




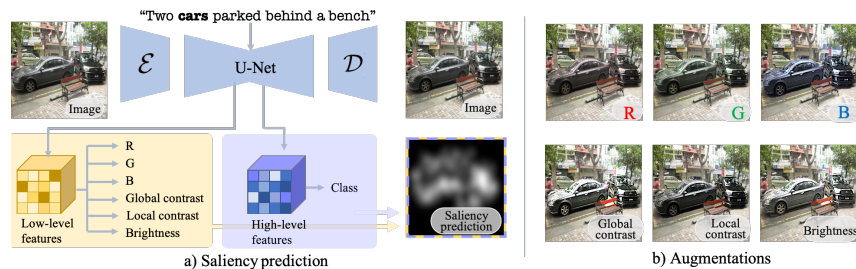


# Data Augmentation via Latent Diffusion for Saliency Prediction

Bahar Aydemir<sup>1</sup>, Deblina Bhattacharjee<sup>1</sup>, Tong Zhang<sup>1</sup>, Mathieu Salzmann<sup>1</sup>, and Sabine Süssstrunk<sup>1</sup>

School of Computer and Communication Sciences, EPFL, Switzerland  
{name}. {surname}@epfl.ch



**Fig. 1: Overview of our data augmentation and saliency prediction method.** We use photometric and semantic properties such as color, contrast, brightness, and class to construct **low-level** and **high-level** features from the intermediate U-Net features. We leverage these features to estimate saliency and generate image edits to augment training data for saliency prediction. We use a saliency-guided cross-attention mechanism between the input image and the text prompt to perform localized edits.

**Abstract.** Saliency prediction models are constrained by the limited diversity and quantity of labeled data. Standard data augmentation techniques such as rotating and cropping alter scene composition, affecting saliency. We propose a novel data augmentation method for deep saliency prediction that edits natural images while preserving the complexity and variability of real-world scenes. Since saliency depends on high-level and low-level features, our approach involves learning both by incorporating photometric and semantic attributes such as color, contrast, brightness, and class. To that end, we introduce a saliency-guided cross-attention mechanism that enables targeted edits on the photometric properties, thereby enhancing saliency within specific image regions. Experimental results show that our data augmentation method consistently improves the performance of various saliency models. Moreover, leveraging the augmentation features for saliency prediction yields superior performance on publicly available saliency benchmarks. Our predictions align closely with human visual attention patterns in the edited images, as validated by a user study. Our code is publicly available on GitHub<sup>1</sup>.

**Keywords:** Visual saliency · latent diffusion · data augmentation.

## 1 Introduction

Visual saliency prediction, with applications in image and video compression [48], and image enhancement [2, 36], aims to identify the regions within an image that attract the human gaze. The taxonomy of saliency estimation covers both bottom-up and deep-learning approaches. Bottom-up methods use hand-crafted features to estimate the saliency of a region, whereas the deep learning approaches aim to benefit from large-scale data. Recent methods incorporate various priors such as object dissimilarity [5], temporal information [6], and inter-object relationships [62] alongside large-scale data to enhance performance. However, the limited size of available datasets poses a significant challenge to these learning-based methods. The most extensive saliency dataset contains only 10k training images [29], in contrast to the millions of images in classic computer vision datasets [16, 55]. The reason is that collecting ground-truth saliency data through psychophysical experiments is both costly and time-consuming. Hence, collecting saliency annotations on the scale of millions is not practically feasible. In the general computer vision context, limited access to data is typically overcome via data augmentation. However, standard data augmentation techniques such as cropping, rotation, and shearing, used in tasks like image classification and segmentation, are not saliency invariant [12]; such manipulations also alter human gaze patterns. This makes the existing data augmentation techniques ineffective, as saliency is not invariant under such transformations. . This raises a crucial question: Is it possible to manipulate image saliency in a *predictable* manner, such as *enhancing the saliency of a region while maintaining the integrity of the rest of the scene*? That is, can we change the saliency of a region in an expected direction? To answer this question, we refer to the factors influencing saliency. Studies in cognitive science [59] have demonstrated that saliency is affected by both low-level features, such as color [7, 44], contrast [47], and brightness [45], and high-level features, such as semantics [11, 20]. In essence, the saliency of a region is scene-dependent and can vary dramatically from one image to another. Although synthetic datasets [8] featuring basic shapes in various colors, sizes, and orientations study the impact of the above factors on saliency, they fall short in complex everyday scenes with real objects. Regarding the question posed earlier, *yes, we can change saliency predictably*. In this work, we build on this idea by selecting a region and increasing its saliency. The edited region’s saliency should be higher than the original. Thus, this transformation, with a known direction of saliency change, is suitable for data augmentation in saliency prediction.

In our work, we address these limitations by introducing an image editing strategy that allows us to alter a single factor in a scene while keeping the others constant. This allows us to control (enhance or decrease) and interpret the saliency of regions within an image while preserving the integrity of the rest of the real-world visual scene. To achieve this, we use readout layers to constrain the image edits within a range for the edited properties while increasing the

---

<sup>1</sup> <https://github.com/IVRL/Augsal>

saliency over the selected region. This approach enables us to alter a single aspect of the scene while modifying the saliency in a desired direction. This process of editing the image and subsequently, modifying the saliency allows our method to generate image-saliency annotation pairs, thereby addressing the lack of large-scale data for learning-based saliency approaches. Moreover, we employ a vision-language-based cross-attention mechanism [56] that localizes the target image regions for editing, thereby eliminating the need for user-provided input masks as required in [1, 43]. Thus, the cross-attention mechanism completely automates our editing method for saliency prediction, making it suitable for data augmentation.

To edit the image in a controllable and interpretable manner for visual saliency, we employ the widely-used diffusion model [17]. Recent text-to-image generation models have proven their ability to generate diverse and creative images. They enable image generation by conditioning on the input text prompts. They have also been adapted for dense prediction tasks such as depth estimation [19, 27, 51] and segmentation [4, 57], showcasing their transferability. However, saliency prediction presents a unique set of challenges, as it is highly dependent on both the semantic content and the low-level details within the images, requiring an understanding of visual importance that goes beyond structural coherence. Since diffusion models [50] are trained on large and diverse data, they are suitable to exploit their pre-trained knowledge in saliency prediction.

We summarize our contributions as follows:

- We introduce a data augmentation method that enriches the training data while predictably modifying image saliency.
- Our approach introduces multi-level feature readouts, allowing us to exploit knowledge from the Stable Diffusion [50] architecture without retraining for image saliency prediction.
- We introduce controllability in saliency by editing photometric properties, namely, contrast, brightness, and color, in a shared latent feature space of generated image edits and saliency using cross-attention mechanisms.
- Our model is interpretable in the type of image edits that lead to enhanced saliency, which, in turn, aligns better with human visual attention, as we show via a user study.

Our experiments evidence that our image editing-based data augmentation strategy consistently improves the state-of-the-art saliency predictors. Furthermore, our novel diffusion-based saliency estimation approach outperforms the existing saliency models.

## 2 Related Work

### 2.1 Saliency Prediction Models

Convolutional neural networks (CNNs) have attained significant popularity in deep saliency prediction, as evidenced by numerous studies [6, 28, 34, 35, 49] and

pioneered by [58]. Notably, Kümmerer et al. [34] have shown that CNNs trained for object classification greatly enhance saliency prediction, aligning with Judd et al.’s [31] earlier findings using bottom-up detectors. This approach has been adopted by leading deep saliency prediction networks like [14, 15, 25, 33, 35] and further developed by EML-Net [28], which integrates features from multiple CNN backbones for object classification. Linardos et al. [38] also explored various object classification backbones for saliency prediction. Recently, [41] has incorporated transformer blocks to encode long-range relationships, although still relying on a CNN encoder. All these methods transfer the learned encodings from *object recognition* (based on ImageNet [16]) to saliency prediction. Differently, in this work, we leverage the encodings from a large *vision-language-based* model to predict saliency. In particular, we use a pre-trained and frozen Stable Diffusion [50] model that encodes information from 2.3 billion images. We create multi-level features from this pre-trained model for generating image edits and predicting saliency. The use of diffusion models allows our approach to be both controllable and interpretable.

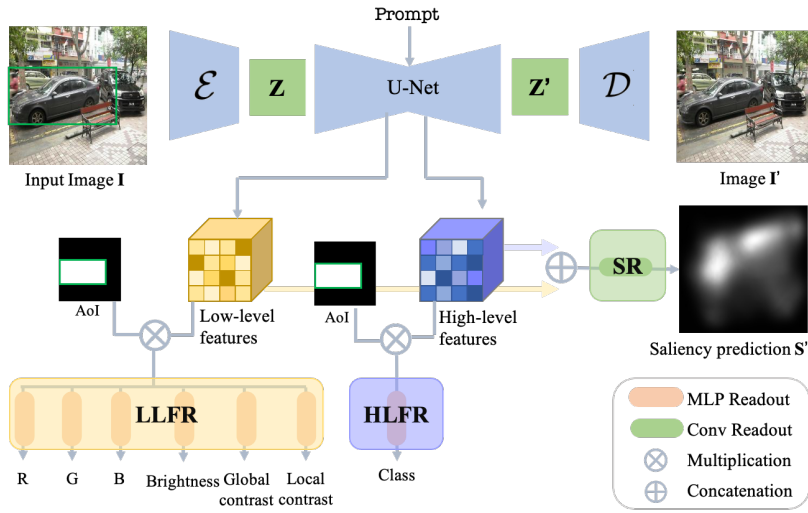
## 2.2 Diffusion Models on Dense Prediction Tasks

Diffusion-based approaches gradually corrupt an image with different levels of noise to train denoiser models. Consequently, the model learns how to denoise a random noise into a high-quality synthetic image [53]. Diffusion models have demonstrated remarkable results in multiple fields including audio processing, natural language processing, and video generation [24, 52, 63], as well as in multiple tasks such as image generation [13, 50, 61], semantic segmentation [4, 57], and depth estimation [19, 27, 51], among others.

All of these methods achieve state-of-the-art performance on their respective benchmarks by exploiting structural coherence in the scene which is encoded within the diffusion model. However, saliency prediction requires understanding the scene beyond structural coherence. Therefore, we exploit the pre-trained diffusion model by generating image edits that control the saliency of a desired target region. Guided by these image edits, our method predicts saliency in alignment with human visual attention. To the best of our knowledge, this is the first work to predict and control saliency in a diffusion framework.

## 2.3 Cross-attention to Generate Image Edits

Generative models are well-known for their capability to edit images. Following the seminal work on cross-attention [56] that exploits language and vision encodings, many works have employed the cross-attention mechanism within the diffusion paradigm to generate image edits [10, 13, 23, 50]. In particular, [13] introduces a semantically guided cross-attention mechanism to generate images. Differently, Hertz et al. [23] presents a text-driven image editing method using cross-attention. Subsequently, Brooks et al. [10] learn to follow image editing instructions given an input image and a text prompt. While these works reason



**Fig. 2: Overview of the proposed architecture in training.** We use an **encoder** ( $\mathcal{E}$ ), a **decoder** ( $\mathcal{D}$ ), and a denoising **U-Net** from Stable Diffusion [50]. We invert the input image to extract encoded representations from the U-Net to construct **low-level** and **high-level** features. We train the Low-Level Feature Readout (LLFR) and High-Level Feature Readout (HLFR) modules using related photometric and semantic properties inside the area of interest (AoI), respectively. Finally, we concatenate the **low-level** and **high-level** features to predict the saliency map ( $\mathbf{S}'$ ) using the Saliency Readout (SR) module.  $\mathbf{Z}$  and  $\mathbf{Z}'$  represent the encoded and denoised image latent vectors respectively.

about the cross-attention between text and image, they do not consider photometric properties such as contrast or salient regions to generate image edits. Since contrast and saliency are significant cues to generate image edits, we introduce a *saliency-guided cross-attention* mechanism within our diffusion paradigm. Using the cross-attention mechanism makes our editing method for saliency prediction completely automated and thus well-suited for data augmentation.

### 3 Methodology

Our approach is depicted in Figure 2. We incorporate the encoder ( $\mathcal{E}$ ), the decoder ( $\mathcal{D}$ ), and the denoising U-Net from Stable Diffusion v1.5 [50], all of which are frozen in our framework. Specifically, we start by encoding the input image and then extracting the intermediate representations from the middle layers of the denoising U-Net into two feature maps: high-level and low-level features. We use Low-Level Feature Readout (LLFR) and High-Level Feature Readout (HLFR) modules to learn desired photometric and semantic properties. The denoising U-Net denoises the latent feature  $\mathbf{Z}_t$  to produce  $\mathbf{Z}'$  and the image itself. Lastly, we concatenate the high and low-level features and decode them with the

Saliency Readout (SR) module to predict saliency. We detail the components of our model in what follows.

### 3.1 Multi-level Features

We start by encoding the input image and then extracting the intermediate representations from the middle layers of the denoising U-Net. We aggregate the intermediate representations with the bottleneck layers, as proposed in [42]. The aggregated representations allow us to construct two feature maps comprising high-level and low-level features, respectively. This multi-level representation allows our model to depend on both low-level and high-level features while predicting saliency. The separation between those features allows them to optimize for the desired tasks without interacting with each other. We learn those features using the LLFR, HLFGR, and SR modules as we explain in the next sections.

### 3.2 Low-Level Feature Readout (LLFR) Module

We employ seven Multi-Level Perceptron based readout networks to access the intermediate representations of the desired photometric properties. Specifically, we utilize small readout networks that share the low-level features to interpret red, green, blue, brightness, and local and global contrast values. The choice of those properties is based on evidence showing that saliency is significantly influenced by color, contrast, and brightness [44,45,47]. To train those networks, we crop a random patch from the image and compute the related photometric properties inside this AoI as follows:

**R, G, B:** For each color channel, we calculate the average color in the patch as

$$\mu_R = \frac{1}{N} \sum_{i=1}^N P_{[0,i]}, \quad \mu_G = \frac{1}{N} \sum_{i=1}^N P_{[1,i]}, \quad \mu_B = \frac{1}{N} \sum_{i=1}^N P_{[2,i]},$$

where  $P_{[j,i]}$  denotes each pixel in the  $j^{th}$  channel of the image patch.

**Brightness:** We first convert the image into grayscale and then compute the brightness as the mean intensity of the patch as

$$\mu_{Br} = \frac{1}{N} \sum_{i=1}^N P_i,$$

where  $P_i$  denotes each pixel in the image patch.

**Local Contrast:** We compute the mean intensity ( $\mu$ ) of the patch pixels, together with their variance ( $\sigma^2$ ) and local contrast ( $c_L$ ) as

$$\mu = \frac{1}{N} \sum_{i=1}^N P_i, \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (P_i - \mu)^2, \quad c_L = \sqrt{\sigma^2}.$$

**Global Contrast:** We compute the mean intensity ( $\mu$ ) of the image pixels, as well as their variance ( $\sigma^2$ ) and the global contrast ( $c_G$ ) as

$$\mu = \frac{1}{N} \sum_{i=1}^N I_i, \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (P_i - \mu)^2, \quad c_G = \sqrt{\sigma^2},$$

where  $P_i$  and  $I_i$  denote each pixel in the patch and image, respectively.

### 3.3 High-Level Feature Readout (HLFR) Module

We use one convolution-based readout network to learn semantic information present in the scene. Specifically, we classify the object present in the area of interest. We aim to learn the semantics since saliency also depends on the semantics of the scene [11, 20].

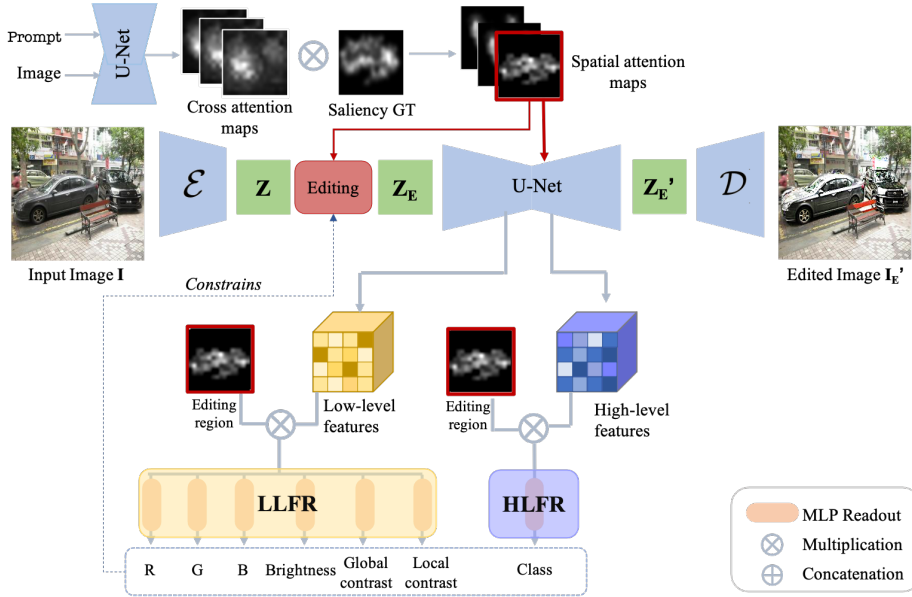
**Classification:** We train the classification readout by using the object bounding boxes from the MS-COCO dataset [37]. We construct high-level features that contain the semantics of the objects in the selected region. Then, we pass these features to the saliency readout (SR) to provide information about the semantics present in the scene.

### 3.4 Saliency Readout (SR) Module

The saliency readout network consists of 7 convolutional layers with SiLU [21] activations. This module takes both the high-level and low-level features as input. We use skip connections to prevent the gradient of the features from vanishing. We decode the concatenated features to predict the final saliency map  $\mathbf{S}'$ . Equipped with high-level and low-level feature representations via readout networks, our approach can achieve both controllability and interpretability for saliency prediction.

## 4 Image Editing for Data Augmentation

Image editing requires two components: 1. Choosing the type of edits, and 2. locating the graphical elements on which the edits are carried out. In this work, we manually choose the type of edits and then apply them to the graphical elements by employing our saliency-guided cross-attention mechanism. We detail the different types of edits that we incorporate in our approach in Section 4.1. As shown in Figure 3, during the editing stage, we extract the cross-attention features between the input image and the text prompt. These features are multiplied with the ground-truth saliency maps from SALICON [29] to create the saliency-guided cross-attention maps. These maps reveal where the salient regions intersect with elements from the prompt. We select the spatial attention map with the highest sum, indicating the most salient area corresponding to a word in the prompt. Thus, we dub this process a saliency-guided cross-attention

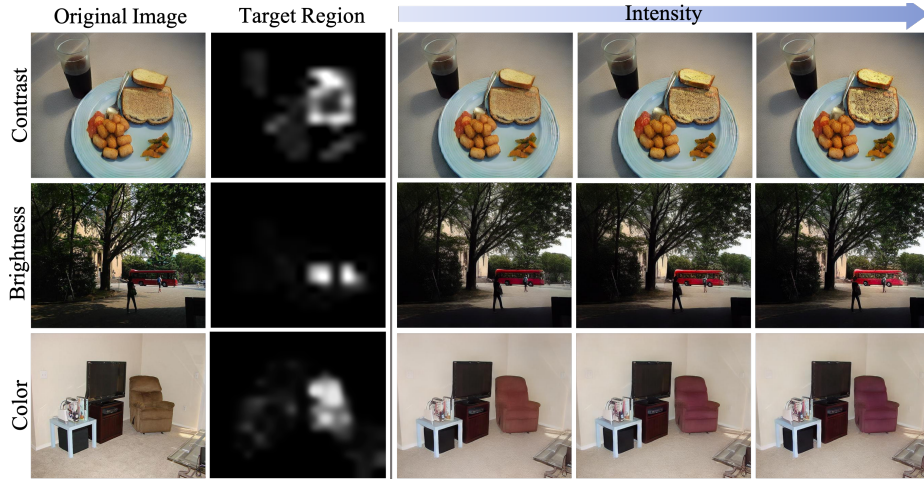


**Fig. 3: Overview of the proposed image editing architecture.** We extract cross-attention maps between the input image and the prompt, and then multiply with the saliency map to create spatial attention maps. These maps highlight the intersections of salient regions with elements from the prompt. We inject this map into the denoising U-Net alongside the edited prompt. This integration modifies the latent features  $Z'$  and the extracted multi-level features, resulting in the generating edits. We use frozen readout layers to constrain the edits in terms of those features.

mechanism. We leverage the obtained saliency-guided cross-attention map to perform localized edits. To this end, we append a word to the prompt depending on the type of edit and inject the selected saliency-guided cross-attention map into the denoising U-Net. This process modifies  $Z'$ . By denoising and decoding this edited  $Z'_E$ , we obtain the generated image edits  $I'_E$ . We refer the reader to the supplementary material for the pseudocode of the editing algorithm.

As evidenced by cognitive studies on human visual attention [59], saliency is influenced by contrast, brightness, and color. Our work involves controlled edits to these photometric properties, allowing us to control and interpret the model’s saliency prediction in response to such edits. We focus on photometric edits that enhance the saliency of the most salient region that corresponds to a word from the prompt. As a result, this region remains salient after the edit. Conversely, when an edit reduces saliency, it is unclear whether the reduction makes the region non-salient or not. We show such edits that diminish contrast and brightness and their effects on saliency in the supplementary material. We describe each edit in detail in the following section.





**Fig. 4:** Original image, selected editing region, and edited images at different intensity levels for contrast, brightness, and color edits. In the first row, we increase the contrast in the bread. The second row shows an increase in the brightness of the red bus. The last row shows a progressive color edit of the chair to purple. These edits aim to increase the saliency of the target region shown in the second column.

#### 4.1 Editing Types

**Contrast Increase.** We increase the contrast of a specific region as

$$Z_E = M \times (\alpha \times (Z - \mu) + \mu) + (1 - M) \times Z, \quad (1)$$

where  $\mu$  denotes the mean inside the selected area  $M$ ,  $Z$  denotes the shared latent vector,  $Z_E$  denotes the edited latent vector, and  $\alpha$  is the scale parameter which denotes the strength of the edit. We inject the selected map into the cross-attention maps of the denoising U-Net. Following this, we insert an empty token before the word token that corresponds to the selected region in the prompt. This approach enables altering the cross-attention while preserving the semantic content of the region. The first row in Figure 4 shows a progressive increase in contrast at varying intensity levels.

**Brightness Increase.** We increase the brightness of a specific region as

$$Z_E = M \times (Z + \alpha) + (1 - M) \times Z, \quad (2)$$

where  $Z$  and  $Z_E$  are the shared latent vector and the edited latent vector, respectively, and  $\alpha$  is the scale parameter which denotes the strength of the edit.

Lastly, similar to the previous section, we insert an empty token into the prompt and inject the selected map into the cross-attention maps of the denoising U-Net. The second row in Figure 4 shows a progressive increase in brightness with varying levels of intensity.

**Color Change.** To modify the color of a region, we add the target color before the word token that corresponds to the targeted region in the prompt. Then, we inject the selected map into the cross-attention maps of the denoising U-Net, thereby associating the target color with the selected regions. We scale the values of the selected map to amplify the intensity of the color change. The last row in Figure 4 shows a gradual alteration in color at different intensity levels.

## 4.2 Scaling and Constraining the Image Edits

In order to use our editing methods, we need to calculate a scale parameter because one unit increase in the RGB space does not correspond to one unit increase in the four-channel diffusion latent space. This can result in oversaturated and high-contrast images [3]. To counter this, we introduce a *channel-wise* scale parameter,  $\gamma$ . We start by approximating the decoder  $\mathcal{D}$  as a matrix, namely  $\mathbf{A}_{4 \times 3}$ , as described in [32]. Subsequently, we compute the pseudo-inverse of the approximated decoder matrix as  $\mathbf{A}^+_{3 \times 4}$ . We solve  $\gamma_{1 \times 4} = \mathbf{1}_{1 \times 3} * \mathbf{A}^+_{3 \times 4}$  to find the value of the scale parameter  $\gamma_{1 \times 4}$ . Consequently, we scale our image edits, specifically  $\mu$  in contrast and  $\alpha$  in brightness edits with this  $\gamma$ . We demonstrate the impact of this scaling in Section 5.6.

Additionally, we employ a mechanism to constrain edits and prevent overly strong edits. We create a vector of photometric properties for each image/patch and calculate the standard deviation across the original images. During editing, if the LLFR module’s output deviates more than two standard deviations from the original image, we reduce the edit strength to maintain image naturalness. This process ensures balanced and realistic edits as we show in Section 5.6.

## 4.3 Loss Functions

We use readout, saliency, and editing losses. The readout losses ensure that extracted features contain the desired image properties. We discuss the hyper-parameters in the supplementary material.

**Readout losses.** We calculate the readout losses as the  $\mathcal{L}_2$  distance between the decoded image properties and their ground-truth values as:

$$\begin{aligned} \mathcal{L}_{\text{readout}}(I) = & \lambda_1 * \mathcal{L}_2(\mu'_{RGB}, \mu_{RGB}) + \lambda_2 * \mathcal{L}_2(\mu'_{Br}, \mu_{Br}) \\ & + \mathcal{L}_2(c'_L, c_L) + \lambda_3 * \mathcal{L}_2(c'_G, c_G) + \lambda_4 * \mathcal{L}_{CE}(c', c), \end{aligned} \quad (3)$$

where  $\mu_{RGB}$ ,  $\mu_{Br}$ ,  $c_L$ , and  $c_G$  are calculated as described in Section 3.2.  $\mu'_{Br}$ ,  $c'_L$ ,  $c'_G$  and  $c'$  are the predicted values shown as outputs of the LLFR and HLFR modules in Figure 2 and 3.  $\mathcal{L}_{CE}$  denotes the cross-entropy loss for classification.

**Saliency loss.** The saliency ground-truth maps are the blurred version of the fixation maps collected by user experiments [31]. Hence, they follow a spatial Gaussian distribution. The MSE denoising loss does not guarantee that the predicted saliency map follows this distribution. Hence, we use the KLD [26] and CC [30] losses to improve the distribution and coverage of our prediction, respectively. Specifically, denoting the saliency ground truth for image  $I$  as  $S$  and

the predicted saliency as  $S'$ , we use the loss

$$\mathcal{L}_{\text{saliency}}(I) = \lambda_5 * \text{CC}(S, S') + \lambda_6 * \text{KLD}(S, S'), \quad (4)$$

where

$$\text{KLD}(S', S) = \sum_i S_i \log \left( \varepsilon + \frac{S_i}{\varepsilon + S'_i} \right), \quad (5)$$

with  $i$  iterating over the image pixels and  $\varepsilon$  being a small constant to avoid numerical instabilities. Furthermore, we have

$$\text{CC} = \frac{\sum(S' - \bar{S}')(S - \bar{S})}{\sqrt{\sum(S' - \bar{S}')^2 \sum(S - \bar{S})^2}}, \quad (6)$$

where  $\bar{S}$  is the mean value over pixels and the summations run over the image pixels. CC directly measures how well the spatial distribution of the predicted saliency map matches the spatial distribution of the ground truth.

**Editing Loss.** For editing, we use the loss

$$\mathcal{L}_{\text{edit}}(I) = \text{BCE}(\mathbf{M} \times S', \mathbf{M} \times S'_E), \quad (7)$$

where BCE is the binary cross-entropy and

$$\mathbf{M} = \mathbf{C}_{i^*} \times \mathbf{S}, \quad i^* = \arg \max_i \langle \mathbf{C}_i, \mathbf{S} \rangle \quad (8)$$

is the target editing region defined as the elementwise product of  $\mathbf{C}_{i^*}$  and  $\mathbf{S}$  where  $i^*$  is the index of the cross-attention map with the largest inner product with  $\mathbf{S}$ .  $\mathbf{M}$  represents the selected spatial attention map. We illustrate this loss in the supplementary material.

## 5 Experiments and Results

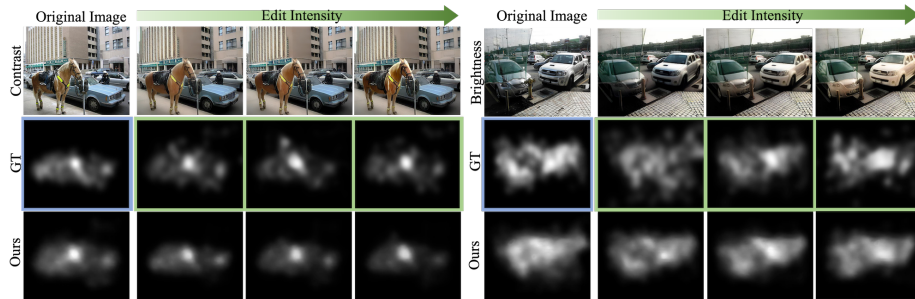
### 5.1 Experimental Setup

We train the HLF and LLFR modules for 20k iterations to be able to reconstruct the respective image properties. We train the SR module for 22k iterations. We use a learning rate of  $5 \times 10^{-5}$  and  $1 \times 10^{-4}$  for the feature extractors and the readout networks, respectively. We use AdamW [40] optimizer for all modules.

### 5.2 Datasets

We report the performance of our methods on three publicly available saliency estimation benchmarks [9, 29, 31]. We train our models on 10,000 images of the SALICON [29] dataset, which consists of diverse context-rich images from the MS COCO dataset [37]. The ground truth of the official SALICON test set is not released but predictions can be submitted for evaluation on the LSUN challenge website<sup>2</sup>. For the prompts, we use the captions from the MS COCO dataset [37].

<sup>2</sup> <https://competitions.codalab.org/competitions/17136>



**Fig. 5:** (**Top Row**): Original and edited images with the (**Middle Row**): ground-truth saliency maps from SALICON [29], shown in **blue** and from our user study, shown in **green**. We report that our generated image edits can shift human attention toward the edited region. For instance, by enhancing the contrast of the horse in the image to the left, we observe that the attention focuses on the horse as the intensity of the edit increases. Similarly, by enhancing the brightness of the cars, they gather more attention as their brightness increases. (**Bottom Row**): Our saliency prediction model can achieve saliency estimations that align with the ground-truth maps.

### 5.3 Augmenting Data

To augment data, we use a sampling parameter  $p$ , which denotes the probability of training with the original image and ground truth. Otherwise, we randomly select one type of edit and train with our editing loss. We use the training images of the SALICON dataset [29] for augmentation. We report results with  $p = 0.5$  in Table 1 and Table 2. We compare different values of  $p$  and illustrate our loss during augmentation in the supplementary material.

### 5.4 User Study

To evaluate our method, we conduct an eye-tracking user study with 8 participants. We randomly select 50 images from each editing category with three intensity levels, giving a total of 150 images. We display the images to the participants in a random sequence for 5 seconds separated by blank intervals of 2-seconds. We constructed the saliency ground truth for these generated image edits following common practice presented in [31]. Figure-5 presents the original and edited images alongside their saliency predictions and the collected ground truth, demonstrating our edits’ effectiveness in increasing the target region’s saliency. Our saliency predictions closely align with human visual attention patterns. A paired samples T-Test between the original and edited images confirms a significantly increased fixation on the edited regions, giving a p-value of  $< 0.05$ .

### 5.5 Quantitative Results

**Saliency Prediction** We evaluate the performance of our saliency prediction model that is trained with our data augmentation method and tested on the

| Model        | AUC $\uparrow$ | KL $\downarrow$ | NSS $\uparrow$ | CC $\uparrow$ | SAUC $\uparrow$ | SIM $\uparrow$ |
|--------------|----------------|-----------------|----------------|---------------|-----------------|----------------|
| DINet        | 0.863          | 0.613           | 1.974          | 0.860         | 0.742           | 0.784          |
| DSCLRCN      | 0.869          | 0.637           | 1.979          | 0.831         | 0.736           | 0.715          |
| SalNet       | 0.860          | 0.674           | 1.766          | 0.730         | 0.711           | 0.696          |
| TempSAL      | 0.869          | 0.195           | 1.967          | 0.911         | 0.745           | 0.800          |
| SAM          | 0.866          | 0.610           | 1.965          | 0.842         | 0.741           | 0.751          |
| SimpleNet    | 0.869          | 0.201           | 1.960          | 0.907         | 0.743           | 0.793          |
| SALICON      | 0.837          | 0.658           | 1.877          | 0.657         | 0.694           | 0.639          |
| DeepGaze IIE | 0.869          | 0.285           | 1.996          | 0.872         | <b>0.767</b>    | 0.733          |
| UNISAL       | 0.864          | 0.350           | 1.952          | 0.879         | 0.739           | 0.775          |
| RINet        | 0.869          | <b>0.189</b>    | 1.982          | 0.911         | 0.746           | 0.803          |
| MDSEM        | 0.868          | 0.568           | <b>2.058</b>   | 0.868         | 0.746           | 0.774          |
| <b>Ours</b>  | <b>0.870</b>   | 0.191           | 1.973          | <b>0.914</b>  | 0.744           | <b>0.805</b>   |

**Table 1:** Evaluation results on the SALICON (LSUN 2017) test benchmark. We compare our model with the state-of-the-art saliency prediction models, namely DINET [60], DSCLRCN [39], SalNet [46], TempSAL [6], SAM-ResNet [15], SimpleNet [49], SALICON [25], DeepGaze IIE [38], RINet [54], MDSEM [22] and UNISAL [18]. The results in bold show the best performance. Our saliency prediction method with our augmentation method outperforms the SOTA methods on 3 out of 6 metrics.

SALICON test data [29]. Table 1 compares standard evaluation metrics for different state-of-the-art saliency models alongside our model. Our model outperforms the SOTA models on 3 out of 6 metrics. We present additional results on MIT1003 and CAT2000 datasets in the supplementary material.

**Data Augmentation** We evaluate the performance of our data augmentation method by choosing the best-performing baseline models and training on the SALICON [29] dataset. We report the saliency prediction performance on the SALICON validation set [29]. Table 2 compares standard evaluation metrics for different baseline saliency models alongside our method with and without our data augmentation technique. Our augmentation method consistently improves the saliency prediction performance of all the baseline models. We present additional results on MIT1003 and CAT2000 datasets in the supplementary material.

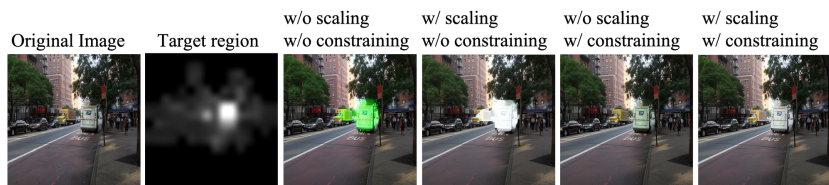
## 5.6 Ablation Studies

**Comparison with Classical Augmentation Methods** We evaluate our augmentation method against 12 classical augmentation methods namely, rotation, vertical flip, horizontal flip, cropping, JPEG compression, motion blur, inversion, noise, contrast, and shearing. We apply the same transformation to the image and its ground truth saliency map. We train all models on the SALICON dataset [29] and the augmented images with these methods. We observe that most of the augmentations decrease performance with the exception of horizontal flip. We provide a table of all parameters and results in the supplementary material.

**Effect of Scaling and Constraining the Image Edits** We show qualitative ablations of the effect of scaling and constraining the image edits in Figure 6. In

| Model        | w/o Augmentation |              |              |              | w/ Augmentation |              |              |              |
|--------------|------------------|--------------|--------------|--------------|-----------------|--------------|--------------|--------------|
|              | KL ↓             | NSS ↑        | CC ↑         | SIM ↑        | KL ↓            | NSS ↑        | CC ↑         | SIM ↑        |
| SimpleNet    | 0.193            | 1.926        | 0.907        | 0.797        | 0.185           | 1.937        | 0.911        | 0.805        |
| TempSAL      | 0.198            | 1.930        | 0.906        | <b>0.798</b> | 0.181           | 1.949        | 0.911        | 0.802        |
| DeepGaze IIE | 0.314            | <b>1.954</b> | 0.872        | 0.733        | 0.258           | <b>1.972</b> | 0.883        | 0.749        |
| UNISAL       | 0.226            | 1.923        | 0.880        | 0.771        | 0.212           | 1.930        | 0.894        | 0.797        |
| <b>Ours</b>  | <b>0.191</b>     | 1.927        | <b>0.908</b> | 0.788        | <b>0.179</b>    | 1.946        | <b>0.915</b> | <b>0.807</b> |

**Table 2:** Evaluation results on the SALICON validation set [29]. We select four existing saliency prediction models as baselines, namely SimpleNet [49], DeepGaze IIE [38], TempSAL [6] and UNISAL [18]. We train all models with the SALICON [29] dataset with and without our augmentation method. Our augmentation method consistently improves the saliency prediction performance of all models, in all metrics. Additionally, our saliency prediction method with data augmentation outperforms the other methods in 3 out of 4 metrics. The results in bold show the best performance.



**Fig. 6:** Scaling ensures that edits within the latent space are equivalent to those in RGB space and constraining the edits prevents the generation of excessively strong edits. For instance, the green coloration in the edited regions indicates the absence of scaling, and without proper constraints, edits can become excessively strong, leading to unnatural results.

the absence of scaling, and without proper constraints, edits can become excessively strong, leading to unnatural results.

We provide additional qualitative and quantitative results, ablation on the augmentation methods, and losses in the supplementary material.

## 6 Conclusion

We have introduced a novel data augmentation approach for deep saliency prediction, addressing the challenge of limited labeled data diversity and quantity. We perform targeted edits on photometric properties such as contrast, brightness, and color while preserving the complexity and variability of real-world visual scenes. Our experiments have concluded that our data augmentation method is suitable for saliency prediction. Moreover, leveraging these features that generate augmentation for saliency prediction yields a better understanding of visual attention patterns as shown by a user study. We hope that this work will contribute to advancing the field of saliency prediction through our data augmentation method, characterized by its ability to generate highly relevant and diverse training examples.

## Acknowledgement

This work was supported in part by the Swiss National Science Foundation via the Sinergia grant CRSII5-180359.

## References

1. Aberman, K., He, J., Gandelsman, Y., Mosseri, I., Jacobs, D., Kohlhoff, K., Pritch, Y., Rubinstein, M.: Deep saliency prior for reducing visual distraction. pp. 19819–19828 (06 2022). <https://doi.org/10.1109/CVPR52688.2022.01923>
2. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: International Conference on Computer Vision Systems (ICVS). pp. 66–75. Springer (2008)
3. Alldieck, T., Kolotouros, N., Sminchisescu, C.: Score distillation sampling with learned manifold corrective (2024)
4. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: SegDiff: Image segmentation with diffusion probabilistic models (2022)
5. Aydemir, B., Bhattacharjee, D., Kim, S., Zhang, T., Salzmann, M., Süsstrunk, S.: Modeling object dissimilarity for deep saliency prediction. Transactions on Machine Learning Research (TMLR) (2022), <https://arxiv.org/abs/2104.03864>
6. Aydemir, B., Hoffstetter, L., Zhang, T., Salzmann, M., Süsstrunk, S.: TempSAL - uncovering temporal information for deep saliency prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
7. Bauer, B., Jolicoeur, P., Cowan, W.B.: Visual search for colour targets that are or are not linearly separable from distractors. Vision Research **36**(10), 1439–1466 (1996). [https://doi.org/https://doi.org/10.1016/0042-6989\(95\)00207-3](https://doi.org/https://doi.org/10.1016/0042-6989(95)00207-3), <https://www.sciencedirect.com/science/article/pii/0042698995002073>
8. Berga, D., Fdez-Vidal, X.R., Otazu, X., Pardo, X.M.: SID4VAM: A benchmark dataset with synthetic images for visual attention modeling. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 8788–8797 (2019), <https://api.semanticscholar.org/CorpusID:204949994>
9. Borji, A., Itti, L.: CAT2000: A large scale fixation dataset for boosting saliency research. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2015)
10. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. CVPR (2023)
11. Chang, K.Y., Liu, T.L., Chen, H.T., Lai, S.H.: Fusing generic objectness and visual saliency for salient object detection. In: Proceedings of the 2011 International Conference on Computer Vision. p. 914. ICCV '11, IEEE Computer Society, USA (2011). <https://doi.org/10.1109/ICCV.2011.6126333>, <https://doi.org/10.1109/ICCV.2011.6126333>
12. Che, Z., Borji, A., Zhai, G., Min, X., Guo, G., Callet, P.L.: How is gaze influenced by image transformations? dataset and model. IEEE Transactions on Image Processing **29**, 2287–2300 (2019), <https://api.semanticscholar.org/CorpusID:204512657>
13. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. arXiv preprint arXiv:2301.13826 (2023)

14. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: IEEE International Conference on Pattern Recognition (ICPR). pp. 3488–3493 (2016). <https://doi.org/10.1109/icpr.2016.7900174>
15. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Transactions on Image Processing (TIP) **27**(10), 5142–5154 (2018). <https://doi.org/10.1109/tip.2018.2851672>
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
17. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), <https://openreview.net/forum?id=AAWuCvzaVt>
18. Droste, R., Jiao, J., Noble, J.A.: Unified Image and Video Saliency Modeling. In: European Conference on Computer Vision (ECCV) (2020)
19. Duan, Y., Guo, X., Zhu, Z.: Diffusiondepth: Diffusion denoising approach for monocular depth estimation (2023)
20. Einhäuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision **8**(14), 18–18 (2008). <https://doi.org/10.1167/8.14.18>, <https://doi.org/10.1167/8.14.18>
21. Elfving, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks **107**, 3–11 (2018). <https://doi.org/https://doi.org/10.1016/j.neunet.2017.12.012>, <https://www.sciencedirect.com/science/article/pii/S0893608017302976>, special issue on deep reinforcement learning
22. 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)
23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
24. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)
25. 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). <https://doi.org/10.1109/iccv.2015.38>
26. Jetley, S., Murray, N., Vig, E.: End-to-end saliency mapping via probability distribution prediction. CoRR **abs/1804.01793** (2018), <http://arxiv.org/abs/1804.01793>
27. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: DDP: Diffusion model for dense visual prediction. arXiv preprint arXiv:2303.17559 (2023)
28. Jia, S., Bruce, N.D.B.: EML-NET: An expandable Multi-Layer NETwork for saliency prediction. Image and Vision Computing **95**, 103887 (2020). <https://doi.org/10.1016/j.imavis.2020.103887>, <http://arxiv.org/abs/1805.01047>
29. Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: Saliency in context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). <https://doi.org/10.1109/cvpr.2015.7298710>



30. Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., Hügli, H.: Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding* **100**(1-2), 107–123 (2005). <https://doi.org/10.1016/j.cviu.2004.10.009>, <http://www.sciencedirect.com/science/article/pii/S107731420500041X>
31. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE (2009). <https://doi.org/10.1109/iccv.2009.5459462>
32. return: Decoding latents to RGB without upscaling. <https://discuss.huggingface.co/t/decoding-latents-to-rgb-without-upscaling/23204/2> (2022), accessed: 2023-03-03
33. Kruthiventi, S.S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing* **26**(9), 4446–4456 (2017). <https://doi.org/10.1109/TIP.2017.2710620>
34. Kümmerer, M., Theis, L., Bethge, M.: Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. In: *International Conference on Learning Representations (ICLR) Workshops* (2015)
35. Kümmerer, M., Wallis, T., Bethge, M.: DeepGaze II: Predicting fixations from deep features over time and tasks. *Journal of Vision (JOV)* **17**(10), 1147 (2017). <https://doi.org/10.1167/17.10.1147>, <http://arxiv.org/abs/1610.01563>
36. Li, Y., Zhang, H., Jia, W., Yuan, D., Cheng, F., Jia, R., Li, L., Sun, M.: Saliency guided naturalness enhancement in color images. *Optik* **127**(3), 1326–1334 (2016). <https://doi.org/10.1016/j.ijleo.2015.07.177>
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: *European Conference on Computer Vision (ECCV)*. pp. 740–755. Springer (2014)
38. 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)
39. Liu, N., Han, J.: A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing (TIP)* **27**(7), 3264–3274 (2018). <https://doi.org/10.1109/tip.2018.2817047>
40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019)
41. Lou, J., Lin, H., Marshall, D., Saupe, D., Liu, H.: TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing* (2022). <https://doi.org/https://doi.org/10.1016/j.neucom.2022.04.080>
42. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In: *Advances in Neural Information Processing Systems* (2023)
43. Miangoleh, S.H., Bylinskii, Z., Kee, E., Shechtman, E., Aksoy, Y.: Realistic saliency guided image enhancement. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 186–194. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). <https://doi.org/10.1109/CVPR52729.2023.00026>, <https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00026>
44. Nagy, A.L., Sanchez, R.R.: Critical color differences determined with a visual search task. *Journal of the Optical Society of America. A, Optics and image science* **7**, 1209–17 (1990), <https://api.semanticscholar.org/CorpusID:32540523>
45. Ochiai, N., Sato, M.: Effects of surrounding brightness on visual search for safety colors. *Color Research & Application* **30**(6), 400–409 (2005). <https://doi.org/https://doi.org/10.1002/col.20152>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/col.20152>

46. Pan, J., Sayrol, E., I-Nieto, X., McGuinness, K., OConnor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (2016). <https://doi.org/10.1109/cvpr.2016.71>
47. Pashler, H., Dobkins, K.R., Huang, L.: Is contrast just another feature for visual selective attention? *Vision Research* **44**, 1403–10 (2004)
48. Patel, Y., Appalaraju, S., Manmatha, R.: Saliency driven perceptual image compression. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 227–236 (2021). <https://doi.org/10.1109/WACV48630.2021.00027>
49. Reddy, N., Jain, S., Yarlagadda, P., Gandhi, V.: Tidying deep saliency prediction architectures. In: International Conference on Intelligent Robots and Systems (IROS) (2020), <https://arxiv.org/abs/2003.04942>
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
51. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models (2023)
52. Schneider, F.: Archisound: Audio generation with diffusion (2023)
53. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf>
54. Song, Y., Liu, Z., Li, G., Zeng, D., Zhang, T., Xu, L., Wang, J.: RINet: Relative importance-aware network for fixation prediction. *IEEE Transactions on Multimedia* **25**, 9263 (July 2023), senior Member, IEEE for Zhi Liu and Dan Zeng
55. Sun, C., Shrivastava, A., Singh, S., Gupta, A.K.: Revisiting unreasonable effectiveness of data in deep learning era. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 843–852 (2017), <https://api.semanticscholar.org/CorpusID:6842201>
56. Tan, H.H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Conference on Empirical Methods in Natural Language Processing (2019), <https://api.semanticscholar.org/CorpusID:201103729>
57. Tan, W., Chen, S., Yan, B.: DiffSS: Diffusion model for few-shot semantic segmentation (2023)
58. 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). <https://doi.org/10.1109/cvpr.2014.358>
59. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience* **5**(June), 1–7 (2004)
60. Yang, S., Lin, G., Jiang, Q., Lin, W.: A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia* **22**(8), 2163–2176 (2020). <https://doi.org/10.1109/tmm.2019.2947352>
61. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
62. Zhang, Y., Jiang, M., Zhao, Q.: Saliency prediction with external knowledge. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 484–493 (2021)
63. Zhu, Y., Zhao, Y.: Diffusion models in NLP: A survey (2023)