COMENIUS UNIVERSITY IN BRATISLAVA

FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

SALIENCY MODELS ANALYSIS FOR PAINTINGS

DIPLOMA THESIS

2020 **Bc. Kristína Miklošová**

COMENIUS UNIVERSITY IN BRATISLAVA

FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

SALIENCY MODELS ANALYSIS FOR PAINTINGS

DIPLOMA THESIS

Study Programme: Cognitive Science

Field of Study: Cognitive Science

Department: Department of Applied Informatics

Supervisor: RNDr. Zuzana Černeková, PhD.

2020 **Bc. Kristína Miklošová**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Kristína Miklošová
Študijný program: kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Saliency models analysis for paintings
Výskum modelov významných oblastí malieb

Anotácia: Naštudovať problematiku určovania významných oblastí v obraze. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Vybrať a otestovať tri metódy prípadne navrhnúť ich modifikácie, za účelom nájsť takú, ktorá najlepšie dokáže zachytiť pohľad umelcov na digitalizované maľby zobrazujúce biblickú scénu Poslednej večere.

Cieľ: Naštudovať problematiku určovania významných oblastí v obraze. Analyzovať existujúce riešenia publikované v dostupnej odbornej literatúre. Vybrať a otestovať tri metódy prípadne navrhnúť ich modifikácie, za účelom nájsť takú, ktorá najlepšie dokáže zachytiť pohľad umelcov na digitalizované maľby zobrazujúce biblickú scénu Poslednej večere.

Literatúra:
R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, Q. Huang (2018). Review of Visual Saliency Detection with Comprehensive Information; IEEE Transactions on Circuits and Systems for Video Technology.
Itti, Koch and Neibur (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE PAMI, 20(11).
J. Kučerová (2015). Saliency map detection and applications. Dissertation thesis, FMFI, Univerzita Komenského, Bratislava.

Vedúci: RNDr. Zuzana Černeková, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 18.10.2018

Dátum schválenia: 27.02.2019

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.....
študent

.....
vedúci práce

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Kristína Miklošová
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Saliency models analysis for paintings

Annotation: Study research area of saliency maps for images. Analyze methods published in scientific literature. Select and test three of them, or propose modifications, in order to find the one which can predict eye movements of the artists during viewing of the digitalized versions of painting depicting biblical scene of The Last Supper.

Aim: Study research area of saliency maps for images. Analyze methods published in scientific literature. Select and test three of them, or propose modifications, in order to find the one which can predict eye movements of the artists during viewing of the digitalized versions of painting depicting biblical scene of The Last Supper.

Literature:
R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, Q. Huang (2018). Review of Visual Saliency Detection with Comprehensive Information; IEEE Transactions on Circuits and Systems for Video Technology.
Itti, Koch and Neibur (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE PAMI, 20(11).
J. Kučerová (2015). Saliency map detection and applications. Dissertation thesis, FMFI, Univerzita Komenského, Bratislava.

Supervisor:	RNDr. Zuzana Černeková, PhD.
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	prof. Ing. Igor Farkaš, Dr.
Assigned:	18.10.2018
Approved:	27.02.2019

prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Declaration

I hereby declare that this thesis is my own and that all sources have been acknowledged.

Bc. Kristína Miklošová

Acknowledgement

I would like to thank my supervisor RNDr. Zuzana Černeková, PhD. for guidance and valuable advice throughout the creation of this thesis. I also would like to thank CReA Laboratory team at University of Vienna for internship opportunity without which this thesis would not have been possible. Last but not least, I would like to thank my family and friends for their support and encouragement.

Bc. Kristína Miklošová

Abstract

The topic of visual saliency spreads across numerous disciplines. In this work we focus on computational saliency models and their performance in the context of art. We examine how saliency models perform on images with specific characteristics. We explore a saliency of 14 different digitized paintings, all representing a biblical scene of The Last Supper. We process the data and evaluate the performance of four saliency models against them. Two of the models are using traditional approach and two are based on deep neural networks. For the evaluation we use AUC, NSS, and CC metric. As a ground truth we use eye-tracking data from 35 participants. Moreover, we incorporate face detection algorithm to one of the models and slightly improve its performance. Our analysis shows that deep-learning models predict the most salient parts of the paintings closest to real eye fixations.

Keywords: visual attention, saliency, saliency modeling, painting

Abstrakt

Téma vizuálnej pozornosti sa rozprestiera mnohými disciplínami. V tejto práci sa sústreďujeme na výpočtové modely významných oblastí a ich výkon v kontexte umenia. Skúmame aký výkon podávajú modely významných oblastí na obrazoch so špecifickými charakteristikami. Skúmame významnosť 14 rôznych zdigitalizovaných malieb reprezentujúcich biblickú scénu Poslednej večere. Spracúvávame dáta a vyhodnocujeme výkon štyroch modelov významných oblastí na nich. Dva z modelov používajú tradičný prístup a dva sú založené na hlbokých neurónových sieťach. Na vyhodnotenie používame AUC, NSS a CC metriku. Ako reálne dáta používame eye-trackingové dáta od 35 účastníkov. Naviac pripojíme algoritmus detekcie tváre k jednému z modelov a mierne vylepšíme jeho výkon. Naša analýza ukazuje, že modely založené na hlbokom učení predpovedajú najviac významné oblasti malieb najbližšie k naozajstným očným fixáciam.

Kľúčové slová: vizuálna pozornosť, významnosť, modelovanie významných oblastí, maľba

Content

Tables

Figures

Introduction

Visual attention helps us to filter out the unnecessary information from the world. Therefore, the visual processing done by our brain, requires less computational power that it otherwise would have to spend to process a great amount of data that we observe through our eyes every second. Movements of our eyes are guided by visual attention that is drawn to the salient stimuli. By recording the eye movements, we can see what attracts our attention. Recent developments in visual attention modeling provide many new approaches and architectures, while also increase the demand for evaluation. These models try to predict where we will look or what the most important parts of visual scene are. Afterwards, these models are usually evaluated against the real eye movements recorded by eye tracking. In the history of art, eye movements also play an important role as means to describe an artwork. Despite this connection, between eye movements and visual attention, there is only few works focusing on visual attention modeling in art domain.

Therefore, in the current work we analyze four saliency models in the context of visual art. Saliency models model bottom-up visual attention. Two of the models that we chose are more traditional and two are based on deep learning. We evaluate the models against our eye-tracking data of paintings. For the evaluation we use metrics that are commonly used in saliency modeling research. These metrics say to what extent the real eye fixations correspond to predictions made by the models. The eye tracking data consist of recordings of 35 people viewing 14 paintings of a biblical scene of The Last Supper. Before the evaluation, the data have to be pre-processed. This includes selecting only information from the data that are important to our evaluation and creating two types of fixation maps: continuous and discrete. Moreover, we slightly improve the performance of one of the traditional models by incorporating face detection algorithm. Although, the results of the improved traditional model were better, it did not achieve the performance of deep-learning models. At the end, the deep-learning models outperform the traditional models.

In the first chapter Visual Saliency, we outline the topic of visual saliency from different disciplines that are relevant to this work. We follow with the chapter Related Work, where we mention similar works to ours and divide them into two general directions. In the third chapter

Saliency Models, we introduce four saliency models that we will evaluate, describe them and show their architectures. The fourth chapter Evaluation Metrics describes and explains metrics that we will use for the evaluation of the models. The next chapter Data, describes the data that we will evaluate the models against, and also in detail describes steps of preprocessing of this data. In the sixth chapter Analysis and Results, we present the results of our analysis of the four models, we discuss the results and factors that may influenced the outcome. In last chapter Incorporating Face Detection, we describe how we combined face detection algorithm with one of the models. We evaluate this altered model with the same metrics as other models. Lastly, we show how the performance improved and compare the results of this altered model to the previous results.

1 Visual Saliency

The topic of this work lies on the intersection of different disciplines. This chapter aims to highlight how visual saliency is relevant in all these areas. We start by basic psychological overview of visual attention. Next, we describe human vision and visual saliency from a neuroscientific point of view, followed by an emphasis on the role of eye movements through art history. We end this introduction chapter by describing the saliency modelling research.

1.1 Visual Attention and Saliency

In order to process the large amount of data that are entering our eyes, we have to decide what information are the most relevant. Processing every information would be inefficient, since it would require enormous computational power that can be spent elsewhere. For example, while watching a movie, where just now two main characters are in a battle, it is better to focus on the fighting technique of your favorite, than being able to tell exactly how many leaves are on the tree in the background. The mechanism that helps us to filter the visual information that are important to us, from the enormous amounts of other available information, is *visual attention*. However, the question is how visual attention selects what are the important information.

Usually we differentiate between:

a) Top-down Attention

   It is goal-directed, we voluntary focus attention according to our intention. It is influenced by our knowledge and memories.

b) Bottom-up Attention

   It is stimulus-driven, involuntary and rapid. Salient stimuli draw our attention.

Visual saliency arises during early bottom-up processing. It characterizes part of a visual scene that stands out from the rest. Visual saliency is a characteristic that makes something stand out from its surroundings and grab our attention. Bottom-up attention can be overridden by top-down attention - our goals or intentions. For example, a red apple in the middle of green grass

would be highly salient and quickly grab our attention in the bottom-up manner. Yet, if we have a goal in mind and are looking for a red apple in the basket of fruits, no particular color grabs our attention at first, until the top-down attention makes all red fruits more salient and therefore stand out from the rest. [17]

1.2 Vision

We are able to see objects thanks to the light that is reflected from their surface and captured by visual receptors in our eyes. The light entering our eyes passes through the lens of the eye, and focused and inverted is projected onto the retina. The retina is made up of 10 layers of neurons, the top layer consists of photoreceptors: rods and cones. They contain protein molecules – photopigments, which are sensitive to light and split apart when come into a contact with light. This in turn alters their membrane potentials and triggers action potentials in subsequent neurons. Rods and cones are connected to bipolar neurons which then synapse with ganglion cells. Ganglion cells are the output layer of the retina and their axons form the optic nerve. Therefore, the retina transforms the incoming light into an electro-chemical signal that is then carried by the optic nerve to the brain.

However, before the information is sent through the optic nerve, it is processed by the retina. It is estimated that humans have approximately 260 million photoreceptors, but only 2 million ganglion cells. Therefore, the information coming from the retina have to be compressed in some way, in order to be passed by the optic nerve to the brain. This also suggests that higher-level visual centers are powerful processors that should be capable of unraveling such information.

Information from the right visual field is projected to the left hemisphere, while information from the left visual field is projected to the right hemisphere. Every optic nerve is divided into multiple pathways. Each of these pathways terminates in different parts of the sub-cortex. The most important one is the *retinogeniculate pathway*, which projects from retina to the *lateral geniculate nucleus* (LGN) of the thalamus. This pathway involves more than 90% of axons in the optic nerve, other 10% of axons use pathways that terminate in different structures – like *pulvinar nucleus* of the thalamus, or the *superior colliculus* of the midbrain. These are very important in visual attention.

The last projection is from the LGN to the visual cortex. It is done by the *geniculocortical pathway*, which consists of almost all of axons from the LGN, and ends in the primary visual cortex (V1) of the occipital lobe as can be seen in Figure 1.

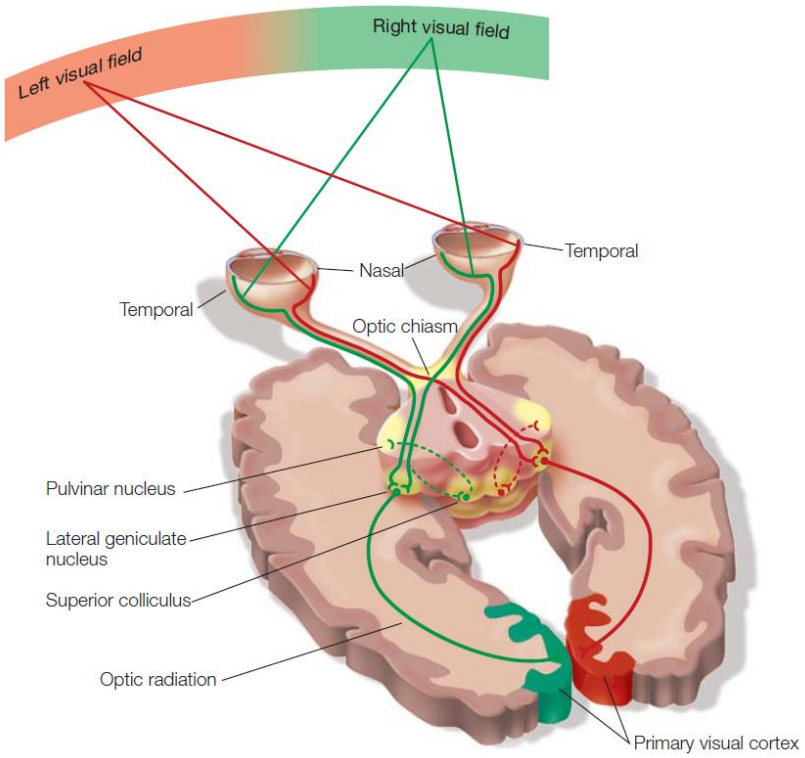

Figure 1 Geniculocortical pathway

When the information from the visual scene enters the V1, it has been processed by (at least) these 4 types of neurons: photoreceptors, bipolar neurons, ganglion cells, and LGN cells. This is where the initial stage of visual processing ends. However, after that the signal is carried to higher order visual areas (V2 – V4), where it is processed even further and eventually interpreted into meaningful percepts. Exact mechanisms of how our brain makes sense of all of this information are to a great extent still a mystery. [14]

Neural correlates of visual saliency are not completely clear, however several brain regions are known to be involved from neurophysiological and imaging studies. Several works suggest that the bottom-up saliency map is constructed in V1 [38] and top-down saliency is associated with V2 and bottom-up and top-down saliency are combined in V4 [24]. Other works showed salience is also represented in parietal cortex [15] or frontal eye field [31].

1.3 Eye Movements and Art

First evidence of describing art in terms of eye movements dates back to the sixth century. The eye movements became a language for describing aesthetic and structural qualities of works of art. It was at the end of the nineteenth century when first attempts were made to measure them empirically. Since then measuring the eye movements became central for psychological and neuroscience research. [28]

The analysis of eye movements distinguishes between fixations and saccades. The act of viewing is composed of fixations, when the eye is relatively stable, and saccades, which are periods between fixations characterized by a rapid movement with a higher magnitude.

The visual attention is guiding the eye movements, and therefore recording the eye fixations can tell us what parts of a painting catch our attention. The first to study eye movement of people viewing paintings was psychologist Buswell in 1935 [6], who demonstrated that some parts of a painting have a higher quantity of fixations, thus are more interesting than others.

1.4 Saliency Modeling

Many attention models derive from the Feature Integration Theory by Treisman and Gelade [32], where they studied how visual features are integrated in order to drive visual attention. Koch and Ullman [20] were first to propose a feed-forward model, where visual features are combined, and processed by winner-take-all neural network, to create a saliency map. First complete implementation of this model was proposed by Itti, Koch and Neibur in 1998 [18] (IttiKoch). We later describe this model in Chapter. 3. To this day many attention models were presented (see [3] for a review). The main distinction is whether a model is modeling the bottom-up or the top-down attention. In this work, we focus on computational bottom-up attention models that detect visually salient regions independent of a task – further we refer to them as “saliency models”. These models are usually evaluated against ground truth eye-tracking data acquired during a free viewing. Further, only first few fixations are considered, as these are more bottom-up [26]. The most salient regions on the image are expected to have the highest density of fixations. A saliency model produces a saliency map, which is a topographical representation of visual saliency of the corresponding image [25].

2 Related Work

Multiple works have benefited from applying the visual saliency concept. The relevancy of the concept spreads over many disciplines such as neuroscience, psychology and computer science. However, in our work we are focusing mostly on the computational approaches to visual saliency in the context of art. In this chapter, we describe some of related works. Of course, the scope of related works could include many research areas. One of such areas could be saliency modeling itself. However, since in this work we are mainly focusing on evaluation of models on paintings and not the creation of models, we divide related works into two general directions: application of saliency in analysis of art and evaluation of saliency approaches.

2.1 Application of Saliency in Analysis of Art

Many authors decided to take use of saliency models in order to confirm their theory, enhance a technique or simplify an analysis, and the results give a promising direction for further research. However, the majority of works focus only one model, in many cases the same one – IttiKoch model (described in Chapter.3). Some of such works we mention here.

In [13] authors investigated an influence of bottom-up attention on eye fixations of paintings. In their experiments, they recorded people viewing paintings and photographs in two conditions: free viewing and search task. To compute the saliency of images, authors used the well-known IttiKoch saliency model that captures bottom-up (stimulus-driven) saliency, based on local feature contrasts in color, orientation, and luminance. They were able to confirm the effect of visual saliency (bottom-up attention) on eye fixations during both free-viewing and search task.

Whether the artistic complexity can be predicted by saliency decided to explore authors in [8]. In this study, an artistic complexity measure based on information theory and visual saliency (calculated by the IttiKoch saliency model) were computed over paintings, and a comparison was performed. The most complex areas of paintings, were shown to be in many cases also the most salient locations. Thus, saliency calculation was revealed to be a relevant qualification of a paintings' complexity.

Building on the previous study, it was shown that saliency can be a predictor of an artistic movement of a painting [29]. Paintings that are characterized by a high level of abstraction (e.g. cubism or expressionism) result in higher variability in fixated areas, because it is difficult to understand the meaning of a painting right away. On the other hand, art movements that are very detailed, propose quite simple interpretation of the painting.

Further, the saliency detection was even applied in classification of professional photos and non-professional snapshots [36]. Authors developed a saliency-enhanced approach to predicting an aesthetic class of photographs. They build upon the fact that aesthetic objects are interesting and can attract attention. Moreover, aesthetically-pleasing photos – professional photos, intentionally direct the attention to the interesting subject. They assume that the aesthetic subject corresponds to salient locations, and taking advantage of the IttiKoch saliency model, were able to improve classification of professional and non-professional photos.

2.2 Evaluation of Saliency Approaches

With a rising number of new saliency models' architectures, a need for evaluation and selection process have emerged. Therefore, various evaluation techniques and many studies that provide elaborate comparison of multiple models are now available. Yet, very little of them specialize on evaluating the performance, or on development of new models, within the art domain. We mention some of such works here.

In [4] authors performed an exhaustive comparison of 35 state-of-the-art saliency models using multiple evaluation scores, as well as multiple datasets – synthetic patterns, natural images, and video datasets.

One of the most influential works is [7], where authors proposed a benchmark data set containing 300 natural mages with eye tracking data, and provided an online platform where new models can be uploaded and evaluated.

Another large comparison was done in [5], where 32 state-of-the-art saliency models underwent extensive analysis with multiple metrics, and were evaluated against numerous popular datasets.

We see the potential of computational saliency research in the context of art, but not many such works try study a variety of saliency models. Further, not many works that focus on variety of models, specialize in art. For this reason, we have decided to take an inspiration from both of these directions, and compare various saliency models against paintings.

3 Saliency Models

In this chapter we introduce saliency models that we decided to test. They are these 4 saliency models: A Model of Saliency-based Visual Attention for Rapid Scene Analysis, 1998 [18]; Random Center Surround Saliency, 2012 [34]; Saliency Attentive Model SAM-VGG, 2015 [10]; and Saliency Attentive Model SAM-ResNet, 2015 [10]. All of them, model the bottom-up (scene-dependent) visual attention and are task-independent. First two are based on traditional approach and last two are based on deep learning. Further, the output of every model is a saliency map with the same dimensions as the input image, and the code for their implementations is publicly available.

3.1 A Model of Saliency-based Visual Attention for Rapid Scene Analysis

The IttiKoch model [18] is inspired by the early primate visual system. The image is analyzed in terms of color, intensity and orientation, and according to these features it is decomposed into topographic feature maps. Every feature is computed in center-surround manner in order to simulate the sensitivity of neurons being higher in the center, and lower in the surround. One group of feature maps is representing the intensity, inspired by the mammals' vision sensitivity to light center and dark surround, and the opposite. The second group of feature maps is for color channels based on the receptive field neurons being excited by one color, and inhibited by a different color. Although in the surround - it is the reverse. The last, third group, is representing the orientation-sensitive neurons in primary visual cortex. These feature maps are then normalized in a way that deals with their initial incomparability, and the problem of strong salient regions in few maps, being masked by less salient regions in a larger number of maps. After that, they are combined into three conspicuity maps, which at the end are summed into the saliency map. At each time, the maximum of saliency map represents the most salient region of the image that captures the attention. This is possible thanks to the winner-take-all neural network that suppresses all locations except for the most salient one. The general architecture of the model is shown in Figure 2.

The IttiKoch model has become a basis for later models, and presently is a standard benchmark for comparison of saliency models. We chose the implementation by Jonathan Harel [16] (part of GBVS toolbox for Matlab).

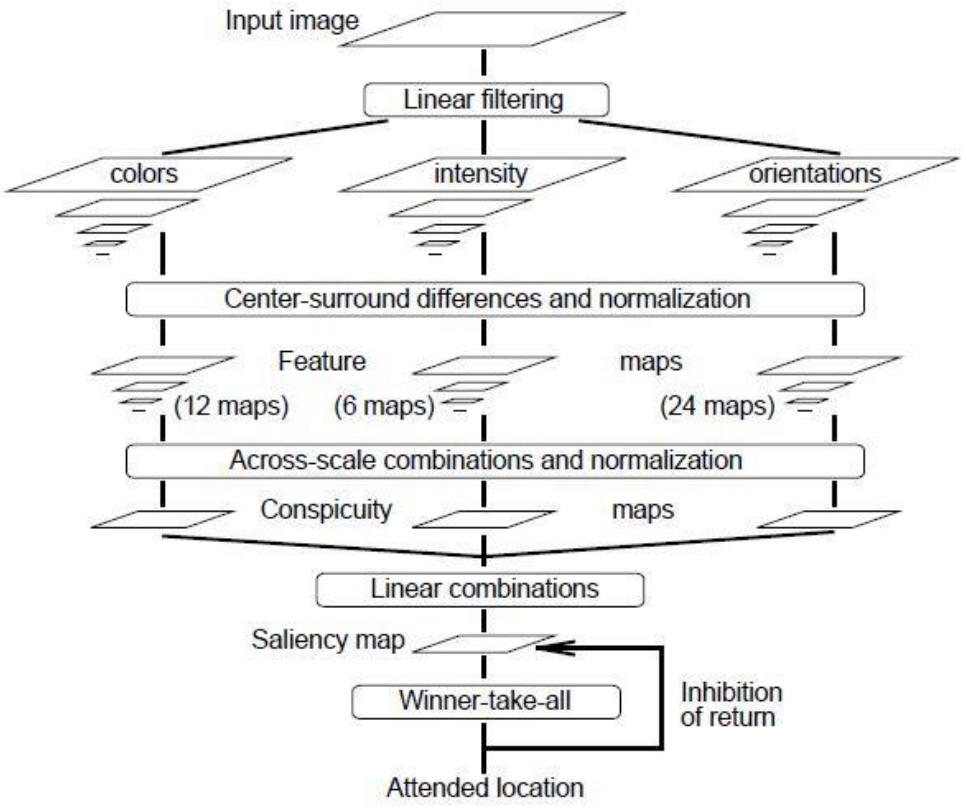

Figure 2 Architecture of IttiKoch model [18]

3.2 Random Center Surround Saliency (RCSS)

Based on findings showing that receptive fields of retina operate randomly at different spatial locations and scales [9], authors have proposed this method that computes saliency over random rectangular regions of interest. The RCSS model [34] is a center-surround method that computes local saliencies at random regions, unlike other models that operate globally. First, Gaussian filter is applied to the input image to remove noise and smooth the image. Then, the image is converted into the CIELAB color space, because of its similarity to human vision. After that, random rectangle windows are generated over L*, a* and b* channels. Local saliencies are computed over these rectangle regions of interests and for each channel

a saliency map is generated. At the end, the final saliency map is created by fusing these channel specific saliency maps using pixel-wise Euclidean norm. The general architecture of the model is shown in Figure 3.

Despite this model being quite simple, it achieves performance comparable to other existing methods and does not require any training. The RCSS model was implemented in MATLAB.

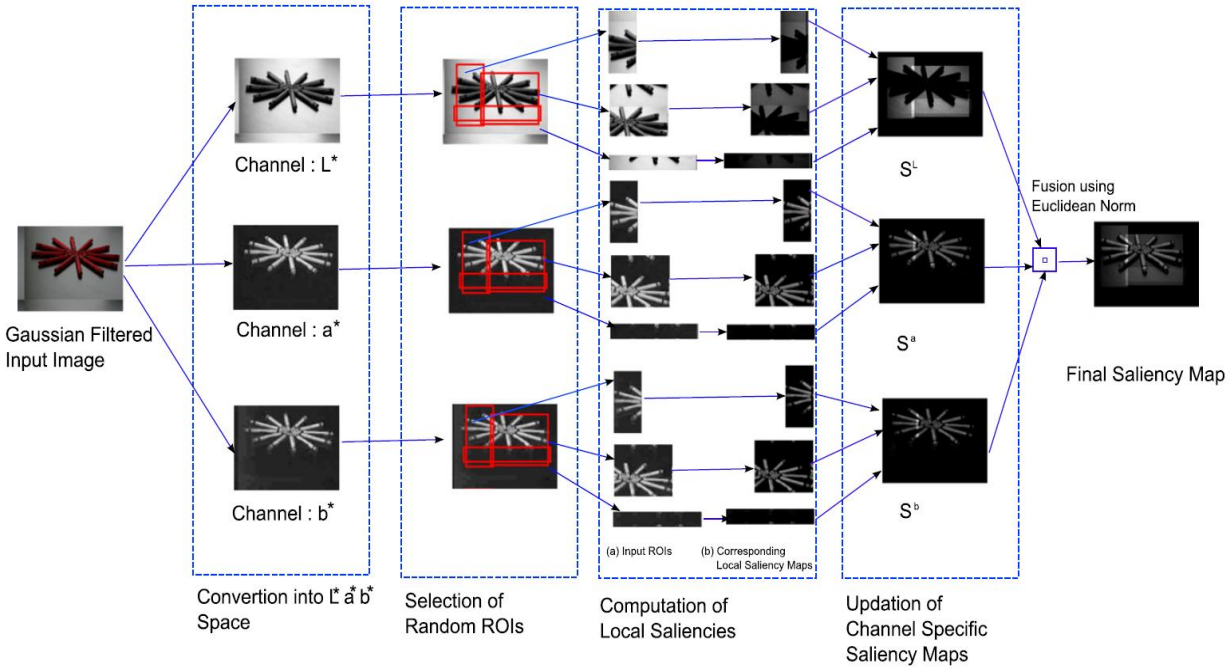

Figure 3 Architecture of RCSS model [34]

3.3 Saliency Attentive Models (SAM-VGG and SAM-ResNet)

Traditional saliency models incorporate low-level features such as color, contrast, orientation, or semantic concepts such as text or faces. These models however, still do not incorporate all of the mechanisms that contribute to visual saliency. How many or which mechanisms are really involved in saliency prediction is remaining an open question. With developments in deep neural networks research and increasing numbers and sizes of datasets, saliency predictions are quickly getting better. Yet, the authors of the SAM-VGG and the SAM-ResNet deep-learning models, were first in the field that decided to explore a machine attention mechanism in the saliency prediction [10].

Their model is composed of these three main components:

a) Dilated Convolutional Network

Usually deep saliency models are constructed with the pre-trained Convolutional Neural Network (CNN) that extracts feature maps from the input image. During this process the input is significantly rescaled, what in consequence worsens the prediction. Therefore, the authors decided to use a Dilated CNN architecture, instead of the standard CNN, thanks to which the predicted saliency maps are less rescaled. They propose two versions of the model based on different Dilated CNNs VGG-16 and ResNet-50, we will refer to them as SAM-VGG model and SAM-ResNet model respectively.

b) Attentive Convolutional LSTM

After the features are extracted from the input image, they are then passed to the Attentive Convolutional LSTM. It is composed of the Attentive model and ConvLSTM (Convolutional Long Short Term Memory) that are adapted in way, so that the feature maps can be processed by the attentive mechanism that at each step, focuses on different locations on the image. These feature maps are then iteratively fed as the input to the ConvLSTM. The result of this process are refined saliency features.

c) Learned Priors

The last component combines the output of the Attentive Convolutional LSTM with learned priors to incorporate the center bias. Usually saliency models include predefined priors, but in this model, they are automatically learned by the network from data.

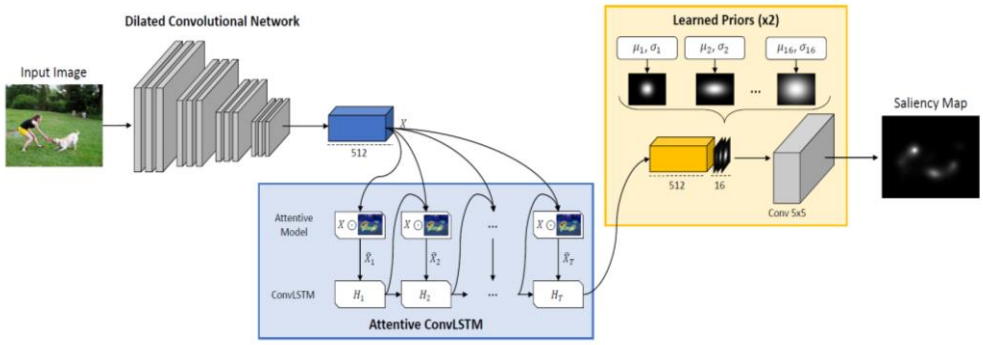

Figure 4 Architecture of SAM models [11]

The SAM model was trained on publicly available saliency prediction dataset SALICON [19] containing 20 000 images, and corresponding saliency maps computed from mouse movements. The general architecture of the model is shown in Figure 4. The model was implemented in Python.

4 Evaluation Metrics

For the evaluation of the saliency models, we decided to use following three metrics: Area Under ROC Curve, Normalized Scanpath Saliency, and Pearson's Correlation Coefficient. All of them differ in how they rank the performance of saliency models and how the ground truth is represented. Following [7] we can categorize them as **location-based** (Area Under ROC Curve, Normalized Scanpath Saliency), where the ground truth is represented as discrete fixation map; and **distribution-based** (Pearson's Correlation Coefficient), where the ground truth is represented as a continuous fixation map. In this chapter we describe the metrics that we chose.

4.1 Area Under ROC Curve (AUC)

Is a location-based metric originating from signal detection theory. A Receiver Operating Characteristic curve (ROC curve) illustrates the tradeoff between true and false positives at different thresholds. The Area Under the ROC curve (AUC) is commonly used evaluation metric in the saliency research. Over the years many different implementations of the classical approach were developed. We decided to use the Judd implementation [7] adapted to python by Dario Zanca et al. [37].

The saliency map is first normalized using Min-Max Feature Scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (1)$$

where x' is a normalized and x the original value.

For each given threshold, the true positive rate (TP rate) and false positive rate (TF rate) is calculated. True positives are salient map values above the threshold at locations of fixations, and the TP rate is then the ratio of the number of true positives to the number of all fixations. False positives are all other saliency map values (not at the locations of the fixations) above the threshold, and the FP rate is then the ratio of the number of false positives to the number of all saliency map pixels minus the number of fixations.

Threshold values are the values of the normalized saliency map at fixation locations.

Sweeping through all of the threshold values sorted in the ascending order, the TP rate is calculated as:

$$\frac{true\ positives}{all\ fixations} \tag{2}$$

and the FP rate is calculated as:

$$\frac{false\ positives}{all\ saliency\ map\ pixels - all\ fixations} \tag{3}$$

After that the ROC curve can be rendered and the final score is the area under it, computed using the Trapezoidal rule which is a technique to approximate the region under the graph function.

Ideally, if the saliency map correctly predicted all of and only the real fixations, the AUC score would be 1. The random classification would provide the score of 0.5 (TP rate = FP rate). Therefore, a positive saliency model score is above 0.5 and below 1.

4.2 Normalized Scanpath Saliency (NSS)

Another location-based metric that we decided to use to compare the accuracy of the models is the NSS [27]. It measures the correspondence of a saliency map and fixation map as the average saliency, computed from standardized saliency map values at fixation locations along a subjects' scanpath. Same as in the AUC case – many different versions of the NSS are available, and we decided to use the Bylinskii implementation [7] adapted to python by Dario Zanca et al. [37].

First, the saliency values are standardized to have zero mean and unit standard deviation:

$$Y_{SM}(x) = \frac{SM(x) - \mu}{\sigma} \tag{4}$$

where SM is the saliency map, Y_{SM} is the standardized saliency map,

μ is the mean expressed as:

$$\mu = \frac{1}{|I|}\sum_{t \epsilon I} SM(x_t) \tag{5}$$

and σ is the standard deviation:

$$\sigma = \sqrt{\frac{1}{|I|}\sum_{t\epsilon I}(SM(x_t) - \mu)^2} \quad (6)$$

where $|I|$ is the number of pixels of the picture.

Value $Y_{SM}(x_i)$ represents a difference of saliency map value x_i from the average saliency value in units of standard deviation. It is positive when saliency value at fixation location is above the mean saliency and negative when it is below the mean saliency.

The NSS score is then given by:

$$NSS = \frac{1}{N}\sum_i Y_{SM}(x_i) \times Q_i \quad (7)$$

where N is the number of fixations and Q is binary map of fixations. It is the mean of all standardized saliency values at fixation locations. For example, the score of 1 means that the saliency values at fixations were 1 unit standard deviation above the average saliency. More generally the correspondence between saliency and fixation map at chance is represented as 0 score, positive score means above chance and negative NSS represents anti-correspondence. Lower and upper limits of the NSS are theoretically$[-\infty, +\infty]$, while empirical limits depend on a particular dataset.

Different from the AUC, the NSS works with the actual saliency values and is more sensitive to false positives.

4.3 Pearson's Correlation Coefficient (CC)

This distribution-based metric measures linear correlation between two variables. In our case: between a continuous fixation map and a saliency map acquired from a model. For the computation of the CC score, we adapted the implementation from [7] to Python.

The CC score is defined as:

$$CC(P, Q^D) = \frac{\sigma(P, Q^D)}{\sigma(P) \times \sigma(Q^D)} \quad (8)$$

where:

- P is the saliency map;
- $Q^D$is the continuous fixation map;
- $\sigma(P, Q^D)$ is the covariance of P and Q^D;
- $\sigma(P)$ is the standard deviation of P; and
- $\sigma(Q^D)$ is the standard deviation of Q^D.

The CC score ranges between -1 and 1. The CC score of value 1 indicates a perfect linear correlation, value 0 indicates that there is no correlation. Values close to zero indicate poor correlation and value -1 means there is a perfect negative correlation.

5 Data

In this chapter we describe the data that saliency models will be evaluated against. We explain their contents, how they were collected, and in detail the process of preparing the data for the evaluation.

5.1 Data Collection

Saliency models are usually evaluated against eye tracking data. Eye fixations represent a ground truth of which regions of a visual scene grabbed viewers' attention. Whether the aim is to model eye movements during a task or not, the data to which the model will be compared, should be recorded during the task or free-viewing respectively. The data we use were originally collected for the purpose of a non-related empirical study. Given that the data were collected during a free-viewing of images, and we aim to model the visual attention without any task involved, we decided to use the data from this experiment.

5.1.1 The Experiment

We originally collected the data while collaborating on an experiment at Laboratory for Cognitive Research in Art History at university of Vienna. During the experiment participants viewed 14 different digitalized paintings, all representing the biblical scene of The Last Supper. The aim of the study was to investigate the relationship of eye movements with the perspectival space construction, and the plane surface composition of these paintings.

5.1.2 Participants and Stimuli

Our data consists of eye tracking recordings from 39 people looking at digitalized paintings. Primarily undergraduate students of art history at university of Vienna were recruited for this study. For stimuli we used high resolution photographs of 14 different paintings [1, 2] from 12^{th} to 16^{th} century, all representing the biblical scene of the Last Supper, chronologically ordered:

1. The Last Supper from the Verdun Alter, Nicholas Von Verdun, 1181 (Verdun)
2. The Last Supper, Giotto, 1306 (Giotto)
3. The Last Supper, Pietro Lorenzetti, 1320 (Lorenzetti)

4. The Last Supper, Andrea del Castagno, 1445-50 (Castagano)
5. The Last Supper from the Altarpiece of the Holy Sacrament, Dieric Bouts, 1465 (Bouts)
6. The Last Supper, Domenico Ghirlandaio, 1480 (Ghirlandaio)
7. Communion of the Apostles, Luca Signorelli, 1512 (Signorelli)
8. The Last Supper, Unknown Netherlandish Painter (Netherlandish)
9. The Last Supper, Lucas Cranach the Elder, 1547 (Cranach)
10. The Last Supper, Juan de Juanes, 1555 - 1562 (Juanes)
11. The Last Supper, Jacopo Tintoretto, 1578 (Tintoretto 1578)
12. The Last Supper, Paolo Veronese, 1585 (Veronese)
13. The Last Supper, Jacopo Tintoretto, 1592 (Tintoretto 1592)
14. The Last Supper, Federico Barocci, 1533-1612 (Barocci)

Participants looked at the paintings displayed on the LCD monitor with 2870 x 2159 pixels size screen. Eye-tracking data of the participants' dominant eyes were recorded on remote eye-tracker EyeLink 1000 Plus at 1000Hz frequency.

5.1.3 Experimental Setup and Procedure

After a colorblindness check, identification of the dominant eye, eyesight check, and calibration, participants viewed 14 painting seated 90 cm in front of the LCD screen. Every picture was displayed for 1 minute and followed by the displayed question: "How did you like this painting?" Participants were asked to answer this question by moving a computer mouse on Likert-type scale from 1 to 5 (1 meaning they did not like it at all and 5 meaning they liked it very much). The question was added in order to strengthen the aesthetical experience, and reinforce the belief that there is no task involved. After the question and before another painting was displayed, also another calibration-check was performed. After the viewing task participants were given the iPad with the same 14 pictures and asked to draw what in their mind are the most important lines of the composition of each painting. At last participants were given printed versions of these painting, and were asked to order them on the table according to perceived depth - from the least deep to the deepest painting in their opinion.

5.2 Data Preprocessing

After excluding the data that were incorrectly recorded, we ended up with data from 35 participants.

The preparation of the data can be divided into these main steps:

1. **Parsing the File**

   The data were in a form of ASC file for every participant. This file contains large amounts of information about the recording session. The eye tracker identifies different events such as fixations, saccades and blinks. Given that we are interested only in fixation events, we parsed the files excluding everything else. For each fixation, the file includes information such as: hundreds of gaze points corresponding to the particular fixation, start and end time, duration, pupil size and coordinates. Again, we excluded all information besides the coordinates of each fixation, to which painting they belong, and we also preserved their order.

2. **Excluding First Fixation**

   Afterwards, we rounded the fixation coordinates to have zero decimal places, and excluded the first fixation from each painting. After the pre-image calibration, in the first moment when image is displayed on the screen, the first fixation that eye tracker records, tends to be still at position of the calibration cross. Therefore it is recommended to ignore the first fixation, as it can wrongly indicate the salient location of the image that many people looked at.

3. **Choosing First 10 Fixations**

   Usually, saliency models are compared against first 3 to 6 fixations, because this amount best represents the early, rapid, scene-dependent saliency. The longer people view an image, the more their gaze differs, since higher cognitive processes start to get involved. We have decided to work with first 3 to 10 fixations from every participant for every painting. With higher quantity of fixations, we expect the effect of bottom-up saliency to lessen.

4. Computing Fixation Maps

From eye fixations we are able to create fixation maps that can then be compared with saliency maps acquired from models. These fixation maps are then taken as an input for saliency metrics. We computed two variants of fixation maps according to [23]:

a) Discrete Maps

A discrete fixation map f^i for the i^{th} observer is defined as:

$$f^i(x) = \sum_{k=1}^{M} \delta(x - x_{f^{(k)}}) \tag{9}$$

where:

- x is the spatial coordinates vector,
- $x_{f^{(k)}}$ is the spatial coordinates of the k^{th} visual fixation,
- M is the total number of i^{th} observers' fixations and
- $\delta(.)$ is the Kronecker delta.

A discrete fixation map f for N observers is defined as:

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f^i(x). \tag{10}$$

b) Continuous Maps

A continuous fixation map is created by convolving Gaussian function over the discrete fixation map f:

$$S(x) = f(x) * G_\sigma(x) \tag{11}$$

where σ is the standard deviation of the Gaussian. Typically, σ is set to 1º of the observers' visual angle as this should estimate the size of the fovea [33]. The actual value depends on the specific experimental setup and is computed from screen dimensions, resolution and the viewing distance.

To blur the fixations, we used AntonioGaussian filter from [7] with σ set to 70. In our setup we computed the 1º of visual angle to be 70 pixels. We computed the visual angle as is purposed in [7].

In total we have computed 1 discrete and 1 continuous map for every painting for every number of fixations from first 3 to first 10. Or in other words: for every number of fixations, we have computed 14 discrete and 14 continuous maps. This has left us with 122 discrete and 122 continuous maps.

6 Analysis and Results

In this chapter, we in detail describe steps of our analysis and the results. To evaluate how well a saliency model is able to predict true eye fixations, we compared previously created fixation maps (ground truth) with computed saliency maps. For the purpose of determining how much a saliency map and fixation map correspond, we have decided to employ three metrics: AUC-Judd, NSS, and CC. Each of them provides a different type of evaluation. We will elaborate more on the concrete results of each metric, possible influencing factors, as well as their advantages and weaknesses.

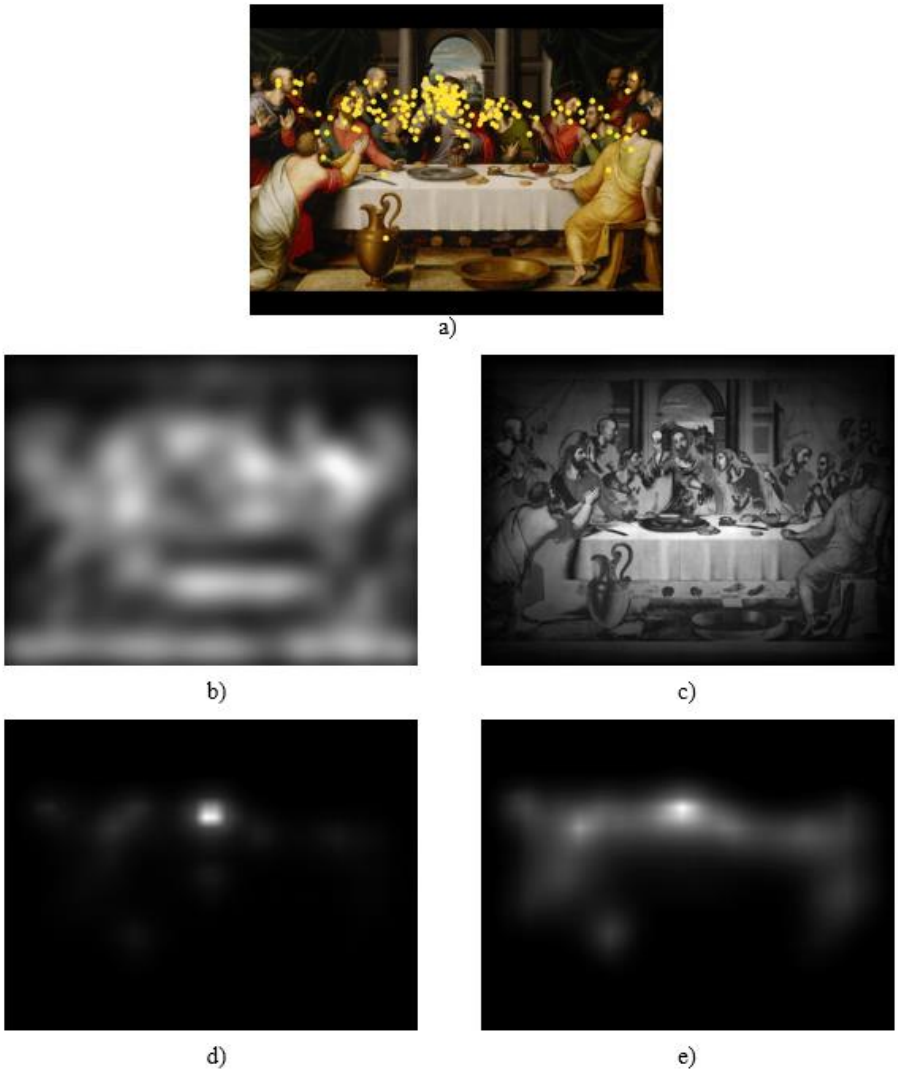

Figure 5 Eye fixations (a)) for the painting by Juanes and corresponding saliency map from b) IttiKoch model, c) RCSS model, d) SAM-VGG model, and e) SAM-ResNet model

The starting point of our analysis was to compute predictions. Each model takes as an input painting and returns prediction in a form of a saliency map. In Figure 5 we can see an example of a saliency map from each model for the painting by Juanes. In Figure 5 a) are shown first 7 eye fixations recorded while viewing the painting, in b) is shown saliency map from IttiKoch model, in c) is saliency map from RCSS model, in d) is saliency map from SAM-VGG model, and in d) is saliency map from SAM-ResNet model. Light parts represent salient locations. The lighter a location is, the more salient it is - the more likely are people to look there according to a model. Of course, everyone is different and no two people would look at a painting in the same way, nor would anyone twice look at a painting in the same way.

6.1 AUC Metric Results

The following section shows detailed results for the AUC metric displayed in separate table for each model. Each table shows AUC score for every painting including first 3, 4, 5, 6, 7, 8, 9, and first 10 participants' fixations. The highest and the lowest scores for each model are marked in green and red respectively. The absolute correspondence between fixations and saliency map would mean AUC score of 1 and a chance is at 0.5.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.68	0.64	0.61	0.61	0.60	0.61	0.61	0.61
Bouts	0.79	0.80	0.81	0.82	0.82	0.82	0.82	0.81
Castagno	0.58	0.58	0.57	0.57	0.55	0.54	**0.53**	**0.53**
Cranach	0.73	0.72	0.71	0.71	0.71	0.71	0.71	0.71
Ghirlandaio	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.55
Giotto	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71
Juanes	0.86	0.85	0.84	0.83	0.82	0.82	0.81	0.80
Lorenzetti	0.62	0.60	0.58	0.56	0.56	0.55	0.55	0.55
Netherlandish	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83
Signorelli	0.83	0.82	0.81	0.80	0.79	0.78	0.77	0.76
Tintoretto 1578	**0.89**	0.87	0.86	0.86	0.85	0.83	0.82	0.82
Tintoretto 1592	0.78	0.75	0.73	0.72	0.71	0.69	0.67	0.66
Verdun	0.58	0.58	0.58	0.59	0.59	0.60	0.60	0.60
Veronese	0.77	0.76	0.74	0.73	0.72	0.71	0.70	0.70

Table 1 AUC metric results for IttiKoch model

In Table 1 we can see that for the IttiKoch model the highest AUC score 0.89 was achieved when considering first 3 fixations of viewing the painting by Tintoretto 1578. On the other hand, the IttiKoch model obtained the lowest AUC score 0.53 considering first 9 and 10 fixations of viewing the painting by Castagno.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.77	0.73	0.70	0.68	0.66	0.66	0.65	0.64
Bouts	0.80	0.79	0.79	0.80	0.80	0.80	0.80	0.79
Castagno	0.68	0.68	0.68	0.68	0.65	0.65	0.64	0.64
Cranach	0.79	0.80	0.79	0.78	0.78	0.77	0.76	0.76
Ghirlandaio	0.69	0.70	0.69	0.68	0.67	0.67	0.66	0.65
Giotto	0.82	0.80	0.80	0.79	0.78	0.78	0.77	0.76
Juanes	0.72	0.71	0.70	0.69	0.68	0.68	0.67	0.67
Lorenzetti	0.81	0.77	0.74	0.72	0.71	0.70	0.69	0.69
Netherlandish	0.83	0.83	0.82	0.81	0.80	0.80	0.80	0.79
Signorelli	0.75	0.73	0.73	0.72	0.71	0.72	0.72	0.72
Tintoretto 1578	**0.84**	0.82	0.81	0.81	0.81	0.80	0.79	0.78
Tintoretto 1592	0.75	0.73	0.72	0.73	0.73	0.72	0.71	0.71
Verdun	0.77	0.76	0.76	0.75	0.74	0.74	0.74	0.74
Veronese	0.71	0.68	0.67	0.64	0.62	0.61	**0.60**	**0.60**

Table 2 AUC metric results for RCSS model

In Table 2 we can see AUC score for the RCSS model. The highest score it obtained is 0.84 what is slightly less than 0.89 in the case of IttiKoch model. However, the lowest AUC score is 0.60 when considering first 9 and 10 fixations of viewing the painting by Veronese, what is a better result than the lowest 0.53 AUC score for IttiKoch model. Both models scored the highest for the Tintoretto 1578 painting.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.81	0.78	0.76	0.75	0.73	0.73	0.72	**0.71**
Bouts	**0.95**	0.94	0.94	0.94	0.93	0.92	0.92	0.91
Castagno	0.88	0.88	0.87	0.85	0.84	0.83	0.82	0.81
Cranach	0.90	0.90	0.91	0.91	0.91	0.91	0.90	0.89
Ghirlandaio	0.83	0.83	0.82	0.80	0.79	0.78	0.77	0.76
Giotto	0.92	0.93	0.92	0.92	0.92	0.91	0.91	0.90

Juanes	**0.95**	0.94	0.93	0.92	0.91	0.89	0.87	0.86
Lorenzetti	0.93	0.91	0.87	0.84	0.81	0.78	0.78	0.76
Netherlandish	0.94	0.94	0.93	0.93	0.92	0.91	0.91	0.91
Signorelli	**0.95**	0.93	0.92	0.91	0.90	0.90	0.89	0.88
Tintoretto 1578	0.93	0.92	0.91	0.91	0.90	0.89	0.88	0.88
Tintoretto 1592	0.87	0.85	0.83	0.81	0.79	0.77	0.76	0.75
Verdun	0.94	0.93	0.92	0.92	0.91	0.91	0.91	0.91
Veronese	0.87	0.88	0.88	0.87	0.86	0.86	0.85	0.85

Table 3 AUC-Judd Metric Results for SAM-VGG Model

In Table 3 we can see that SAM-VGG model obtained the highest score *0.95* when considering first 3 fixations of viewing the painting by Bouts, Juanes and Signorelli and the lowest score *0.71* when considering first 10 fixations of viewing the painting by Barocci.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.84	0.82	0.79	0.77	0.75	0.75	0.74	**0.73**
Bouts	0.95	0.94	0.94	0.94	0.93	0.93	0.92	0.91
Castagno	0.91	0.91	0.90	0.90	0.89	0.89	0.88	0.88
Cranach	0.90	0.90	0.90	0.90	0.91	0.91	0.90	0.89
Ghirlandaio	0.87	0.88	0.87	0.86	0.84	0.83	0.81	0.80
Giotto	0.91	0.92	0.92	0.92	0.91	0.90	0.90	0.89
Juanes	0.95	0.95	0.94	0.94	0.92	0.90	0.89	0.88
Lorenzetti	0.94	0.92	0.88	0.86	0.83	0.82	0.80	0.79
Netherlandish	0.93	0.93	0.92	0.92	0.92	0.91	0.91	0.91
Signorelli	**0.96**	0.95	0.94	0.93	0.92	0.91	0.91	0.91
Tintoretto 1578	0.94	0.93	0.92	0.92	0.91	0.90	0.89	0.89
Tintoretto 1592	0.89	0.87	0.85	0.84	0.83	0.81	0.80	0.79
Verdun	0.93	0.92	0.91	0.91	0.90	0.90	0.90	0.89
Veronese	0.90	0.90	0.90	0.89	0.88	0.87	0.87	0.86

Table 4 AUC metric results for SAM-ResNet model

In Table 4 we can see that the SAM-ResNet model obtained the highest AUC score *0.96* when considering first 3 fixations of viewing the painting by Signorelli. The lowest score *0.73* this model obtained when considering first 10 fixations of viewing the painting by Barocci. On average SAM-ResNet model achieved the best results for the AUC metric than all the other models.

Limitation of the AUC metric is low-valued false positives. During the first few threshold values (at the peaks of the saliency map) the false positives are penalized the most, however while approaching the lower thresholds they are penalized less. Models that have a lot of low-valued false positives are not largely penalized. Therefore, a saliency map with many modest predictions and correctly predicting fixation locations with lower certainty, would score similarly, as saliency map with less, but more confident predictions in the same number of correctly predicted locations. For example, the saliency map in Figure 6 has many false positives and AUC score of *0.89*. Whereas the saliency map in Figure 7, has considerably less false positives, however scores in AUC metric only about 5% better with score of *0.94*. For comparison, the saliency map in Figure 6 has NSS score of *1.57* and the saliency map in Figure 7 scores *2.61*, which is a difference of about 39%.

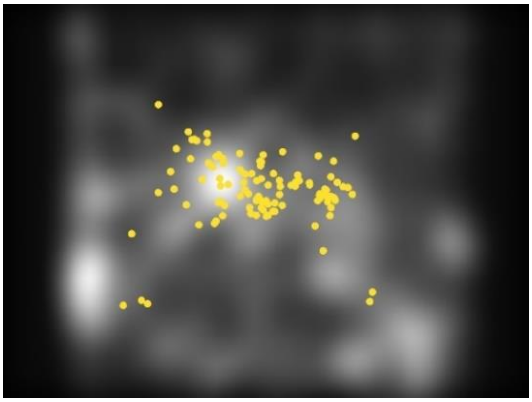

Figure 6 IttiKoch saliency map for painting by Tintoretto 1578 with overlaid first three fixations from every participant

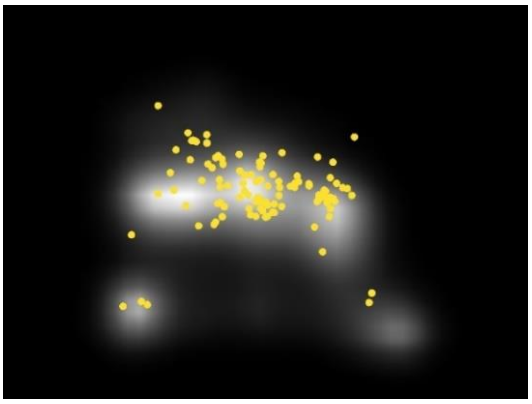

Figure 7 SAM-ResNet saliency map for painting by Tintoretto 1578 with overlaid first three fixations from every participant

Note: fixations shown in images are post-processed for illustration purposes and are not the exact data that we worked with.

6.2 NSS Metric Results

The following section shows detailed results of the NSS metric displayed in separate tables for each model. The highest and the lowest scores of each model are marked in red and green respectively. The higher NSS score, the more correspondence is between saliency map and ground truth fixations. The score of *0* represents a chance.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.66	0.52	0.42	0.39	0.38	0.41	0.41	0.41
Bouts	0.96	1.04	1.12	1.12	1.14	1.14	1.14	1.10
Castagno	0.26	0.27	0.24	0.23	0.21	0.18	0.16	0.15
Cranach	0.63	0.59	0.57	0.54	0.54	0.56	0.55	0.54
Ghirlandaio	**0.02**	0.03	0.03	0.03	0.04	0.05	0.07	0.07
Giotto	0.59	0.60	0.61	0.61	0.61	0.61	0.61	0.61
Juanes	1.30	1.30	1.27	1.25	1.19	1.16	1.14	1.12
Lorenzetti	0.30	0.23	0.21	0.16	0.14	0.14	0.13	0.12
Netherlandish	1.15	1.20	1.23	1.24	1.25	1.25	1.26	1.27
Signorelli	1.15	1.06	1.01	0.97	0.94	0.92	0.88	0.84
Tintoretto 1578	**1.57**	1.46	1.42	1.38	1.34	1.28	1.23	1.20
Tintoretto 1592	1.44	1.32	1.18	1.12	1.04	0.92	0.82	0.78
Verdun	0.30	0.29	0.30	0.31	0.33	0.34	0.35	0.34
Veronese	0.98	0.93	0.85	0.79	0.75	0.73	0.69	0.66

Table 5 NSS metric results for IttiKoch model

In Table 5 we can see results for the IttiKoch model. The highest NSS score of *1.57* this model obtained when considering first 3 fixations of viewing the painting by Tintoretto 1578. NSS score at almost a chance *0.02* obtained the IttiKoch model when considering first 3 fixations of viewing the painting by Ghirlandaio.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	1.10	0.87	0.74	0.66	0.63	0.65	0.62	0.59
Bouts	1.07	1.01	1.03	1.04	1.07	1.07	1.03	1.00

Castagno	0.48	0.49	0.48	0.46	0.39	0.38	0.34	0.36
Cranach	1.09	1.14	1.03	0.98	0.97	0.93	0.90	0.88
Ghirlandaio	0.57	0.61	0.61	0.55	0.53	0.50	0.48	0.45
Giotto	1.27	1.18	1.16	1.11	1.06	1.04	1.02	0.95
Juanes	0.58	0.57	0.54	0.49	0.47	0.47	0.46	0.43
Lorenzetti	1.15	1.01	0.87	0.80	0.76	0.73	0.69	0.68
Netherlandish	1.29	1.28	1.30	1.28	1.22	1.24	1.23	1.20
Signorelli	0.83	0.74	0.75	0.75	0.71	0.75	0.78	0.77
Tintoretto 1578	**1.32**	1.27	1.20	1.21	1.24	1.19	1.15	1.12
Tintoretto 1592	0.91	0.88	0.78	0.81	0.78	0.72	0.68	0.65
Verdun	0.94	0.89	0.86	0.83	0.79	0.79	0.77	0.77
Veronese	0.67	0.58	0.52	0.41	0.37	0.34	0.30	**0.28**

Table 6 NSS metric results for RCSS model

In Table 6 we can see results for the RCSS model. The highest NSS score that the model obtained is *1.32* what is worse compared to *1.57* for IttiKoch model. However, on average the RCSS model achieved better NSS results than the IttiKoch model. The worst this model performed when considering first 10 fixations of viewing the painting by Veronese with score of *0.28*.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	1.28	1.14	1.00	0.96	0.91	0.86	0.80	0.76
Bouts	2.73	2.66	2.55	2.43	2.37	2.35	2.23	2.10
Castagno	4.26	3.92	3.43	3.29	3.06	2.86	2.67	2.57
Cranach	1.23	1.48	1.60	1.59	1.76	1.83	1.70	1.61
Ghirlandaio	0.81	0.91	0.92	0.86	0.80	0.76	0.70	**0.66**
Giotto	2.62	2.67	2.67	2.61	2.48	2.38	2.31	2.24
Juanes	**4.36**	3.73	3.53	3.33	3.12	3.03	2.81	2.66
Lorenzetti	3.88	3.75	3.28	2.94	2.64	2.40	2.25	2.17
Netherlandish	2.77	2.69	2.66	2.65	2.67	2.60	2.57	2.50
Signorelli	4.35	3.41	3.03	2.70	2.55	2.53	2.38	2.24
Tintoretto 1578	2.23	2.07	1.99	1.98	1.95	1.85	1.74	1.69
Tintoretto 1592	2.08	1.90	1.66	1.58	1.39	1.22	1.11	1.05
Verdun	2.79	2.63	2.50	2.36	2.24	2.28	2.31	2.22
Veronese	1.31	1.40	1.49	1.44	1.41	1.36	1.33	1.33

Table 7 NSS metric results for SAM-VGG model

In Table 7 we can see that the best NSS score SAM-VGG model obtained is *4.36* when considering first 3 fixations of viewing the painting by Juanes, and the worst NSS score is *0.66* when considering first 10 fixations of viewing the painting by Ghirlandaio.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	1.45	1.26	1.11	1.04	0.99	0.98	0.95	**0.92**
Bouts	2.70	2.62	2.58	2.55	2.50	2.45	2.36	2.27
Castagno	**3.76**	3.64	3.31	3.18	3.03	2.89	2.74	2.66
Cranach	1.66	1.83	1.93	1.93	2.05	2.06	1.97	1.90
Ghirlandaio	1.65	1.77	1.74	1.63	1.54	1.46	1.37	1.29
Giotto	2.33	2.41	2.38	2.35	2.27	2.20	2.13	2.08
Juanes	3.46	3.24	3.12	2.99	2.84	2.74	2.59	2.47
Lorenzetti	2.68	2.53	2.27	2.06	1.89	1.75	1.67	1.60
Netherlandish	2.21	2.22	2.20	2.14	2.12	2.09	2.07	2.04
Signorelli	3.66	3.18	2.93	2.71	2.63	2.58	2.52	2.43
Tintoretto 1578	2.61	2.43	2.34	2.32	2.24	2.14	2.06	2.03
Tintoretto 1592	1.93	1.79	1.66	1.61	1.51	1.37	1.27	1.22
Verdun	2.01	1.95	1.90	1.86	1.81	1.82	1.82	1.80
Veronese	1.81	1.90	1.89	1.81	1.74	1.67	1.62	1.58

Table 8 NSS metric results for SAM-ResNet model

In Table 8 we can see the results for SAM-ResNet model. The best NSS score that this model obtained is *3.76* when considering first 3 fixations of viewing the painting by Castagno and the worst is *0.92* when considering first 10 fixations of viewing the painting by Barocci.

On average SAM-VGG model scored better than SAM-ResNet model and also better than all of the other models according to NSS metric.

One of the advantages of NSS metric is that it is sensitive to false positives. Many low false positives in NSS result in lowering the final score, unlike it is in the AUC metric. This is reflected in the difference among scores. Models with many false positives like IttiKoch and RCSS have an average NSS score of *0.72* and *0.82* respectively. While models with less false positives like SAM-VGG and SAM-ResNet have NSS score of *2.13* and *2.18* respectively (as can be seen in sTable 14).

6.3 CC Metric Results

The following section shows detailed results for the CC metric displayed in separate tables for each model. The highest and the lowest scores of each model are marked in red and green respectively. The best score would be 1 and values closer to 0 characterize poor performance.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.20	0.17	0.14	0.14	0.15	0.17	0.19	0.19
Bouts	0.25	0.29	0.32	0.33	0.36	0.38	0.40	0.41
Castagno	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06
Cranach	0.25	0.26	0.26	0.25	0.25	0.26	0.26	0.27
Ghirlandaio	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	0.03	0.04	0.05
Giotto	0.15	0.16	0.18	0.20	0.22	0.23	0.23	0.24
Juanes	0.34	0.37	0.39	0.41	0.41	0.42	0.44	0.44
Lorenzetti	0.11	0.10	0.10	0.09	0.08	0.08	0.08	0.07
Netherlandish	0.26	0.28	0.30	0.33	0.35	0.37	0.39	0.41
Signorelli	0.27	0.28	0.28	0.29	0.29	0.30	0.31	0.31
Tintoretto 1578	0.50	0.50	0.52	0.51	0.52	0.52	0.52	**0.53**
Tintoretto 1592	0.44	0.44	0.43	0.43	0.42	0.40	0.38	0.37
Verdun	0.12	0.13	0.14	0.15	0.16	0.17	0.17	0.17
Veronese	0.35	0.35	0.35	0.34	0.34	0.34	0.34	0.34

Table 9 CC metric results for IttiKoch model

In Table 9 we can see that the worst CC score for the IttiKoch model is *0.02* when considering first 3 to 7 fixations of viewing the painting by Ghirlandaio. As can be seen in Table 5 the IttiKoch model scored worst for Ghirlandaio painting also for the NSS metric. The best CC score *0.53* this model obtained when considering first 10 fixations of viewing the painting by Tintoretto 1578.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.32	0.30	0.29	0.29	0.30	0.32	0.33	0.33
Bouts	0.25	0.28	0.31	0.32	0.34	0.36	0.38	0.39
Castagno	0.22	0.22	0.24	0.23	0.22	0.22	0.23	0.23
Cranach	0.51	**0.53**	0.51	0.49	0.48	0.48	0.49	0.50
Ghirlandaio	0.31	0.30	0.31	0.31	0.32	0.33	0.34	0.35

Giotto	0.36	0.39	0.41	0.42	0.44	0.46	0.46	0.46
Juanes	0.18	0.18	0.17	0.17	**0.16**	0.17	0.17	0.18
Lorenzetti	0.39	0.39	0.40	0.40	0.40	0.41	0.41	0.40
Netherlandish	0.31	0.32	0.33	0.34	0.36	0.38	0.39	0.40
Signorelli	0.28	0.29	0.30	0.32	0.34	0.35	0.37	0.38
Tintoretto 1578	0.43	0.44	0.45	0.45	0.47	0.47	0.48	0.48
Tintoretto 1592	0.39	0.40	0.42	0.43	0.43	0.44	0.44	0.44
Verdun	0.28	0.29	0.30	0.32	0.33	0.33	0.33	0.33
Veronese	0.24	0.23	0.22	0.19	0.18	0.17	0.17	0.17

Table 10 CC metric results for RCSS model

In Table 10 we can see results for CC metric for the RCSS model. The worst obtained CC score is *0.16* when considering first 7 fixations of viewing the painting by Juanes, and the best CC score is *0.53* when considering first 4 fixations of viewing the painting by Cranach.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.39	0.36	0.36	0.37	0.37	0.38	0.38	0.37
Bouts	0.54	0.58	0.60	0.61	0.64	0.67	0.69	0.68
Castagno	0.83	0.82	0.81	0.82	0.80	0.79	0.78	0.77
Cranach	0.46	0.56	0.64	0.66	0.70	0.71	0.69	0.68
Ghirlandaio	**0.32**	0.35	0.37	0.37	0.37	0.37	0.37	0.38
Giotto	0.63	0.70	0.77	0.81	0.84	0.86	0.86	0.86
Juanes	0.81	0.81	0.82	0.82	0.82	0.84	0.83	0.82
Lorenzetti	**0.88**	**0.88**	0.87	0.85	0.83	0.82	0.81	0.80
Netherlandish	0.56	0.56	0.59	0.62	0.66	0.68	0.69	0.71
Signorelli	0.79	0.71	0.65	0.63	0.63	0.65	0.65	0.65
Tintoretto 1578	0.64	0.64	0.64	0.65	0.66	0.65	0.65	0.65
Tintoretto 1592	0.61	0.61	0.60	0.61	0.58	0.55	0.54	0.53
Verdun	0.76	0.77	0.77	0.77	0.75	0.78	0.80	0.79
Veronese	0.48	0.55	0.62	0.63	0.64	0.65	0.67	0.69

Table 11 CC metric results for SAM-VGG model

In Table 11 we can see results of the CC metric for the SAM-VGG model. The worst score *0.32* this model obtained when considering first 3 fixations of viewing the painting by Ghirlandaio and the best *0.88* when considering first 3 and 4 fixations of viewing the painting by Lorenzetti.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.46	**0.42**	**0.42**	**0.42**	0.43	0.46	0.47	0.47
Bouts	0.58	0.62	0.66	0.68	0.71	0.73	0.76	0.76
Castagno	0.83	0.85	0.87	0.89	0.89	0.90	**0.91**	0.90
Cranach	0.65	0.74	0.80	0.82	0.84	0.85	0.84	0.84
Ghirlandaio	0.58	0.61	0.62	0.63	0.63	0.64	0.66	0.66
Giotto	0.60	0.68	0.74	0.78	0.82	0.84	0.84	0.85
Juanes	0.76	0.82	0.84	0.86	0.87	0.87	0.88	0.88
Lorenzetti	0.72	0.72	0.72	0.71	0.71	0.71	0.72	0.71
Netherlandish	0.51	0.52	0.54	0.57	0.60	0.62	0.64	0.66
Signorelli	0.75	0.75	0.73	0.73	0.74	0.76	0.78	0.79
Tintoretto 1578	0.77	0.78	0.79	0.80	0.79	0.79	0.80	0.81
Tintoretto 1592	0.59	0.61	0.64	0.67	0.68	0.67	0.66	0.67
Verdun	0.62	0.64	0.66	0.67	0.67	0.69	0.70	0.71
Veronese	0.65	0.72	0.77	0.77	0.78	0.78	0.8	0.81

Table 12 CC metric results for SAM-ResNet model

In Table 12 we can see CC results for the SAM-ResNet model. The worst score is *0.42* when considering first 4 to 6 fixations of viewing the painting by Barocci and the best CC sore is *0.91* when considering first 9 fixations of viewing the painting by Castagno.

On average the SAM-ResNet model obtained the best results for the CC metric.

CC penalized false positives and negatives equally. This metric compares a continuous fixation map with a saliency map. High positive CC score is at locations where fixation map and saliency map have values of a similar intensity.

On the contrary to AUC and NSS metrics, where the best scores where reached always for first 3 fixations and had tendency to decrease with a number of fixations. For CC metric, in many cases, score increases with a number of fixations.

6.4 Summary of Results

In this section we summarize the results of our analysis and discuss influencing factors. Table 13 shows the results for the three metrics we used. The table shows scores for the first 3 fixations of viewing a painting. Colors from the lightest to darkest represent the scores from the worst to the best, for the particular metric. The darkest color, representing the best scores,

is mostly pointing to the SAM-ResNet for the AUC and CC metrics and to SAM-VGG for the NSS metric.

	AUC				NSS				CC			
	Itti Koch	RCSS	SAM-VGG	SAM-ResNet	Itti Koch	RCSS	SAM-VGG	SAM-ResNet	Itti Koch	RCSS	SAM-VGG	SAM-ResNet
Barocci	0.68	0.77	0.81	0.84	0.66	1.10	1.28	1.45	0.20	0.32	0.39	0.46
Bouts	0.79	0.80	0.95	0.95	0.96	1.07	2.73	2.70	0.25	0.25	0.54	0.58
Castagno	0.58	0.68	0.88	0.91	0.26	0.48	4.26	3.76	0.07	0.22	0.83	0.83
Cranach	0.73	0.79	0.90	0.90	0.63	1.09	1.23	1.66	0.25	0.51	0.46	0.65
Ghirlandaio	0.54	0.69	0.83	0.87	0.02	0.57	0.81	1.65	0.02	0.31	0.32	0.58
Giotto	0.71	0.82	0.92	0.91	0.59	1.27	2.62	2.33	0.15	0.36	0.63	0.60
Juanes	0.86	0.72	0.95	0.95	1.30	0.58	4.36	3.46	0.34	0.18	0.81	0.76
Lorenzetti	0.62	0.81	0.93	0.94	0.30	1.15	3.88	2.68	0.11	0.39	0.88	0.72
Netherland.	0.82	0.83	0.94	0.93	1.15	1.29	2.77	2.21	0.26	0.31	0.56	0.51
Signorelli	0.83	0.75	0.95	0.96	1.15	0.83	4.35	3.66	0.27	0.28	0.79	0.75
Tint.1578	0.89	0.84	0.93	0.94	1.57	1.32	2.23	2.61	0.50	0.43	0.64	0.77
Tint. 1592	0.78	0.75	0.87	0.89	1.44	0.91	2.08	1.93	0.44	0.39	0.61	0.59
Verdun	0.58	0.77	0.94	0.93	0.30	0.94	2.79	2.01	0.12	0.28	0.76	0.62
Veronese	0.77	0.71	0.87	0.90	0.98	0.67	1.31	1.81	0.35	0.24	0.48	0.65

Low Score  **High Score**

Table 13 Summary of results for first three fixations

Overall our analysis favors the SAM-ResNet saliency model according to AUC and CC metric. According to NSS metric the best performing model is SAM-VGG. The worst results achieved the IttiKoch saliency model in all metrics (sTable 14).

Model	AUC			NSS			CC		
	Average	Best	Worst	Average	Best	Worst	Average	Best	Worst
IttiKoch	0.70	0.89	0.53	0.72	1.57	0.02	0.26	0.53	0.02
RCSS	0.73	0.84	0.60	0.82	1.32	0.28	0.34	0.53	0.16
VGG	0.87	0.95	0.71	2.18	4.36	0.66	0.66	0.88	0.32
ResNet	0.89	0.96	0.73	2.13	3.76	0.92	0.71	0.91	0.42

sTable 14 Average scores of saliency models

6.4.1 Other Factors

It is important to first recognize, and then consider how results could have been influenced by various factors that could have affected our findings. Most important factors, that we have noticed, we will describe in this section.

6.4.1.1 Dataset

Several datasets are publicly available for either a comparison or training of saliency models. Most of them consist of natural images, but many of them contain images from various different categories such as Cartoon, Sketch, Low Resolution, Object, Pattern or Art. Some of the datasets are recorded by an eye tracker and some are computed from mouse movements. Depending on a dataset, models provide different results.

The dataset that we used is also different from the others. Apart from images having the same topic – the scene of Last Supper, our dataset consists of images with specific characteristics. Paintings are not an accurate representation of the world, rather are a product of painters' skills, painting materials and individual style. This was important to take into consideration. None of the models was designed specifically for paintings, however SAM models were trained on eye-tracking datasets that contained art stimuli.

6.4.1.2 Center Bias

Human fixations are biased to be located near the center of the image [30]. On a center-biased dataset a model that for example makes its predictions based only on the center bias, does not even have to consider the image content, and will score higher. Our dataset consists of paintings with intentional and well-considered compositions designed by the painter. Many of the paintings (not all) depict Jesus, as the most important part of the scene, at the center – making our dataset quite center biased. In addition, all of our models incorporate center bias in their predictions.

6.4.1.3 Bottom-Up Approach vs. Deep Learning

As our analysis pointed out, bottom-up models IttiKoch and RCSS performed worse than deep-learning SAM models. We will expand more on what are the possible reasons for our analysis favoring deep-learning models.

Visual attention is a result of combining both bottom-up and top-down information. However, the main approach in modelling visual attention, was to implement only bottom-up or only top-down information in a model. Bottom-up models, based on features, were extensively studied. But there was less research done on top-down models, which are based on learning. With an arrival of deep learning, focus shifted to deep-learning saliency models that inherently incorporate top-down information. These models in many cases improved saliency detection, by using features trained on object recognition, since top-down information as objects or texts, are known to be highly salient. Research suggests that objects (top-down) may be even more important for saliency detection than bottom-up information [12].

Yet, authors in [22] showed that in some cases bottom-up models outperformed deep-learning models. Mainly, when pictures consisted of less top-down information, suggesting that deep-learning models neglect the bottom-up element of visual attention. Maybe these models were trained to rather detect objects than the real saliency.

Whether the deep learning changed saliency models into simple weighted object detectors, decided to investigate authors in [21]. They showed that while an importance of objects in saliency detection is high, bottom-up information are important too. If deep-learning saliency models are very accurate in detecting top-down features, they tend to miss bottom-up information that are salient. For example, if the image is crowded and complex, it may be difficult to identify top-down information as faces, so bottom-up features become more important.

Thanks to the aforementioned findings, we suspect that one of the reasons for the superior performance of the deep-learning models on our dataset, is the high number of objects (persons, faces, etc.) in the paintings. None of the paintings is abstract, each one of them portray characters and objects. If a strong facet of deep-learning models is object detection, then our paintings are a great match.

7 Incorporating Face Detection

We hypothesize that if object detection is a significant part of the reason for SAM models' superior performance, if we add such mechanism to RCSS model, it will improve its results. Since, the theme of our images revolves around people, the prevalent objects are faces. Therefore, we decided to incorporate a face detection algorithm to the RCSS model.

7.1 The Viola-Jones Algorithm

In 2001 Paul Viola and Michael Jones developed a general object detection framework that is able to process images very quickly and with high detection rates [35]. Although, it can be used to various detection problems, it was motivated by, and demonstrated on, the task of face detection. This machine learning approach is to this day one of the most popular face detection algorithms. It is based on three kinds of Haar-like features (see Figure 8): two-rectangle (A, B), three-rectangle (C) and four-rectangle features (D). The value of a feature is calculated as a difference between the sum of all pixels in the dark rectangle and the sum of all pixels in the light rectangle.

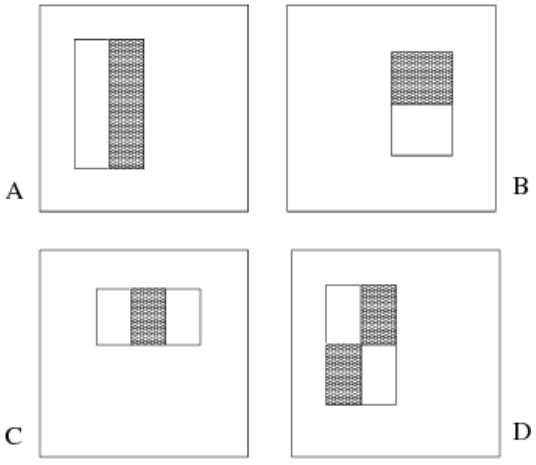

Figure 8 Types of Haar-like features used in Viola-Jones algorithm [35]

Moreover, the speed of evaluation of features is improved by converting the image into an integral image representation. Given, that the image is converted to integral image, the Haar-like features can then be evaluated in constant time at any location or scale. The training phase consists of a selection of Haar-like features and combining them into a cascade classifier. This is done by using AdaBoost algorithm that selects only the most important Haar-like feature

from all possible features and creates a strong classifier. This classifier consists of weak classifiers that are arranged in a way that many negative computations can be rejected sooner with simpler classifiers, and therefore make the algorithm computationally efficient.

7.2 Integrating Viola-Jones into RCSS

Firstly, we used the Viola-Jones algorithm to detect faces on paintings. We chose an implementation from OpenCV library with pre-trained cascade classifiers from file "*haarcascade_frontalface_default.xml*". To detect the faces we used function detectMultiScale() and we set these parameters:

- **scaleFactor** – defines how much is size of an image reduced at each scale. By resizing the image, we make the detection scale-invariant and smaller and larger faces can be detected by the same detection window.
  We chose a value 1.1, which means the size is reduced by 10% at each step. The higher the value, the faster a detection is, but also the chance of matching a face is reduced.
- **minNeighbours –**defines a minimal number of neighboring rectangles for a candidate rectangle to preserve it.
  We chose a value 5, what results in less but more accurate detections.

The output of this function is an array of detected rectangles' coordinates from which we created mask image. Afterwards, we blurred the image with Gaussian with kernel size set to 101x101 and standard deviation in X direction of 70. We created such images for every painting.

Given, that our data are portraying faces slightly different from reality, it was expected that the algorithm may give worse results than if would on regular images. Moreover, the classifier we used, was designed to detect frontal faces. But of course, our paintings portray faces in many different angles.

The overall accuracy of the face detection can be seen in Table 15. True positives are instances when the algorithm correctly detected a face, and false positives are instances when the algorithm detected face in a place where there was no face.

	Viola-Jones	
Painting	**True Positives**	**False Positives**
Barocci	4	7
Bouts	9	11
Castagno	2	0
Cranach	3	5
Ghirlandaio	7	4
Giotto	4	1
Juanes	7	5
Lorenzetti	3	2
Netherlandish	4	12
Signorelli	4	9
Tintoretto 1578	4	5
Tintoretto 1592	1	7
Verdun	2	3
Veronese	2	6

Table 15 Accuracy of Viola-Jones algorithm on paintings

In the RCSS model, after local saliencies of random rectangles are computed, channel-specific saliency maps are created. These are consequently fused into the final saliency map. At this point, we added our mask image with detected faces to the final saliency map. Our image has pixel values 255 (white) at rectangles and 0 (black) everywhere else. We added these to pixel values of saliency map and at the end normalized values of the saliency map again to be in a range from 0 to 255. Example of a final saliency map can be seen in Figure 9.

Figure 9 Example of RCSS-Viola saliency map

Afterwards, we evaluated these new saliency maps against the eye-tracking data, the same way as we did for other models' saliency maps. Detailed results for every metric can be seen in the next section in Tables 16 to 19.

7.3 RCSS-Viola Model Results

This section contains detailed results for the RCSS-Viola model displayed in separate table for each metric. Each table shows score for every painting including first 3, 4, 5, 6, 7, 8, 9, and first 10 participants' fixations. The best and the worst scores for each metric are marked in green and red respectively.

In Table 16 we can see AUC results for the RCSS-Viola model. The highest score is *0.87* when considering first 3 fixations of viewing the painting by Bouts and the lowest AUC sore is *0.60* when considering first 10 fixations of viewing the painting by Veronese.

Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.81	0.77	0.74	0.72	0.70	0.69	0.68	0.67
Bouts	**0.87**	0.86	0.86	0.86	0.86	0.86	0.85	0.84
Castagno	0.68	0.68	0.69	0.68	0.66	0.65	0.64	0.64
Cranach	0.78	0.79	0.78	0.77	0.77	0.77	0.76	0.76
Ghirlandaio	0.72	0.73	0.74	0.73	0.72	0.71	0.70	0.68
Giotto	0.86	0.84	0.82	0.82	0.80	0.80	0.79	0.78
Juanes	0.84	0.83	0.82	0.80	0.79	0.78	0.77	0.76
Lorenzetti	0.78	0.75	0.72	0.70	0.69	0.68	0.67	0.67
Netherlandish	0.81	0.81	0.80	0.79	0.78	0.79	0.79	0.78
Signorelli	0.76	0.76	0.76	0.75	0.74	0.75	0.75	0.75
Tintoretto 1578	0.83	0.81	0.80	0.80	0.81	0.79	0.78	0.78
Tintoretto 1592	0.76	0.74	0.73	0.74	0.73	0.72	0.72	0.71
Verdun	0.77	0.77	0.76	0.76	0.75	0.75	0.75	0.75
Veronese	0.71	0.69	0.67	0.64	0.63	0.62	0.61	**0.60**

Table 16 AUC metric results for RCSS-Viola model

In Table 17 we can see NSS results for the RCSS-Viola model. The highest score is *1.62* when considering first 3 fixations of viewing the painting by Giotto and the lowest NSS sore is *0.37* when considering first 10 fixations of viewing the painting by Veronese. Next, in Table 18 we can see that the highest CC score is *0.50* when considering first 10 fixations of viewing the

painting by Cranach and the lowest CC sore is *0.19* when considering first 8 to 10 fixations of viewing the painting by Veronese.

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	1.24	1.10	0.97	0.89	0.85	0.81	0.77	0.72
Bouts	1.41	1.36	1.35	1.37	1.37	1.34	1.29	1.28
Castagno	0.63	0.65	0.66	0.63	0.56	0.54	0.51	0.52
Cranach	0.95	1.02	0.96	0.95	0.95	0.94	0.93	0.93
Ghirlandaio	0.84	0.87	0.90	0.86	0.82	0.78	0.74	0.70
Giotto	**1.62**	1.55	1.46	1.39	1.32	1.30	1.25	1.21
Juanes	1.52	1.40	1.39	1.34	1.29	1.26	1.2	1.17
Lorenzetti	0.96	0.86	0.74	0.67	0.65	0.62	0.60	0.60
Netherlandish	0.98	1.02	1.00	0.97	0.96	0.97	0.98	0.97
Signorelli	0.94	1.02	1.03	1.01	0.98	0.99	1.02	1.02
Tintoretto 1578	1.13	1.06	1.04	1.05	1.05	1.01	0.98	0.96
Tintoretto 1592	0.91	0.84	0.82	0.85	0.83	0.80	0.78	0.75
Verdun	1.06	1.04	1.04	1.01	0.95	0.98	1.01	1.00
Veronese	0.74	0.65	0.58	0.49	0.45	0.41	0.38	**0.37**

Table 17 NSS metric results for RCSS-Viola model

	Number of first fixations							
Painting	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Barocci	0.37	0.36	0.35	0.35	0.35	0.36	0.37	0.38
Bouts	0.29	0.32	0.34	0.34	0.38	0.40	0.42	0.43
Castagno	0.23	0.23	0.25	0.25	0.25	0.25	0.25	0.26
Cranach	0.45	0.48	0.48	0.48	0.47	0.48	0.49	**0.50**
Ghirlandaio	0.34	0.33	0.34	0.34	0.37	0.38	0.40	0.41
Giotto	0.38	0.40	0.43	0.43	0.46	0.47	0.47	0.48
Juanes	0.36	0.38	0.39	0.39	0.40	0.41	0.42	0.42
Lorenzetti	0.32	0.32	0.33	0.33	0.33	0.34	0.34	0.34
Netherlandish	0.24	0.24	0.25	0.25	0.28	0.30	0.31	0.32
Signorelli	0.27	0.31	0.33	0.33	0.36	0.38	0.40	0.41
Tintoretto 1578	0.36	0.37	0.38	0.38	0.40	0.41	0.42	0.42
Tintoretto 1592	0.33	0.35	0.38	0.38	0.41	0.43	0.45	0.46
Verdun	0.32	0.34	0.36	0.36	0.36	0.38	0.39	0.39
Veronese	0.25	0.25	0.24	0.24	0.20	**0.19**	**0.19**	**0.19**

Table 18 CC metric results for RCSS-Viola model

Overall, the incorporated face detection improved RCSS models' results. In some cases, more than in others. How much the performance improved depends also on how many faces were detected. For example, in the case of Tintoretto 1592 painting, we can see in Table 15, that the least number of faces (one) was detected. Moreover, seven false positive faces were detected in this painting. This even worsened the results for this painting for the CC metric for the first 3 to 8 fixations. Generally, the results were very slightly improved (see Table 19). However, the performance of deep-learning models was not reached.

Model	AUC			NSS			CC		
	Average	**Best**	**Worst**	**Average**	**Best**	**Worst**	**Average**	**Best**	**Worst**
IttiKoch	0.70	0.89	0.53	0.72	1.57	0.02	0.26	0.53	0.02
RCSS	0.73	0.84	0.60	0.82	1.32	0.28	0.34	0.53	0.16
RCSS-Viola	0.75	0.87	0.60	0.95	1.62	0.37	0.35	0.50	0.19
VGG	0.87	0.95	0.71	2.18	4.36	0.66	0.66	0.88	0.32
ResNet	0.89	0.96	0.73	2.13	3.76	0.92	0.71	0.91	0.42

Table 19 Average scores of saliency models with RCSS-Viola model

Conclusion

Despite the topic of visual saliency being relevant in many different fields of research, in our work, we focused mainly on the intersection of the two – saliency modelling and art. Models of visual saliency try to predict where people look, or what the most salient parts of a visual scene are. In the current work, we outlined the topic of saliency in different disciplines, described related works, pre-processed eye-tracking data, explained evaluation metrics, analyzed saliency models, and discussed the results.

We decided to evaluate the performance of these four saliency models: IttiKoch, RCSS, SAM-VGG and SAM-ResNet. These models take as input an image and return a corresponding saliency map with the same dimensions. For the evaluation we used AUC, NSS and CC metric. First, we had to pre-process the raw eye-tracking data consisting of eye fixations recorded while free-viewing 14 versions of digitized paintings, all portraying the biblical scene of The Last Supper. From these data, we created discrete and continuous fixation maps. Then, using the metrics we compared the saliency maps obtained from the saliency models with fixation maps that we created. Our analysis shows that according to AUC and CC metrics the SAM-ResNet model best predicts where people look on paintings in our dataset. According to NSS metric, the SAM-VGG model makes the best predictions. Overall, the deep-learning models (SAM-VGG and SAM-ResNet) showed better results than models based on traditional approach (IttiKoch and RCSS). Moreover, we achieved slightly better performance of RCSS model by incorporating Viola-Jones face detection algorithm. Despite that our alteration of the RCSS model improved its performance, it did not reach the performance of deep-learning models.

We suspect that one of the reasons for deep-learning models' superiority might be the implicit object detection, and the fact that our dataset consists entirely of figurative art. For further research we suggest expanding the dataset to other types of paintings, for example abstract paintings. As these tend to contain fewer distinct objects, the deep-learning models may no longer be in such an advantage. In addition, dataset with greater variety of paintings and with more non-center-biased paintings, could provide different results.

We consider research in this direction valuable, as the impact of computational methods spreads into every field, providing new perspectives or making tasks easier. Employing saliency models in art domain has proven to be beneficial, what we also pointed out by mentioning few interesting works. Research in this area will need more datasets with such specific images as paintings or in general visual art. Our findings suggest that next direction of research could yield interesting results by focusing on deep-learning saliency approaches within the art.

References

[1] Artstor. http://www.artstor.org. Accessed: November, 2018.

[2] prometheus. https://prometheus-bildarchiv.de. Accessed: November, 2018.

[3] 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.

[4] Ali Borji, Dicky N Sihite, and Laurent Itti. Quantitative analysis of human model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing, 22(1):55-69, 2012.

[5] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. In Proceedings of the IEEE international conference on computer vision, pages 921-928, 2013.

[6] Guy Thomas Buswell. How people look at pictures: a study of the psychology and perception in art. Univ. Chicago Press, 1935.

[7] 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.

[8] Matteo Cacciola, Gianluigi Occhiuto, and Francesco Carlo Morabito. Artistic complexity and saliency: Two faces of the same coin? International Journal of Information Acquisition, 9(02):1350010, 2013.

[9] Carol L. Colby and Michael E. Goldberg. Space and attention in parietal cortex. Annual Review of Neuroscience, 22(1):319-349, 1999.

[10] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing, 27(10):5142-5154, 2018.

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Sam: Pushing the limits of saliency prediction models. 06 2018.

[12] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. Journal of Vision, 8(14):18-18, 2008.

[13] Isabella Fuchs, Ulrich Ansorge, Christoph Redies, and Helmut Leder. Salience in paintings: bottom-up influences on eye fixations. Cognitive Computation, 3(1):25-36, 2011.

[14] Michael Gazzaniga, Richard Ivry, and George Mangun. Cognitive Neuroscience: The Biology of the Mind, Fourth Edition. New York: W. W. Norton & Company, Inc., 01 2014. ISBN: 978-0-393-91348-4.

[15] Jacqueline P Gottlieb, Makoto Kusunoki, and Michael E Goldberg. The representation of visual salience in monkey parietal cortex. Nature, 391(6666):481-484, 1998.

[16] J. Harel. A saliency implementation in matlab. 2012.2

[17] L. Itti. Visual salience. Scholarpedia, 2(9):3327, 2007. revision #72776.

[18] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1254-1259, Nov 1998.

[19] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. 06 2015.

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

[21] Phutphalla Kong, Matei Mancas, Nimol Thuon, Seng Kheang, and Bernard Gosselin. Do deep-learning saliency models really model saliency? In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 2331-2335. IEEE, 2018.

[22] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In Proceedings of the IEEE International Conference on Computer Vision, pages 4789-4798, 2017.

[23] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behavior research methods, 45(1):251-266, 2013.

[24] Lucia Melloni, Sara van Leeuwen, Arjen Alink, and Notger G. Muuller. Interaction between Bottom-up Saliency and Top-down Control: How Saliency Maps Are Created in the Human Brain. Cerebral Cortex, 22(12):2943-2952, 01 2012.

[25] E. Niebur. Saliency map. Scholarpedia, 2(8):2675, 2007. revision #147400.

[26] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. Vision research, 42(1):107-123, 2002.

[27] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. Vision research, 45(18):2397-2416, 2005.

[28] Raphael Rosenberg and Christoph Klein. The moving eye of the beholder: Eye tracking and the perception of paintings. In Joseph P. Huston, Marcos Nadal, Francisco Mora, Luigi F. Agnati, and Camilo J. Cela-Conde, editors, Art, Aesthetics, and the Brain, chapter 5, pages 79-108. Oxford University Press, 2015.

[29] Bogdan Stoica, Laura Florea, Alexandra Badeanu, Andrei Racoviteanu, Iulian Felea, and Corneliu Florea. Visual saliency analysis in paintings. In 2017 International Symposium on Signals, Circuits and Systems (ISSCS), pages 1-4. IEEE, 2017.

[30] Benjamin 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):4-4, 2007.

[31] Kirk G Thompson and Narcisse P Bichot. A visual salience map in the primate frontal eye field. Progress in brain research, 147:249-262, 2005.

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

[33] Boris Velichkovsky, Marc Pomplun, and Johannes Rieser. Attention and communication: Eye-movement-based research paradigms. In W.H. Zangemeister, H.S. Stiehl, and C. Freksa, editors, Visual Attention and Cognition, volume 116 of Advances in Psychology, pages 125-154. North-Holland, 1996.

[34] Tadmeri Narayan Vikram, Marko Tscherepanow, and Britta Wrede. A random center surround bottom up visual attention model useful for salient region detection. In 2011 IEEE Workshop on Applications of Computer Vision (WACV), pages 166-173. IEEE, 2011.

[35] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, volume 1. IEEE, 2001.

[36] Lai-Kuan Wong and Kok-Lim Low. Saliency-enhanced image aesthetics class prediction. In 2009 16th IEEE International Conference on Image Processing (ICIP), pages 997-1000. IEEE, 2009.

[37] Dario Zanca, Valeria Serchi, Pietro Piu, Francesca Rosini, and Alessandra Rufa. FixaTons: A collection of human fixations datasets and metrics for scanpath similarity. CoRR, abs/1802.02534, 2018.

[38] Xilin Zhang, Li Zhaoping, Tiangang Zhou, and Fang Fang. Neural activities in v1 create a bottom-up saliency map. Neuron, 73(1):183-192, 2012.