360-aware Saliency Estimation with Conventional Image Saliency Predictors

Mikhail Startsev[a,*], Michael Dorr[a]

[a] *Technical University of Munich, Institute for Human-Machine Communication,*
Arcisstr. 21, Munich, Germany, 80333

Abstract

This work explores saliency prediction for panoramic 360°-scenes stored as equirectangular images, using exclusively regular "flat" image saliency predictors. The simple equirectangular projection causes severe distortions in the resulting image, which need to be compensated for sensible saliency prediction in all viewports. To address this and other arising issues, we propose several ways of interpreting equirectangular images and analyse how these affect the quality of the resulting saliency maps. We perform our experiments with three popular conventional saliency predictors and achieve excellent results on the "Salient360!" Grand Challenge data set (ranked 1st among the blind-test submissions in the Head-Eye Saliency Prediction track).

Keywords: Saliency prediction, Equirectangular projection, Panoramic images

1. Introduction

Even though we seemingly perceive our entire surrounding as a whole, this is impossible because of the physical constraints of our visual system. Only a small part of our visual field is projected onto a high-resolution part of the retina – the area called *fovea*. This foveation reduces the computational load on the visual cortex and bandwidth requirements on the optic nerve, but forces our eyes to constantly scan the scene to obtain the "full picture". This means that from such fragmented input our brain has to reconstruct a comprehensive model of what surrounds us. The strategy of visual exploration is therefore an important factor of human adaptation, which had both social and environmental factors impact its development.

Being able to predict or model the process of this "biologically-approved" attention allocation can aid various computer vision-related areas in the struggle for sparsity [1, 2], help action recognition [3, 4] and semantic segmentation [5], or even potentially shed light on and aid diagnosis of mental disorders [6, 7]. With 360°-content becoming more and more widespread on popular image- and video-sharing platforms, as well as with the rise of consumer-oriented virtual reality applications and 360-camera setups, the saliency models for such stimuli can facilitate its analysis and compression, for example in order to enhance user immersion.

Working with the panoramic image scenario is generally beneficial for understanding attention. First of all, whereas conventional 2D image saliency data sets are often recorded under restrictive laboratory conditions, the free head motion of 360°-recordings means this scenario is much closer to real-life viewing behaviour.

Just as regular image or video saliency, this scenario does not yet introduce the social aspects of attention, such as avoiding either prolonged eye contact with strangers [8] or even looking at people when they are close-by in a genuine social context altogether [9], or seeking out familiar faces in crowds. But the prioritisation of observers' attention has a different component to it, making it two levels deep: first the head rotation, and then the eye gaze direction.

Compared to fully-unconstrained complex recording scenarios, static 360°-stimuli allow us to analyse common objects and regions of interest for multiple observers without having to match the contents of the foveated patches with one another, or deal with depth perception or occlusions. This eases the transition from numerous readily available 2D image saliency predictors, which have much larger data sets that could be used for training and evaluation. This work explores the possibilities and needed image transformations to perform this very transition.

In this work we have, therefore, proposed a range of transformations of the input equirectangular images, which we call "interpretations", that allow us to predict 360° saliency using any existing 2D attention model. In our experiments, we used three publicly available saliency prediction algorithms that model different levels of the visual processing hierarchy. Our approach demonstrated excellent results on a data set of omnidirectional images without any training or parameter adjustments.

In contrast to the work in [10, 11], for example, which presents a CNN-based approach, where the network is fitted for the available set of the equirectangular images, and several strategies to prevent overfitting had to be applied

[*] Corresponding author
Email address: `mikhail.startsev@tum.de` (Mikhail Startsev)

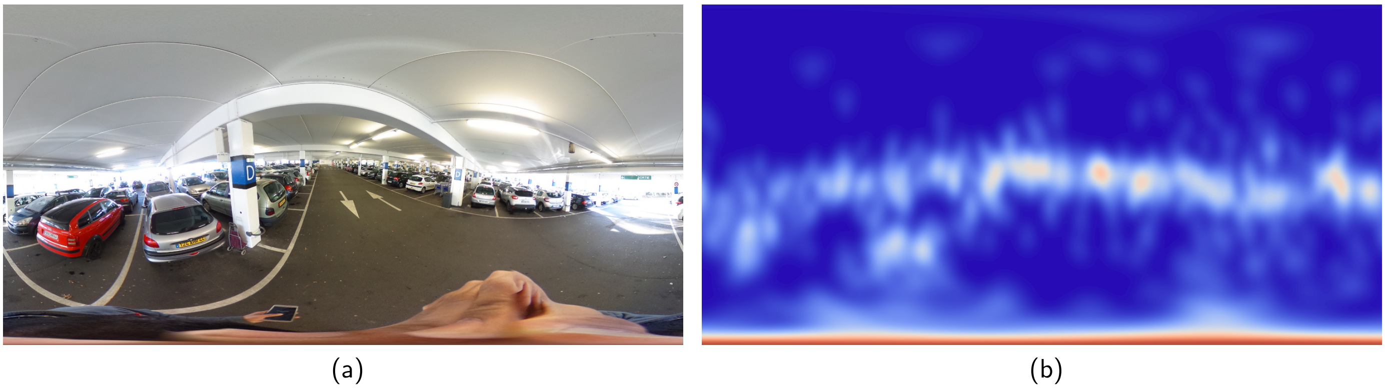


Figure 1: Equirectangular image example (1a) and its ground-truth saliency map (1b).

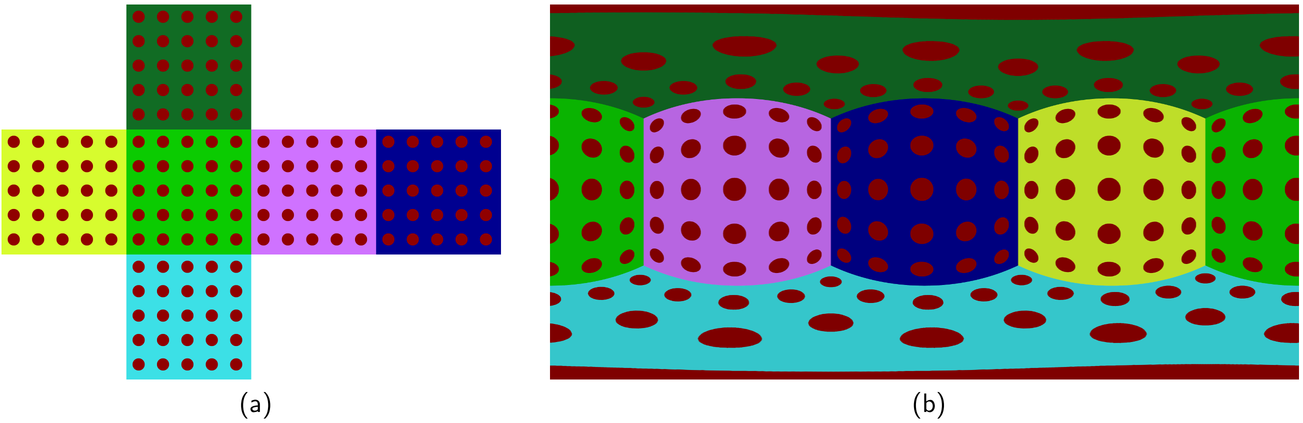


Figure 2: Distortion visualization for equirectangular projection (2b) and the set of corresponding cube map faces (2a). Note that the red bottom and top stripes in (2b) each represent just one disk on top and bottom faces of (2a).

as a consequence of the data set size, our approach does not require any additional training and can be used with any conventional pre-trained saliency model. In [12], an approach involving an idea similar to what we here call "interpretations" was applied for predicting salient viewports as a post-processing step for conventional saliency predictors' outputs, but it does not get rid of all the issues that arise for the eye gaze-based saliency prediction, originally only addressing the centre bias.

2. Proposed approach

Dealing with omnidirectional images is a challenge on its own, as the "perfect" way to store and process them is yet to be developed: So far, there is always a trade-off between efficiency, visual interpretability, and convenience of use. The data set that we use in this work (described in Section 3.1) employs equirectangular projection, so we first examine its artefacts, and then describe how they can be mitigated for a better saliency prediction via proposed interpretations.

2.1. Motivation

Aside from the obvious unnatural visual stretching of the objects at the top and the bottom of any equirectangular image (for an example, see Figure 1a), there are several issues that are particularly prominent when such an image is being processed automatically, for instance by attention predictors (for an example of an empirical ground truth saliency map, see Figure 1b). In [13], a similar data set to the one used here was introduced, and the authors reported some preliminary findings regarding the equirectangular image peculiarities in the context of subjective and objective quality evaluation. In [12], the authors investigated the prediction of head rotation-based saliency and examined the artefacts occurring in such "head saliency maps".

A regular saliency predictor expects its input to be a 2D image, and does not rely on any additional information about it. Below we describe several reasons why directly applying a saliency prediction models to equirectangular images might not be wise. First, the already mentioned image structure distortions might result in irregular feature responses. A significant part of an image produced through equirectangular projection suffers little to mod-

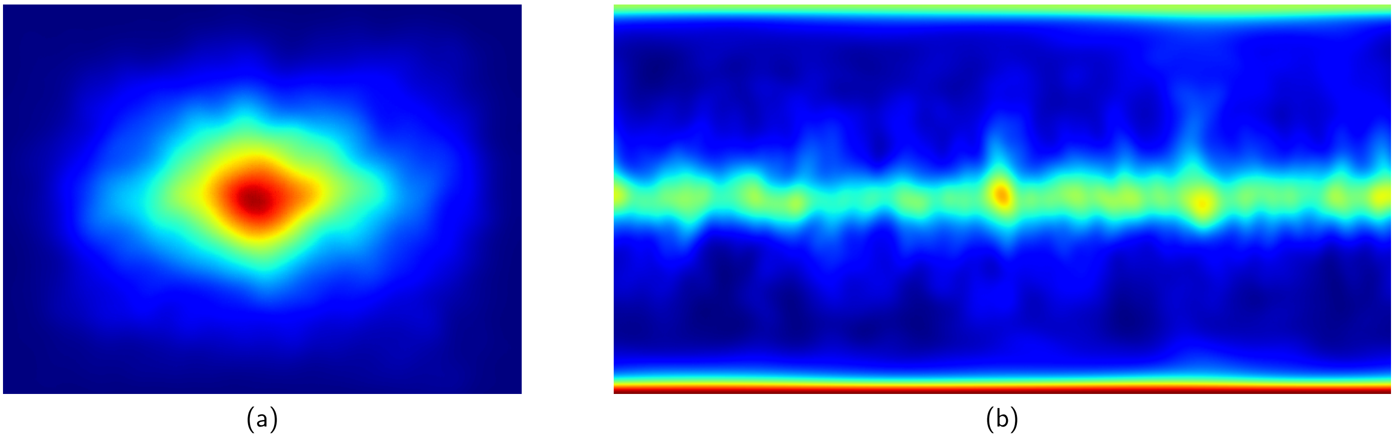


Figure 3: "Centre bias" visualized as empirical mean saliency maps for the MIT1003 data set [19] of regular 2D images (3a) and for the "Salient360!" [18] training set (3b).

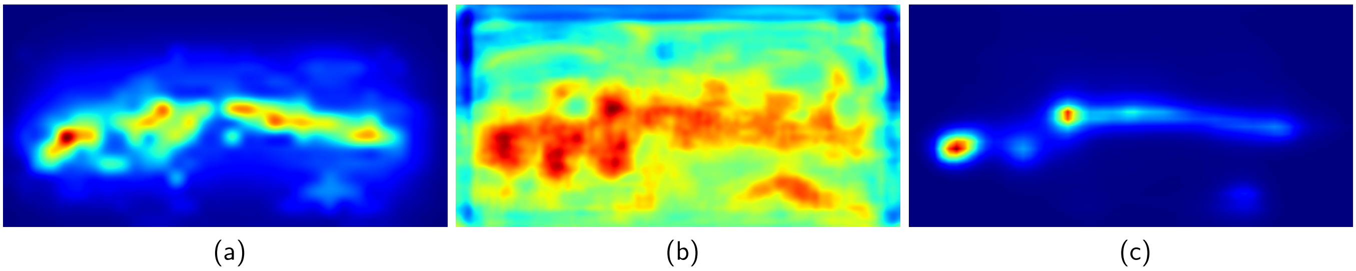


Figure 4: Example saliency map predictions directly on an equirectangular image with the three existing predictor models we use in Section 3.3: GBVS [20] (4a), eDN [21] (4b) and SAM-ResNet [22] (4c). Here we take the image in Figure 1a as input. The ground-truth saliency map in Figure 1b has its highest values along the bottom border, and the vertical borders neither on the left nor on the right side affect the continuity of the central "saliency strip". Both these observations do not hold for either of the directly predicted saliency maps.

erate shape distortion, but the parts close to its top and bottom are noticeably malformed, enough for a human not to recognise a shape right away (see an example image pair for a set of simplest shapes in Figure 2; also, can you recognise a human head in Figure 1?).

The second issue is related to the well-known centre bias effect, observed at least as early as 1935 [14], which is very noticeable in regular image saliency data sets (see Figure 3a), and is extremely persistent across different data sets, tasks, image feature distributions, or forced first fixation location for static images [15], as well as for videos of dynamic natural scenes [16, 17].

This effect is very different for 360° images (see Figure 3b). Instead, we see attention bias along the vertical axis, with the central, the top-most, and the bottom-most locations of the equirectangular images all accumulating significant portions of the overall saliency distribution. This was also observed in [18], as well as in [12], in that case even more prominently so for the head-only saliency. The term "equator bias" was used in the latter to describe this effect, and a general way to overcome the centre bias tendency in regular saliency predictions was introduced.

The two issues described above lead in turn to a third problem. The border artefacts that could be neglected for regular image saliency prediction, in part due to the centre bias (on average, only a small part of saliency is allocated close to the image borders), can be neglected no more. From the theoretical standpoint, there were no actual borders in the stimulus, the viewport never contained a discontinuous image during recording. Now from the practical point of view, directly applying a regular saliency model to an equirectangular stimulus will most likely generate some border effects, both vertical (i.e. neglecting horizontal continuity; the object right behind the starting point of the observation is basically cut in half and is not seen as a set of closely located pixels by the saliency predictor) and horizontal (which means that the most prominent parts of the average ground truth empirical saliency map in Figure 3b are likely to fall into the border effect zone). Example saliency maps produced by the three saliency predictors we use in our experiments (see Section 3) can be found in Figure 4.

2.2. Outline

We propose to deal with these issues with what we call "interpretations" of the equirectangular image format. In our approach (see Figure 5 for the overview of its stages),

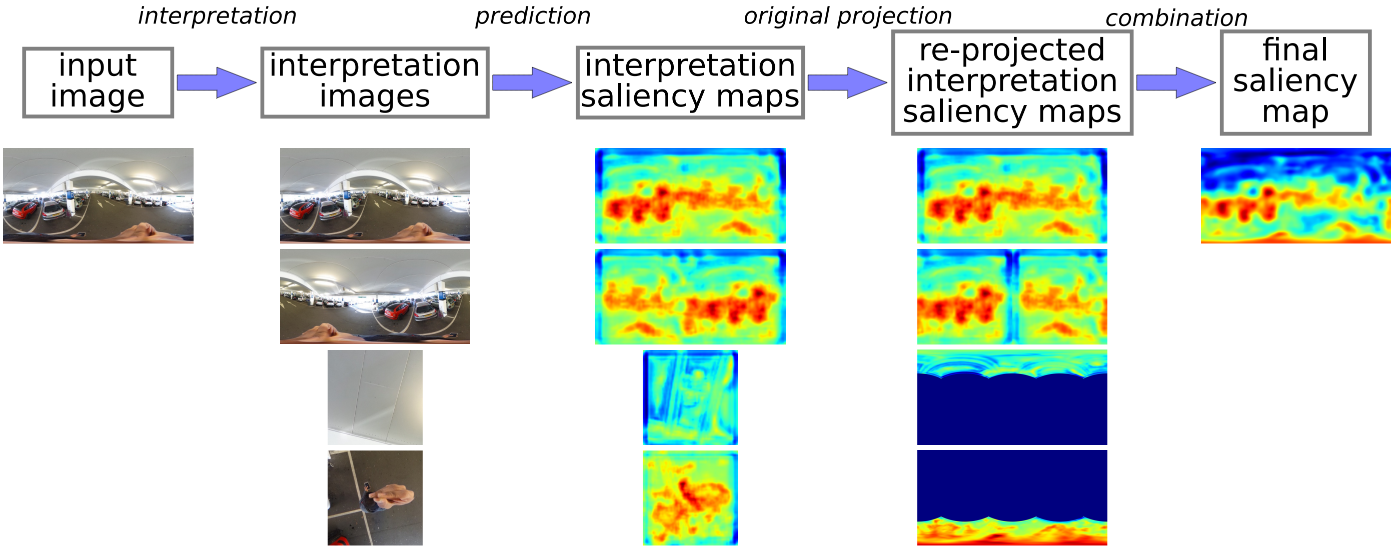


Figure 5: The sequence of steps in our approach (top row). Example data samples from each stage are presented towards the bottom of the figure.

we first create a set of intermediary images derived from the input image (this derivation is what we mean by "interpretation"). For these, respective saliency maps are predicted and subsequently re-projected into the equirectangular space corresponding to the input image, before they are combined into a final saliency map.

In order to combine several overlapping saliency maps into one, the final map is produced by taking the greatest predicted value in each individual pixel (i.e. applying the pixel-wise maximum operation). If we were to use the mean of predicted values, the pixels that were affected by border-related effects at least in one of the intermediary saliency maps would be at great disadvantage, compared to pixels that were never close to saliency map borders. Since we cannot guarantee the uniformity of the individual, interpretation-, model-, and content-dependent border effects across all pixels of the final saliency map, a reasonable solution would be to ignore the saliency values that were affected by being too close to borders. Using pixel-wise maximum achieves just that, discarding the very low intermediate saliency scores along the borders, provided that the respective values have been re-computed in any of the other saliency maps with a higher estimated saliency score.

The resulting saliency map is always smoothed with a Gaussian filter (σ proportional to the image size, $\sigma = 16\,\mathtt{px}$ for input image height of $1024\,\mathtt{px}$), and normalized to contain only non-negative values that sum to 1 over the entire map.

The following sections provide a detailed description of the several interpretation techniques we have explored.

2.3. Continuity-aware interpretation

To address the artefacts occurring at the left and the right borders of the input equirectangular images, we can use the knowledge that those edges can be seamlessly stitched. We therefore compute the saliency maps both for the original image without any preprocessing, and an image that has its left and right halves swapped (this is equivalent to looking in the direction opposite to the starting gaze direction, i.e. backwards). The reverse transformation is applied to the respective saliency map (the "original projection" step in Figure 5). The idea is graphically explained in Figure 6.

This is similar to the Fused Saliency Map post-processing method in [12], where the equirectangular input was translated horizontally several times before saliency maps were predicted, and weighted averaging was applied to the prediction results in order to cancel out the centre bias effect of individual predictions. Here we need fewer rotations (2 instead of 4), since we attempted to switch off the centre bias for our models, where possible, so we mostly needed the rotation just to deal with the border artefacts of our saliency predictors, i.e. help preserve local scene context for feature computation near the borders.

2.4. Cube map-based interpretations

The continuity-aware interpretation only deals with left and right input image borders. Projection distortions, as well as the top and bottom border artefacts are not addressed. To remove the distortions of the equirectangular projection, we can convert the input 360-image to six faces of the cube centred around the camera position. The reverse projection brings the saliency maps from the cube map domain back into the equirectangular one.

Another benefit of this interpretation can be inferred from Figure 2. For example, since the entire bottom stripe of the equirectangular image is produced from just one disk in the centre of the bottom cube map face, the saliency values in this stripe will be extracted from the middle of the

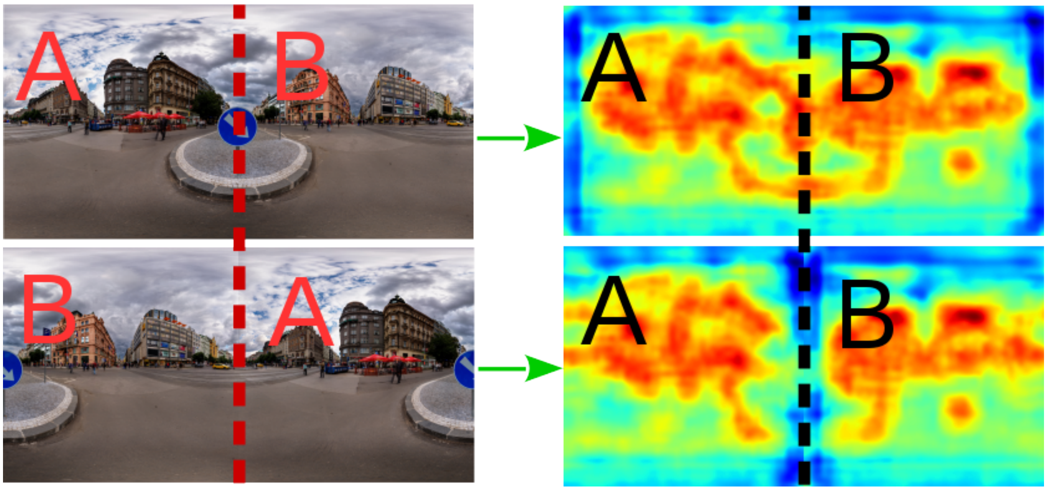


Figure 6: Saliency map computation via continuity-aware interpretation. The final saliency map is obtained with a pixel-wise maximum operation on the two saliency maps on the right, which counteracts the artefacts seen on each of the two maps as dark to light blue vertical stripes either at the left and right borders, or near the dotted cut line.

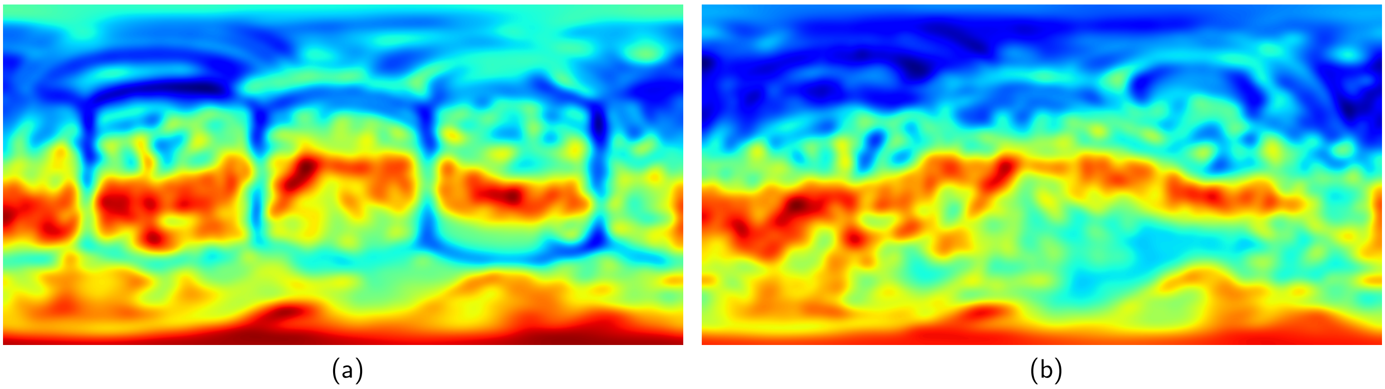


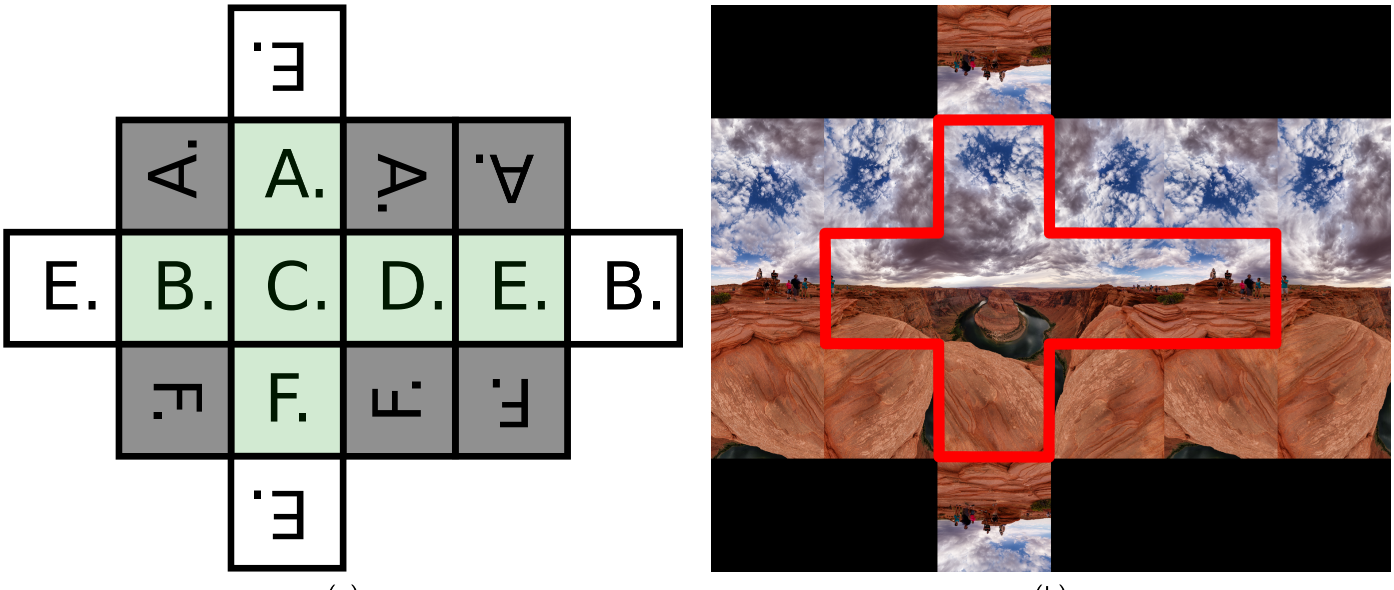
Figure 7: An example of an equirectangular saliency map assembled from the individual saliency maps for the cube faces (7a) and a combination of five such maps, produced at different cube rotation angles (7b).

respective cube face, which is unaffected by any potential border effects. As a result, the equirectangular saliency map produced with this interpretation in mind will be devoid of the top and the bottom border artefacts (for an example, see Figure 7a).

The use of cube maps for omnidirectional scenes is not novel: In [23], several sphere-to-planar projections were examined in search for alternatives to the equirectangular format, in order to reduce bitrate or increase video quality at a given bitrate. Even though the cube map was not the best one overall, it was still an improvement over the equirectangular projection, while being natively supported by modern software. The authors of [24] also looked at a set of projections in the context of using the geometric structure of the projection layouts to select the "Quality Emphasized Regions" (QOR) for full-quality rendering. The quality of the respective spherical video presented to the observer was evaluated at a fixed bit-rate. The cube map layout yielded the best results in this study. Using saliency maps to prioritize different viewports was also suggested there (for selecting the QORs, adapted to scene content). This generally indicates that the cube map "interpretation" is not foreign to the field of 360°-scenes.

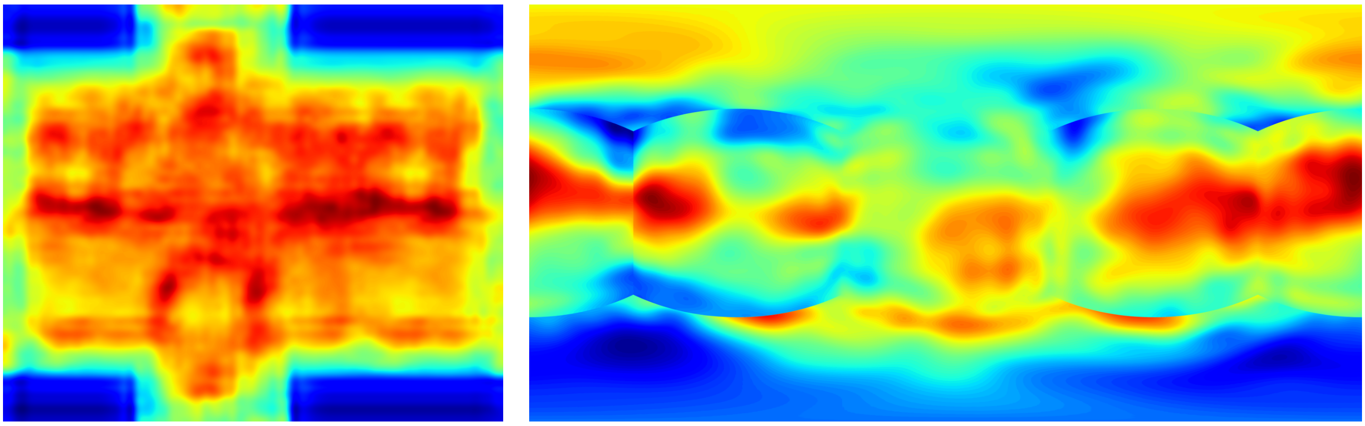
We explored multiple ways of leveraging this particular interpretation of the scene. First, we directly generated the saliency maps for all the cube faces and assembled them into an equirectangular saliency map (an example can be seen in Figure 7a). This approach, however, loses the global context and introduces as many as 24 smaller border artefacts (4 for each face) that greatly deteriorated the quality of the final saliency prediction.

To compensate for these borders, one can generate a larger set of intermediary images and respective saliency maps by extracting the faces at several different rotations of the underlying cubic representation. This way we shift the borders between the stitched faces around the equirectangular saliency map (after the re-projection step), thus lessening the effect of these borders on the final map (see Figure 7b). We take five different cube orientations: its original orientation, rotated by 45° relative to each axis



(a) (b)

Figure 8: Extended cutout construction scheme (8a) and an image example (8b). The "main", not-extended cutout is highlighted in light green on 8a and in red on 8b. A filled cutout consists of all the shaded cube faces in 8a.



(a) (b)

Figure 9: An example extended cutout raw saliency map (9a) and its respective equirectangular projection (9b).

separately, and rotated by 45° relative to the first two axes at the same time.

We can observe that the resulting saliency map does not exhibit any artefacts around its borders (e.g. the lower border accumulates significant amount of saliency, just as in the ground truth saliency map for this input in Figure 1b). The context of the full scene is, however, still lost for the saliency predictors, since they only process one individual cube face at a time.

To preserve all the original image information and context in one image, one can assemble a cube map cutout, which will look similar to that in Figure 2a, with the faces replaced with pixels from the "main" cutout – the highlighted part – of Figure 8b. This does not fully get rid of the border artefacts, since five out of the six faces have at least two problematic edges either at the image border or due to bordering with an empty part of the cutout (only face "C" in Figure 8a would have no discontinuities at its borders). A *filled cutout*, which is an image consisting of a grid of 3 × 4 cube faces stitched together (see the shaded areas of Figure 8a), just like the central rectangle around the main cutout in Figure 8b, resolves only part of the border issues (four of the six main cutout faces are still at the image border). To further minimize these, we introduce an *extended cutout*, which augments the "main" and the "filled" cutouts in such a way that all of the six original cube map faces share all their borders with another face (see the additional "E" and "B" faces to the left and right of the centre row, and the inverted "E"-faces at top and bottom in Figure 8a).

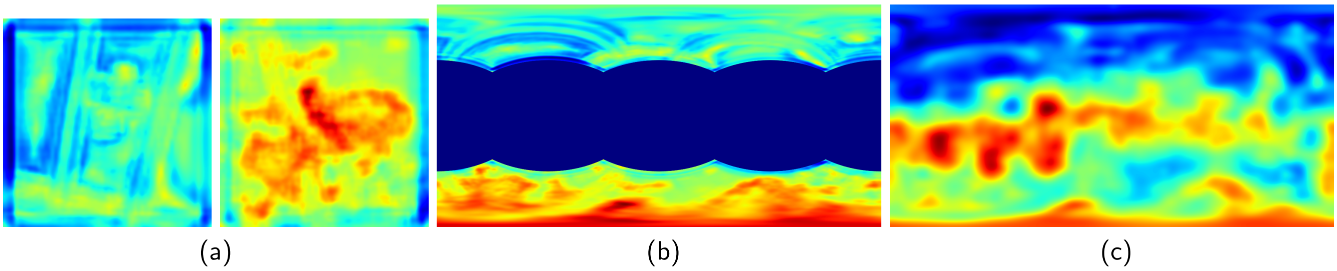


Figure 10: For the input image in Figure 1a: saliency maps for its top and bottom cube faces (10a), their combined projection onto the partial equirectangular map (10b), and the final saliency map (10c), achieved by taking a pixel-wise maximum of the maps in Figures 6 and 10b, plus blurring. Note that value ranges for Figures 6 and 10b are different.

We then compute the saliency map for the whole extended cutout at once (see Figure 9a), extract the maps for all the cube faces of the "main" cutout and project them back onto the equirectangular map (see Figure 9b).

This approach preserves the global context of the scene, even though it over-represents parts of the panorama (in particular, the top and the bottom faces are repeated more than the rest; if these contain highly salient objects, this can have noticeable effects on the final prediction). Distortions are cancelled out, but the stitching in Figure 8b is not perfect (e.g. "A" in Figure 8a wrongly borders on rotated versions of itself in order to fulfil continuity constraints for "B" and "D"). This interpretation also has the scene continuity information built into the cutout, since the objects at the borders of the main cutout are now augmented with the scene parts from the neighbouring cube faces, thus preserving local context. These trade-offs and limitations can be partly visually observed in the saliency maps produced with this interpretation (see Figure 9).

We experimentally concluded that the extended cutout was the best cube map-based interpretation we considered (see Section 4.2).

2.5. Combined interpretation

With this interpretation, we try to combine the benefits of both ideas above: the **continuity-aware** interpretation makes use of all the available contextual information in an equirectangular image without any artificial over-representation, while the **cube map** interpretation helps undo the distortions introduced by the projection, as well as does away with border effects at the top and the bottom of the input image.

The idea here is to now use the cube map interpretation for the two most distorted cube faces only: the top and the bottom ones ("A" and "F" in Figure 8a). The two resulting saliency maps (see Figure 10a) are projected onto the partial equirectangular map (see Figure 10b), and then combined (see Figure 10c) with the full saliency map produced by the continuity-aware approach (as in Figure 6). This interpretation was used to give example visualizations for the pipeline of our approach in Figure 5, so it can be consulted for a better overview.

This way, the resulting map (in Figure 10c) has no left or right vertical border artefacts due to the continuity-awareness, and no horizontal border artefacts due to the top and the bottom cube map faces being processed separately. The distortions are addressed where it is needed the most, and the scene context was not disbalanced during prediction.

3. Experimental methods

In this section we outline the experiments we performed and the evaluation procedures employed in the context of this work.

3.1. Data set

The data sets used in this work were provided by the "Salient360!" Grand Challenge at the IEEE International Conference on Multimedia & Expo (ICME) 2017 [18, 25]. For head-eye saliency (i.e. for each viewport, the direction of eye gaze was considered; this is a natural extension of regular 2D saliency for the 360°-image domain), a training set of 40 images and corresponding scanpaths and fixation heat maps were provided. During the eye tracking recordings, the images were presented for 25 seconds with identical starting observation direction for all observers (at least 40 for each image). The stimuli were presented with an HMD Oculus-DK2 at 75 Hz and with a resolution of 960 × 1080 px per eye. Gaze data was recorded binocularly with an SMI tracker at 60 Hz.

The test set consisted of 25 spherical images, with their respective ground truth collected under conditions identical to those of the training set. Both the test image set and its ground-truth empirical saliency maps were hidden at the time of submission to the Grand Challenge.

All the 360° images and heat maps were represented as flat 2D images through the equirectangular projection. Scanpath coordinates were also given relative to this projection. An example image of the data set that visualizes this projection is shown in Figure 1, along with its empirical saliency map.

3.2. Evaluation

For evaluation, the Grand Challenge used four saliency map metrics [18, 25]: i) two *density-based* metrics, which compare the entire saliency map to the empirical "ground truth" map: Kullback-Leibler divergence (KLD) and Correlation Coefficient (CC), and ii) two *location-based* metrics, which consider only a set of selected locations on the saliency map: Normalized Scanpath Saliency (NSS) and Area Under the Curve (AUC, no class balancing; it technically considers the entire set of pixels of the saliency map by sampling all the possible locations, but the thresholds for building the Receiver Operating Characteristic (ROC) only iterate through the values at fixated locations).

3.3. Saliency predictors

As for this work we focused on already existing pretrained models for image saliency prediction, we took three different, well-performing open-source models from the MIT300 image saliency benchmark [26, 27] (probably the most widespread and established benchmark for image saliency; the ground truth saliency maps are not publicly available, and each submitted model is evaluated by the benchmark organizers, after which the scores with respect to eight popular quality metrics are published on the website). No additional training was performed.

Small modifications were applied to all the models (where possible and necessary) in order to i) support varying image ratios by implementing adaptive downscale parameter choice (since the original images are 1:2, and our input interpretations in Section 2 additionally produce 1:1, 3:4 and 5:6 images, scaling all of them to one size would impede accurate saliency prediction); ii) yield saliency maps without any post-processing, such as blurring and normalization (which would otherwise make the saliency values incomparable when combining several saliency maps into one); and iii) store saliency maps to disk using matrix-based formats instead of images to avoid 8-bit quantization.

Below we describe the three literature models that were used in this work, in chronological order. Graph-based visual saliency (GBVS) was introduced in 2006 [20]. This approach uses a set of Gabor filter responses, local contrast, and luminance maps as features on several spatial scales. The feature maps are heavily downsampled, after which sophisticated activation and normalization steps are applied.

Ensemble of deep networks (eDN), introduced in 2014 [21], was a precursor of the deep learning methods for saliency prediction that have afterwards become very popular. The model's architecture can be represented as a combination of six multilayer structures (one to three layers) of operations that were inspired by their biological counterparts that take place in the visual cortex. Both the final combination and each individual layered structure of the richly-parameterized operations were obtained through hyper-parameter optimisation. A simple linear classifier is used to distinguish salient and non-salient image locations.

Saliency Attentive Model (SAM) is a recently (in 2016) introduced model [22] that extracts image features via a dilated ResNet architecture [28] (in the version used for this work; the framework also includes an option to use dilated VGG-16 [29] for feature extraction). It then employs a convolutional Long Short-Term Memory (LSTM) network, which recurrently attends to different locations of the feature tensor.

As saliency prediction is a multifaceted problem, there is no one definitive metric for model evaluation, and hence no one best model. If we use the well-established MIT300 benchmark [26, 27] to compare the three models listed above, each of them comes out on top of the others according to at least one metric. Table 1 contains an overview of the models' performance in the form of their ranks (out of 74 models) with respect to several metrics [30] (the ranking snapshot was taken on the date of the Grand Challenge submission deadline, May 2017). It can be seen that all the models have their strengths and weaknesses, but SAM-ResNet is probably the more consistently well-performing one.

	KLD* rank	CC* rank	NSS* rank	AUC* rank	balanced AUC rank
GBVS	**9**	27	28	22	14
eDN	43	35	39	18	**7**
SAM-ResNet	59	**4**	**2**	**5**	30

Table 1: The overview of the used 2D image saliency models' performance, as the rank of each respective model in the MIT300 benchmark [27] (metrics marked with * were also used in the "Salient360!" Grand Challenge [18, 25]).

To enhance the performance of our saliency prediction, we also combined the final saliency maps generated by the three models above. The benefits of combining several saliency predictions into one have been thoroughly discussed in [31], as well as earlier in [32]. Taking the mean of the predicted saliency maps falls under the category of non-learning based approaches described in [31], and was shown to outperform all of the baseline saliency models, especially when averaging only over a small set of best performers. The work in [32] only considered summation (with different weighting schemes) and multiplication approaches, concluding that the simple mean performed best. We therefore computed the average of the final saliency maps produced with all three base saliency predictors (after the normalization step).

3.4. Experiments

In our work we tested various combinations of interpretations (see Section 2) and saliency predictors (see Section 3.3). Most of the preliminary experiments were performed with eDN, whereas the final selection of interpretations was tested with all the models. We selected a subset

Table 2: Training-set performance of the cube map interpretation variations (with eDN as saliency predictor). The symbol $\prec$ indicates inferiority of the number on the left to the number on the right (i.e. greater for KLD and lower for the rest of the metrics).

Metric	Filled cutout		Cube faces		Cube faces (5 rotations)		**Extended cutout**
KLD	0.76	$\prec$	0.74	$\prec$	0.71	$\prec$	**0.69**
CC	0.28	$\prec$	0.33	$\approx$	0.33	$\prec$	**0.35**
NSS	0.30	$\prec$	0.31	$\prec$	0.40	$\prec$	**0.50**
AUC	0.59	$\approx$	0.59	$\prec$	0.61	$\prec$	**0.64**

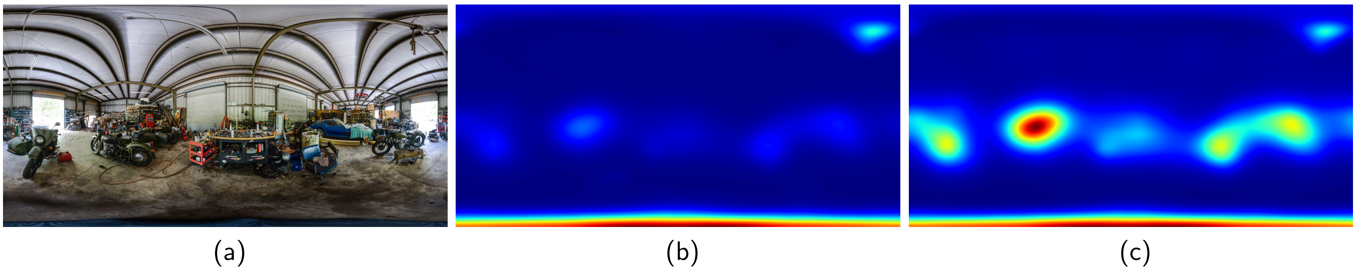


Figure 11: The input image (11a), the respective SAM-ResNet saliency maps produced with the **combined interpretation** without (11b) and with (11c) the rescaling factor for the partial saliency map.

of interesting combinations for submission to the Grand Challenge.

4. Results and discussion

First, we here discuss the limitations and related preliminary experiment of each interpretation group. Section 4.4 summarises the performance figures of all evaluated saliency predictors.

4.1. Continuity-aware interpretation

This is the simplest approach of the ones we have used, which essentially changes the location of the vertical border in the equirectangular image by rotating the spherical image representation by 180° in the horizontal plane. Since we do not know whether any objects happen to be located at the stitching line, neither before nor after the rotation, we simply combine the saliency maps produced for the original image and the shifted one.

Another approach here could be finding such a stitching point on the image, where no object would be bisected, and only predicting the saliency map for one equirectangular image. It is, however, not guaranteed that such a point always exists, and the resulting saliency map would still have noticeable visually unnatural artefacts near the stitching line.

A similar approach could be additionally applied to eliminate vertical borders, but this requires more complex spherical image manipulations (e.g. converting to a cube map, rotating by 90° in the respective plane, and projecting back onto the equirectangular surface, with corresponding reverse transformations taking place after saliency prediction), whereas this interpretation was intended as the simplest way of incorporating additional information into the prediction process.

4.2. Cube map interpretations

As described in Section 2.4, there are multiple ways to use a cube map to produce equirectangular saliency maps. We evaluated (on the training set) four of them to find the best one: individual cube faces (as in Figure 7a), individual cube faces at five different rotations of the spherical image (same as in Figure 7b), filled cutout (the shaded areas in Figure 8a), and extended cutout (all cube faces in Figure 8a). Their performance figures are summarised in Table 2. The trend is the same for all the four metrics: a filled cutout is inferior to using the individual cube map faces, which is in turn improved by using several rotated versions of the cube map, and the extended cutout outperforms the rest (marked in bold in the table).

4.3. Combined interpretation

For this interpretation, we have additionally experimented with the way of computing the saliency maps for the top and the bottom cube faces: either separately, or as part of an extended cutout. The former approach proved to outperform the latter with big margins (on the training set, with eDN used for saliency prediction): 0.65 vs. 0.57 AUC, 0.36 vs. 0.29 CC, 0.68 vs. 0.75 KLD, 0.53 vs. 0.24 NSS, respectively.

One adjustment we had to make for this approach was concerning one of the saliency predictors (namely SAM-ResNet), which in this set-up tended to over-represent the top and bottom cube planes (see Figure 11b), probably because of the lacking context. We therefore attempted to quantitatively examine this disbalance. To this end, we split each of the resulting saliency maps in two parts: part A – the middle third (horizontally) – and part B – the rest of the map. We then computed the ratio of the

Table 3: Saliency maps evaluation results, depending on the equirectangular image interpretation and the saliency predictor model. Best results for each metric are boldified.

Metric	Predictor	Continuity-aware	Extended cutout	Combined
KLD	GBVS	0.67	0.76	0.66
	eDN	0.67	0.64	0.62
	SAM-ResNet	0.55	0.74	0.48
	average	0.50	0.58	**0.45**
CC	GBVS	0.35	0.29	0.35
	eDN	0.41	0.40	0.43
	SAM-ResNet	0.54	0.31	0.56
	average	0.55	0.41	**0.58**
NSS	GBVS	0.73	0.46	0.64
	eDN	0.75	0.63	0.67
	SAM-ResNet	0.84	0.56	0.70
	average	**0.92**	0.69	0.81
AUC	GBVS	0.71	0.64	0.70
	eDN	0.72	0.68	0.69
	SAM-ResNet	**0.75**	0.67	0.71
	average	**0.75**	0.69	0.73

maximal saliency value in part B to that in part A for each individual saliency map.

It turned out that the ground truth maps and both the eDN and the GBVS saliency maps (produced via the combined interpretation) all had the mean of these ratios around 1 (0.73 for the ground truth to 1.16 for GBVS). For the SAM-ResNet saliency maps it was, however, 4.51. We therefore divided all the values in the partial (for the top and the bottom cube map faces, see Figure 10b) equirectangular SAM saliency map by this coefficient prior to combining it with the continuity-aware saliency maps (see Figure 11c). The improvement of this rescaling is again quantitatively noticeable: 0.68 vs. 0.62 AUC, 0.53 vs. 0.4 CC, 0.51 vs. 0.7 KLD, 0.48 vs. 0.1 NSS, with and without this modification, respectively.

4.4. All results

For a more complete evaluation of our approach, we can consider using each of the selected 360°-image interpretations (i.e. *continuity-aware*, *extended cutout* and *combined*) with each of the employed saliency predictors (i.e. GBVS, eDN, SAM-ResNet, and their average) in turn. The full table for all results of our predictor-interpretation pairs can be found in Table 3.

Additionally, to any of the resulting saliency maps we can optionally add the mean ground truth saliency map (of the training set) with a certain weight. We empirically determined 0.2 to be a good choice. This way, we explicitly take into account the "vertical centre bias" that was observed in Figure 3b. This gives us a total of 24 models.

To better analyse the evaluation results, we can differently group them: If we group the entire set of models by the saliency predictor, we can see that the "newer" model's performance is consistently superior to that of an "older" one, while the average model outperforms all of the individual models (see Figure 12).

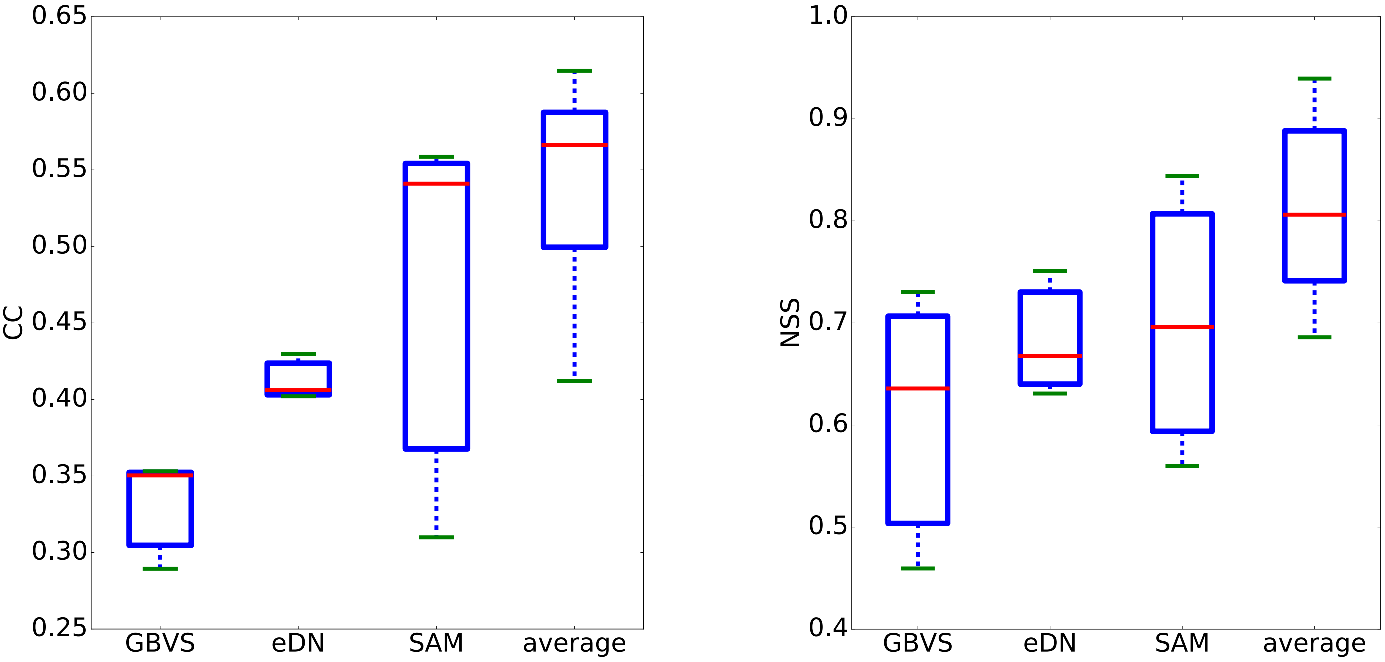
If we now group by the interpretation method, the conclusions become less clear-cut. For both density-based metrics, the combined interpretation performs best, followed by the continuity-aware interpretation (see Figure 13a). For both location-based metrics, the continuity-aware interpretation is now the one in the lead, closely followed by the combined interpretation (see Figure 13b).

For the average saliency predictor, however, it turned out that some of these differences were not statistically significant, and so *the combined interpretation with the average saliency predictor* was ranked 1^{st} for all the metrics in the "Salient360!" Grand Challenge [25], for some metrics tied in the first place with several other approaches, including *the continuity-aware interpretation with the average saliency predictor* (see Table 4).

It is also interesting to note that the worst (on average) saliency predictor – GBVS – in combination with the best (on average) interpretation – combined – performs better than the best (on average) predictor – SAM-ResNet – with the worst (on average) interpretation – the extended cutout.

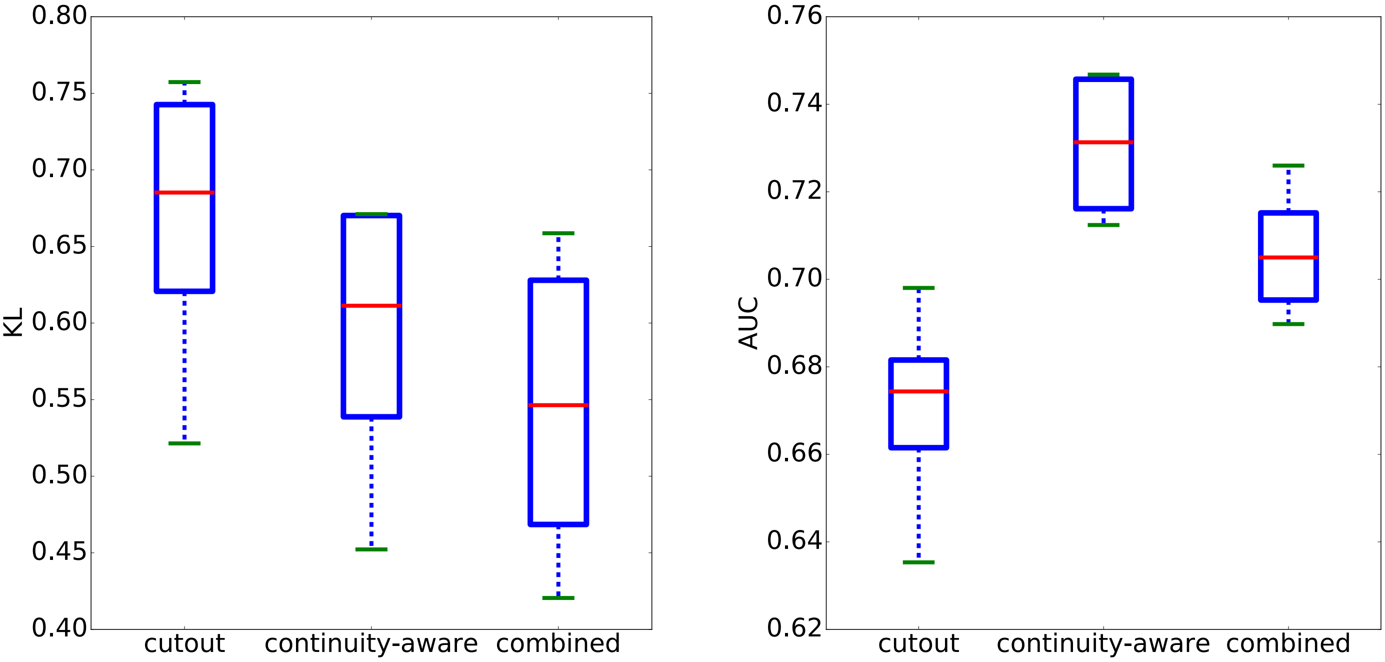
All the qualitative results were reproduced both on the training and the test set. We see that the optimal choice of the interpretation can depend on the metric choice, but the combined interpretation generally fares rather well, delivering the best-ranked results (out of the models submitted before the test set was released) for all metrics at the "Salient360!" Grand Challenge in the "Head-Eye saliency prediction" track [18, 25]. It also yields the best (in terms of absolute values) average scores for KLD and CC metrics across all submitted models.

Naturally, saliency prediction can benefit from specialized models, which were trained with the information about the equirectangular format of the images and the 360° nature of the scenes in mind, so training a dedi-



(a) (b)

Figure 12: Performance summary of all the models, split by the saliency predictor: correlation coefficient (12a) and normalized scanpath saliency (12b). A similar trend is observed for the other metrics as well.



(a) (b)

Figure 13: Performance summary of all the models, split by the input image interpretation: Kullback-Leibler divergence (13a, similar results for correlation coefficient) and area under the curve (13b, similar results for normalized scanpath saliency).

cated model for this kind of stimuli is still worthwhile. It seems, however, that using pretrained state-of-the-art image saliency predictors to tackle the 360°-scene saliency prediction problem could suffice, at least as a first approximation, for some applications. For a minimal-effort model, one can therefore focus on an appropriate stimulus interpretation rather than on developing and training a whole new prediction model. Combining input interpretations

	KLD (rank)	CC (rank)	NSS (rank)	AUC (rank)	mean rank
Combined interp. + avg. saliency model + centre bias	0.42 **(1)**	0.62 **(1)**	0.81 **(1)**	0.72 **(1)**	**1**
Combined interp. + avg. saliency model	0.45 **(1)**	0.58 **(1)**	0.81 **(1)**	0.73 **(1)**	**1**
Zhu et al. [33]	0.48 **(1)**	0.53 (6)	0.92 **(1)**	0.74 **(1)**	2.25
Ling et al. [34]	0.51 (5)	0.54 (6)	0.94 **(1)**	0.74 **(1)**	3.25
Continuity-aware interp. + avg. saliency model	0.50 (5)	0.55 (6)	0.92 **(1)**	0.75 **(1)**	3.25
...	...	...	...	...	...
Extended cutout interp. + avg. saliency model	0.58 (5)	0.41 (12)	0.69 (8)	0.69 (6)	7.75

Table 4: The "Salient360!" Grand Challenge official unbiased results for the Head-Eye Saliency track, top-5 snippet and our extended cutout interpretation-based model. The rank (within each metric) was only increased if the difference between the respective sets of performance figures was statistically significant. 16 models were submitted to the challenge in total, with the worst average rank of 14.25.

and dedicated training procedure may yield even better results.

The source code of our approach is publicly available at `http://www.michaeldorr.de/salient360`.

5. Conclusion

In this work we have explored the applicability of regular image saliency models for the panoramic image case with a full 360° field of view. To this end we proposed several ways of "interpreting" the input equirectangular image, which would deal with the projection-related issues. We used three well-performing regular 2D image saliency predictors (and their combination via averaging). Our best-performing input interpretation is a combination of the continuity-aware and the cube map approach, and requires computing four saliency maps: one for the frontal equirectangular view, one for the "rear view" (i.e. looking backwards from the starting viewing position), and one saliency map for each of the top and the bottom cube map faces. Combined with the average saliency predictor, this took the first prize at the Head-Eye Saliency Prediction track of the "Salient360!" Grand Challenge.

Acknowledgements

This research was supported by the Elite Network Bavaria, funded by the Bavarian State Ministry for Research and Education.

References

[1] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, F. Pellandini, Adaptive color image compression based on visual attention, in: Proceedings 11th International Conference on Image Analysis and Processing, 2001, pp. 416–421. doi:10.1109/ICIAP.2001.957045.

[2] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Transactions on Image Processing 19 (1) (2010) 185–198. doi:10.1109/TIP.2009.2030969.

[3] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1454–1461. doi:10.1109/CVPR.2009.5206525.

[4] E. Vig, M. Dorr, D. Cox, Space-variant descriptor sampling for action recognition based on saliency and eye movements, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 84–97. doi:10.1007/978-3-642-33786-4_7.
URL https://doi.org/10.1007/978-3-642-33786-4_7

[5] Y. Wei, X. Liang, Y. Chen, X. Shen, M. M. Cheng, J. Feng, Y. Zhao, S. Yan, Stc: A simple to complex framework for weakly-supervised semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (11) (2017) 2314–2320. doi:10.1109/TPAMI.2016.2636150.

[6] S. Wang, M. Jiang, X. Duchesne, E. Laugeson, D. Kennedy, R. Adolphs, Q. Zhao, Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking, Neuron 88 (3) (2015) 604 – 616. doi:http://dx.doi.org/10.1016/j.neuron.2015.09.042.
URL http://www.sciencedirect.com/science/article/pii/S0896627315008314

[7] J. E. Silberg, I. Agtzidis, M. Startsev, T. Fasshauer, K. Silling, A. Sprenger, M. Dorr, R. Lencer, Free visual exploration of natural movies in schizophrenia, European Archives of Psychiatry and Clinical Neuroscience (2018) 1–12doi:10.1007/s00406-017-0863-1.
URL https://doi.org/10.1007/s00406-017-0863-1

[8] E. Goffman, Behavior in public places, Simon and Schuster, 2008.

[9] T. Foulsham, E. Walker, A. Kingstone, The where, what and when of gaze allocation in the lab and the natural environment, Vision Research 51 (17) (2011) 1920 – 1931. doi:http://dx.doi.org/10.1016/j.visres.2011.07.002.
URL http://www.sciencedirect.com/science/article/pii/S0042698911002392

[10] M. Assens, K. McGuinness, X. Giro-i-Nieto, N. E. O'Connor, SaltiNet: Scan-path prediction on 360 degree images using saliency volumes, ArXiv e-prints (2017) 1–8arXiv:1707.03123.

[11] A. Reina, Marc, K. McGuinness, X. Giro-i-Nietro, E. O'Connor, Noel, Scanpath and saliency prediction on 360 degree images, Signal Processing: Image Communication ?? (2018) ??–??

[12] A. D. Abreu, C. Ozcinar, A. Smolic, Look around you: Saliency

maps for omnidirectional images in VR applications, in: Ninth International Conference on Quality of Multimedia Experience (QoMEX), 2017, pp. 1–6. doi:10.1109/QoMEX.2017.7965634.

[13] E. Upenik, M. Řeřábek, T. Ebrahimi, Testbed for subjective evaluation of omnidirectional visual content, in: Picture Coding Symposium (PCS), 2016, pp. 1–5. doi:10.1109/PCS.2016.7906378.

[14] G. T. Buswell, How people look at pictures: a study of the psychology of perception in art, University of Chicago Press Chicago, 1935.

[15] B. W. Tatler, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions, Journal of Vision 7 (14) (2007) 4. arXiv:/data/journals/jov/932846/jov-7-14-4.pdf, doi:10.1167/7.14.4.
URL +http://dx.doi.org/10.1167/7.14.4

[16] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, L. Itti, Quantifying center bias of observers in free viewing of dynamic natural scenes, Journal of Vision 9 (7) (2009) 4. arXiv:/data/journals/jov/932863/jov-9-7-4.pdf, doi:10.1167/9.7.4.
URL +http://dx.doi.org/10.1167/9.7.4

[17] M. Dorr, T. Martinetz, K. R. Gegenfurtner, E. Barth, Variability of eye movements when viewing dynamic natural scenes, Journal of Vision 10 (10) (2010) 28. arXiv:/data/journals/jov/932797/jov-10-10-28.pdf, doi:10.1167/10.10.28.
URL +http://dx.doi.org/10.1167/10.10.28

[18] Y. Rai, J. Gutiérrez, P. Le Callet, A dataset of head and eye movements for 360 degree images, in: Proceedings of the 8th ACM on Multimedia Systems Conference, MMSys'17, ACM, New York, NY, USA, 2017, pp. 205–210. doi:10.1145/3083187.3083218.
URL http://doi.acm.org/10.1145/3083187.3083218

[19] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE 12th International Conference on Computer Vision, 2009, pp. 2106–2113. doi:10.1109/ICCV.2009.5459462.

[20] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems, 2007, pp. 545–552.

[21] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2798–2805.

[22] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an lstm-based saliency attentive model, CoRR abs/1611.09571 (2016) 1–13. arXiv:1611.09571.
URL http://arxiv.org/abs/1611.09571

[23] M. Yu, H. Lakshman, B. Girod, A framework to evaluate omnidirectional video coding schemes, in: IEEE International Symposium on Mixed and Augmented Reality, 2015, pp. 31–36. doi:10.1109/ISMAR.2015.12.

[24] X. Corbillon, G. Simon, A. Devlic, J. Chakareski, Viewport-adaptive navigable 360-degree video delivery, ArXiv e-prints (2016) 1–7arXiv:1609.08042.

[25] J. Gutiérrez, E. David, Y. Rai, P. Le Callet, Toolbox and dataset for the development of saliency and scanpath models for omnidirectional / 360° still images, Signal Processing: Image Communication ?? (2018) ??–??

[26] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, in: MIT Technical Report, 2012.

[27] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, A. Torralba, MIT saliency benchmark.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, ArXiv e-prints (2014) 1–14arXiv:1409.1556.

[30] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models?, CoRR abs/1604.03605 (2016) 1–24. arXiv:1604.03605.
URL http://arxiv.org/abs/1604.03605

[31] J. Wang, A. Borji, C. C. J. Kuo, L. Itti, Learning a combined model of visual saliency for fixation prediction, IEEE Transactions on Image Processing 25 (4) (2016) 1566–1579. doi:10.1109/TIP.2016.2522380.

[32] A. Borji, D. N. Sihite, L. Itti, Salient object detection: A benchmark, in: Proceedings of the 12th European Conference on Computer Vision - Volume Part II, ECCV'12, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 414–429. doi:10.1007/978-3-642-33709-3_30.
URL http://dx.doi.org/10.1007/978-3-642-33709-3_30

[33] Y. Zhu, G. Zhai, X. Min, The prediction of head and eye movement for 360 degree images, Signal Processing: Image Communication ?? (2018) ??–??

[34] J. Ling, K. Zhang, Y. Zhang, D. Yang, Z. Chen, A saliency prediction model on 360 degree images using color dictionary based sparse representation, Signal Processing: Image Communication ?? (2018) ??–??