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

Abstract

We present a novel approach for saliency prediction in images, leveraging parallel decoding in transformers to learn saliency solely from fixation maps. Models typically rely on continuous saliency maps, to overcome the difficulty of optimizing for the discrete fixation map. We attempt to replicate the experimental setup that generates saliency datasets. Our approach treats saliency prediction as a direct set prediction problem, via a global loss that enforces unique fixations prediction through bipartite matching and a transformer encoder-decoder architecture. By utilizing a fixed set of learned fixation queries, the cross-attention reasons over the image features to directly output the fixation points, distinguishing it from other modern saliency predictors. Our approach, named Saliency TRansformer (SalTR), achieves metric scores on par with state-of-the-art approaches on the Salicon and MIT300 benchmarks.

1. Introduction

In recent years, deep learning models have achieved significant progress in saliency prediction [3, 52], leveraging large-scale annotated datasets [5, 7, 29]. A saliency dataset collection requires a set of images serving as visual stimuli, along with recorded eye movements data captured through eye-tracking devices. Generally, each stimuli is observed by several human subjects. During the observation, the eye positions of the subjects are continuously tracked in relation to the coordinates of the images to obtain the fixation maps. Then, the individual fixations are aggregated and blurred with a Gaussian filter to generate a continuous saliency map [65].

Most existing saliency prediction models rely on continuous saliency maps, where each pixel represents the probability of attending at that location in the image [38, 52, 53]. However, these models face challenges in optimizing for the discrete fixation maps, which indicate the specific locations of saccades by human observers. While successful in producing high quality saliency maps, the common architecture does not replicate the data collection pipeline, but rather tries to implicitly learn through the post-processed saliency map. To overcome this limitation, and inspired from [10], we propose a novel approach that learns saliency prediction solely from fixation maps, without relying on continuous saliency annotations. Our method, named Saliency TRansformer (SalTR), leverages parallel decoding in transformers to directly predict the fixation points directly.

In our approach, we treat saliency prediction as a direct set prediction problem, where the goal is to predict a set of spatial fixation points. To achieve this, we employ a transformer encoder-decoder architecture [62], with a fixed set of learned fixation queries. The cross-attention mechanism in the transformer decoder reasons over the image features using these fixation queries to directly output the fixation points. This distinguishes our approach from other modern saliency predictors, which typically rely on continuous saliency maps. In summary, our contributions are:

- We propose a novel approach for saliency prediction, leveraging parallel decoding in transformers to learn saliency solely from fixation maps.
- We demonstrate the effectiveness of our approach on the Salicon benchmark, achieving remarkable performance compared to state-of-the-art methods.
- Furthermore, we extend the approach to the scanpaths prediction problem, and demonstrates its effectiveness.

2. Related works

The seminal work of the feature integration theory [61], is a cornerstone in identifying the visual features that guide human attention. This foundational theory has served as a launchpad for the development of various computational models. In the realm of computer vision, the emphasis is primarily placed on the selective mechanism when modeling attention. Saliency, in this context, is subtly defined in relation to the gaze policy on a scene – characterizing the particular subsets of space where a human observer would likely concentrate their focus. The term "salient" surfaced in the sphere of bottom-up computations [27, 34], while the

Code: https://github.com/YasserdahouML/SalTR

concept of attention spans a broader spectrum. Furthermore, the last decade has witnessed the remarkable progress of saliency prediction, and many methods have been presented and achieved remarkable performances on the recently introduced benchmarks, especially the deep learning based methods have yielded a boost in performance. Researchers tend to typically repurpose existing Convolutional Neural Networks (CNN) architectures [25, 51, 58] to make predictions about saliency. These models involve architectural enhancements tailored to the specific demands of the saliency downstream task. These models are trained end-to-end on saliency datasets, framing saliency as a regression problem. A common challenge faced in this area is the scarcity of annotated fixation data. To mitigate this issue, the model's encoder is usually pretrained on extensive image recognition datasets, such as ImageNet [56]. This pretraining step allows for the acquisition of valuable representations at the level of latent space, which are then fine-tuned on saliency datasets. Prior to that, handcrafted approaches attempted in modelling the human visual attention.

Heuristic approaches. The prediction of saliency for images has been a major focus of academic research over the past few decades. A seminal study by [27] introduced a bottom-up approach to visual saliency, utilizing center-surround differences across multiple scales of image features. This method generates conspicuity maps by linearly combining and normalizing feature maps, with color (C), orientation (O), and intensity (I) serving as the three primary features:

$$C_I = f_I, \quad C_C = \mathcal{N}(\sum_{l \in L_C} f_l), \quad C_O = \mathcal{N}(\sum_{l \in L_O} f_l). \quad (1)$$

In this equation, $\mathcal{N}(.)$ represents the map normalization operator. The ultimate saliency map is an average of the three conspicuity maps: $S = \frac{1}{3}\sum_{k \in I,C,O} C_k$. A more complex bottom-up saliency model, taking into account additional Human Visual System (HVS) features such as contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions, was later proposed in [43]. Additional static saliency models, such as those developed by [6, 20–22, 35, 48, 49, 57], are predominantly cognitive-based models. These models utilize various visual features, including color, edge, and orientation, at numerous spatial scales to construct a saliency map.

In addition, Bayesian models have been employed to supplement these cognitive models, introducing a layer of prior knowledge (e.g., scene context or gist) through a probabilistic approach such as Bayes' rule for combination [18, 50, 60, 66]. These models demonstrate the capacity to integrate various factors in a principled manner.

Deep learning approaches. are rooted in data-driven approached from recorded eye-fixations or labeled salient maps. Authors from [33] pioneered a non-parametric bottom-up method to learn saliency from human eye fixation data. Their model utilizes a support vector machine (SVM) [26] to determine saliency based on local intensities, marking the first method that did not rely on any assumptions about Human Visual System (HVS) features to encode saliency. Similarly, Judd et al. [32] used a linear SVM to train on 1003 labeled images, leveraging a range of low, mid, and high-level image features.

Recent deep learning-based static saliency models, such as those proposed by [11, 12, 38, 39, 52, 53], have made notable advancements, leveraging the success of deep neural networks and the availability of large-scale saliency datasets for static scenes, such as those described in [5, 7].

The works of [38, 53] were pioneers in the application of Convolutional Neural Networks (CNNs) for saliency prediction, culminating in the creation of the eDN and DeepFix models respectively. Specifically, DeepFix employs a unique approach in its initial phase, utilizing the weights from the first five convolution blocks of the VGG-16 model [58]. Furthermore, it introduces two Location Based Convolutional (LBC) layers, adept at capturing semantics at various scales. Subsequently, Pan et al. [52] leveraged Generative Adversarial Networks (GANs) [23] to devise the SalGAN model. The SalGAN architecture comprises a generator model, the weights of which are learned through back-propagation. This learning is driven by a binary cross-entropy (BCE) loss computed over pre-existing saliency maps. The ensuing prediction generated by the model is further processed by a discriminator network.

This discriminator network is trained to perform a binary classification task, tasked with distinguishing between the saliency maps generated by the generator and the ground truth maps. The process follows a min-max game format, utilizing the ensuing adversarial loss:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{BCE}} + \mathcal{L}(D(q, p_{model}), 1). \quad (2)$$

In this equation, the aim is to optimize the value of $\mathcal{L}(D(I,S),1)$. Here, $D(I,S)$ represents the probability that the discriminator network is deceived, meaning that the generated saliency maps closely mimic those from the data distribution ground truth. Authors from [30] proposed SALICON, the model optimizes an objective function based on the saliency evaluation metrics, from two parallel streams at different image scales. [45] trained the deep spatial contextual long-term recurrent convolutional network (DSCLRCN), incorporating both global spatial interconnections and scene context modulation. EML-NET proposed by [28] consists of a disjoint encoder and decoder trained separately. Furthermore, the encoder can contain many networks extracting features, while the decoder learns to combine many latent variables generated by the encoder networks. Unisal [17] introduced four domain

adaptation techniques aimed at addressing the significant challenge of disparity in datasets for effective joint modeling. These strategies encompass Domain-Adaptive Priors, Domain-Adaptive Fusion, Domain-Adaptive Smoothing, and Bypass-RNN. The model also proposes a refined formulation of learned Gaussian priors. These techniques were incorporated into a streamlined, lightweight network constructed in the encoder-RNN-decoder style. SalFBNet [16], builds on the benefits of feedback connections by linking higher-level feature blocks to low-level layers. The study also proposes a novel Selective Fixation and Non-Fixation Error loss, a large-scale Pseudo-Saliency dataset, and shows that SalFBNet performs competitively on public benchmarks with fewer parameters, indicating the effectiveness of their approach. Through principled combination of multiple backbones, they developed a the current SoTA model, "DeepGaze IIE [44]", that achieves the best performance on the MIT/Tuebingen Saliency [40] Benchmark across all metrics, highlighting the model's ability to maintain good confidence calibration on unseen datasets. Overall, these deep models achieve results closer to the human baseline results on the SALICON [30], MIT300 [8], and CAT2000 [5] datasets.

DEtection TRansformers. DETR [10] shifted the object detection paradigm, casting the problem as a set-based prediction one, eliminating the hand-designed components, and maintaining comparable performance against the well-established FasterRCNN [55]. DETR [10] combines several techniques such as bipartite matching loss, transformer encoder-decoder with parallel decoding to design DEtection Transformer (DETR). The approach formulates the object detection task as an image-to-set problem. The model outputs a fixed-length unordered set of classes and bounding boxes of all possible objects present in the image. The bipartite matching forces unique one-to-one predictions. Intuitively, the decoder queries can be interpreted as humans saccading at various spatial locations of an image; each human hence observes others before making its prediction.

It can be seen from the above review that saliency prediction models mostly use the continious saliency maps for learning. Inspired by DETR, we attempt to learn saliency from fixations only.

3. Method

Following DETR [10], our model is an end-to-end saliency predictor which includes a ResNet backbone [25], Transformer encoder and decoder [62], and a fixations prediction head. We adapt the decoder part to solve the saliency task, as shown in Figure 1. Given an image, we extract the latent representations using a CNN backbone followed with a Transformer encoder to enrich the CNN features. Then, the fixation queries are fed to the Transformer decoder to search for fixation locations given the image information through cross attention.

3.1. SalTR components

Given an image dataset, $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_{|\mathcal{D}|}\}$ where $\mathbf{x}_i \in \mathcal{R}^{C\times H\times W}$, the goal is to predict a set of spatial fixation points $\bar{f}_i = (\bar{x}_i, \bar{y}_i)$, that will serve as the basis to build the fixation map $\mathcal{M}_p$. We then smooth this map with a Gaussian filter with a standard deviation σ to obtain the final continious saliency map. However, there are some significant differences that distinguish our approach from previous works. As the model predicts the fixation points only, we do not leverage the continious ground truth saliency map at training time. Thus, we aim at mimicking the way visual attention datasets are created, with the use of both the fixation queries and parallel decoding. The key modules are outlined in the following.

CNN Encoder ($\mathbf{f}_\theta$). The encoder is a network $\mathbf{f}_\theta : \mathbf{x} \mapsto \mathbf{\Gamma}$ parameterised by $\boldsymbol{\theta_e}$. $\mathbf{f}_\theta$ is implemented as a backbone ResNet50 [25], followed by a 2-layer 1×1 convolutional projection head with batch normalization and ReLU activation, that reduces the channel dimension from 2048 to 256.

Transformer encoder. $\boldsymbol{\Omega_\omega}$ maps $\mathbf{\Gamma}^{c\times h\times w}$ to $\boldsymbol{\gamma}^{c\times h\times w}$. $\mathbf{\Gamma}$ is first wrapped to a sequence of size $c \times hw$, then augmented with 2D positional encodings [2]. The multi-head self-attention layers perform message parsing across $\mathbf{\Gamma}$ channels, in order to capture the contextual information. This acts as a smoothing prior, hence, pixels sharing the same semantic class repulsively attend irrespective of their position in the image, ignoring all structural information. Furthermore, $\boldsymbol{\gamma}$ is the smoothed transform of $\mathbf{\Gamma}$, ensuring coherence both spatially in the neighborhood of a given pixel i and semantically for pixels j further away but sharing the same class.

Transformer decoder. The decoder transforms a fixed number of embeddings (i.e. fixation queries) of size 256 using multi-headed cross attention mechanisms, where the keys and values are sourced from the image features. The output embeddings are then mapped to a fixation point $\bar{f}_i = (\bar{x}_i, \bar{y}_i)$ using a 3-layer MLP.

3.2. Learning from fixations

Following the experimental setup used in creating saliency datasets, we denote the decoder fixation queries as $\mathcal{F}_q = F_{q0}, ..., F_{qN}$. These queries simulate the viewer's attention, i.e. where they attend to the image features, and saccade to a spatial position that maximizes the attention task. Furthermore, we do not want the queries to predict the same ground truth fixations, thus, the loss assigns a unique matching using the Hungarian algorithm between the prediction and ground truth fixations, and then minimizes the l_1 distance for the respective matched positions.

Denote $f_i = (x_i, y_i)$ as a sample from the ground truth set, and the corresponding prediction as $\bar{f}_i = (\bar{x}_i, \bar{y}_i)$. To

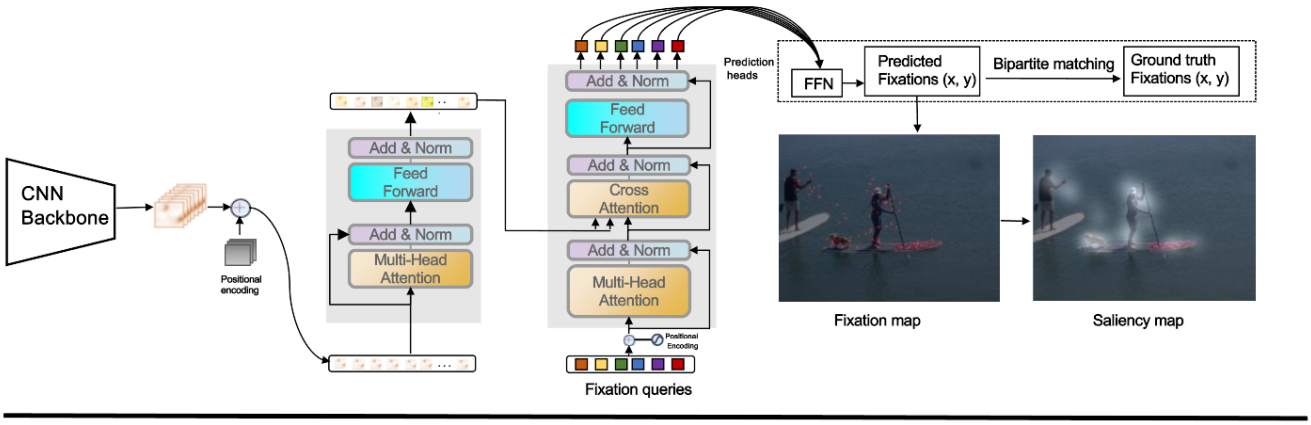

Figure 1. Complete pipeline for training. The CNN backbone produces the latent representation given an input image. The transformer encoder enhances this representation for a suitable decoding. The queries in the transformer decoder cross-attention are a fixed number of fixation queries, that attend to the image features. The prediction head maps the output embedding to a spatial fixation location.

find a bipartite matching between these two sets, as in [10], we search for a permutation of N elements $\phi \in S_N$ with the lowest cost:

$$\hat{\phi} = \arg\min_{\phi \in S_N} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(f_i, \hat{f}_{\phi(i)}), \tag{3}$$

where $\mathcal{L}_{\text{Match}}(f_i, \hat{f}_{\phi(i)})$ is a pairwise matching cost between the ground truth f_i and a prediction with index $\phi(i)$. The final loss is:

$$\mathcal{L} = \sum_{i=1}^{N} \left|\left| f_i - \hat{f}_{\hat{\sigma}(i)} \right|\right|_1 + \alpha \mathcal{L}_{\text{NSS}}(\mathcal{M}_p, \mathcal{M}_{gt}), \tag{4}$$

where $\mathcal{M}_{gt} \in \{0,1\}^{H \times W}$ is the ground truth fixation, and the predicted fixation map $\mathcal{M}_p$ is obtained using:

$$\mathcal{M}_{p_{ij}} = \begin{cases} 1 & \text{if location } (i,j) \text{ is a fixation} \\ 0 & \text{otherwise,} \end{cases}$$

The normalized scanpath saliency loss (NSS) is defined as follows:

$$\mathcal{L}_{\text{NSS}}(\mathcal{M}_p, \mathcal{M}_{gt}) = \frac{1}{S} \sum_{i} \hat{\mathcal{M}}_{p_i} \times \mathcal{M}_{gt_i}, \tag{5}$$

where $S = \sum_i \mathcal{M}_{gt_i}$ and $\hat{\mathcal{M}}_p = \frac{\mathcal{M}_p - \mu(\mathcal{M}_p)}{\sigma(\mathcal{M}_p)}$, and i refers to the i-th pixel, and S represents the total count of fixated pixels. A score of 0 indicates chance, while a positive NSS value indicates agreement between the maps beyond chance. Thus, we aim at minimizing the negative value of $\mathcal{L}_{\text{NSS}}$, by setting α to a negative value (-0.2).

4. Experimental setting

Training. To evaluate the proposed framework, we train SalTR on the 10k/5k train/validation splits of the image saliency dataset Salicon [29].

Sampling the target fixations. The ground truth fixation map $\mathcal{M}_{gt}$ contains a large number of fixations points (i.e., up to 500) gathered across a batch of subjects (16 viewers per image for the Salicon dataset). Hence, it's necessary to select N points to match the number of fixation queries in the transformer decoder. Moreover, N should be highly smaller than 500, for computational efficiency. For simplicity, we adopt a uniform sampling of N samples out of S points of $\mathcal{M}_{gt}$ to obtain $f_i = (x_i, y_i)$, where $i \in \{1, \ldots, N\}$. This guarantees a diverse set of samples from the different viewers.

Accelerating the training. We observed that SalTR is difficult to optimize, and suffers from slow convergence, i.e., more than 100 epochs are needed to obtain comparable performance to baselines. Following the object detection literature [14, 59, 64, 64], several hypotheses can be proposed to account for this. First, the attention weights are uniformly assigned to all pixels in the feature maps at initialization, hence attending to meaningless locations that do not contribute to the feature propagation mechanism. Second, the discrete bipartite matching is unstable under stochastic optimization, as the same query is matched with different objects across epochs. Lastly, the decoder cross attention is under optimized in the early training, resulting in noisy contextual information for the queries. Inspired by the concept of deformable convolution [13], the approach of [67] is to add a translation term into the formula of the transformer attention, allowing a sparse spatial sampling by

attending to a smaller set of locations (reference points). Consequently, this gating mechanism approximates the full self-attention via the locality inductive bias excluding potential long-term dependencies from the calculation. We adapt the deformable attention mechanism to our settings, termed Deformable SalTR.

Evaluation setting. We compare against the SoTA methods listed in [63] and add newer models with available implementations [42]. Moreover, we test on the MIT300 benchmark [31], which is more challenging than the Salicon test set. As suggested in [9, 42], we use the following evaluation metrics: Similarity Metric (SIM), shuffled AUC (s-AUC), Linear Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), and the Kullback Leibler Divergence (KLD) [9]. We adopt the more recent metrics formulation from [42].

Technical details. SalTR is implemented in PyTorch [54] and trained using a single NVidia RTX3090 24GB GPU. We consider two configurations: SMALL with 3 transformer encoder-decoder layers and BASE with 6 layers, both with a ResNet-50 backbone. The number of fixation queries is set to 100, hence, 100 fixation points aresampled from the ground truth fixation map. SalTR variants are trained for 100 epochs using the AdamW [46] optimizer, whereas the Deformable SalTR is trained for 40 epochs. We employ a warmup of 10 epochs and a Cosine learning rate scheduler with maximum lr set to 10^{-3}.

4.1. Results

State-of-the-art comparison. Here we compare the proposed approach to SoTA image saliency models on both the Salicon and MIT300 validation sets. Table 1 shows the performance comparison in terms of the five metrics for the respective validation sets. We observe that our method performs favorably against existing approaches.

Salicon. SalTR-Base achieves similar scores to UNISAL, whereas the Small variant showed a slightly worse performance highlighting the importance of the transformer decoder depth to optimize appropriately for the unstable bipartite matching. Using the same amount of compute, Deformable SalTR-Small matches the UNISAL performance due to the effective gated attention mechanism allowing for a smoother optimization. Conversely, Deformable SalTR-Base exhibits SoTA scores across the five metrics on par with the best model on these test examples (i.e. EML-NET [28])

MIT300. The model was not trained on this dataset, making this an out-of-distribution test. The proposed models produces a reasonable improvement in accuracy compared to other models, except UNISAL and DeepGaze, which were trained on this specific dataset.

Figure 2 illustrates the predictions on a sample of the Salicon validation set. It can be seen that the fixation maps generated by our model (SalTR-Base) correlate well with the ground truth fixation maps in terms of fixation distribution. Also, we smooth our fixation maps with a Gaussian filter to obtain the continuous saliency map, giving the approach the flexibility to set the standard deviation parameters to fit the downstream application. Also, the effectiveness of the model in saccading to the main objects in the scene can be observed (see Figure 1,2 in supplementary). This demonstrates the effectiveness of the fixation queries in playing the role of the human subjects.

Furthermore, we visualize the decoder self-attentions maps for a set of a randomly selected queries with their respective fixation predictions in Figure 3. We notice that each query's attention is highly local, where it attends to granular details of the image (such as: the Giraffe's mouth for query 13, and the person's leg for query 36). We believe that given well structured keys and values from the encoder latent representations, the queries focus on the refined granularities to predict the fixation points. This design is intuitive and is the closest to the experimental setting in obtaining the saliency datasets.

Saliency for low-level features. SoTA saliency models capture high-level features such as cars, humans, etc. However, these kinds of approaches may fail to adequately capture a number of other crucial features that describe aspects of human visual attention that have been extensively investigated in psychology and neuroscience. Visual search, is one of the most prominent processes shaping human attention [36, 61]. This is where a subject's brain parallel processes regions that differ significantly in one feature dimension i.e. color, intensity, orientation. These correspond to low-level features, which operate as the basic mechanisms of the human visual system. We conducted evaluations of the performance of UNISAL, and our SalTR on samples of low-level attention using images from a recently proposed dataset [36]. The aim is to understand the main differences on how saliency exploration is performed when the self-attention mechanism promotes global connectivity between the image patches. See Section A.3 in the supplementary materials for more details.

As shown in Figure.2 in supplementary, UNISAL produces high-quality saliency maps consistent with the ground truth maps for natural images from Salicon. High-level features such as: human faces, bus, monument, etc; are dominant in these images. The human visual system combines the bottom-up with top-down features to solve the attention task. This behaviour might not be reflected in the fixation/saliency datasets. Hence, end-to-end deep learning based models might learn a good saliency mapping, but actually violate the subtleties of its true definition. Early computational approaches for the visual human system e.g. [6, 20–22, 35, 48, 49, 57] were mostly cognitive based models relying on computing multiple visual features

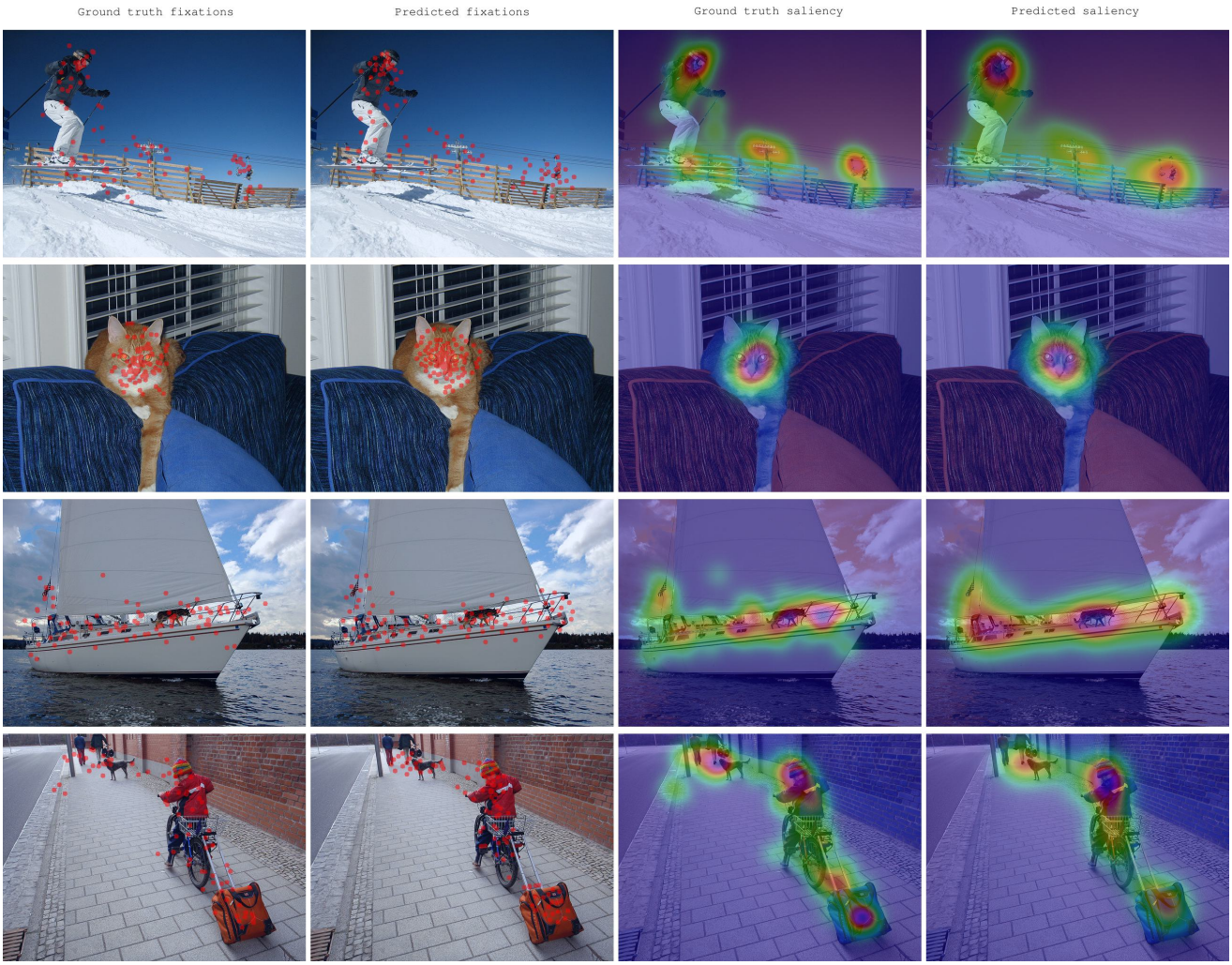

Figure 2. Qualitative results of our model on sample images from SALICON. It can be observed that the proposed approach is able to handle various challenging scenes well and produces consistent fixation/saliency maps.

such as color, edge, and orientation at multiple spatial scales to produce a saliency map. Moreover, could the fixation queries in SalTR bridge this gap, by reasoning over the recurring features and drawing dependencies/similarities?

In fact, UNISAL fails to respond to simple features. For example, considering colour (Figure.3 in supplementary), UNISAL [17] did not capture the penguin as the most salient object, whereas the SalTR succeeded in doing so, as this pattern is solved with the global nature of the self-attention mechanism. This suggests that SalTR is better at incorporating characteristics of the visual system as important priors induced by the self-attention mechanism and the Hungarian matching. Clearly however, as shown in Figure.4 in supplementary, SalTR severely fails when given synthetic images, the model does not respond well to low-level features from the O3 dataset, and nearly produces random fixations around the center.

4.2. Ablation study

Losses importance. The bipartite matching loss appears intuitive to avoid fixation queries collapse. We validate this hypothesis by eliminating the Hungarian matching loss, and only use the final loss in Eq. 4 without any assignments. As expected, the model learns the saliency dataset center bias. In other words. the fixation queries tend to all focus on the most salient features of the input image, while ignoring the rest. This is an artifact of the cross-attention optimization, where a shortcut over the features dominates learning more diverse and rich predictions.

Object detection vs saliency prediction. We attempted to initialize SalTR with DETR weights trained on COCO for object detection. We only train the MLP head on the saliency dataset. The aim is to measure the alignment between the two downstream tasks. The model was un-

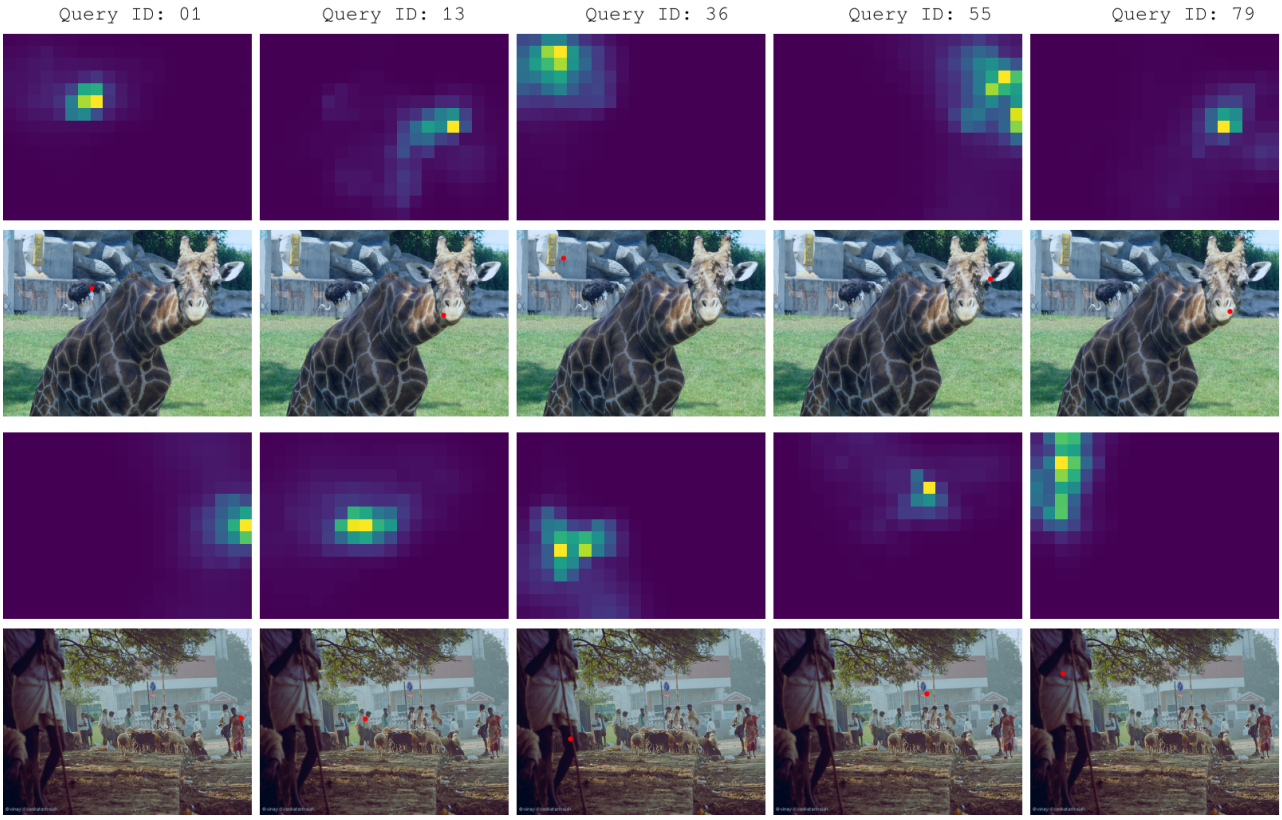

Figure 3. Decoder self-attention for a set of fixation queries. It can be seen that queries attend to spatial locations consistent with their predicted fixation points. Predictions are made with SalTR-Base on a validation set of images.

Table 1. Comparative performance study on Salicon and MIT300.

Models	Salicon					MIT300				
	SIM	s-AUC	CC	NSS	KLD	SIM	s-AUC	CC	NSS	KLD
ITTI [27]	0.37	0.61	0.20	–	–	0.46	0.13	0.44	1.11	0.95
GBVS [24]	0.44	0.63	0.42	–	–	0.48	0.62	0.47	1.24	0.88
Salicon [29]	–	–	–	–	–	0.51	0.73	0.56	1.70	0.78
CASNet [19]		–	–	–	–	0.58	0.73	0.70	1.98	0.58
EML-NET [28]	0.79	0.74	0.89	2.05	0.52	0.74	0.67	0.78	2.48	0.84
MSI-Net [37]	0.80	0.74	0.90	2.01		0.67	0.74	0.77	2.30	0.42
TranSalNet [47]	–	–	–	–	–	0.68	0.74	0.80	2.41	1.01
SalGAN [52]		0.75	0.76	2.47	-	0.63	0.73	0.67	1.86	0.75
UNISAL [17]	0.77	0.73	0.87	1.95	-	0.67	0.78	0.78	2.36	0.41
DeepGaze [41]	–	–	–	–	–	0.66	0.77	0.77	2.33	0.42
SalTR-Small	0.75	0.71	0.84	1.79	0.98	0.61	0.70	0.70	1.98	0.59
SalTR-Base	0.78	0.75	0.87	1.97	0.74	0.65	0.74	0.75	2.22	0.46
Deformable SalTR-Small	0.77	0.73	0.86	1.88	0.82	0.64	0.75	0.76	2.09	0.49
Deformable SalTR-Base	0.79	0.77	0.89	2.12	0.62	0.69	0.79	0.80	2.45	0.36

able to achieve the baseline scores (i.e. KLD: 2.5, NSS: 0.9). We observed that the fixations were mostly overestimated around the objects center (see Figure 3 in supplementary). We hypothesize that the salient regions in an image may correspond to the objects of interest within the image. Clearly, however, the most salient regions are

Table 2. Impact of the number of fixation queries. Performance comparison when it is varied. 100 is the optimal number of queries for the saliency task. SalTR-Base is showed.

SalTR		Salicon				
		SIM	s-AUC	CC	NSS	KLD
Number of queries	50	0.75	0.70	0.82	1.55	1.04
	100	0.78	0.75	0.87	1.97	0.74
	150	0.79	0.74	0.88	1.92	0.71
	200	0.78	0.74	0.86	1.94	0.78

not necessarily the objects in the image, but could rather be other features or patterns that catch the viewer's attention [4]. This potentially explains the failure of this setting.

Varying the number of fixation queries. We investigate the optimal number of target fixations for better results. As shown in Table 2, we vary the number of queries N, hence the number of target fixations S. We observe that 100 is the optimal value and higher values result in a diminishing returns in the performance since higher number of queries make the optimization harder.

The impact of the Gaussian smoothing. Table 3 represents the impact of the standard deviation (σ) of a Gaussian filter applied to the fixation map for the SalTR-Base model. It can be observed that the $\sigma = 19.0$ generally offers the best performance across most metrics. Notably, the model achieves the highest SIM, CC, and NSS scores at $\sigma = 19.0$. Furthermore, we can observe that the KLD is the metric that is highly affected by the smoothing parameters, whereas the CC is more or less the same after the 19.0 value.

All vs single subject. To the best of our knowledge, SalTR is the first approach for learning saliency prediction from fixations only. However, it still does not replicate the the experimental setting fully, as the fixation queries predict a single locations, whereas human subjects may attend to multiple locations. We attempted to sample a single subject as targets so all the queries simulates a single human with multiple predicted fixations. This design resulted in visually appealing saliency maps (see Figure 6 in supplementary), but the scores did not match the baselines because the predictions were mostly sparse. This potentially highlights an issue with the metrics, that require an over-estimated saliency map to match the dense ground truth map aggregated over a batch of subjects.

4.3. SalTR for scanpath prediction

Scanpath prediction refers to the process of estimating the timely trajectory of fixation points for a human subject when viewing a visual stimulus. The SalTR design allows for manipulating the Transformer decoder mask. Indeed, we used the full mask for parallel decoding previously; clearly, however, we can use a causal mask, and add an end-of-sequence (EOS) (i.e. position $(0, 0)$) to the target fixations, for the auto-regressive decoding. By doing so, SalTR can naturally handle ordered fixation predictions (i.e. scanpaths). The prediction head outputs the fixations points and their duration in seconds. We train SalTR-Base using the 10k Salicon images; for each image, we select all the human subject's scanpaths, each serving as its own separate training example. The model should converge on the correct distribution over scanpaths given the image. We finetune the transformer decoder only for 20 epochs.

Table 3. Impact of the number standard deviation. Performance comparison when the Gaussian smoothing parameter is varied. SalTR-Base is showed.

	Salicon				
	SIM	s-AUC	CC	NSS	KLD
5.0	0.42	0.50	0.54	1.33	1.63
10.0	066	0.70	0.77	1.77	0.99
19.0	0.78	0.75	0.87	1.97	0.74
30.0	0.76	0.72	0.87	1.78	1.16

Results. The Multi-Match (MM) [15] measure is used as the main metric for ranking scanpath prediction models. Our approach obtains an MM score of 0.93 on the 5k Salicon examples (PathGAN [1]: 0.96). This is achieved by averaging the individual results against all human viewer scanpaths for a single image. Figure 4 in the supplementary material shows the quantitative results on sample images from Salicon. Our model demonstrates a clear and consistent tracking of the ground truth, indicating its ability to accurately capture human visual attention.

5. Conclusion

We present a novel approach for saliency prediction in images, named Saliency TRansformer (SalTR), which leverages parallel decoding in transformers to learn saliency solely from fixation maps. Unlike existing models that rely on continuous saliency maps, our approach directly predicts fixations by treating saliency prediction as a set prediction problem. We conducted experiments on the Salicon and MIT300 benchmarks and achieved remarkable performance compared to SoTA methods. Our approach not only replicates the data collection pipeline used in generating saliency datasets but also eliminates the need for continuous saliency annotations. Furthermore, we extended our approach to the scanpaths prediction problem and demonstrated its effectiveness. Overall, our approach offers a promising direction for saliency prediction, focusing on the discrete fixation maps and directly predicting fixation points. It opens up possibilities for application-guided saliency prediction, as per the flexibility offered in our design.

References

[1] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 8

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019. 3

[3] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):679–700, 2019. 1

[4] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. 8

[5] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 1, 2, 3

[6] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006. 2, 5

[7] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. 1, 2

[8] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. 3

[9] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016. 5

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1, 3, 4

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE, 2016. 2

[12] Yasser Dahou, Marouane Tliba, Kevin McGuinness, and Noel O'Connor. Atsal: An attention based architecture for saliency prediction in 360 videos. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 305–320, Cham, 2021. Springer International Publishing. 2

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4

[14] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 4

[15] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44:1079–1100, 2012. 8

[16] Guanqun Ding, Nevrez İmamoğlu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing*, 120:104395, 2022. 3

[17] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 419–435. Springer, 2020. 2, 6, 7

[18] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009. 2

[19] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7521–7531, 2018. 7

[20] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Advances in neural information processing systems*, pages 481–488, 2005. 2, 5

[21] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil. Decorrelation and distinctiveness provide with human-like saliency. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 343–354. Springer, 2009. 2, 5

[22] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1915–1926, 2012. 2, 5

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[24] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006. 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[26] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. 2

[27] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998. 1, 2, 7

[28] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, page 103887, 2020. 2, 5, 7

[29] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 1, 4, 7

[30] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 809–824. Springer, 2016. 2, 3

[31] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 5

[32] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 2

[33] Wolf Kienzle, Matthias O Franz, Bernhard Schölkopf, and Felix A Wichmann. Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of vision*, 9(5):7–7, 2009. 2

[34] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. 1

[35] Gert Kootstra, Arco Nederveen, and Bart De Boer. Paying attention to symmetry. In *British Machine Vision Conference (BMVC2008)*, pages 1115–1125. The British Machine Vision Association and Society for Pattern Recognition, 2008. 2, 5

[36] 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. 5

[37] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 7

[38] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. 1, 2

[39] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014. 2

[40] Matthias Kummerer, Thomas SA Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–787, 2018. 3

[41] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 4789–4798, 2017. 7

[42] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing. 5

[43] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):802–817, 2006. 2

[44] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021. 3

[45] Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing*, 27(7):3264–3274, 2018. 2

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[47] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 2022. 7

[48] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*, pages 433–440. IEEE, 2011. 2, 5

[49] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2049–2056. IEEE, 2006. 2, 5

[50] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003. 2

[51] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 2

[52] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 1, 2, 7

[53] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016. 1, 2

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner,

Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, 2019. 5

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

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

[57] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009. 2, 5

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[59] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021. 4

[60] Antonio Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003. 2

[61] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 1, 5

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[63] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903, 2018. 5

[64] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based object detection. *arXiv preprint arXiv:2109.07107*, 2021. 4

[65] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 1

[66] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008. 2

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4