UNETRSal: Saliency Prediction with Hybrid Transformer-Based Architecture

Azamat Kaibaldiyev, Jérémie Pantin, Alexis Lechervy, Fabrice Maurel, Youssef Chahir, and Gaël Dias

Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR6072, F-14000 Caen, France
azamat.kaibaldiyev@unicaen.fr, jeremie.pantin@unicaen.fr, alexis.lechervy@unicaen.fr, fabrice.maurel@unicaen.fr, youssef.chahir@unicaen.fr, gael.dias@unicaen.fr

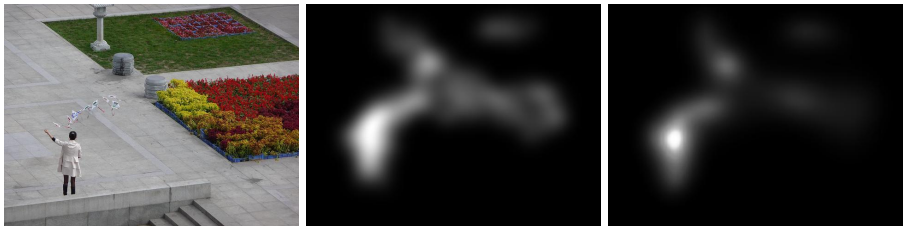
Abstract. Saliency prediction plays a critical role in understanding visual attention as it is a cornerstone for both natural scene understanding and automated document analysis. In this work, we propose the UNETRSal model for saliency prediction. Based on UNETR transformer-based model, we introduce a new decoder to increase efficiency on 2D images. Comprehensive evaluations on benchmark datasets, such as SALICON and CAT2000, demonstrate that UNETRSal achieves state-of-the-art performance across multiple saliency metrics, surpassing both conventional CNN-based and transformer-based methods. These results not only underscore the strengths of hybrid transformer architectures in modeling visual attention but also highlight the potential impact on advancing document representation modeling and layout analysis.

Keywords: Saliency Prediction · Visual Attention · Hybrid Architecture · 2D Image Processing

1 Introduction

Saliency prediction [3,43,49], or salient object detection, is the task of identifying the most visually important regions in an image. As it follows the human attention, it plays a crucial role in various computer vision applications, including object detection [47], image captioning [15], and visual scene understanding. Early works heavily relied on handcrafted features like color contrast, edge density or center bias [16,20]. Advanced deep learning models such as transformers-based architectures [10,30] have considerably improved saliency maps estimation, as opposed to previous approaches using Convolutional Neural Networks (CNN) with attention mechanism [33]. While recent models have improved performances on image and vision data, saliency detection still requires dedicated mechanisms.

Despite recent successes, saliency prediction faces several critical challenges, particularly CNNs that are limited by their inherent focus on local features [33]. Additionally, the scarcity of large scale, high quality annotated datasets for saliency prediction leads to overfitting and reduced generalization of trained



(a) Input Image (b) Ground Truth (c) TranSalNet [32]

Fig. 1: Example of saliency prediction limitations: (a) Original image, (b) Human ground truth saliency map (SALICON dataset), and (c) Prediction from a transformers-based approach, which fails to fully capture the salient regions.

models. Unlike CNNs that naturally incorporate strong local inductive biases, transformer-based approaches [44,46] often struggle to capture fine-grained spatial details [31]. The Figure 1 illustrates this problem. We note that such issues are even more critical on document image analysis [12].

In this work, we introduce UNETRSal, a transfomer-based approach for saliency prediction. Based on the UNETR [17] architecture, which is an extension of the UNET architecture [35], we design a novel decoder and remove batch normalization from specific convolutional blocks in order to take into account spatial information more efficiently. Conducted experiments demonstrate state-of-the-art performance on both SALICON [23] and CAT2000 [5] datasets.

2 Related Work

2.1 Salient Object Detection Models.

Saliency prediction started with handcrafted features approaches [20,24] and simple data-driven methods [4,16,48]. Early CNN-based models, for instance Ensemble of Deep Networks (eDN) [41] and SALICON [19], fused features from early layers and exploited two-stream architectures to capture multi-scale information. However, these methods were hampered by limited receptive fields and a tendency to overfit on small datasets. Subsequent advances focused on end-to-end architectures using deeper CNNs such as VGG [37] and ResNet [18]. SalGA [33] introduced a Generative Adversarial Network (GAN) framework to predict saliency maps. MLNet [7] merged pretrained multi-level features with learnable center bias. Attention mechanisms have been integrated into saliency models to better address spatial precision. For example, MSAGNet [39] uses HRNet [38] for improving localization using backbones with attention-gated multi-scale fusion. CASNet [13] uses channel attention for adaptive feature weighting. DeepGaze II [26] further demonstrated the benefits of transfer learning. Despite these advances, CNN-based methods often still emphasize local features, limiting their ability to model global scene layouts effectively [21].

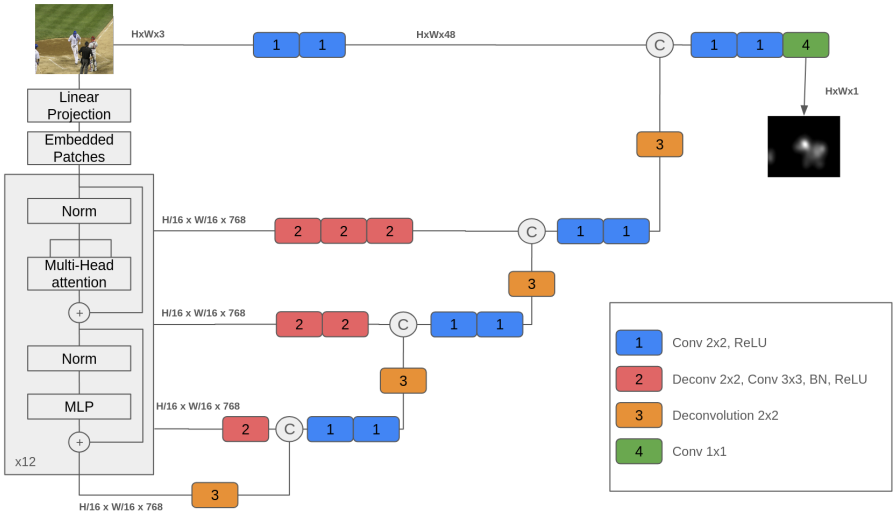


Fig. 2: Architecture of UNETRSal model.

2.2 Transformer-Based Approaches.

Transformer-based architectures have gained prominence in saliency prediction tasks due to their capability to capture global context effectively. Visual Saliency Transformer (VST) [29] uses a pure transformerbased framework leveraging multi-level token fusion and a novel token upsampling. TranSalNet [32] replaced CNN backbones with Visual Transformers (ViT) [10]. SATSal [40] introduced self-attention modules on skip connections to fuse multi-level features. In medical imaging, UNETR model [17] showed that transformers excel at modeling 3D spatial relationships in segmentation tasks. However, the existing transformer-based saliency models mostly focus on natural images and overlook contributions where global context and local precision is balanced through hybrid designs.

3 Methodology

3.1 3-D to 2-D UNETR

Backbone Model. Transformers model that handle spatial characteristics, like UNETR, processes 3D volumetric images with dimensions $H \times W \times D$, where H,W and D respectively represent the height, width and depth of the volume. Each voxel in the volume is typically represented by a single intensity value, leading input tensor of shape $(H, W, D, 1)$. In contrast, saliency prediction operates on 2D RGB images with dimensions $H \times W \times 3$ where the three channels correspond to the red, green, and blue color components. Therefore, the input tensor shape is $(H, W, 3)$. Generally, volumetric data dimensions are represented such as (H, W, D, C), with C the number of input channels (e.g. grayscale here).

Based on [17], volume is divided into non-overlapping patches of size (P, P, P), with a total of patches $N = \frac{H \times W \times D}{P^3}$. Each patch $x_v \in \mathbb{R}^{N \times (P^3 C)}$ is linearly projected into an embedding space of dimension K, and a learnable positional embedding $E_{\text{pos}} \in \mathbb{R}^{N \times K}$ is added:

$$z_0 = [x_v^1 E; x_v^2 E; \ldots; x_v^N E] + E_{\text{pos}} \tag{1}$$

This 1D sequence representation is then processed through transformers layers, following the standard multi-head self-attention (MSA) mechanism.

Architecture and Adaptation to 2D. The architecture is composed of 2 main components: a transformer-based encoder and hierarchical convolutional decoder, which can be seen in Figure 2. Saliency prediction operates on 2D RGB images, where the input dimension is $x \in \mathbb{R}^{H \times W \times 3}$. Instead of dividing the input into (P, P, P) volumetric patches, we extract 2D patches of size (P, P), leading to $N = \frac{H \times W}{P^2}$. Each 2D patch $x_v \in \mathbb{R}^{N \times (P^2 .3)}$ is linearly embedded in a feature space of size K. Before passing to the decoder, we reshape the latent representation using the following transformation:

$$z_i \longrightarrow \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times K} \tag{2}$$

allowing the decoder to reconstruct the 2D saliency map $\hat{y} \in \mathbb{R}^{H \times W \times 1}$.

Removing Batch Normalization. Batch normalization is commonly used to stabilize and accelerate training by normalizing layer inputs [2,36]. We observed that the convolutional blocks, the ones that perform 3x3 convolutions, batch normalization, and ReLU activation, work better when the intermediate batch normalization is removed. Based o [42], we remove batch normalization layers in these specific blocks helping to stabilize the gradient flow in regards to the saliency task and reducing overfitting.

3.2 Loss Functions

The model is optimized by using a combined loss function which is utilized to capture both the distributional similarity and correlation between predicted and ground truth saliency maps. It is comprised of three distinct terms, and each one serves to capture various aspects of the difference between the predicted saliency map P and the groundtruth saliency map G. In particular, we use the Kullback-Leibler Divergence (KLD) to measure distrubution differences, the Pearson's Correlation Coefficient (CC) to measure linear correlation, and the Similarity (SIM) to assess the overlap between the two maps.

Kullback-Leibler Divergence. The KLD measures the divergence between the predictied probability distribution and the ground truth distribution. It is

defined as:

$$L_{\text{KLD}} = \sum_i Gi \log\left(\frac{G(i)}{P(i)+\epsilon}\right) \tag{3}$$

where $G(i)$ is the ground truth saliency value at pixel i, $P(i)$ is the predicted saliency value, and ϵ is a small constant which is added for stability purposes. A lower value of L_{KLD} means that the predicted and the ground truth distributions are close to each other.

Correlation Coefficient. The Pearson's Correlation Coefficient is used to measure the linear correlation between predicted saliency map P and ground truth saliency map G. It is calculated as:

$$L_{\text{CC}} = -\frac{cov(G,P)}{\sigma_G \sigma_P} \tag{4}$$

where $cov(G,P)$ is the covariance between G and P, and σ_G and σ_P are the standard deviations of G and P, respectively. The negative sign in the equation is used so that an increase in correlation leads to a reduction in a loss.

Similarity. The Similarity metric is used to evaluate the similarity between the predicted saliency map P and ground truth saliency map G. Its definition is given by:

$$L_{\text{SIM}} = 1 - \sum_i \min(G(i), P(i)) \tag{5}$$

where a higher overlap leads to a lower L_{SIM}, which means that there is a closer match between the prediction and the ground truth.

Complete Loss Function. The final loss which is used in training the model is a weighted combination of three losses, and is defined as:

$$Loss = \lambda_1 L_{KLD} - \lambda_2 L_{CC} - \lambda_3 L_{SIM} \tag{6}$$

where λ_1, λ_2 and λ_1 are the hyperparameters. This combination was established for balancing different saliency map quality aspects.

It is crucial to note that the selection of loss functions for saliency prediction should be based on underlying model assumptions and its application. As presented in [32], the metric CC provides a fair comparison in terms of perceptual quality, while KLD is recommended for probabilistic models due to its ability to quantify the divergence between two distributions. Other studies [27,45] showed that CC and SIM metrics align most closely with human perception and are the most suitable saliency evaluation metrics for application involving image quality assessment.

4 Experiments

4.1 Experimental Setup

We benchmark our model on two widely used saliency benchmark datasets: SALICON [23] and CAT2000 [5]. Evaluation is performed on Area Under the ROC Curve (AUC), Shuffled AUC (sAUC), CC, SIM, KLD, Information Gain (IG) and Normalized Scanpath Saliency (NSS). These methods are used to comprehensively assess how accurately a model predicts human visual attention and saliency. Each metric highlights different aspects of the prediction, such as alignment with human behavior (AUC, NSS), similarity to ground truth (SIM), and the divergence of predicted and actual distributions (KLD, IG).

Training Parameters. We trained UNETRSal using the Adam optimizer with initial learning rate of 1×10^{-4}, which was decayed linearly during the training epochs. The batch size was set to 16 due to memory constraints of training transformer-based architectures. The training ran for 10 epochs with early stopping based on validation loss. The model is trained end-to-end using a combination of loss functions, including KLD, CC, and SIM. To balance their contributions, we introduce loss weights λ_1, λ_2, and λ_3, which are respectively set to 10.0, -1.0 and -1.0 for SALICON and 2.0, -1.0 and -1.0 for CAT2000 dataset based on preliminary experiments. These values provided stable gradients and led to improved convergence.

Datasets and Evaluation Process. The benchmark datasets used for saliency prediction are SALICON and CAT2000. SALICON [23] is produced from the Microsoft COCO dataset and provides bigger saliency annotations in comparison to other datasets. The annotations are collected via mouse-tracking, thus simulating human attention patterns. It includes 10,000 training, 5,000 validation and 5,000 test images. CAT2000 [5] is comprised of 2,000 training and 2,000 test images across 20 different categories. Each category (Art, Cartoon, Pattern, etc.) is designed to challenge different aspects of saliency models. Evaluation protocols for both datasets involve external benchmarking. For SALICON, the predictions on the test set should be submitted to the official challenge website to obtain performance results. For CAT2000, the test set prediciton should be emailed to the dataset maintainers for evaluation. To ensure fair and transparent comparison, we obtained performance results of the state-of-the-art models through official challenge leaderboards and originally published papers.

Transfering Knowledge on Benchmark Datasets. We first trained our model on the SALICON dataset, which provides highest high-quality human labeled saliency maps. The model was optimized using the combined loss function. For the CAT2000, we trained UNETRSal using only human-annotated data, as synthetic data augmentation did not show significant improvements. Moreover, since the CAT2000 dataset contains significantly fewer samples in comparison

Method	AUC ↑	CC ↑	KLDiv ↓	sAUC ↑	NSS ↑	SIM ↑	IG ↑
SAM-Resnet [8]	0.865	0.899	0.610	0.741	1.990	0.793	0.538
MSI-Net [25]	0.865	0.899	0.307	0.736	1.931	0.784	0.793
GazeGAN [6]	0.864	0.879	0.376	0.736	1.899	0.773	0.720
MDNSal [34]	0.865	0.899	0.221	0.736	1.935	0.790	0.863
UNISAL [11]	0.864	0.879	0.354	0.739	1.952	0.775	0.780
MD-SEM [14]	0.864	0.868	0.568	0.746	**2.058**	0.774	0.660
TranSalNet [32]	0.868	0.907	0.373	0.747	2.014	0.803	0.788
SimpleNet [34]	0.869	0.907	0.201	0.743	1.960	0.793	0.880
DeepGaze IIE [28]	0.869	0.872	0.285	**0.767**	1.996	0.733	0.766
TempSAL [1]	0.869	0.911	**0.195**	0.745	1.967	0.800	**0.896**
UNETRSal (ours)	**0.870**	**0.914**	0.316	0.745	1.985	**0.808**	0.821

Table 1: Comparison of saliency prediction performance on SALICON test set.

to the SALICON dataset, we first pretrained the model on SALICON and then finetuned it separately on CAT2000. Fine-tuning allowed the model to adapt to different distributions of human gaze data, while also using the pretraining benefits from the larger dataset. We used the same loss function and optimizer settings, but with a reduced learning rate of 1×10^{-5} during finetuning process.

4.2 Quantitative Results

SALICON. Table 1 compares the performance of our modified UNETRSal model with SOTA methods on the SALICON dataset. Our model achieves superior performance across most metrics, specially AUC, CC and SIM, meaning that it succeeds to find more salient areas than other models. In particular, it outperforms in almost all metrics the TranSalNet architecture that it is also based on transformer architectures.

CAT2000. Our model also achieves improved results on the CAT2000 dataset by beating most SOTA models in all metrics. In particular, our model outperforms all compared methods in four out of six metrics, which are AUC, NSS, CC and SIM (see Table 2). These results underscore the robustness of our approach in capturing global and local saliency features across different image categorie. Indeed, CAT2000 includes images from 20 diverse categories. The improved performance in these key metrics demonstrates that our architecture uses transformer-based global context modeling to output accurate saliency maps.

4.3 Qualitatitive Results

In addition to increased quantitative performance, UNETRSal shows noticeable improvements in spatial alignment in qualitative evaluation, which is illustrated in Figure 3. The existing transformer-based models, such as TranSalNet [32], often focus on a small amount of pixels, the central and most prominent object

Method	AUC ↑	CC ↑	KLDiv ↓	sAUC ↑	NSS ↑	SIM ↑
SalGAN [33]	0.8085	0.5668	0.9392	0.6354	1.4624	0.5441
EML-NET [22]	0.8310	0.6209	1.6914	0.5853	1.5649	0.5840
SalFBNet [9]	0.8549	0.7027	1.1983	0.6330	1.8789	0.6425
UNISAL [11]	0.8604	0.7399	0.4703	0.6684	1.9359	0.6633
DeepGaze II [26]	0.8640	0.7950	0.3815	0.6498	1.9619	0.6865
DeepGaze IIE [28]	0.8692	0.8189	**0.3448**	**0.6677**	2.1122	0.7060
UNETRSal (ours)	**0.8801**	**0.9012**	0.6135	0.6040	**2.4071**	**0.7750**

Table 2: Comparison of saliency prediction performance on CAT2000 test set.

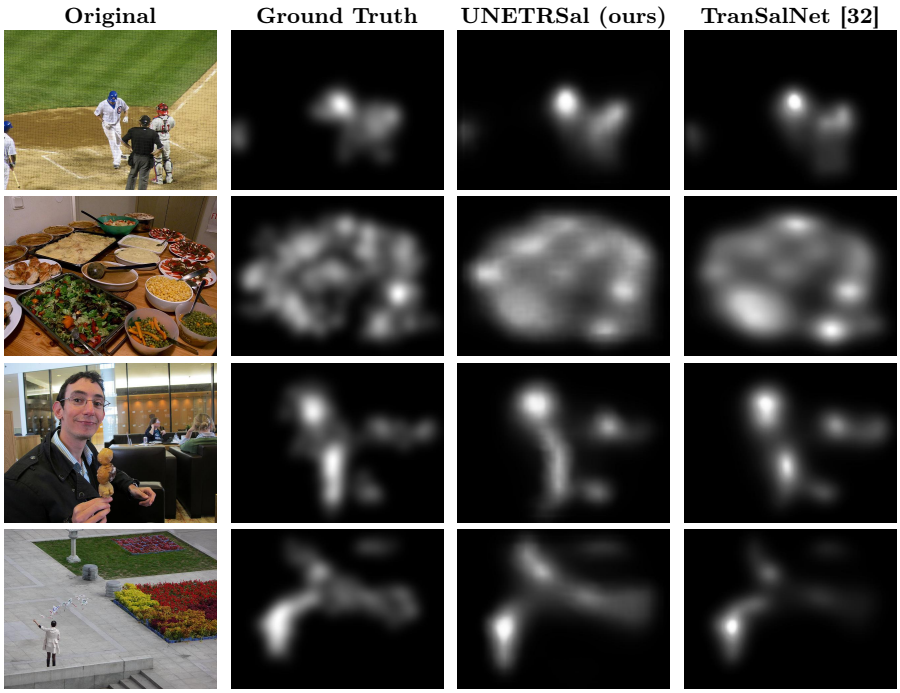


Fig. 3: Qualitative comparison: (1) First row: Outdoor scene, (2) Second row: Complex indoor scene, (3) Third row: Human-centric image, (4) Fourth row: Outdoor scene. All saliency maps are compared against SALICON ground truth.

regions, while UNETRSal distributes attention more smoothly across the scene. For example, in the last row example, UNETRSal successfully captures the entire body of a person, while TranSalNet mostly focuses on the head region. Moreover, UNETRSal attends more to contextual cues such as background objects, while TranSalNet typically neglects it. These results suggest that UNETRSal better mimics human visual attention with more aligned saliency maps.

5 Conclusion

In this paper, we present UNETRSal an hybrid transformer-based model for saliency prediction. By adapting a 3D medical image segmentation model for 2D saliency tasks, we show that our hybrid transformer-based approach effectively captures local and global dependencies, which gives accurate saliency predictions. UNETRSal's modifications, such as decoder adjustment and batch normalization removal, greatly improve the performance in saliency prediction, by helping to achieve more stable training and improved accuracy. Experimental results across the CAT2000 and SALICON gold-standard datasets demonstrate that the UNETRSal achieves state-of-the-art results across most metrics such as AUC, NSS, CC and SIM, and produces qualitatively better saliency maps.

References

1. Aydemir, B., Hoffstetter, L., Zhang, T., Salzmann, M., Süsstrunk, S.: Tempsal-uncovering temporal information for deep saliency prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6461–6470 (2023)
2. Bjorck, N., Gomes, C.P., Selman, B., Weinberger, K.Q.: Understanding batch normalization. Advances in neural information processing systems (NeurIPS) **31** (2018)
3. Borji, A.: Saliency prediction in the deep learning era: Successes and limitations. IEEE transactions on pattern analysis and machine intelligence **43**(2), 679–700 (2019)
4. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Transactions on Image Processing **24**(12), 5706–5722 (2015). `https://doi.org/10.1109/TIP.2015.2487833`
5. Borji, A., Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581 (2015)
6. Che, Z., Borji, A., Zhai, G., Min, X., Guo, G., Le Callet, P.: How is gaze influenced by image transformations? dataset and model. IEEE Transactions on Image Processing **29**, 2287–2300 (2019)
7. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: 23rd International Conference on Pattern Recognition (ICPR). pp. 3488–3493. IEEE (2016)
8. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing **27**(10), 5142–5154 (2018)
9. Ding, G., İmamoğlu, N., Caglayan, A., Murakawa, M., Nakamura, R.: Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks. Image and Vision Computing **120**, 104395 (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Droste, R., Jiao, J., Noble, J.A.: Unified image and video saliency modeling. In: 16th European Conference on Computer Vision (ECCV). pp. 419–435. Springer (2020)

12. Eglin, V., Bres, S.: Document page similarity based on layout visual saliency: application to query by example and document classification. In: 7th International Conference on Document Analysis and Recognition (ICDAR). pp. 1208–1212. Citeseer (2003)
13. Fan, S., Shen, Z., Jiang, M., Koenig, B.L., Xu, J., Kankanhalli, M.S., Zhao, Q.: Emotional attention: A study of image sentiment and visual attention. In: IEEE Conference on computer vision and pattern recognition (CVPR). pp. 7521–7531 (2018)
14. Fosco, C., Newman, A., Sukhum, P., Zhang, Y.B., Zhao, N., Oliva, A., Bylinskii, Z.: How much time do you have? modeling multi-duration saliency. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 4473–4482 (2020)
15. Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. ACM Computing Surveys **56**(3), 1–39 (2023)
16. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. Advances in neural information processing systems **19** (2006)
17. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: IEEE/CVF winter conference on applications of computer vision (ICCV). pp. 574–584 (2022)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: IEEE international conference on computer vision (ICCV). pp. 262–270 (2015)
20. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence **20**(11), 1254–1259 (2002)
21. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.: Learn to pay attention. arXiv preprint arXiv:1804.02391 (2018)
22. Jia, S., Bruce, N.D.: Eml-net: An expandable multi-layer network for saliency prediction. Image and vision computing **95**, 103887 (2020)
23. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: Saliency in context. In: IEEE conference on computer vision and pattern recognition (CVPR). pp. 1072–1080 (2015)
24. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence: Conceptual structures in cognitive neuroscience, pp. 115–141. Springer (1987)
25. Kroner, A., Senden, M., Driessens, K., Goebel, R.: Contextual encoder–decoder network for visual saliency prediction. Neural Networks **129**, 261–270 (2020)
26. Kümmerer, M., Wallis, T.S., Bethge, M.: Deepgaze ii: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563 (2016)
27. Li, J., Xia, C., Song, Y., Fang, S., Chen, X.: A data-driven metric for comprehensive evaluation of saliency models. In: IEEE international conference on computer vision (ICCV). pp. 190–198 (2015)
28. Linardos, A., Kümmerer, M., Press, O., Bethge, M.: Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12919–12928 (2021)

29. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: IEEE/CVF international conference on computer vision (ICCV). pp. 4722–4732 (2021)
30. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z.: A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems (2023)
31. Liu, Z., Tan, Y., He, Q., Xiao, Y.: Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4486–4497 (2021)
32. Lou, J., Lin, H., Marshall, D., Saupe, D., Liu, H.: Transalnet: Towards perceptually relevant visual saliency prediction. Neurocomputing **494**, 455–467 (2022)
33. Pan, J., Sayrol, E., Nieto, X.G.i., Ferrer, C.C., Torres, J., McGuinness, K., OConnor, N.E.: Salgan: Visual saliency prediction with adversarial networks. In: CVPR scene understanding workshop (SUNw) (2017)
34. Reddy, N., Jain, S., Yarlagadda, P., Gandhi, V.: Tidying deep saliency prediction architectures. In: IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 10241–10247. IEEE (2020)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: 18th international conference on Medical image computing and computer-assisted intervention (MICCAI). pp. 234–241. Springer (2015)
36. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? Advances in neural information processing systems (NeurIPS) **31** (2018)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
38. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 5693–5703 (2019)
39. Sun, Y., Zhao, M., Hu, K., Fan, S.: Visual saliency prediction using multi-scale attention gated network. Multimedia Systems **28**(1), 131–139 (2022)
40. Tliba, M., Kerkouri, M.A., Ghariba, B., Chetouani, A., Çöltekin, A., Shehata, M.S., Bruno, A.: Satsal: A multi-level self-attention based architecture for visual saliency prediction. IEEE Access **10**, 20701–20713 (2022)
41. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: IEEE conference on computer vision and pattern recognition (CVPR). pp. 2798–2805 (2014)
42. Wang, H., Zhang, A., Zheng, S., Shi, X., Li, M., Wang, Z.: Removing batch normalization boosts adversarial training. In: International Conference on Machine Learning (ICML). pp. 23433–23445. PMLR (2022)
43. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3239–3259 (2021)
44. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: IEEE/CVF international conference on computer vision (ICCV). pp. 22–31 (2021)
45. Yang, X., Li, F., Liu, H.: A measurement for distortion induced saliency variation in natural images. IEEE Transactions on Instrumentation and Measurement **70**, 1–14 (2021)

46. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: IEEE/CVF international conference on computer vision (ICCV). pp. 558–567 (2021)
47. Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B.: A survey of modern deep learning based object detection models. Digital Signal Processing **126**, 103514 (2022)
48. Zhang, J., Sclaroff, S.: Saliency detection: A boolean map approach. In: IEEE international conference on computer vision (ICCV). pp. 153–160 (2013)
49. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 12546–12555 (2020)