Rethinking 360° Image Visual Attention Modelling with Unsupervised Learning.

Yasser Abdelaziz Dahou Djilali*, Tarun Krishna*, Kevin McGuinness and Noel E. O'Connor
Insight Centre for Data Analytics, Dublin City University (DCU)
{yasser.dahoudjilali2, tarun.krishna2}@mail.dcu.ie

Abstract

Despite the success of self-supervised representation learning on planar data, to date it has not been studied on 360° images. In this paper, we extend recent advances in contrastive learning to learn latent representations that are sufficiently invariant to be highly effective for spherical saliency prediction as a downstream task. We argue that omni-directional images are particularly suited to such an approach due to the geometry of the data domain. To verify this hypothesis, we design an unsupervised framework that effectively maximizes the mutual information between the different views from both the equator and the poles. We show that the decoder is able to learn good quality saliency distributions from the encoder embeddings. Our model compares favorably with fully-supervised learning methods on the Salient360!, VR-EyeTracking and Sitzman datasets. This performance is achieved using an encoder that is trained in a completely unsupervised way and a relatively lightweight supervised decoder (3.8 × fewer parameters in the case of the ResNet50 encoder). We believe that this combination of supervised and unsupervised learning is an important step toward flexible formulations of human visual attention. The results can be reproduced on GitHub

1. Introduction

Unlike traditional media, omni-directional images (ODIs) provide users with the ability to explore different regions of the viewing sphere. The average person's head movements (HM) are typically a good prediction of the most probable viewport localized within the sphere, while eye movements (EM) reflect regions-of-interest (RoIs) inside the predicted viewports. Thus, when predicting the most salient pixels for 360° images, it is necessary to predict both HM and EM [68]. Despite remarkable advances in the field of visual attention [34, 6, 68], existing approaches for 360° saliency prediction are still limited in scope/power for two main reasons.

First, all previous state-of-the-art 360° saliency static approaches are trained end-to-end in a supervised manner. This limits their capacity to leverage unlabelled data. Compared to the large-scale 2D video/image saliency datasets [6] (i.e., up to 10000 images / 1000 video sequences), 360° video/image HM/EM datasets are rather small. This is due to the complex annotation process, which limits the capacity of the fully supervised approaches. Therefore, exploiting unlabelled data for learning better features is critical, and intuitively a good design to follow. Second, most previous approaches apply a CNN on each patch/cube resulting from the equi-rectangular (ERP) and cube map (CMP) projections. The former suffer from geometric distortions near the poles, whereas the latter stretch the salient regions into different cube faces, forcing the model to lose the global contextual information. These methods are also of high computational complexity, which may limit their applicability.

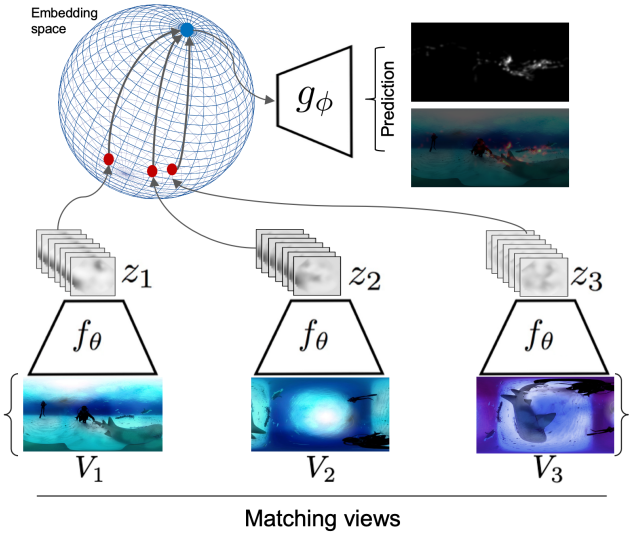

Figure 1. Given a set of 360° images and associated projections, a deep representation is learnt by maximizing the mutual information between views of the same scene in the embedding space, while discarding views of different scenes.

Modeling visual attention in ODIs using a representation learning function (an encoder) has the core objective of discovering useful representations conditioned by the spherical

*Equal contribution.

domain definition of the input. Using a simple convolutional encoder applied on non-euclidean data is often insufficient for learning good quality representations. Indeed the filters will produce a weak response to the signal associated with distorted regions, mostly in the poles (i.e. the Zenith and Nadir), decreasing the prediction ability. We leverage the mutual information (MI) maximization approach and show that maximizing average MI between the representation and local regions of the input (e.g. projections related to the poles) improves the expressive power of the encoding function for the downstream task of saliency prediction.

A powerful recent paradigm for estimating MI is contrastive learning based on noise contrastive estimation (NCE) [28], where multiple views of the same scene are brought together in embedding space, while pushing apart views from different scenes. Additionally, as the choice of the views is important for contrastive learning, 360° data offers a new set of choices for more effective MI estimation. The projections used in the scope of this work are task-relevant, but also, make the optimization problem harder, since they are not as susceptible to optimization short-cuts [22] as simple augmentations like color jitter and horizontal flips. This improves the expressive power of the encoder. This also motivates us to discard the use of CMP at training/inference, as we argue that the encoder is inherently sensitive to the signal coming from the Zenith and Nadir regions.

Our goal in this paper is to learn representations that capture signal shared between a support image V_1 and its corresponding projections V_2 and V_3 as shown in Figure 1. This is achieved by maximizing the agreement between global and local representations of support images and its projections respectively. The approach is inspired from the notion of mutual information (MI) maximization as proposed in Deep Info Max (DIM) [32] and Augmented Multi-scale DIM [3]; however, we introduce some important differences. First, we add self-attention to induce a soft feature selection mechanism over local representations (i.e. intermediate activation maps). Second, we formulate the total loss (Section 3.2) in a way to induce invariance to projections as in [49] and maximize the MI across different augmented (projected) views. Finally, unlike (AMDIM), instead of relying on batch sizes for negative samples, a memory bank is adopted for computational efficiency. Our contributions are as follows:

- We propose a framework to extend the idea of contrastive/self-supervised learning to a new data domain, specifically 360° images, and show how it can be effectively used for a regression downstream task rather than a simple recognition task.
- Through extensive evaluation as shown in Table 1, we show that contrastive learning can be exploited for saliency prediction, and furthermore that it performs on par with fully supervised methods.
- Our approach addresses one of the key challenges encountered when predicting 360° saliency by excluding any use of CMP. The design implicitly embeds the geometric specifications in the model weights.
- A single subsequent stream of learning on the equirectangular projection (ERM) images significantly reduces the computational cost (8× faster than the most efficient model among other 360° saliency approaches).

2. Related Work

This section reviews important works related to attention modelling for 360° images, and contrastive learning in general. For the former, we focus on works related to the prediction of the HM/EM saliency maps in 360° images, which can be grouped into heuristic and data-driven approaches.

Visual attention modelling for ODIs. The authors of [21] introduced the fused saliency map (FSM) approach for HM saliency prediction to ODIs, where the input 360° image is rotated by several angles and then projected as a set of 2D patches using the ERP. SALICON [36] (a SoTA 2D image saliency prediction model) is applied to each patch separately and the FSM approach fuses local saliency maps to generate the final prediction. The motivation of the approach introduced in [47] is to reduce the border artifact after sphere-to-plane projection. The authors applied a 2D saliency model on two CMPs:

$$w_{\text{face}}(i_{\text{face}}, j_{\text{face}}) = \frac{1}{1 + \left(\frac{\max\{i_{\text{face}}, j_{\text{face}}\}^2}{0.3 L_{\text{face}}}\right)^{10}}, \qquad (1)$$

where $(i_{\text{face}}, j_{\text{face}})$ is the pixel coordinate representing the origin at the center of the cube face and L_{face} is the width of the cube face. The final prediction is a weighted average of the saliency maps produced from each cube.

Unlike previous approaches, [59] combined both the ERP and CMP, for better reducing the negative impact of border artifacts. The former swaps the left and right halves of the image to reduce the distortions on the vertical sides. 2D saliency prediction approaches are applied to obtain two saliency maps, corresponding to the top and bottom faces of the cube, after incorporating CMP. The final saliency map is obtained by pixel-wise maximum multiplication as a method for fusing the two ERP and CMP generated maps.

Other approaches [39, 44, 19], adjusted predictions on the extracted view-ports rather than ERP/CMP projections, assuming that view-ports feature fewer geometric distortions. The main challenge is how to project several view-ports back into the final spherical saliency map. Rather than adapting 2D saliency prediction approaches on ODIs, some works [1, 61, 74, 26, 4] proposed the extraction of handcrafted low-level features such as hue, saturation, luminance, texture, color channels, boundary connectivity, but also high-level

features such as skin, faces, and cars. The low- and high-level maps are integrated to obtain the final saliency map.

Few end-to-end learning models have been proposed for 360° saliency. The SaltiNet [2] model is initialized with the pre-trained parameters of SalNet [54], and then trained over the Salient360! dataset using the binary cross entropy (BCE) loss. The SalNet360 [50] approach trained SalNet on the cube faces of a 360° image under CMP. Then, a fully convolutional network (FCN) is adapted to fuse the spherical coordinates of the cube faces with the extracted saliency maps. The work in [12] proposed rotating the 360° image at different CMPs with several angles, then SalGAN [53] is fine-tuned on the Salient360! dataset using these projections. Unlike previous DNN approaches, [60] explicitly learns the equator bias with a layer in the proposed CNN architecture, which acts on the viewports for generating the final saliency map of the 360° image. ATSAL [20] combines a latent attention mechanism that allows the network to focus on the most relevant parts of the input space, with expert instances of SalEMA [42] for each patch location produced by the CMP, to learn effective features for saliency modelling.

It is clear from the above review that the models targeting HM/EM 360° image visual attention modeling share the same core concept of applying a CNN on patches from ERP/CMP projections. As outlined above, this design is conceptually limited, and is computationally demanding at the inference stage. The contributions of this paper attempt to better address these limitations.

Contrastive learning has become prominent in recent years due to its ability to exploit large-scale unlabelled data. Contrastive learning [38, 35] refers to learning by comparison where the final objective is based on some variations of Noise Contrastive Estimation [28, 29]. Based on this idea of contrastive similarity, the authors in [17] learn a face embedding for a facial verification task, which was then referred to as *pair* loss as it required distance between negative pairs to be larger than a fixed margin (m). The authors in [14, 65] proposed the *triplet* loss to further fuse similarity and dissimilarity among positive and negative pairs forming a triplet. Later, Exemplar CNNS [24] introduced surrogate labels derived through heavy augmentation (distortion) where the pretext task was to discriminate between a set of surrogate labels i.e. enforcing invariance to specified transformations. Similarly, NPID [66] formulated an instance classification task via discriminative learning using *non-parametric softmax* to encode instance similarity while increasing the number of contrastive *negative* samples by introducing memory banks. Later CPC [52] showed that minimizing an NCE objective is equivalent to maximizing mutual information (MI), which they termed InfoNCE. CMC [62] builds upon this notion of MI maximization and extends it to an arbitrary collection of views. Independently, DIM [32, 3] formulated contrastive learning as a MI maximization problem between local and global representations. Taking cues from previous approaches, authors in [15] proposed a much simpler framework (SimCLR), which relies on maximizing the agreement between augmented views of the same data. To reduce the reliance on offline-representations (memory banks), MoCo [30, 16] looked at contrastive learning as a dynamic dictionary with a queue and a moving-averaged encoder (for offline representation). In summary, all these models try to enforce invariance to geometric distortions (augmentation) through instance classification (or maximizing agreement) and, in doing so, they exploit semantic (context) similarity and spatial structure among different variations of data samples to learn better representations.

There has been tremendous progress in this direction of unsupervised learning [11, 27, 10, 49, 5] in recent times. However, the scope of most of these methods has been limited to recognition as a downstream task (refer to [38, 35] for an extensive review). In this work, we take a step forward and extend the approach of MI estimation to regression, specifically to the task of saliency prediction, which is more fine-grained than recognition. Given that this domain of 360^o data conceptually provides new sets of choices i.e. signals for Zenith and Nadir regions, makes it particularly suited for contrastive learning (for MI maximization).

3. Method

Our algorithm takes advantage of the geometric flexibility of the 360° data definition domain i.e. the spherical representation, where the different projections represent robust views for training a differentiable parametric function $f_\theta : \mathbf{x} \mapsto \mathbf{\Lambda}$, with parameters θ (e.g. neural network) to maximize the mutual information among the views without any further supervision. The encoder is optimized to detect the polar regions, i.e. views, pushing the convolution filters to exploit larger groups of symmetries, including spherical transformations and rotations, because the translation symmetry preserved by a CNN is not enough to detect the distorted objects in the polar regions. We argue that the only information shared between the views is task-relevant, and there is no irrelevant noise, as the three views can fully reconstruct the sphere. Furthermore, we rely on exploiting contrastive learning-based approaches [38] to learn optimal and robust representations for 360° data. To further measure the quality of the latent representations, a separate parametric function $g_\phi : \mathbf{\Lambda} \mapsto \mathbf{y}$ (decoder), is able to decode good quality saliency maps for the downstream task. It is worth mentioning that the two stages are asynchronous.

3.1. Overview of the Approach

Suppose we are given a 360° image dataset, $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_{|\mathcal{D}|}\}$ where $\mathbf{x}_i \in \mathcal{R}^{3\times H\times W}$, and a set of transformations $\mathcal{T}$ and projections $\mathcal{P}$, with empirical probability distribution $p(\mathbf{X})$. The set $\mathcal{T}$ contains standard transfor-

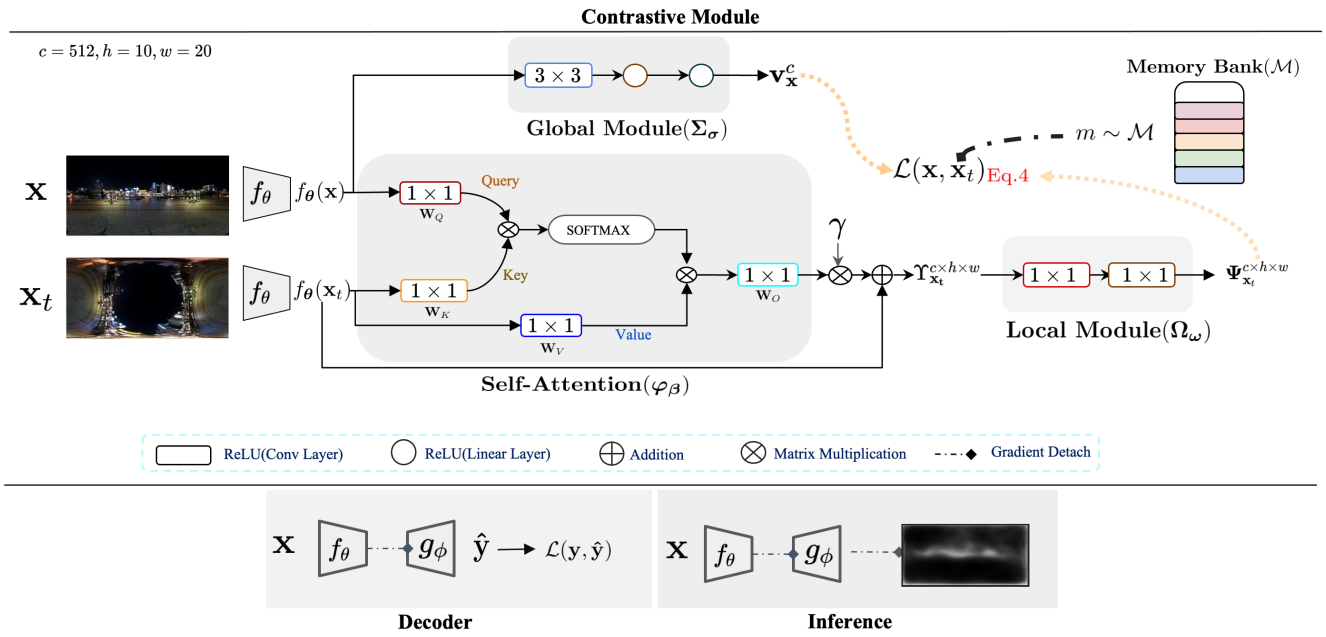

Figure 2. Complete pipeline for training. The **contrastive module** is composed of sub-functions which consist of an encoder f_θ, a global module Σ_σ, self-attention φ_β, and local module Ω_ω parameterised by $\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\beta}, \boldsymbol{\omega}$ respectively. **Decoder** g_ϕ is trained to optimize $\boldsymbol{\phi}$, keeping the encoder fixed (no gradient flow). **Inference** can be performed to predict the saliency on unseen test data. Different **colored** frames ([Conv2D $\rightarrow$ ReLU]) and circles ([Linear $\rightarrow$ ReLU]) represents different weight instances i.e. they don't share weights.

mations and specifically small random crops (<5% of the image size; large crops can affect saliency [13]), random jitter in color space, random conversion to gray-scale, random horizontal flips. The set $\mathcal{P}$ specifically contains projections top-to-front (*tf*) and bottom-to-front (*bf*), from the sphere-to-plane, using the ERP projection. The aim is to learn representations that maximize the agreement between global representations of $\mathbf{x}$ (source view) and local spatial patches of $\mathbf{x}_t \sim \mathcal{T}(\mathcal{P}(\mathbf{x}) \sim \mathcal{U}(tf(\mathbf{x}), bf(\mathbf{x})))$ (augmented view) as in [32, 3]. However, there are some significant differences that distinguish our approach from previous works. As we are mainly inferring for a regression downstream task, we are not concerned with the exact value of the MI, as minimizing further the contrastive objective encourages clusters to form in the representation space. Thus, we aim at optimizing the feature maps across spatial locations to capture enough symmetries about the input data, with the use of both the local to global approach and the self-attention module.

3.2. Unsupervised Contrastive Module

Base encoder ($\mathbf{f}_{\boldsymbol{\theta}}$). Learns a network $\mathbf{f} : \mathbf{x} \mapsto \boldsymbol{\Lambda}$ parameterised by $\boldsymbol{\theta}$, where $\mathbf{x} \in \mathbb{R}^{3\times160\times320}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{512\times10\times20}$. To be precise $\mathbf{x}$[1] is the whole panorama and $\mathbf{x}_t$ is perspective image with augmentations as depicted in Figure 2 with $f_\theta(\mathbf{x})$ and $f_\theta(\mathbf{x}_t)$ representing their local latent representations[2]. We report findings for VGG16 [56] and ResNet50 [31] as encoders. For complete architectural details see Section A.1 and A.2 in supplementary.

Global module ($\boldsymbol{\Sigma}_{\boldsymbol{\sigma}}$). Learns a mapping $\boldsymbol{\Sigma} : f_\theta(\mathbf{x}) \mapsto \mathbf{v}_{\mathbf{x}}$[3] parameterised by $\boldsymbol{\sigma}$, where $\mathbf{v}_{\mathbf{x}} \in \mathbb{R}^{512}$. This module provides a compact/global representation of $\mathbf{x}$ as shown in Figure 2[4]. This module can also be understood as a projection layer often used in self-supervised literature but in this case it is asymmetric[5]. More details can be found in Section A.1 and A.2 in supplementary.

Self-attention module ($\boldsymbol{\varphi}_{\boldsymbol{\beta}}$). This serves as a medium to build spatial relationship between local representations $f_\theta(\mathbf{x})$ and $f_\theta(\mathbf{x}_t)$. Architecture is similar to [64, 72] but unlike *query* and *key* are derived from $f_\theta(\mathbf{x})$ and $f_\theta(\mathbf{x}_t)$ respectively[6]. Refer Section A.3 in supplementary for more intuition and details.

Local module ($\boldsymbol{\Omega}_{\boldsymbol{\omega}}$) is again a non-linear mapping $\boldsymbol{\Omega} : \boldsymbol{\Upsilon}_{\mathbf{x}_t} \mapsto \Psi_{\mathbf{x}_t}$ parameterised by $\boldsymbol{\omega}$, where $\Psi_{\mathbf{x}_t} \in \mathbb{R}^{512\times\times10\times20}$. The architecture of the local module consist of $2\times$[Conv2d $\rightarrow$ ReLU] followed by a BatchNorm2D and is fixed for both VGG16 and ResNet50.

Loss function. We minimize a NCE [28] based objective

[1] Source view
[2] $f_\theta(\mathbf{x}_t), f_\theta(\mathbf{x}) \in \boldsymbol{\Lambda}$
[3] $\mathbf{x}$ always represents a panorama image
[4] In Figure 2 global module corresponds to VGG16 encoder
[5] i.e not applied to $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_t)$
[6] Usually *key* and *query* are linear projections of same representations.

as in [62]:

$$\mathcal{L}_{\text{NCE}}(\mathbf{x}, \mathbf{x}_t) = -\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\Big\{\mathop{\mathbb{E}}_{\mathbf{x}_t\sim p(.|\mathbf{x})}\big[\log\big(P(D=1|\mathbf{x}_t,\mathbf{x})\big)\big] + m\mathop{\mathbb{E}}_{\mathbf{x}_n\sim p_n(.|\mathbf{x})}\big[\log(P(D=0|\mathbf{x}_n,\mathbf{x}))\big]\Big\}. \tag{2}$$

Optimizing $\mathcal{L}_{\text{NCE}}(\mathbf{x}, \mathbf{x}_t)$ is simply minimizing the negative log-posterior probability of label D to distinguish "positive pair" $(\mathbf{x}, \mathbf{x}_t)$ $(D = 1)$ from "negative pair" $(\mathbf{x}, \mathbf{x}_n)$ $(D = 0)$ where $\mathbf{x}_n$ is often referred to as a negative sample. A negative sample is any sample that is not typically derived from $\mathbf{x}$ and its distortion/augmentation. $p(\mathbf{x})$ and $p_n(.)$ in Equation 2 is the empirical data distribution and distribution of noisy samples respectively. The posterior distribution with *m* noise sample is given by:

$$P(D=1|\mathbf{x}_t,\mathbf{x}) = \frac{p(\mathbf{x}_t|\mathbf{x})}{p(\mathbf{x}_t|\mathbf{x}) + m\, p_n(\mathbf{x}_n|\mathbf{x})}, \tag{3}$$

with $p(\mathbf{x}_t|\mathbf{x})$ being the true unknown distribution, which is approximated by a score function $s(\mathbf{x}_t, \mathbf{x}) = \exp\left(\mathbf{x}_t^T\mathbf{x}/\tau\right)$, where τ is the temperature hyper-parameter (fixed to 0.07) that modulates the distribution. This function assumes L_2 normalized vectors. Refer to [62, 29] for further details on the derivation of the NCE loss.

Memory bank $(\mathcal{M})$**.** Following [66, 49], we maintain a memory bank to retrieve $m \sim \mathcal{M}$ negative samples. These samples are exponential moving average of feature representations $\mathbf{v}_\mathbf{x}$ that were computed in prior epochs. A sample from memory bank corresponding to $\mathbf{v}_\mathbf{x}$ is denoted by $\mathbf{m}_\mathbf{x}$.

The **final objective** is defined as a convex combination of both the global($\mathcal{L}_G$) and local($\mathcal{L}_L$) NCE losses:

$$\mathcal{L}(\mathbf{x}, \mathbf{x}_t) = \lambda\mathcal{L}_L(\mathbf{m}_\mathbf{x}, \Psi_{\mathbf{x}_t}) + (1-\lambda)\mathcal{L}_G(\mathbf{m}_\mathbf{x}, \mathbf{v}_x). \tag{4}$$

Note we do not directly minimize the NCE between global and local representations but instead rely on representations from memory bank $\mathcal{M}$ $(\mathbf{m}_x)$. Firstly, this encourages similarity to memory representations encoding invariances as shown in [49], and secondly it directly maximizes the MI between global and local representations via memory representations.

$\mathcal{L}_G$ is the global NCE between two feature vectors $\mathbf{m}_x$ and $\mathbf{v}_\mathbf{x}$ (each $\in \mathbb{R}^{512}$), while $\mathcal{L}_L$ is the local NCE between a vector $\mathbf{m}_\mathbf{x}$ and feature map $\Psi_{\mathbf{x}_t} \in \mathbb{R}^{512\times10\times20}$. In this later case the dot product in $s(\mathbf{m}_x, \Psi_{\mathbf{x}_t})$ is calculated as $\frac{1}{hw}\sum_{i=0}^{h}\sum_{j=0}^{w}\mathbf{m}_x^T\Psi(:,i,j)$, which is referred to as *local-dot* encode in [32]. Recall that the dot-product in the scoring function assumes L_2 normalized vectors, so $\Psi_{\mathbf{x}_t}$ is L_2 normalized along each location i.e. across $\Psi(:,i,j)$. The dimensions $(c,h,w) = (512, 10, 20)$ stay fixed across all settings.

3.3. Supervised Module

Problem formulation. Visual attention modelling for ODIs is the downstream task chosen to measure the quality of the representations. The motivation lies with the difficulty of the task, and the availability of benchmarks. It consists of predicting an (head+eye) based ERP-saliency map from the input 360° image. In this setting, the ground truth saliency maps are computed by convolving each fixation or trajectory points (for all observers of one image), defined as:

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

with a Gaussian or Kent kernel. The resulting saliency map $P \in [0,1]^{W\times H}$ can be treated as a multivariate Bernoulli distribution where each pixel is Bernoulli distributed, with a probability p to be attended, and $(1-p)$ to be discarded.

Decoder module. Human attention is driven by both global and local features. In ODIs, the CMP forces the model to lose the global contextual information while considering each cube face separately. Through the contrastive encoder, more explicit global features are learned inherently as a super-position in the encoding function weights [23]. Thus, the convolution filters are more responsive to the signal connected to the poles. Therefore, we argue that the latent representations $\Lambda \in \mathbb{R}^{512\times10\times20}$ lie within a feasible manifold to be decoded into saliency maps. The decoder architecture is inspired from SalGAN; however, we only kept one single convolution layer per block, rather than three layers as in the original SalGAN. The main motivation for this is to avoid over-parametrization, and to show that a less complex function is able to decode the representations and provide evidence of the generality and robustness of Ψ.

Saliency loss function. The saliency task can be seen as a distance measure between the predicted saliency distribution $Y \in [0,1]^{W\times H}$, and the continuous ground truth $P \in [0,1]^{W\times H}$. The objective function must be designed to maximise the in-variance of predictive maps and give higher weights to locations with higher fixation probability. Thus, the decoder is trained to minimize the Kullback-Leibler Divergence (KLD), widely adopted for benchmarking saliency models [9], the KLD between Y and P is given by:

$$\mathcal{L}_{\text{KLD}}(Y,P) = \sum_{i=1}^{W\times H} P_i \log\left(\epsilon + \frac{P_i}{\epsilon + Y_i}\right), \tag{5}$$

4. Experimental Setup and Results

Training. We first train the encoder following the unsupervised scheme. Contrastive learning requires a large amount of unlabelled data to be trained effectively. Due to the unavailability of large-scale 360° images datasets, we had to gather a new one with 90K ODIs from multiple sources. The dataset comprises the following:

Table 1. Comparative performance study on: Salient360! and VR-EyeTracking datasets. Training setting (i): trained w/o self-attention, Training setting (ii): trained with self attention. The best scores are marked in **bold** and second best in blue.

Model		Salient360!					VR-EyeTracking+Sitzman				
		AUC-J ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓	AUC-J ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓
Baseline: infinite humans		0.788	2.09	1.0	1.0	0.0	0.985	3.421	1.0	1.0	0.0
2D models	UNISAL [25]	0.701	1.404	0.389	0.435	2.519	0.783	1.918	0.276	0.242	9.044
	SalGAN [53]	0.701	1.398	0.377	0.483	1.544	0.718	1.023	0.145	0.152	10.195
360° models	ATSAL [20]	0.777	**1.638**	**0.642**	**0.639**	**0.761**	0.822	1.613	0.239	0.191	9.796
	SalGAN360 [12]	**0.831**	1.598	0.639	0.611	0.798	0.704	1.267	0.226	0.218	7.938
	SaltiNet [2]	0.702	1.057	0.536	0.541	1.098	0.674	0.967	0.186	0.198	9.938
Training setting (i)	VGG	0.758	1.557	0.553	0.585	0.909	0.841	1.583	0.246	0.221	7.965
	ResNet	0.756	1.524	0.520	0.54	1.039	0.833	1.545	0.232	0.203	8.574
Training setting (ii)	VGG	0.760	1.548	0.538	0.569	0.922	0.867	1.880	0.308	0.234	7.583
	ResNet	0.769	1.601	0.584	0.591	0.849	**0.869**	**2.089**	**0.329**	**0.248**	**7.110**

- PVS: HMEM [69] contains 76 panoramic videos, images were sampled at a rate of 1 frame per second (fps).
- 360-Indoor [18] contains 3024 complex indoor scenes containing common objects.
- VR-VQA [67], a quality assessment dataset, comprises 48 ODVs from 8 classes: sport, movies, etc.
- Videos gathered from YouTube playlists[7] (1 fps).

Furthermore, evaluation of these representations relies on the downstream task. Often, classification accuracy is used as a proxy for correlation between representations and class labels. Clearly, as the symmetries and invariances encoded in the representation are abstract, a more granular regression proxy task such as saliency prediction reveals more insights. Thus, we evaluate the unsupervised module's representational properties on three saliency datasets:

- Salient360! images [55]: a small-scale dataset, consisting of (80/23) images for training and validation respectively, each recorded for at least 40 observers. It provides the head-eye saliency map obtained jointly from eye tracking and head positions in the ERP format.
- Due to the small amount of labelled static data (103 ODIs), we sampled at a rate of 1 fps from the large-scale video dataset VR-EyeTracking [70]. The resulting set contains (4700/1300) 360° images.
- Sitzmann [57] containing a total of (14/11) training/validation ODIs; the authors captured and analyzed gaze and head orientation data of 169 users.

It is worth mentioning that the unsupervised encoder weights were not fine-tuned on the downstream task; only the randomly initialized decoder is trained on top of the frozen encoder. The main motivation for this is to set a robust evaluation procedure and to prevent the encoder adapting its parameters to saliency specific requirements.

Our approach is experimentally compared to five models, two state-of-the-art 2D static saliency models, UNISAL [25] and SalGAN, and three 360° specific models: ATSAL [20], SalGAN360 [12] and SaltiNet [2]. This choice is motivated by the availability of the source code. All approaches were evaluated according to five different saliency metrics: Normalized Scanpath Saliency (NSS), Kullback-Leibler Divergence (KLD), Similarity (SIM), Linear Correlation Coefficient (CC), and AUC-Judd (AUC-J). Please refer to [9] for an extensive review of these metrics.

Technical details. Both the contrastive encoder and the supervised decoder were implemented in PyTorch, and trained using two GPUs (RTX 3090 & RTX 2080). The contrastive encoder was optimized using SGD with a learning rate of 10^{-2}. The encoder was trained for 250 epochs using the max batch size of 80[8], with negative samples fixed to 16000, $\tau = 0.07$ and $\lambda = 0.7$. Choices about total epochs and number of negative samples are based on computation and time constraints. Training for more epochs or with large negative samples may provide further boost in the performance [37, 49]. Adam with a learning rate of 10^{-4} was used to train the supervised decoder for 100 epochs.

Table 1 shows the comparative study with the aforementioned models according to the different saliency metrics on Salient360! and (VR-EyeTracking+Sitzman) datasets 25/1300 test 360° images. Our model is very competitive in the two datasets, and exhibits the top score for all metrics on VR-EyeTracking+Sitzman. As expected, 2D SoTA approaches fail to generalize on ODIs, which questions the effectiveness of the direct transfer of visual attention features from 2D to 360° data.

[7] Playlist 1, Playlist 2, Playlist 3

[8] maximum batch size that could be achieved given our constraints

Table 2. Comparative performance study on: VR-EyeTracking datasets. VGG_(i)/ VGG_(ii) following training setting (i)/(ii)

Model		VR-EyeTracking+Sitzman				
		AUC-J ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓
VGG_(i)	$\lambda = 0.5$	0.852	1.872	0.306	0.237	**7.540**
	$\lambda = 0.7$	0.841	1.583	0.246	0.221	7.965
	$\lambda = 0.9$	0.849	1.825	0.278	0.226	7.875
VGG_(ii)	$\lambda = 0.5$	0.860	**1.894**	0.307	0.231	7.648
	$\lambda = 0.7$	0.867	1.880	**0.308**	0.234	7.583
	$\lambda = 0.9$	0.860	**1.894**	0.301	**0.241**	7.648

VR-EyeTracking & Sitzman. The 1300 validation/test images were sampled from the 75 diverse test ODVs of the first dataset, mixed up with the 11 images from Sitzman, making the prediction task very challenging. It can be seen that both VGG and ResNet-based models outperform 2D saliency models by a good margin, with a significant improvement over 360° specialized models, trained in an end-to-end scheme on supervised data. The ResNet-based model trained with self-attention achieves the best results following: (KLD ↓ = 7.110).

Salient360! We used the 25 360° validation images for testing the model[9]. The model was not trained on this dataset, making this an out-of-distribution test. The proposed model produces a reasonable improvement in accuracy compared to other models, except SalGAN360 and ATSAL, which were trained on this specific dataset.

Figure 3 illustrates the prediction task on a sample of 360° images from two datasets: Salient360! and VR-EyeTracking. It can be seen that the saliency maps generated by our model (ResNet-based with self-attention) correlate well with the ground truth maps in terms of fixation distribution. Other competitors shown in the same figure overestimate saliency in general, or overly bias it to the equator/center. Furthermore, the effectiveness of the predictor in capturing the main objects in the scene can be observed. Another key point is the model's capacity to accurately detect saliency in the Zenith and Nadir without using any form of projections at inference time (see Figure 2,3 in supplementary). This demonstrates the effectiveness of the contrastive encoder in embedding the views as a superposition in the function weights and biases.

Computational load. As model efficiency is a key factor for real-world ODIs applications, Table 3 shows a GPU runtime comparison (processing time per 360° image) of the different competitors on the 4K Salient360! ODIs. Compared with other 360° specialized models, our model exhibits a remarkable improvement, being over $8\times$ faster than ATSAL, which is the fastest model in this category.

[9]The reserved test set was unavailable due to COVID-19

Table 3. GPU inference time comparison of video saliency prediction methods (NVIDIA RTX 3090). All methods are reported based on the Salient360! benchmark [70]. The best computational performance among dedicated 360° models is shown in bold. (*) 2D models.

Model	Runtime (s)
SalGAN360 [12]	14.330
ATSAL [20]	0.230
SaltiNet [2]	0.450
(*) SalEMA [25]	0.020
(*) UNISAL [53]	0.010
Ours (ResNet-based)	**0.025**

Table 4. Results on Salient360! validation images for a model based on a contrastive encoder trained with/without projections.

	Salient360!				
	AUC-J ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓
ResNet w/o	0.736	1.524	0.479	0.536	0.999
ResNet w/	**0.769**	**1.601**	**0.584**	**0.591**	**0.849**

5. Ablations

In this section we justify the choices by ablating key features of the procedure.

What is the effect of λ in the loss functions? The total loss is a convex combination with a hyper-parameter λ, that trade-offs between the two NCE losses namely global ($\mathcal{L}_{\mathrm{G}}$) and local ($\mathcal{L}_{\mathrm{L}}$) NCE. As depicted in Table 2 we varied λ to 0.5, 0.7, and 0.9. As the results suggests increasing λ improves the performance on the downstream task. Intuitively if we look closely in Equation 4, giving more emphasis to $\mathcal{L}_{\mathrm{G}}$ biases the function to learn trivial solution as $\mathbf{m_x}$ is moving average of $\mathbf{v_x}$, as result this leads to much easier classification task. However, more emphasis on $\mathcal{L}_{\mathrm{L}}$ makes the classification task (NCE optimization) more difficult because the optimizer is maximizing the agreement from different view of the same scene (object) and that too locally. This leads to better expressive power and generalisation capability. We chose $\lambda = 0.7$ for all our previous evaluations.

Training with/without self attention: We evaluate the models to observe the effect of self-attention in the learning regime. Table 2 compares the VGG-based model trained with and without self-attention. A performance boost is observed irrespective of λ when using self-attention. Table 1 also shows results for models trained with and without attention. Self-attention/attention is intuitively motivated by how we humans pay attention to specific regions or parts of images and try to correlate among them. At the same time, this correlation can be extended to different images/patches of same view. In summary, it helps to infer a patch/region in an image based on this correlation (importance vectors). Given that the objective is to maximize the agreement between two

Figure 3. Qualitative results of our model and four other competitors on sample images from VR-EyeTracking and Salient360! datasets. It can be observed that the proposed approach is able to handle various challenging scenes well and produces consistent saliency maps.

views, the self-attention module serves as a mode of finding the best correlation in the two views in terms of information shared among them by performing this soft feature selection mechanism across the views.

Training with/without projections: To further showcase the importance of projections, we trained a model without any projections (only augmentations). One could also argue training without augmentations and projections but this will lead us to NPID [66] ($\lambda = 0$) and this is not our objective. Table 4 depicts the results for this setting. Performance drops when projections are removed, validating the hypothesis that using different projections in addition to augmentation is natural for 360° images and produces representations that are more effective in downstream tasks. Intuitively by removing projections we make the feature extractor f_θ prone to exploiting low-level visual features such as color aberrations as observed in [22, 51], and not learning useful semantic representations, resulting in a performance drop on the downstream task. This phenomenon of relying on low-level features given large unlabelled datasets is often referred to as short-cuts in the unsupervised learning literature [48, 15]. Finally, this experiment further validates our hypotheses of exploiting the top/bottom to equator projections for contrastive learning.

6. Discussion

2D vs 360° models. Deep learning based saliency models [2, 54, 12, 60] trained end-to-end on 360° datasets show remarkable improvements over early models adapting 2D approaches on ODIs [21, 47, 59, 39, 19, 44]. This demonstrates the new constraints imposed by the spherical domain when modelling visual attention.

Supervised vs Unsupervised learning. To the best of our knowledge, we are the first to design an unsupervised learning approach for saliency prediction. Our approach is an opening for a new line of research exploring the subtle definition of gaze policy naturally embedded in the brain. In fact, the early research into how the human visual system functions, produced many interesting results, demonstrating that visual attention could be influenced by regions that maximize a reward in a task-driven scenario [58], which are typically the most informative regions [33, 8]. This suggests that saliency can be disentangled into low-level (e.g. color, intensity, etc.) and high-level (e.g. human faces) features [34, 7]. Research in cognitive science (e.g. [71, 41, 63]) indicates that low-level saliency in both humans and animals happens early in the primary visual cortex, suggesting that it can potentially be learned without supervision. We believe this could be an important future research direction.

Generalization to other 360° downstream tasks. The last decade has witnessed many works on 360° video/image processing including, visual attention, visual quality assessment (VQA), and compression [68]. Visual attention can serve as a tool for compression approaches (e.g. saliency-aware adaptive coding [73, 44, 43]). Perceptual approaches for VQA require predicted saliency maps as weight maps [45, 46, 40]. Generally, the spherical characteristics of the input data means that many of these tasks face the challenges addressed in this work. Thus, we suspect that the contrastive encoder can be indeed exploited for VQA and compression.

7. Conclusion

We introduced a method for modelling human visual attention with contrastive self-supervised learning, which improves the generalization and expressive power of the model. The approach exploits the geometric flexibility of the spherical data to learn representations that contain locally-consistent information across the views. The qualitative and quantitative results on the downstream saliency task have demonstrated the competitiveness of the approach. We believe this is an important step towards better human visual attention modelling using unsupervised methods.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. Tarun Krishna also acknowledges support from Xperi FotoNation Ltd.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Scanpath and saliency prediction on 360 degree images. *Signal Processing: Image Communication*, 69:8–14, 2018.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[4] Federica Battisti, Sara Baldoni, Michele Brizzi, and Marco Carli. A feature-based approach for saliency estimation of omni-directional images. *Signal Processing: Image Communication*, 69:53–59, 2018.

[5] Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Sutter, and Björn Ommer. Cliquecnn: Deep unsupervised exemplar learning. *arXiv preprint arXiv:1608.08792*, 2016.

[6] Ali Borji. Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*, 10, 2018.

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

[9] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[12] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 01–04. IEEE, 2018.

[13] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 29:2287–2300, 2019.

[14] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[17] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[18] Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan-Ting Hsu, Min Sun, and Jianlong Fu. 360-indoor: Towards learning real-world objects in 360deg indoor equirectangular images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 845–853, 2020.

[19] 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.

[20] 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.

[21] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. Look around you: Saliency maps for omnidirectional images in vr applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017.

[22] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[23] Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.

[24] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

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

[26] Yuming Fang, Xiaoqiang Zhang, and Nevrez Imamoglu. A novel superpixel-based saliency detection model for 360-degree images. *Signal Processing: Image Communication*, 69:1–7, 2018.

[27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[28] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[29] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(2), 2012.

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[31] 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.

[32] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[33] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[34] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[35] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021.

[36] 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.

[37] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

[38] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 2020.

[39] Pierre Lebreton and Alexander Raake. Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images. *Signal Processing: Image Communication*, 69:69–78, 2018.

[40] Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, and Xiaoming Tao. Viewport proposal cnn for 360° video quality assessment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10169–10178. IEEE, 2019.

[41] Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, 2002.

[42] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Noel E O'Connor, Xavier Giro-i Nieto, and Kevin McGuinness. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*, 2019.

[43] Yufan Liu, Li Yang, Mai Xu, and Zulin Wang. Rate control schemes for panoramic video coding. *Journal of Visual Communication and Image Representation*, 53:76–85, 2018.

[44] Guilherme Luz, João Ascenso, Catarina Brites, and Fernando Pereira. Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment. In *2017 IEEE 19th international workshop on multimedia signal processing (MMSP)*, pages 1–6. IEEE, 2017.

[45] Qi Ma and Liming Zhang. Image quality assessment with visual attention. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[46] Qi Ma and Liming Zhang. Saliency-based image quality assessment criterion. In *International Conference on Intelligent Computing*, pages 1124–1133. Springer, 2008.

[47] Thomas Maugey, Olivier Le Meur, and Zhi Liu. Saliency-based navigation in omnidirectional image. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2017.

[48] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pages 6927–6937. PMLR, 2020.

[49] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[50] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Processing: Image Communication*, 69:26–34, 2018.

[51] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[53] 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.

[54] 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.

[55] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 205–210, 2017.

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

[57] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018.

[58] Nathan Sprague and Dana Ballard. Eye movements for reward maximization. In *Advances in neural information processing systems*, pages 1467–1474, 2004.

[59] Mikhail Startsev and Michael Dorr. 360-aware saliency estimation with conventional image saliency predictors. *Signal Processing: Image Communication*, 69:43–52, 2018.

[60] Tatsuya Suzuki and Takao Yamanaka. Saliency map estimation for omni-directional image considering prior distributions. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2079–2084. IEEE, 2018.

[61] Sayantan Thakur, Sayantanu Paul, Ankur Mondal, Swagatam Das, and Ajith Abraham. Face detection using skin tone segmentation. In *2011 World Congress on Information and Communication Technologies*, pages 53–60. IEEE, 2011.

[62] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019.

[63] Richard Veale, Ziad M Hafed, and Masatoshi Yoshida. How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160113, 2017.

[64] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[65] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[66] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[67] Mai Xu, Chen Li, Yufan Liu, Xin Deng, and Jiaxin Lu. A subjective visual quality assessment method of panoramic videos. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 517–522. IEEE, 2017.

[68] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020.

[69] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2693–2708, 2018.

[70] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018.

[71] Yin Yan, Li Zhaoping, and Wu Li. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, 115(41):10499–10504, 2018.

[72] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[73] Chunbiao Zhu, Kan Huang, and Ge Li. An innovative saliency guided roi selection model for panoramic images compression. In *2018 Data Compression Conference*, pages 436–436. IEEE, 2018.

[74] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15–25, 2018.