Learning Saliency From Fixations

Yasser Abdelaziz Dahou Djilali[1,2] Kevin McGuinness [1] Noel O'Connor[1]
[1]Dublin City University, Ireland [2]Technology Innovation Institute, UAE

A. Additional results

Extra qualitative results on saliency prediction are shown in Figure A.1, it can be seen that SalTR produces high quality fixation and saliency maps compared to the Ground Truth ones. Furthermore, Unisal [1] predicts good saliency maps as well, hence, both models demonstrate their robustness across a variety of image complexities.

Failure cases. In an effort to delve deeper into the SalTR capabilities, we have identified certain anomalies within the Salicon validation set, specifically images marked by the highest Kullback-Leibler Divergence (KLD) value of 1.51 and the lowest Normalized Scanpath Saliency (NSS) score of 1.09. As visualized in Figure A.2, we have so far pinpointed two areas where the model did not perform as expected, which we refer to as *failure modes*. In the top image, the model fails to sharpen the prediction around the human, that can be considered as a small object. The lack of hierarchical decoding in our framework may be the reason behind this failure, it is important to consider decoding multiple resolutions of the latent representations to allow the SalTR to adapt to multiple scale fixation regions. A second failure we observe in the last image, where the Ground Truth fixations are all concentrated around the rabbit in the center. The Hungarian matching forces distinct predictions, thus, it favors more spread out predictions. This behaviour is overcomed by the decoder when the input image does not account any other candidate salient regions. Clearly however, when the decoder picks other informative areas in this scenario, it tends to uniformly assign fixations.

No bipartite matching. Figure A.4 illustrates the results of the SalTR model trained without the incorporation of the Hungarian matching loss. As can be observed, the predictions lack diversity and are all centered around the same region, akin to a duplicated query trained for 100 epochs. The lack of variety in the predicted regions suggests that the model is picking a specific feature or set of features in the training data. In the absence of the Hungarian matching loss, which usually serves to optimize assignment between predictions and ground truth, the SalTR model seems to struggle with providing unique, diversified predictions. The model appears to 'latch on' to a particular set of features or patterns, consequently producing a narrowed range of output. This behavior strongly suggests the crucial role of the Hungarian matching loss in facilitating SalTR's ability to make diverse predictions. Without it, the model's ability to generalize well across varying input data seems to be significantly hindered. The training duration of 100 epochs, in this case, does not appear to alleviate the observed issue.

Low-level features. We evaluate our model's performance using images from both the P3 and O3 datasets [2]. The model's performance on real images, while not perfect, is quite commendable, as illustrated in Figure A.5. It successfully identifies the feature in question and assigns a higher saliency density to it. Concurrently, it maintains attention to other regions in the image, demonstrating an acceptable level of distribution in its focus. However, defining an ideal saliency map for such images is an intricate endeavor. It is generally agreed that humans tend to initially focus on the most visually prominent or 'bottom-up' features within the first moments of observation. Following this, exploratory eye movements typically begin, covering other areas of interest in the image. Capturing this dual-phase attention process is a challenging task that current models are yet to master comprehensively.

On the other hand, our model's performance on synthetic images, as displayed in Figure A.6, is noticeably weaker. The model appears to struggle with understanding and interpreting the abstract features typically presented in synthetic imagery. This shortcoming is possibly attributed to the learning process. Specifically, the Salicon training set, which our model was trained on, does not include any synthetic images. The model's ability to generalize from the real images in the training set to synthetic images in the test set seems to be a significant hurdle.

In light of these observations, future efforts might be geared towards including a more diverse range of image types in the training set, particularly synthetic images. Furthermore, more research is needed to improve the model's ability to mimic the human attention process more accurately, especially concerning the transition from bottom-up to exploratory attention.

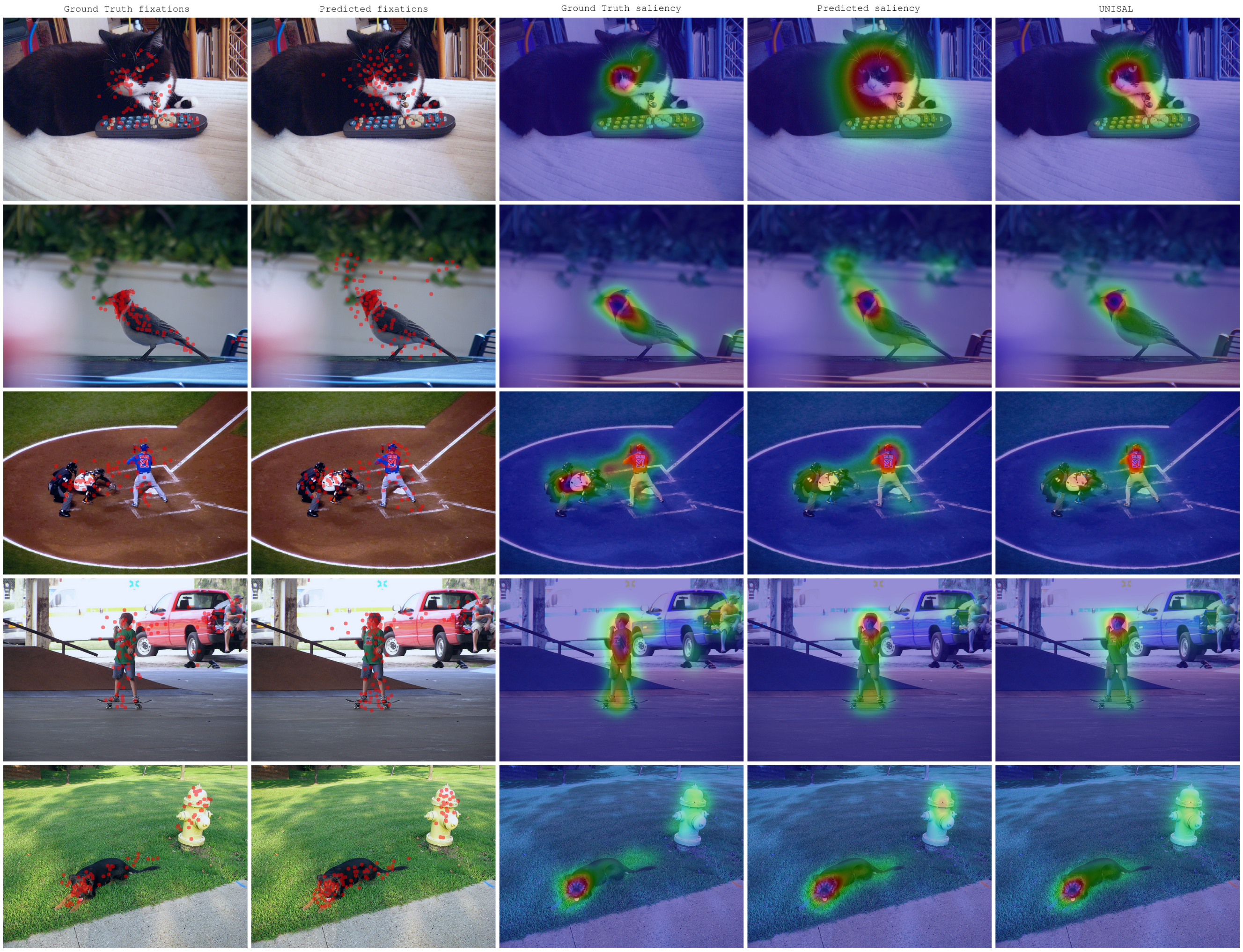

Figure A.1. Qualitative results on the Salicon validation images against Unisal.

References

[1] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020. 1

[2] Iuliia Kotseruba, Calden Wloka, Amir Rasouli, and John K Tsotsos. Do saliency models detect odd-one-out targets? new datasets and evaluations. *arXiv preprint arXiv:2005.06583*, 2020. 1

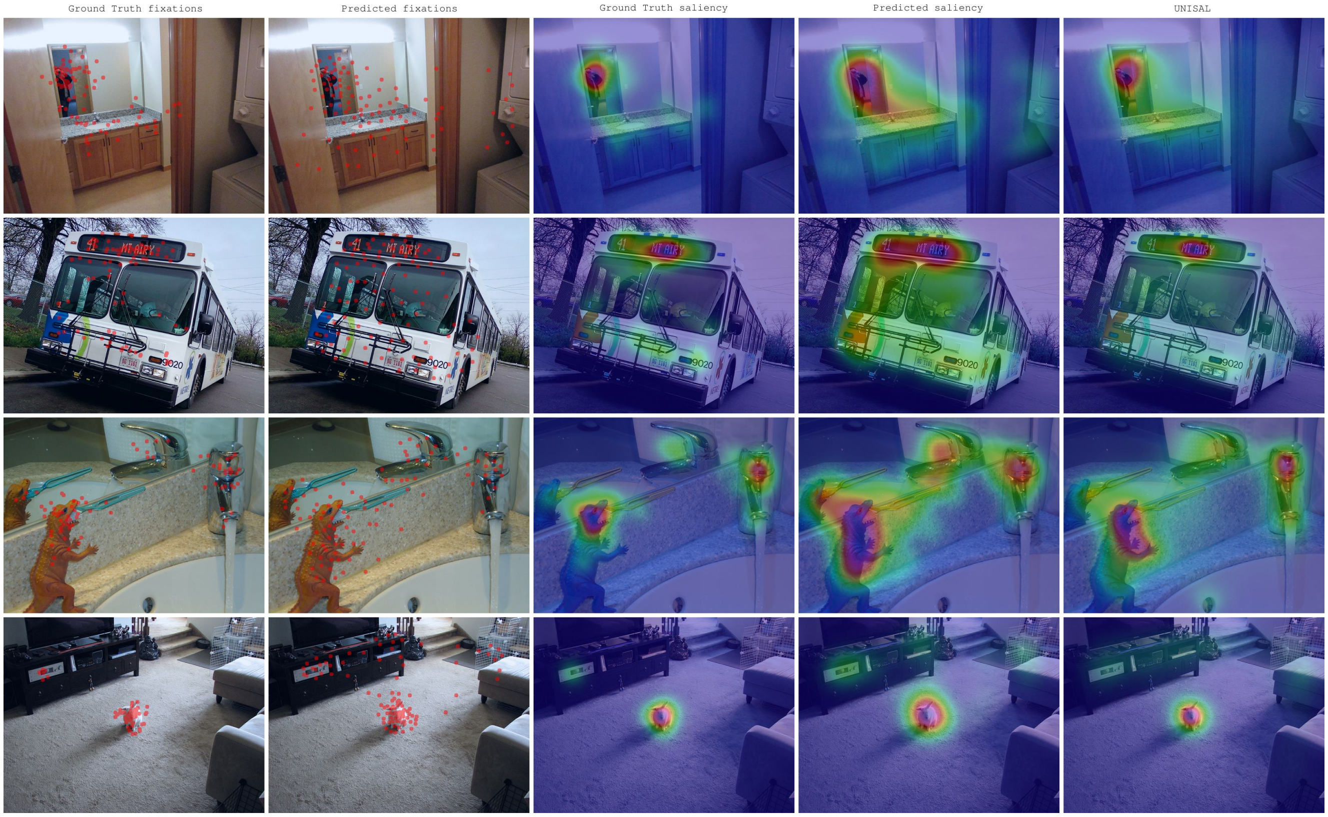

Figure A.2. Images from the Salicon validation set where SalTR obtains high KLD scores.

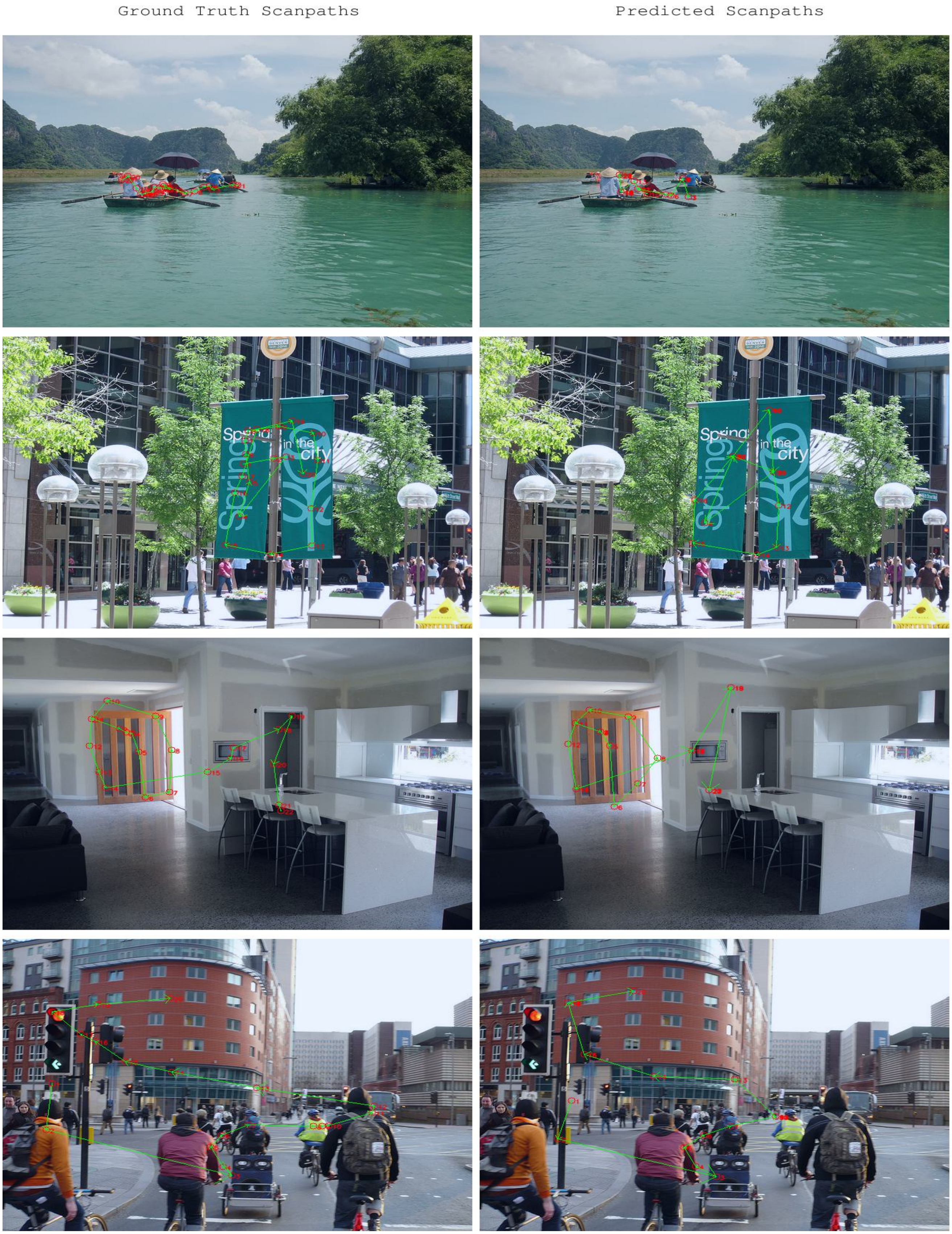

Figure A.3. Scanpaths for a set of images from the Salicon validation set.

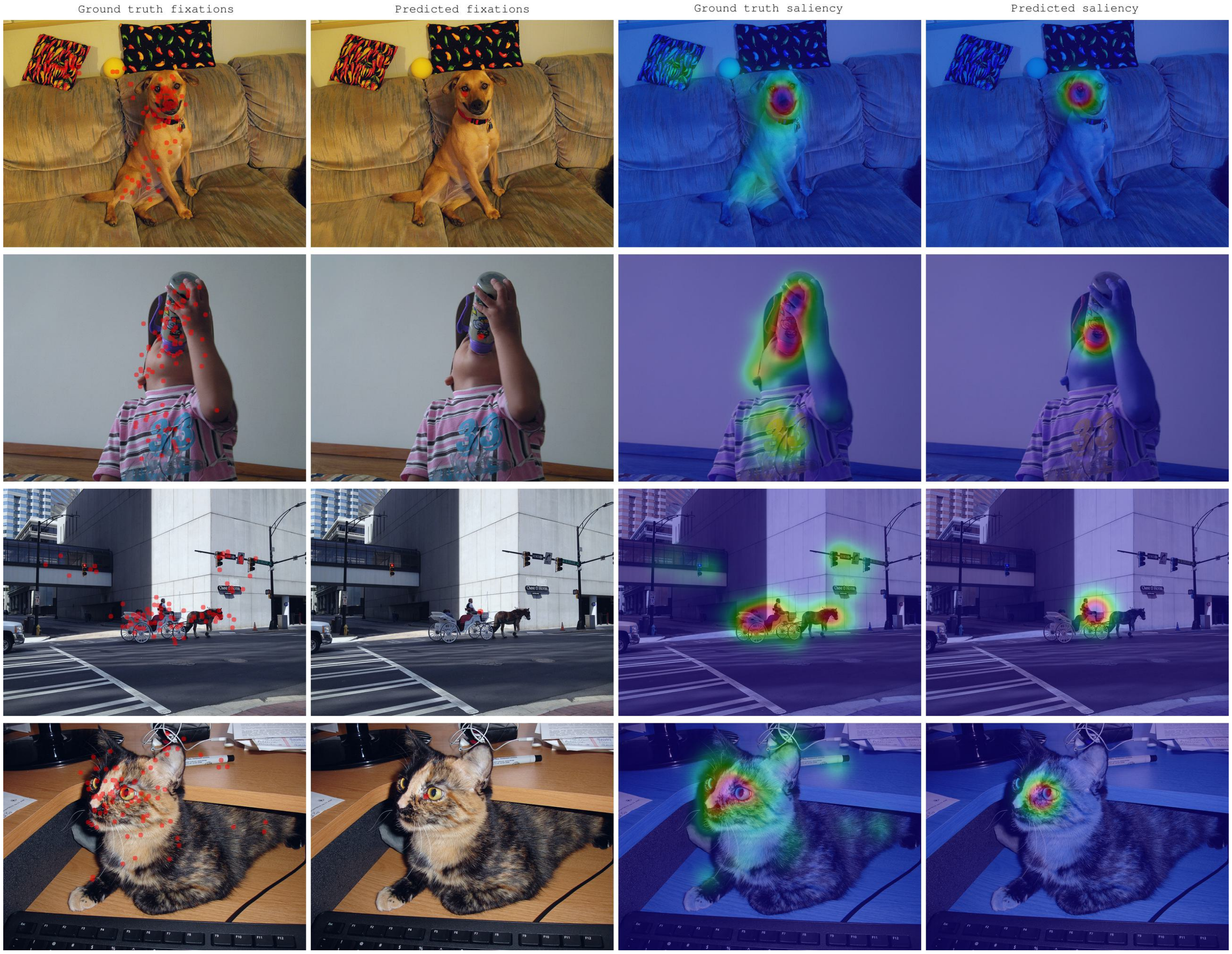

Figure A.4. Images from the Salicon validation set where the model was trained without any Hungarian matching loss.

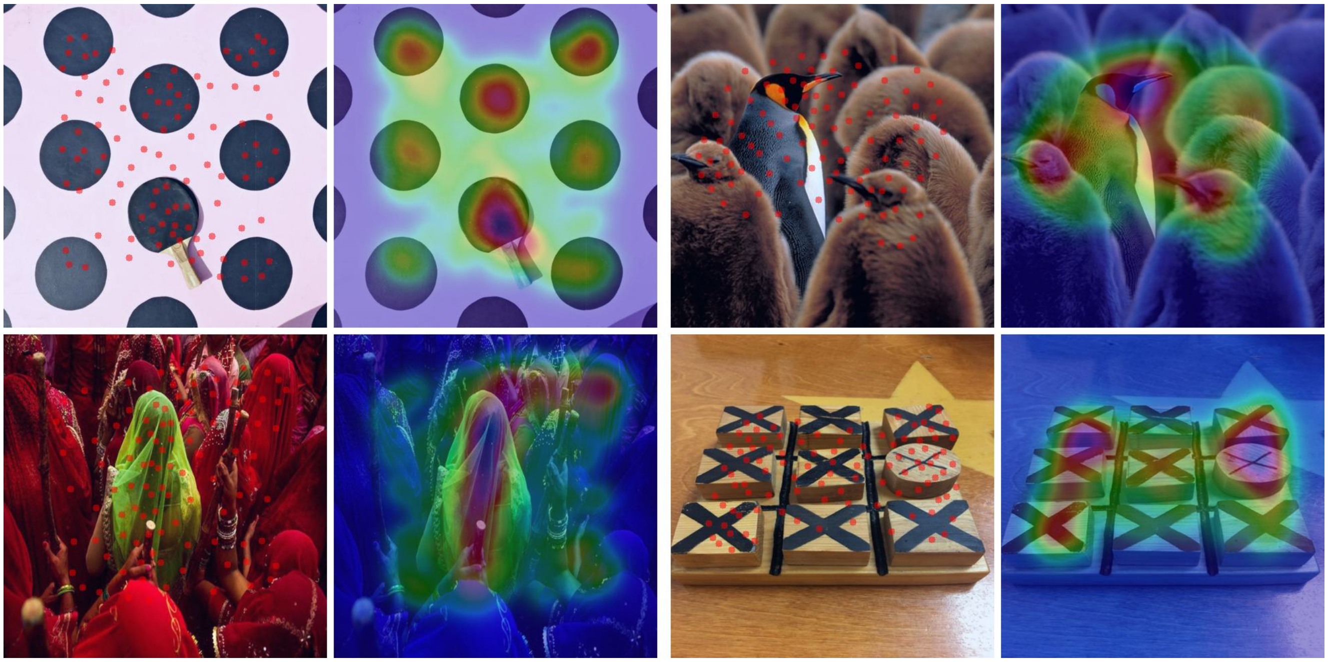

Figure A.5. Predictions on the low-level features P3 dataset.

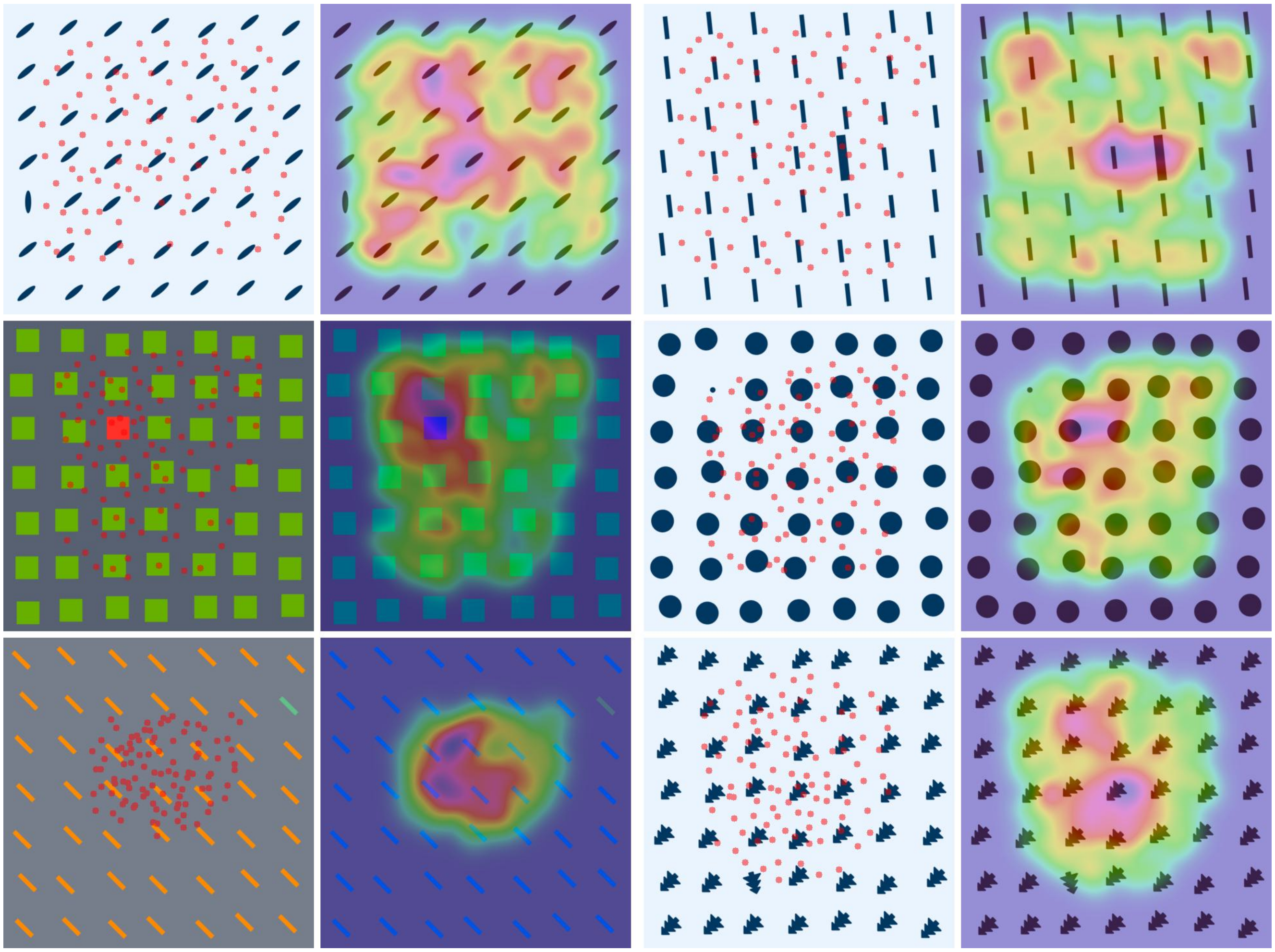

Figure A.6. Predictions on the low-level features O3 dataset.