WYSIWYG: What You See Is Where Your Gaze

Raphaëlle Lemaire*
raphaelle.lemaire@unicaen.fr
Normandie Caen Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000 Caen, France

Azamat Kaibaldiyev*
azamat.kaibaldiyev@unicaen.fr
Normandie Caen Univ, GREYC
14000 Caen, France

Eléonore Mariette
eleonore.mariette@unicaen.fr
Normandie Caen Univ, phIND - Inserm - UMR-S 1237, Caen, France

Débora Viglieri
debora.viglieri@unicaen.fr
Normandie Caen Univ, NIMH - Inserm - UMR-S 1077
14000 Caen, France

Alexis Lechervy
alexis.lechervy@unicaen.fr
Normandie Caen Univ, GREYC
14000 Caen, France

Fabrice Maurel
fabrice.maurel@unicaen.fr
Normandie Caen Univ, GREYC
14000 Caen, France

Gaël Dias
gael.dias@unicaen.fr
Normandie Caen Univ, GREYC
14000 Caen, France

Jérémie Pantin
jeremie.pantin@unicaen.fr
Normandie Caen Univ, GREYC
14000 Caen, France

Gaëtane Blaizot
gaetane.blaizot@unicaen.fr
Normandie Caen Univ, BB@C, GIP Cyceron
14000 Caen, France

Véronique Agin
veronique.agin@unicaen.fr
Normandie Caen Univ, phIND - Inserm - UMR-S 1237
14000 Caen, France

Nicolas Poirel
nicolas.poirel@u-paris.fr
René Descartes Univ, LaPsyDÉ UMR CNRS 8240
75000 Paris, France

Eric Bui
bui-th@chu-caen.fr
Normandie Caen Univ, phIND - Inserm - UMR-S 1237
14000 Caen, France

Hervé Platel
herve.platel@unicaen.fr
Normandie Caen Univ, NIMH - Inserm - UMR-S 1077
14000 Caen, France

Denis Vivien
vivien@cyceron.fr
Normandie Caen Univ, phIND - Inserm - UMR-S 1237
14000 Caen, France

Youssef Chahir
youssef.chahir@unicaen.fr
Normandie Caen Univ, GREYC
14000 Caen, France

Abstract

In the words of Picasso, *a painting lives only through the one who looks at it.* To materialize this thought, we propose to automatically produce artworks, which present visual transformations of paintings, where the most observed areas (by human viewers) are amplified and distorted. Our work is grounded in a study conducted at the Caen Museum of Fine Arts in France, which aims to assess the perceived well-being associated with museum visits. During the study, 151 participants were equipped with eye-tracking glasses, and observed various paintings, first alone and then in pairs. Based on the fixation and gaze path stored data, we first generate saliency maps that reflect the visual attention given to each painting of the Museum. These maps are then used to fine-tune the UNETRSal model, a neural network designed to predict saliency maps, in order to align its outputs with human visual patterns observed during the experiment. The saliency maps generated by UNETRSal are subsequently used to create deformations of the original painting. This overall process gives rise to a new artwork born from the interaction between human gaze and artificial intelligence prediction.

Keywords

Eye-tracking, Saliency prediction, Visual attention, Gaze-driven generation, Art-based transformation.

ACM Reference Format:
Raphaëlle Lemaire*, Azamat Kaibaldiyev*, Eléonore Mariette, Débora Viglieri, Alexis Lechervy, Fabrice Maurel, Gaël Dias, Jérémie Pantin, Gaëtane Blaizot, Véronique Agin, Nicolas Poirel, Eric Bui, Hervé Platel, Denis Vivien, and Youssef Chahir. 2025. WYSIWYG: What You See Is Where Your Gaze. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* , 5 pages. https://doi.org/10.1145/3746027.3756140

1 Introduction

Art is a lie that makes us realize truth, said Picasso, capturing the tension between perception and interpretation that lies at the heart of visual experience. This interplay becomes especially compelling when computational models attempt to understand, or emulate, how humans engage with visual art [7]. Our study explores this

*Raphaëlle Lemaire and Azamat Kaibaldiyev contributed equally to this work.

MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3756140

idea by projecting human perception into paintings to produce new artworks through targeted deformations of the original images, guided by simulated visual attention patterns. The project originates from a scientific study conducted at the Caen Museum of Fine Arts in France, which aims to investigate the cognitive and emotional benefits of museum visits on humans. To this end, in addition to completing neuropsychological questionnaires, 151 visitors were equipped with eye-tracking glasses to record their visual trajectories and gaze points, first while viewing paintings alone and then in pairs [2, 3]. The paintings included various artistic movements and genres such as portraits, surrealism, pointillism and genre scenes.

Based on the eye-tracking stored data, we first generated saliency maps that model the visual attention paid to each pixel of each of the 11 paintings included in the study. These maps were then used to fine-tune a saliency map predictor, called UNETRSal [6], enabling to anticipate which areas of a painting are most likely to attract human gaze. The new artwork emerges from visually transforming the original painting based on the simulated saliency map generated by UNETRSal. In particular, the most observed zones are amplified, revealing the collective visual attention of all study participants as illustrated in Table 1. The demonstration video can be viewed here.

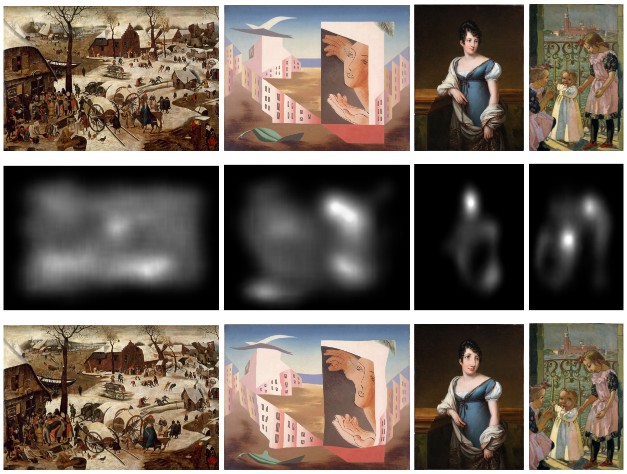

Table 1: Four out of eleven paintings: the original painting (first row); the predicted saliency maps (second row) and the visual transformations (third row).

2 Data Acquisition Protocol

We collected data from 151 participants equiped with eye-tracking glasses[1], during two different visits to the Caen Museum of Fine Arts from September 2024 to May 2025. For the first visit, each participant viewed six paintings for 2 minutes each, followed by a 30-second screening test where they had to focus on their feelings about the painting. In particular, two groups of visitors were randomly formed: a first group (A) with 74 people and another one (B) with 77 visitors, each group looking at three paintings in common and three other paintings exclusive to their group. Only group B received an explanation of the painting during the 2 minute observation phase, provided by a member of the museum. During the second visit, participants were paired, each pair composed of two visitors from group A and group B. These pairs jointly observed eight paintings, including four in common with the first visit. The pairs were formed by respecting a maximum age difference of six years whenever possible. Technically, a calibration of the eye-tracking glasses was performed before each visit in order to ensure the reliability of the recorded coordinates. This step consisted of following the appearance of fixed points projected on a screen, without moving the head. This allowed the recording of gaze paths and fixation points for each participant during 2 minutes in front of the painting and 30 seconds to record the emotions felt on each of the 11 paintings.

3 Saliency Map Prediction

We use the UNETRSal [6] model to predict generic saliency maps from the paintings viewed in the museum. UNETRSal is a hybrid model combining a Transformer-based encoder derived from UNETR [4] and a hierarchical convolutional decoder, which is specifically adapted to 2D image saliency prediction. While UNETR was initially designed for 3D medical image segmentation, its backbone was modified by replacing the 3D patch embeddings with 2D patch extraction, and by introducing a reshaping mechanism to reconstruct 2D saliency maps from Transformer outputs. The decoder is composed of multi-scale deconvolution and convolution blocks, where batch normalization layers were removed from specific blocks. This removal helped to stabilize training and improved saliency prediction. The model was trained using a composite loss function that combines Kullback-Leibler Divergence, Pearson's Correlation Coefficient, and Similarity metrics. This combination ensures alignment in distribution, linear correlation and structural similarity between predicted and ground-truth saliency maps.

UNETRSal has demonstrated strong performance on eye-tracking saliency benchmark datasets like SALICON [5] and CAT2000 [1]. We used the pretrained model on SALICON to fine-tune it using the eye-tracking data collected during museum visits in order to better align its predictions with the human gaze museum-based behavior. We chose the SALICON pretrained model as this is the largest available eye-tracking dataset, making it a strong foundation for pretraining. Indeed, it is commonly used as a basis for fine-tuning on smaller and task-specific datasets [6].

In order to fine-tune the SALICON-based UNETRSal model, we built a dataset that associates each painting with all participants individual ground-truth saliency maps. We also add a collective saliency map to each painting, by averaging the participants individual saliency maps. The dataset is split into training, validation and test sets, and covers the 11 paintings viewed by participants during the visit at the museum. To ensure that the training set is representative, it includes at least one painting from each major category, i.e. portrait, genre scene with crowds, genre scene with few characters, pointillism, and surrealism. The test and validation sets are composed of the most frequently occurring types, i.e. genre scenes (with crowds and few characters), pointillism, and portrait to counterbalance the lack of paintings diversity. During model training, the paintings and their corresponding saliency maps are

[1] The Core model from Pupil Labs available at https://pupil-labs.com/products/core.

resized to 480×640 pixels to fit the UNETRSal input size. Fine-tuning the UNETRSal model on museum-based data enables to generate individual and collective saliency maps that synthesize the different gaze patterns recorded on each painting.

4 Transformation Based on Saliency Maps

In order to create the new artworks, each painting is first normalized so that its pixel values range between 0 and 1. The saliency map simulated from UNETRSal is rescaled and resized to match the exact dimensions of the painting. To identify the key points of visual attention, the saliency map is smoothed using a Gaussian filter with $\sigma = 5$, and the local maxima of key regions are detected. To affect each pixel in the painting to a local maximum, an attention center is determined among these local maxima. This pixel-center association is based on both the spatial distance between the pixel and the various attention centers, and the saliency value of each attention center. Once the pixel-center association have been computed, the artwork transformation is applied to each pixel. In particular, its coordinates are shifted towards its associated center in proportion to the local saliency value. As such, in highly salient regions, pixels are more strongly pulled towards their local center, creating an impression of expansion around these areas of interest. The applied deformation factor depends on a global parameter controlling the overall zoom intensity. This mechanism allows for precise enhancement of salient regions while maintaining visual coherence across the painting. The final artwork is obtained by using bilinear interpolation, which estimates color values at the new pixel positions.

5 Conclusion

This work introduces a novel interactive pipeline that bridges human gaze and machine vision to create perceptually grounded transformations of fine art paintings. The accompanying demo will allow visitors to observe paintings using eye-tracking glasses, generate personalized saliency-driven transformations, and compare them with the UNETRSal predictions. This setup highlights the interpretability, creativity, and emotional resonance of gaze-guided art generation, opening new perspectives for cognitive interaction in digital art installations.

Acknowledgments

This work was performed using the computing resources of CRIANN, and funded by GIP Millénaire Caen 2025 as part of the ABC project (https://artbienetrecerveau.fr/).

References

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

[2] Serena Castellotti, Ottavia D’Agostino, Angelica Mencarini, Martina Fabozzi, Raimondo Varano, Stefano Mastandrea, Irene Baldriga, and Maria Michela Del Viva. 2023. Psychophysiological and behavioral responses to descriptive labels in modern art museums. *PLoS One* 18, 5 (2023), e0284149.

[3] Michael Garbutt, Scott East, Branka Spehar, Vicente Estrada-Gonzalez, Brooke Carson-Ewart, and Josephine Touma. 2020. The embodied gaze: Exploring applications for mobile eye tracking in the art museum. *Visitor Studies* 23, 1 (2020), 82–100.

[4] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. 2022. UNETR: Transformers for 3D Medical Image Segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1748–1758. doi:10.1109/WACV51458.2022.00181

[5] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1072–1080.

[6] Azamat Kaibaldiyev, Jérémie Pantin, Alexis Lechervy, Fabrice Maurel, Youssef Chahir, and Gaël Dias. 2025. UNETRSal: Saliency Prediction with Hybrid Transformer-Based Architecture. *22nd Advanced Concepts for Intelligent Vision Systems (ACIVS)* (2025).

[7] Hironori Takimoto, Tatsuhiko Kokui, Hitoshi Yamauchi, Mitsuyoshi Kishihara, and Kensuke Okubo. 2015. Image modification based on a visual saliency map for guiding visual attention. *IEICE TRANSACTIONS on Information and Systems* 98, 11 (2015), 1967–1975.

Appendix A: Installation Information

We present an interactive installation where visitors engage with an eye-tracking setup to explore a fine art painting (see figure **??**). As participants observe the painting through wearable eye-tracking glasses, their gaze data is captured and used to generate a personalized visual transformation map. This map highlights the regions they focused on and is subsequently compared to a predicted saliency map generated by our SALICON-based UNETRSal model. The installation fosters a unique dialogue between human perception and computational interpretation, offering a rich, real-time experience at the intersection of cognitive science, digital art, and artificial intelligence. At the end of the interaction, each visitor will receive a QR code that allows them to download their personalized artwork, turning the installation into a tangible and memorable creative experience.

To support the demo, we require the following technical and logistical setup:

- **Furniture**: One table and three chairs for participants and equipment;
- **Display support**: A hanging grid for the painting and a large screen so that the audience can visualize the participant’s gaze in real time (with consent of the participant);
- **Power**: A standard 220 V electrical outlet to power the screen and the laptop;
- **Lighting**: A well-lit space to ensure accurate gaze tracking;
- **Space**: Approximately 9m^2 (3m × 3m) to allow unobstructed interaction around the installation.

After a short calibration phase, participants will view the painting while wearing the eye-tracking glasses. Their gaze data is processed in real time to produce a transformation that visually reflects their attention. A comparison with the system’s predicted saliency map is displayed, and participants receive a QR code to download their unique, gaze-driven artwork.