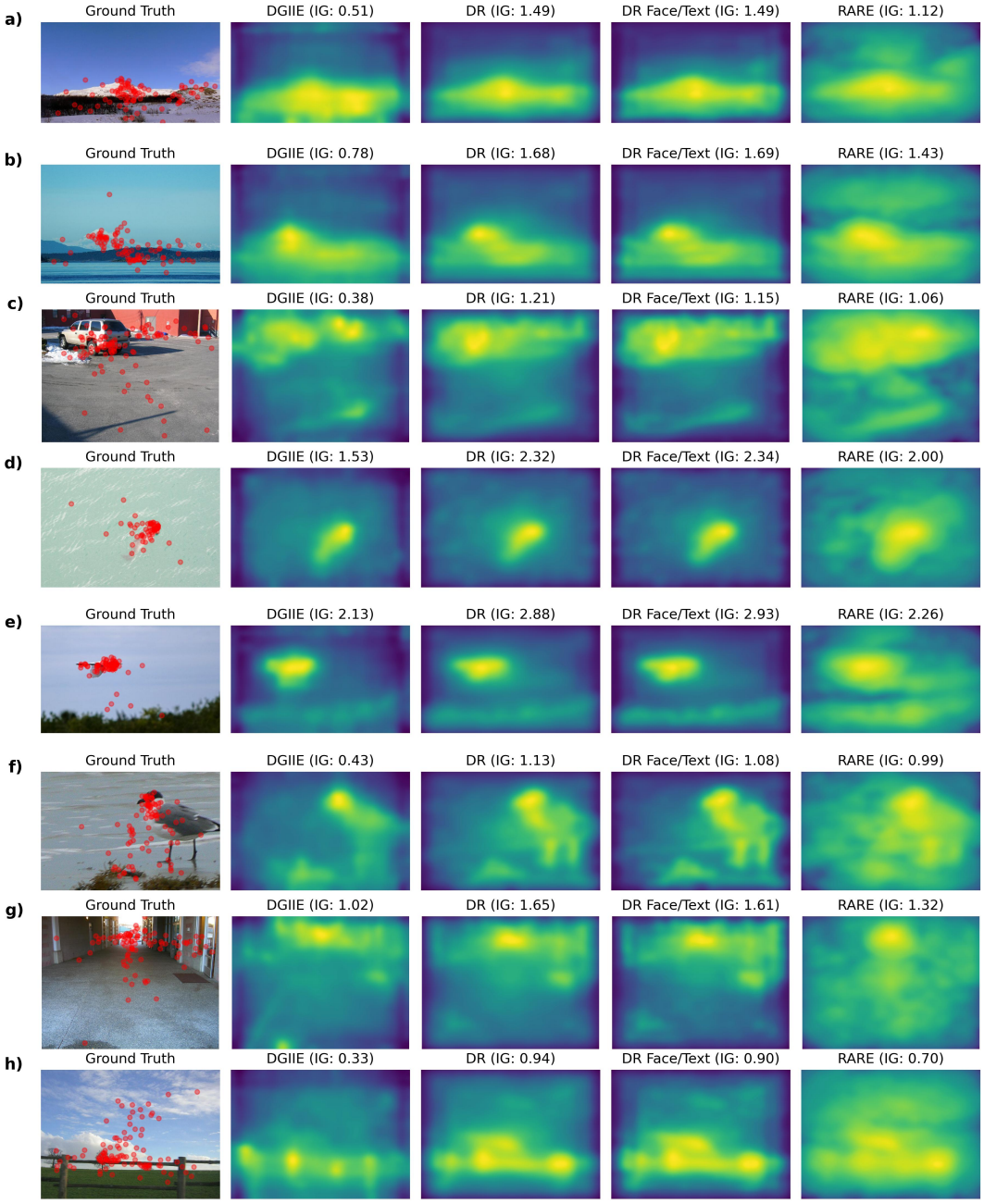


Figure 21: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of DeepRARE compared to DeepGaze IIE. More examples of Figure 5

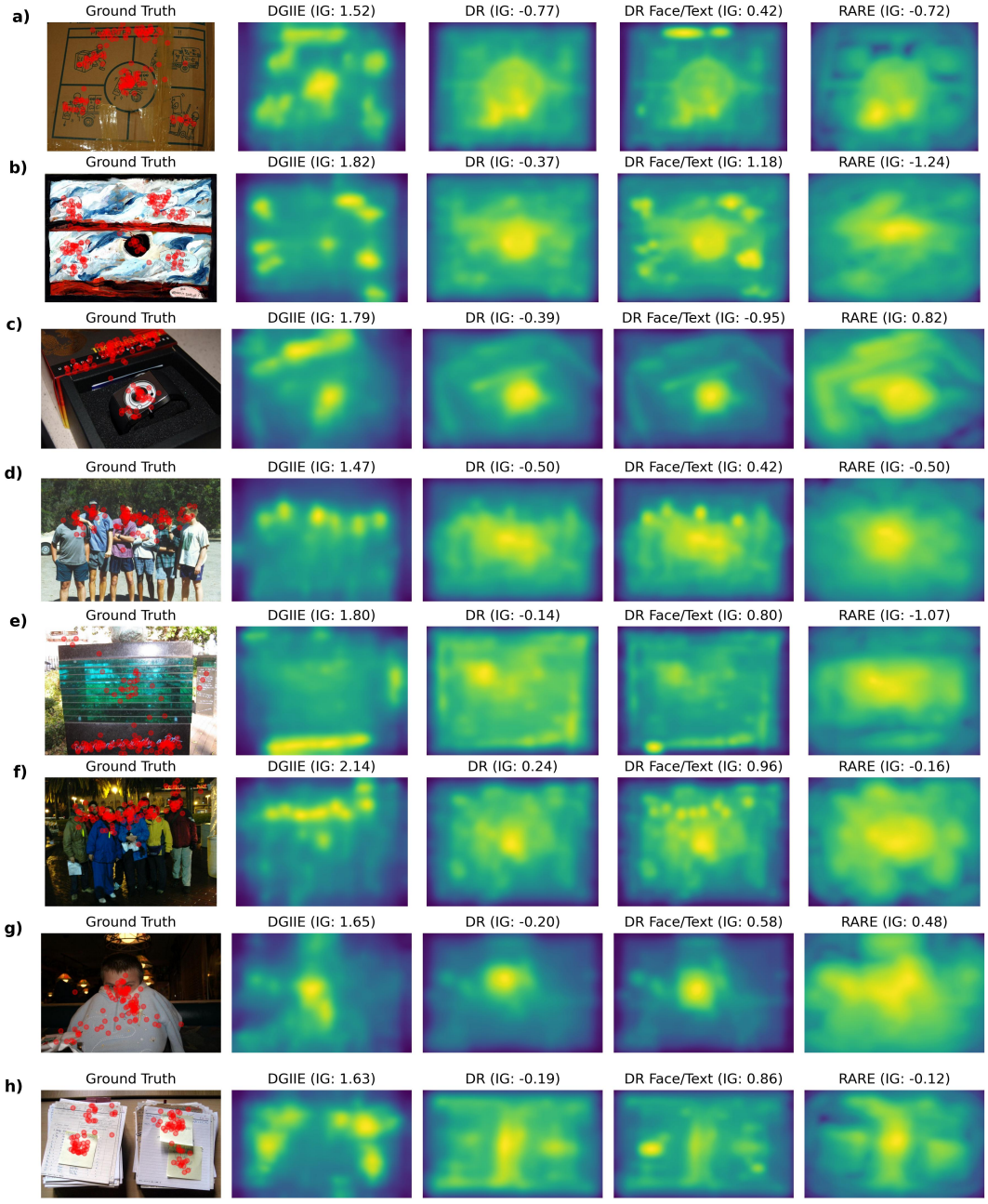


Figure 22: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with lowest IG scores of DeepRARE compared to DeepGaze IIE. More examples of Figure 6

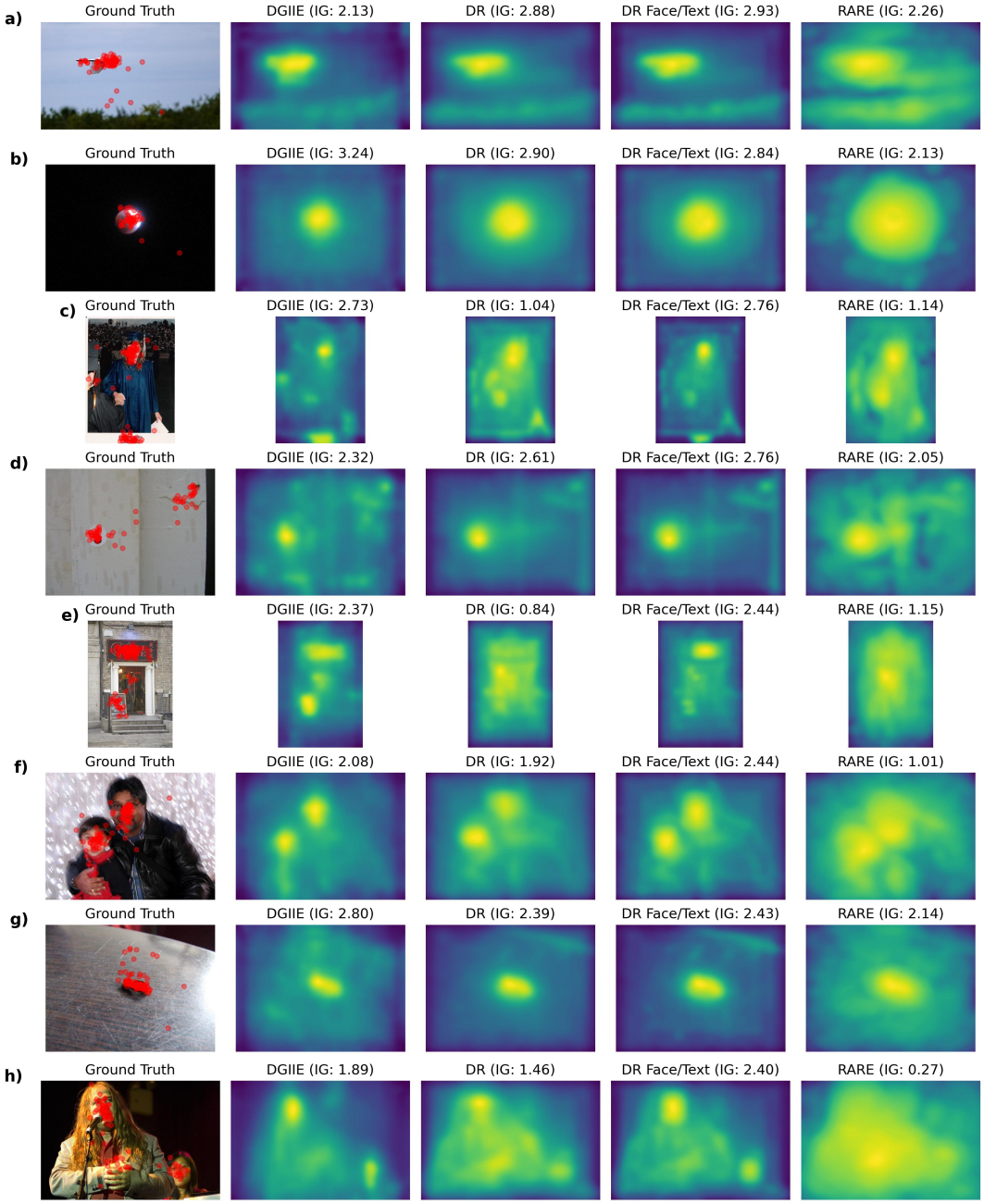


Figure 23: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of DeepRARE with face and text detector

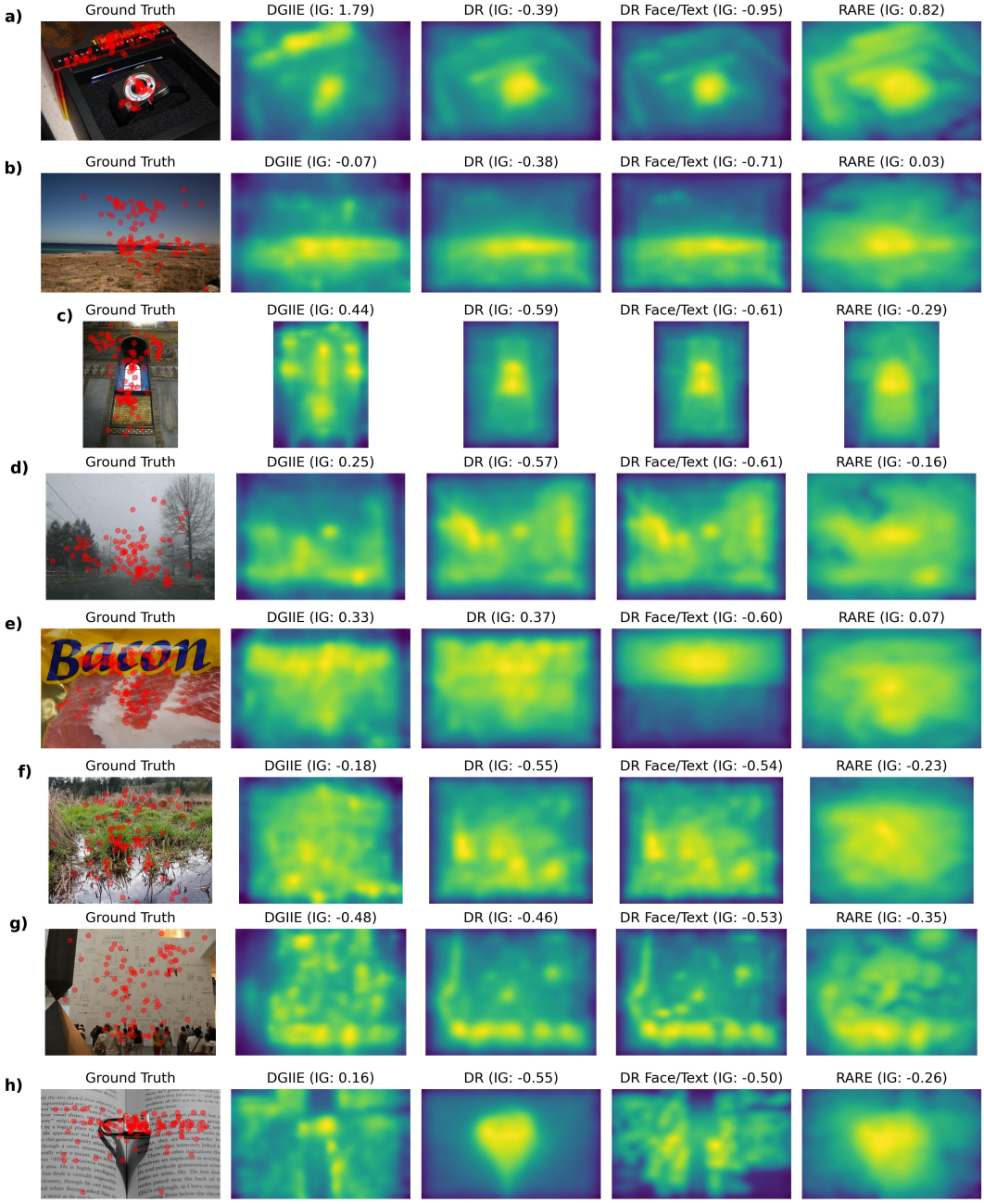


Figure 24: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with lowest IG scores of DeepRARE with face and text detector

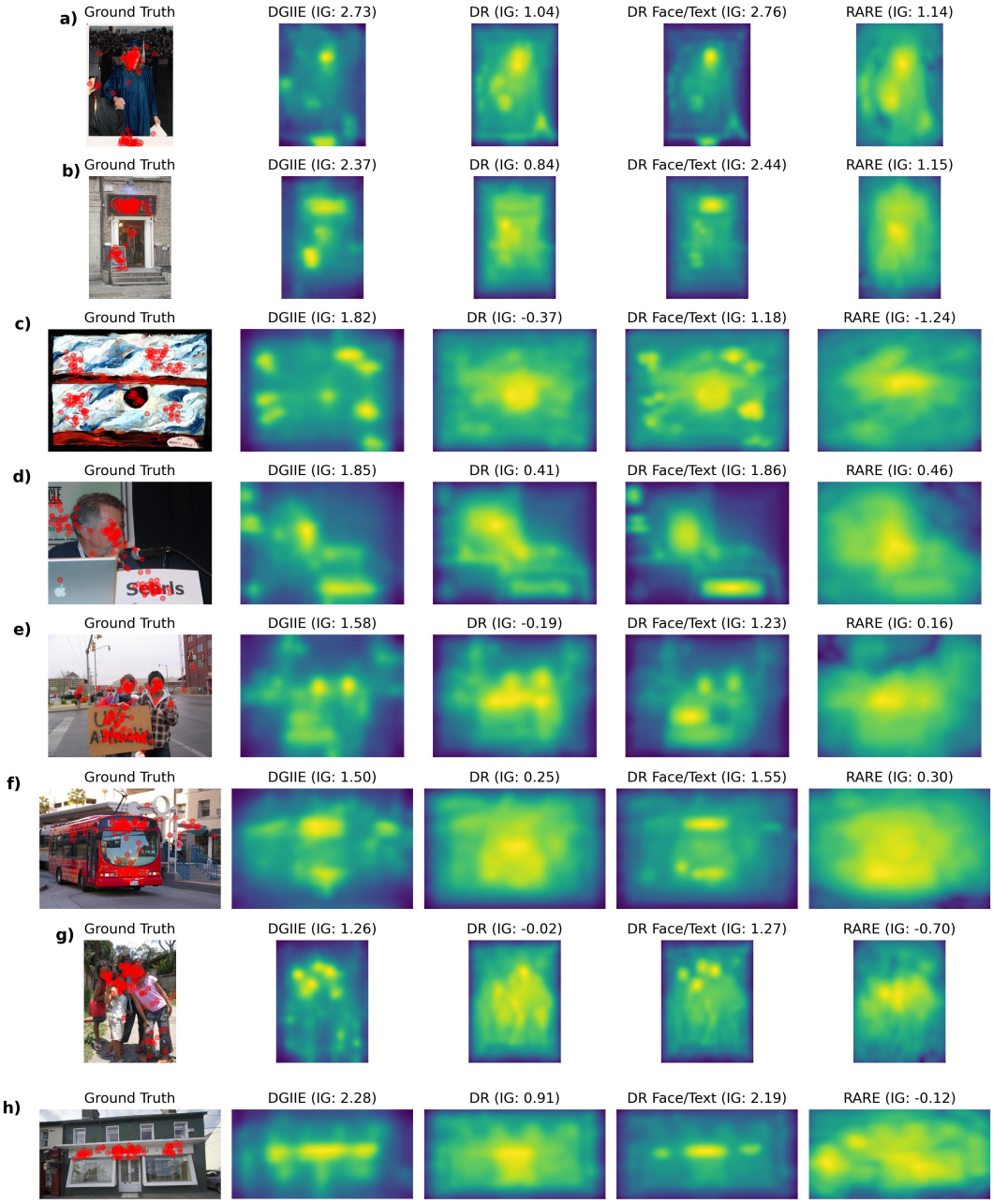


Figure 25: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of DeepRARE with face and text detector compared to DeepRARE. More examples of Figure 8

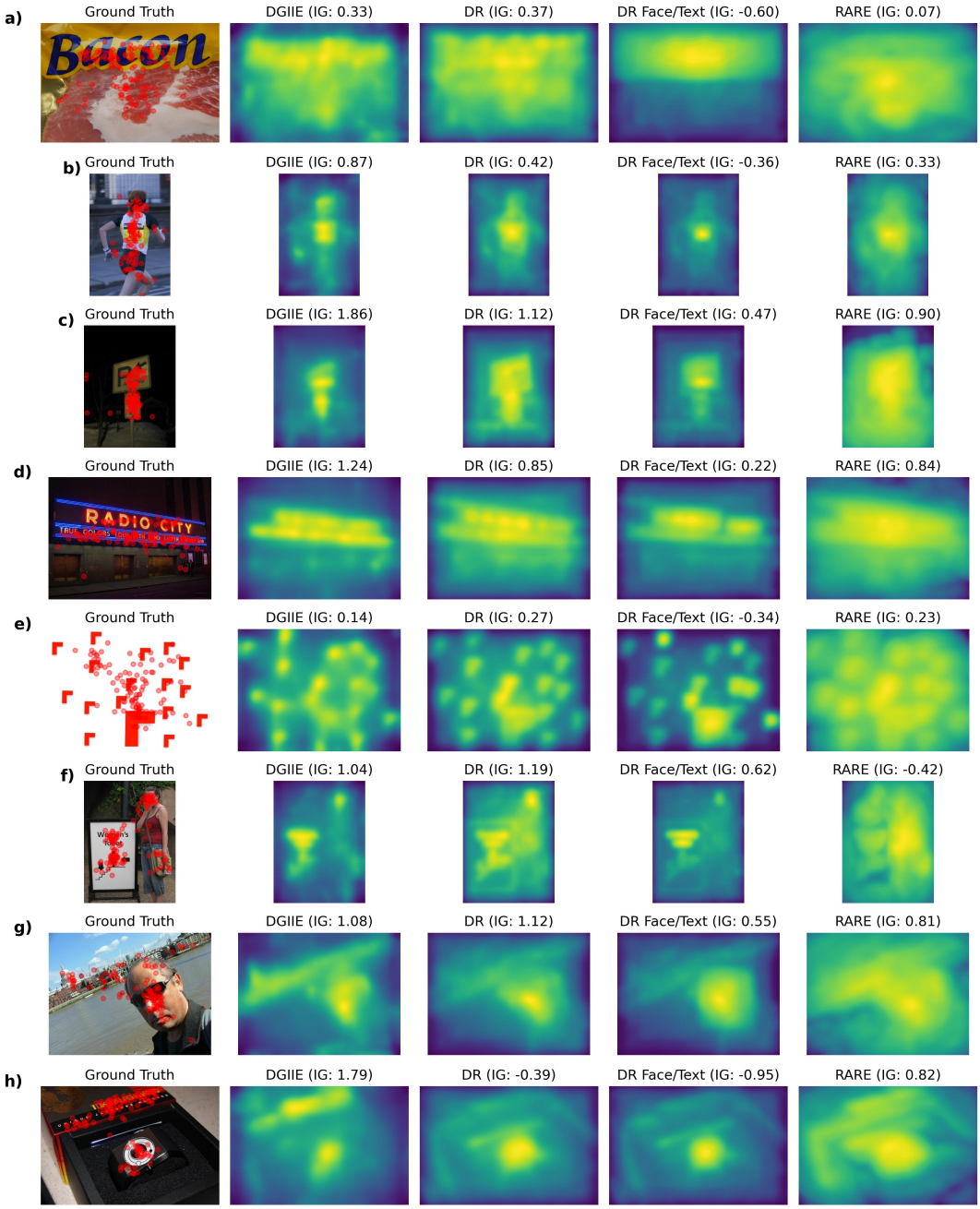


Figure 26: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with lowest IG scores of DeepRARE with face and text detector compared to DeepRARE. More examples of Figure 9

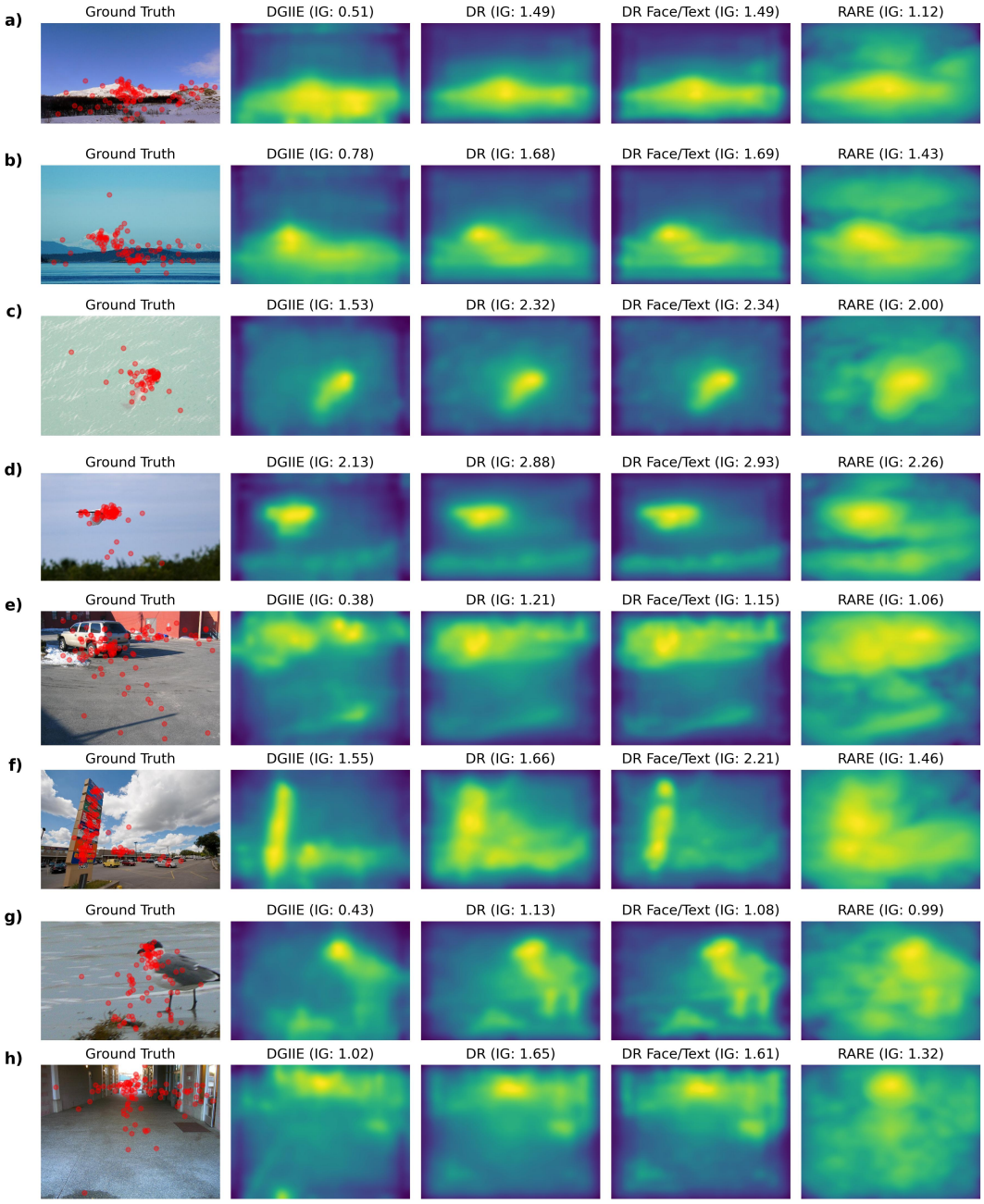


Figure 27: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with highest IG scores of DeepRARE with face and text detector compared to DeepGaze IIE. More examples of Figure 10

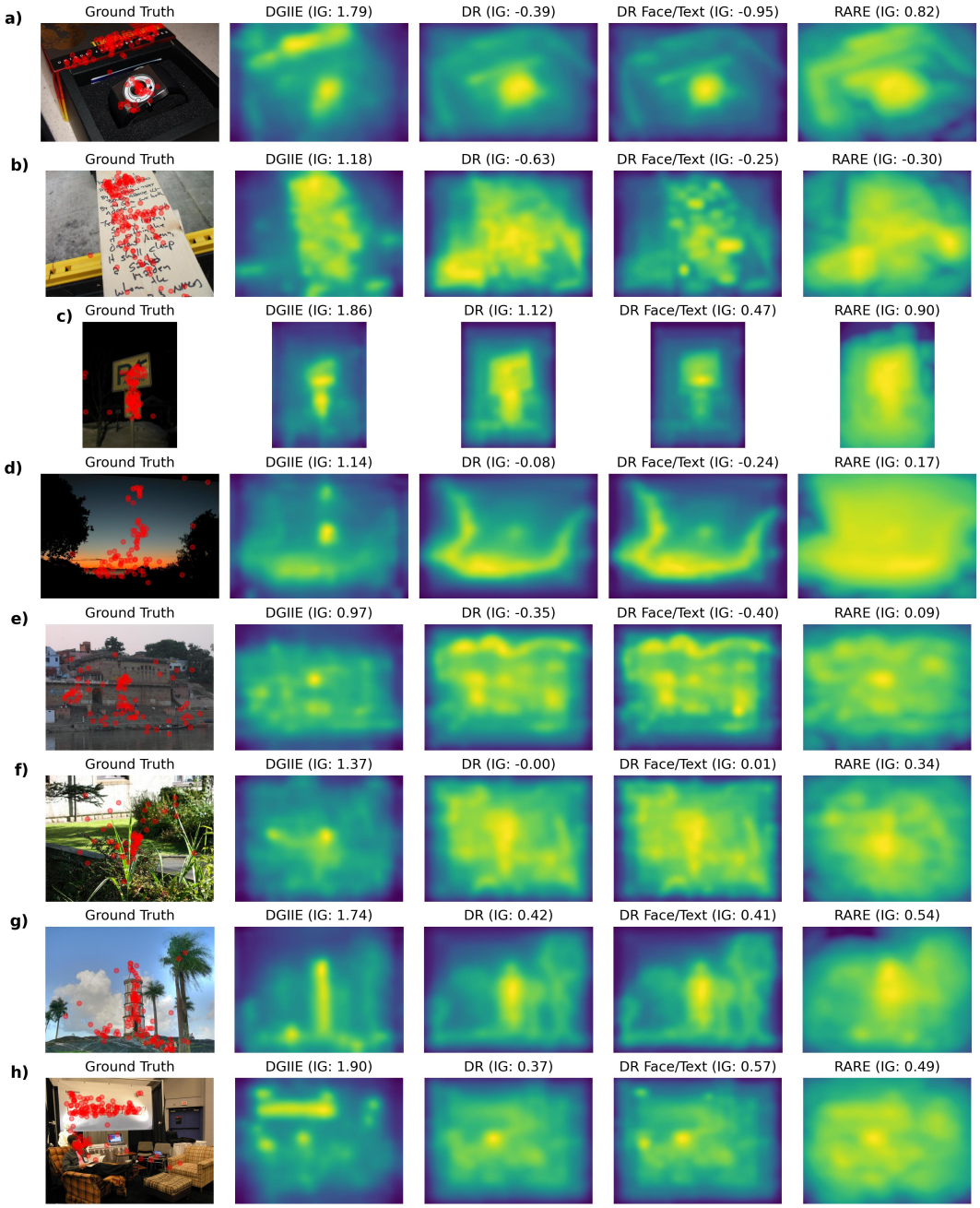


Figure 28: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE with additional face and text detector (DR Face/Text) Section 5.3 and RARE (Riche, Mancas, Gosselin, et al. 2012). Selection: Top 8 images with lowest IG scores of DeepRARE with face and text detector compared to DeepGaze IIE. Related to Figure 10

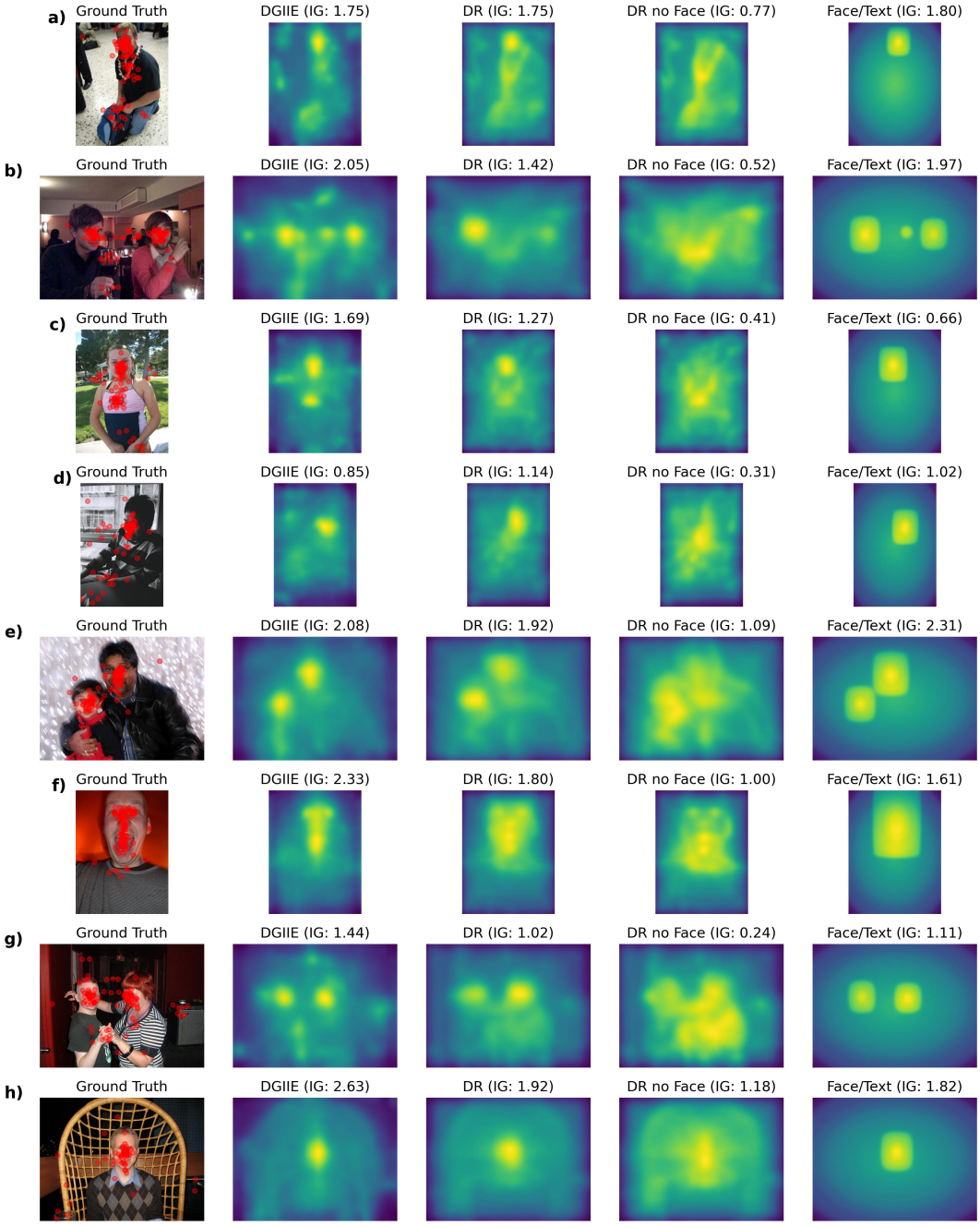


Figure 29: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE without built-in face feature (DR no Face) Section 5.3 and face and text detector with center bias instead of saliency model (Face/Text). Selection: Top 8 images with highest IG scores of DeepRARE compared to DeepRARE without built-in face feature. More examples of Figure 11

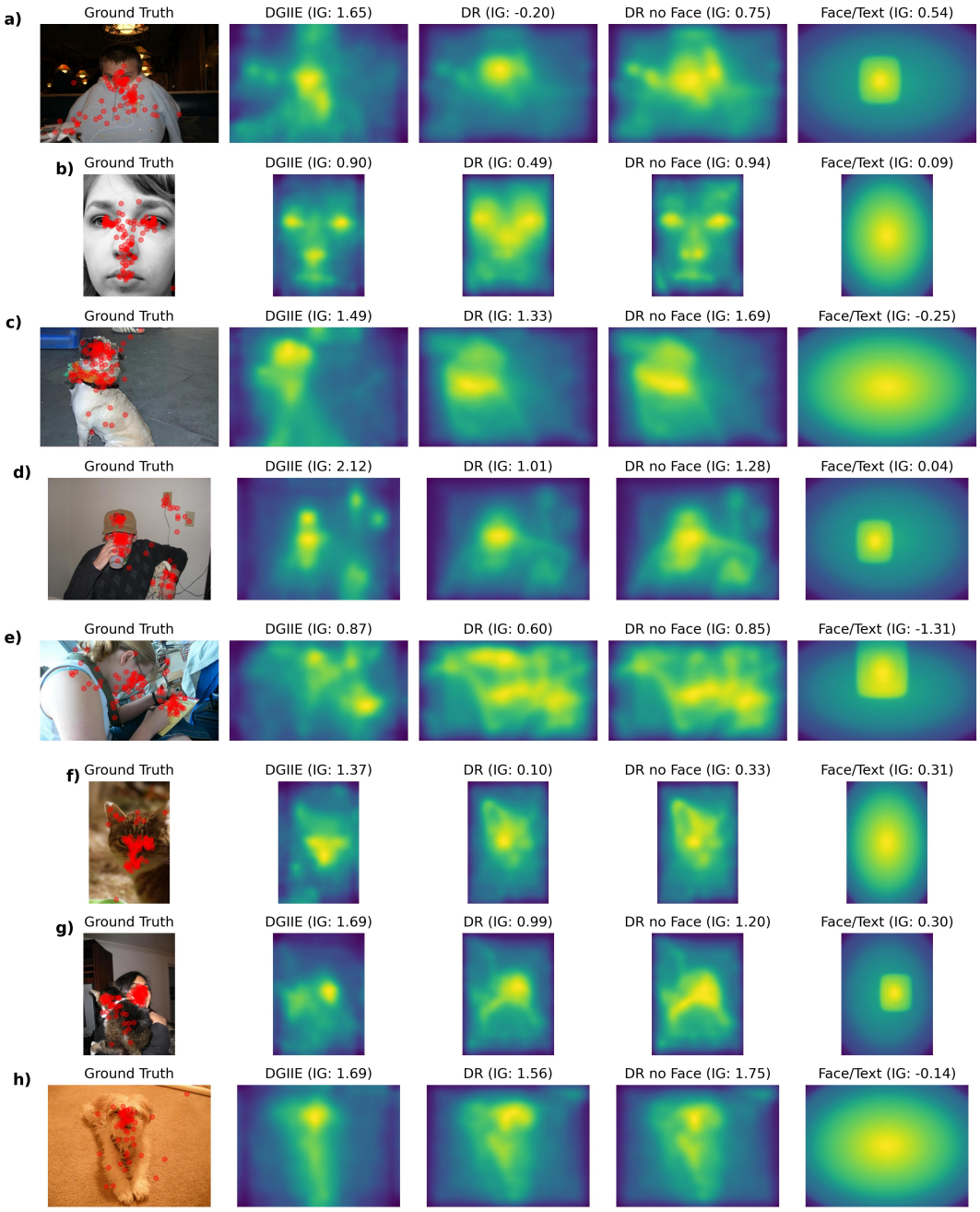


Figure 30: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE without built-in face feature (DR no Face) Section 5.3 and face and text detector with center bias instead of saliency model (Face/Text). Selection: Top 8 images with lowest IG scores of DeepRARE compared to DeepRARE without built-in face feature. More examples of Figure 12

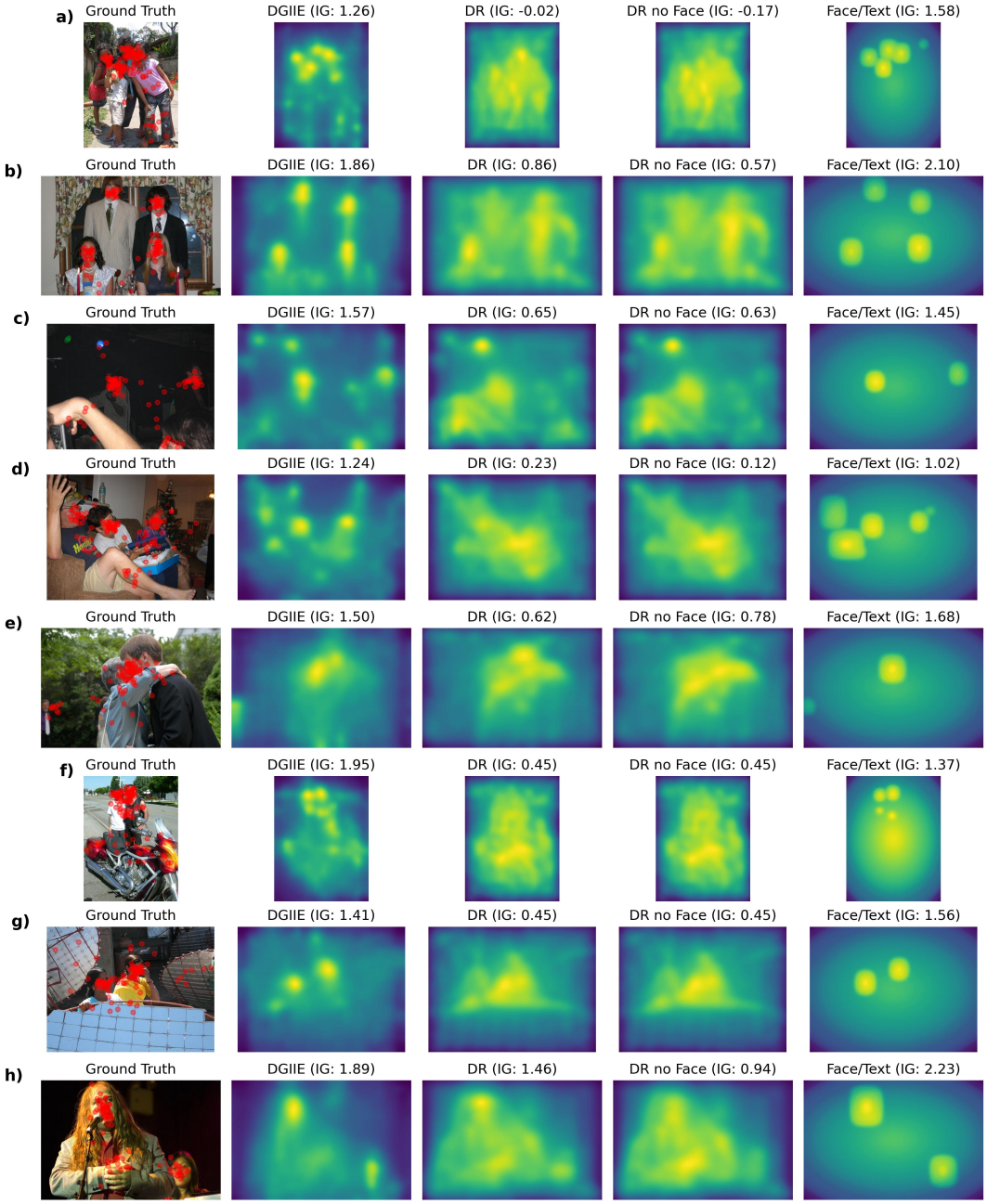


Figure 31: Image with ground truth fixations as red dots, predicted posteriors of DeepGaze IIE (DGIIE) (Linardos et al. 2021), DeepRARE (DR) (Mancas, Kong, and Gosselin 2020), DeepRARE without built-in face feature (DR no Face) Section 5.3 and face and text detector with center bias instead of saliency model (Face/Text). Selection: Top 8 images with highest IG scores of DeepRARE with additional face detector compared to DeepRARE. More examples of Figure 13