Uncovering Aesthetic Preferences of Neural Style Transfer-Generated Images with the Two-Alternative-Forced-Choice Task

Chaehan So
Information & Interaction Design
Humanities, Arts & Social Sciences Division, Yonsei University
Seoul, South Korea
Email: cso@yonsei.ac.kr

Abstract

Neural style transfer is a popular deep learning algorithm to generate images to mimic human artistry. This work applies the psychological method of the two-alternative forced choice (2afc) task to measure aesthetic preferences for neural style generated images.
Portrait photos of three popular celebrities were generated by varying three parameters of neural style transfer in five configuration levels. Participants had to choose the image they preferred aesthetically from all pairwise combinations of configurations per style. The rate of being chosen was calculated for each neural style transfer configuration level. The findings show a differentiated picture of aesthetic preferences. On the one side, they indicate that people prefer images rendered with 500 iterations and a learning rate of 2e1, i.e. configurations that allow them to recognize the structure of the portrait image despite the stylization. On the other side, aesthetic preferences peak for two distinctly different content-to-style weight ratios. Whereas the medium-high configuration (100:100) may be favored by people who like abstract arts, the high configuration (300:100) may be chosen by people who prefer realistic art.
These results indicate that aesthetic preferences for neural style transfer-generated images can be characterized by unique patterns, and their optimal configuration levels can be captured by the 2afc task.

Keywords: Neural style transfer, aesthetic preferences, 2afc task, parameter configuration

Introduction

Neural style transfer has become a popular generative deep learning algorithm for generating creative images by merging the structure of a visual input with the style of another image. Although the output is often called "creative", the literature lacks a differentiated view on how visually pleasing humans perceive this stylized image.
Historically, the development of the neural style transfer algorithm built on substantial progress of convolutional neural networks (CNNs) and generative adversarial networks (GANs). Gatys, Ecker, and Bethge (2015) introduced the former algorithm to show how the artistic style of painters can be transferred to another image. The neural style algorithm extracts the structural information of the input image, learns the color and texture information in the style image, and then renders the semantic structure of the input image in the color and texture of the style image (Gatys, Ecker, and Bethge 2016).
Originally, neural style transfer was demonstrated with common photo motifs like houses or landscapes (Gatys, Ecker, and Bethge 2016). Neural style transfer has later been applied to doodles (Champandard 2016), videos (Huang et al. 2017), artistic improvisation (Choi 2018), and fashion (Jiang and Fu 2017; Zhu et al. 2017). Recent developments of neural style transfer include using pretrained models for stylization in so-called feedforward networks (Chen and Schmidt 2016). One such approach (Johnson, Alahi, and Fei-Fei 2016) leverages the use of the pretrained models to calculate losses on high-level features (so-called perceptual losses) instead of per-pixel losses.
The generation of neural style transfer is configurable by many parameters. Therefore, it may be insightful which of these parameters configurations lead to aesthetically pleasing results.

Aesthetic Preferences

What image looks good? Psychological research has yielded manifold insight into this question from the investigation of interindividual differences in aesthetic taste. For example, aesthetic preferences are not stable across the human lifespan but follow an inverted U-shape peaking around early to middle adulthood (Pugach, Leder, and Graham 2017). The preference for different art genres depends on personality traits, e.g. people with high scores on trait openness tend to prefer abstract arts over renaissance art (Pelowski et al. 2017). People prefer a size of a displayed object which is larger (smaller) relative to the frame for larger (smaller) objects, and the preferred displayed object size is proportional to the logarithm of its physical size (Linsen et al. 2011).

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

Two-alternative forced choice

This study investigates people's perception of artistry in an image with the *two-alternative-forced choice (2afc)* task. The 2afc task is an experimental method in psychology first introduced by Thurstone (1927).

The 2afc task is frequently used in cognitive psychology to detect perception thresholds or evaluate the psychological differentiation of stimulus variation. Applying signal detection theory, the Thurstonian model of item-response theory identifies the detection thresholds of perceived stimuli in a 2afc task by fitting a maximum-likelihood estimator, the psychometric function, to the averaged 2afc task responses. In such 2afc applications, the presented image pairs contain a baseline image, i.e. an image without any effect, to which the other image is compared.

The 2afc task is a *paired comparison test*, i.e. it specifies the assessment of two samples. This simple design has shown the advantage of significantly reducing fatigue, carryover and memory effects encountered when assessing more samples (Yang and Ng 2017), and reducing the required sample size. Importantly, both samples in a 2afc task must be presented at the same time (Macmillan & Creelman, 2004, p. 148). If this condition is not met (one stimulus is shown), it represents a *yes/no task* (often used for lexical decisions like word/non-word) not to be confused with a 2afc task.

The 2afc task has also been used in psychological research to investigate aesthetic preferences, e.g. the classification of images as art vs. non-art (Pelowski et al. 2017), the spatial composition in multi-object pictures (Leyssen et al. 2012) (Leyssen et al. 2012), or the size of images for real-world objects (Linsen et al. 2011).

According to Palmer, Schloss, and Sammartino (2013), the common procedure for the 2afc for testing aesthetic preferences is to present the participant with all possible pairs of stimuli instead of all comparison with a baseline image. For each pairwise combination, participants are asked which they *"like better"* which corresponds to their aesthetic preference. The global measure of an image's relative preference is calculated by the average probability (or actual count) of selecting it over all other images.

Research Question

Based on the increasing proliferation of neural style transfer applications, it becomes interesting from a design and art perspective how it can produce aesthetically pleasing results. Understanding related mechanisms leads to the investigation of available parameter configurations of neural style transfer. Such an investigation can unveil important underlying factors and relationships of human aesthetics perception. It may reveal for example that for certain parameters of neural style transfer, the preference curve may have a curvilinear (inverted-U or V) shape rather than a monotonous positive or negative gradient.

The preceding considerations lead to the following research question.

RQ: Which parameter configurations of neural style transfer reflect the highest aesthetic preference?

Method

Participants

This study recruited a sample of 18 participants undergraduate and graduate students of interaction design, as well as professional UX designers.

Participants were 50% male and 50% female, and on average 30.5 (SD = 8.52) years old.

The academic status was 33.3 % *Bachelor student*, followed by 27.8% *Master student*, 5.56 % *professional with Bachelor's degree*, 27.8% *professional with Master's degree*, and 5.56% *professional with Ph.D. degree*.

Participants' nationality was majorly *South Korean* (77.8%) with one person each from *Canada, Germany, Singapore*, and the *United States*.

Neural Style Transfer Algorithm

The neural style algorithm merges the structural information of an input image with the color and texture of a style image. This ensures that the input image's structural information (like face and body line structure) can be recognized in the output image.

The present work uses the implementation of neural style transfer provided by the Github repository *jcjohnson/fast-neural-style* (Johnson 2016). It is implemented in torch (Collobert et al. 2018) and provides an improved version of the neural style transfer algorithm of the Github repository *jcjohnson/neural-style* (Johnson 2015). The latter version implements the original optimization-based algorithm introduced by Gatys, Ecker and Bethge (2015).

The optimization-based algorithm also provided in the *jcjohnson/fast-neural-style* Github repository by the script "slow_neural_style.lua" and allows many configuration options. One can configure the options *num_iterations* (number of iterations processed), *save_every* (image generation after save_every iterations), *GPU* (use GPU or CPU).

To modify the stylization outcome, one can configure many detailed options of the loss network. The interface allows determining the content layers, style layers, style target (gram matrices vs. spatial average) and the choice of the optimizer (commonly LBFG-S or Adam algorithm). All these options require a lot of expertise to understand the direction and magnitude of parameter changes. Configuration parameters with directly perceivable output impact include the number of iterations, the learning rate and the ratio between content weights and style weights.

All results of this paper were generated on an Ubuntu 16.04 LTS virtual machine in the Google Cloud. Image processing becomes impractical in CPU mode due to slow processing speed. Therefore, the present work used an Nvidia Tesla P100 GPU with CUDA 9.1 and CUDNN 8.0 libraries that offered a processing speed of a factor of approximately 100 times faster than a contemporary laptop with 8 CPUs.

Input Images. The stimuli were based on three faces of popular celebrities (Charlie Puth, Jessica Alba, Ellie Goulding). Figure 1 shows these input images.

Preliminary experiments provided the insight that pictures with high variance in the background were evaluated as part of the foreground by the algorithm and thus stylized the same way as the foreground. To avoid this effect, portraits were only selected if they showed a clear separation to the background.

Figure 1: Input Images

Style Images. The style images were chosen from three different categories: black & white patterns, cloth design, and abstract arts. The preliminary experiments revealed that style images have the most impact on the output image if they contain texture information of finer granularity that separates well from the background. Figure 2 shows the set of style images used in this study.

Figure 2: Style Images

Parameters

The neural style transfer was applied to the set of input images to create variations that were subsequently tested by 2afc tasks with the following parameters and configuration levels.

Number of iterations. The neural transfer algorithm by Gatys, Ecker and Bethge (2015) applies the L-BFGS optimizer on forward and backward iterations through the VGG-16 loss network. Johnson, Alahi and Fei-Fei (2016) found that the optimization is successful within 500 iterations in most cases. Therefore, this study compared generated visual results with 100, 200, 300, 400, and 500 iterations.

Learning Rate. The framework allows changing the optimizer from LBFGS to Adam. The Adam optimizer can be configured by the learning rate. This parameter specifies the step size in which weights are updated during the optimization. As Ruder (2016) points out, a too small learning rate slows down the convergence to the minimum, whereas an overly high learning rate can cause fluctuation around the minimum and thus hinder convergence.

The default learning rate is set to 1e-3. Therefore, this study explored the impact of learning rates for the Adam optimizer for 0.5e1, 1e1, 2e1, 4e1, and 6e1.

Content-to-Style-Weight Ratio. The content-to-style-weight ratio determines the degree of importance for the input image vs. the style image for rendering the output image. The default ratio is 1:5, i.e. the style weights are five times larger than the content weights. The parameter content weight is set as a relative value to style weight. The content-to-style-weight ratio defines the degree of importance given to the input image vs. style image for rendering the output image.

The default setting is 1:1 or 100:100. Therefore, this study explored results for the content-to-style-weight ratios 10:100, 50:100, 100:100, 200:100, and 300:100.

2afc Task

Stimuli. The preceding specifications generated the following three sets of stimuli.

1. Figure 4 shows Charlie Puth in variations of *the number of iterations* with values 100, 200, 300, 400, 500.
2. Figure 5 shows Jessica Alba in variations of the *learning rate* with values 0.5e1, 1e1, 2e1, 4e1, and 6e1.
3. Figure 6 shows Ellie Goulding in variations of the content-to-style-weight ratio with values 10:100, 50:100, 100:100, 200:100, and 300:100.

Condition Counterbalancing. The position of the two images in the 2afc task (either left or right) was counterbalanced within participants so that each image was evaluated in the same frequency in the left and right position.

Pairwise Comparison Reduction. The number of pairwise comparisons was reduced to reduce survey fatigue by using each image pair comparison only once, and balancing the image position only in absolute numbers rather than for each image pair combination.

Procedure. Before starting the survey, participants were asked to not use a smartphone but a laptop, desktop or tablet for larger image display. They were instructed to make the choice between the two images intuitively rather than by objective criteria. Each participant assessed 5 configurations of 3 parameters for 5 styles. Each configuration was assessed in 4 pairwise comparisons per style and participant, hence 72 times by all participants. Participants were encouraged to take a break every 50 trials for 10-20 seconds for stretching their upper body and relax their eyes.

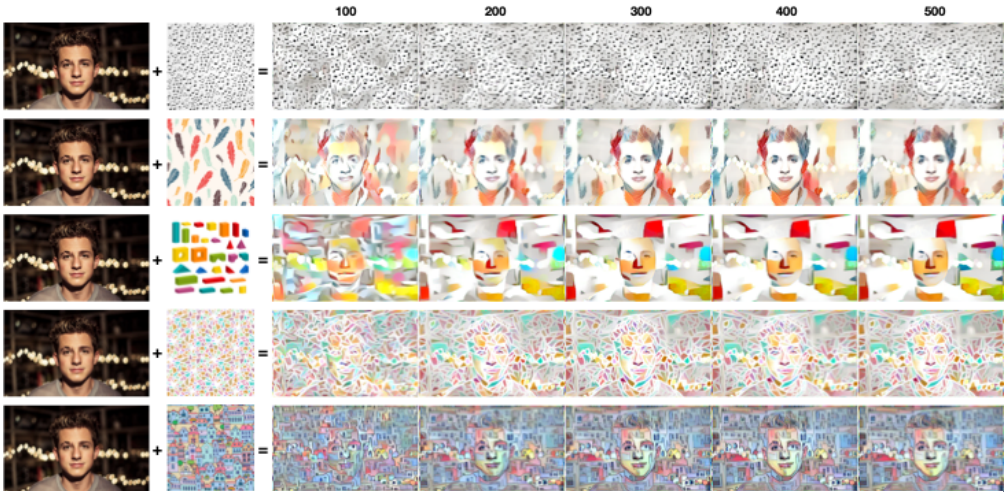

Figure 4: Neural style transfer-variations of parameter A: Number of iterations

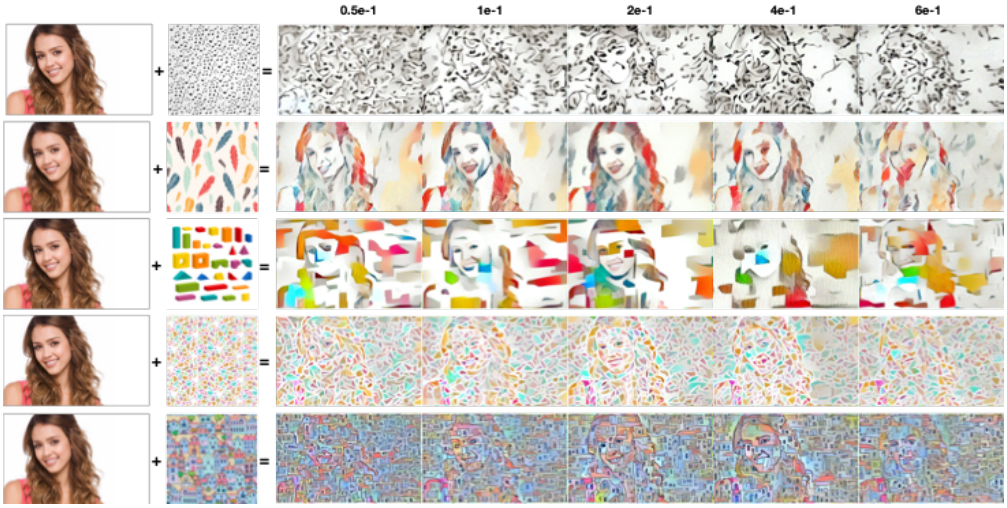

Figure 5: Neural style transfer-variations of parameter B: Learning rate

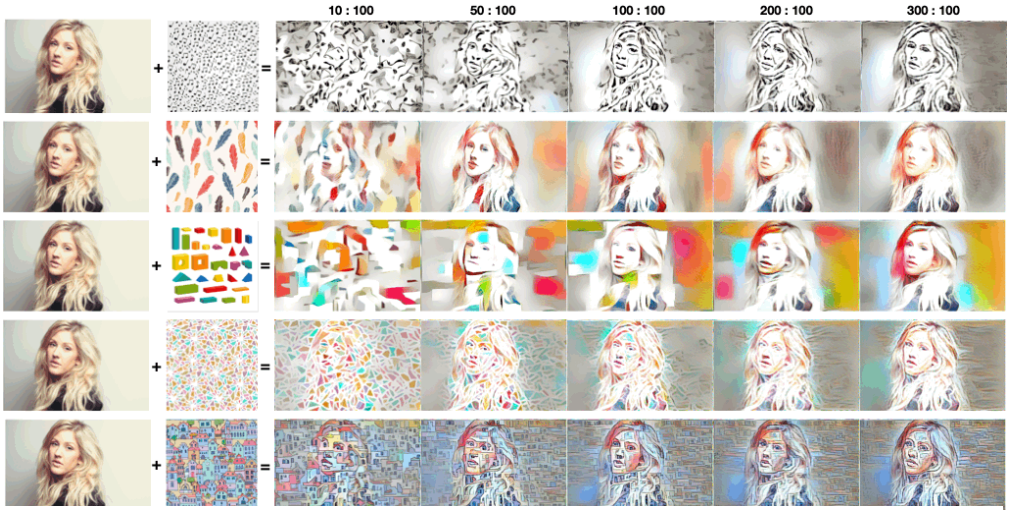

Figure 3: Neural style transfer-variations of parameter C: Content-to-style-weight ratio

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

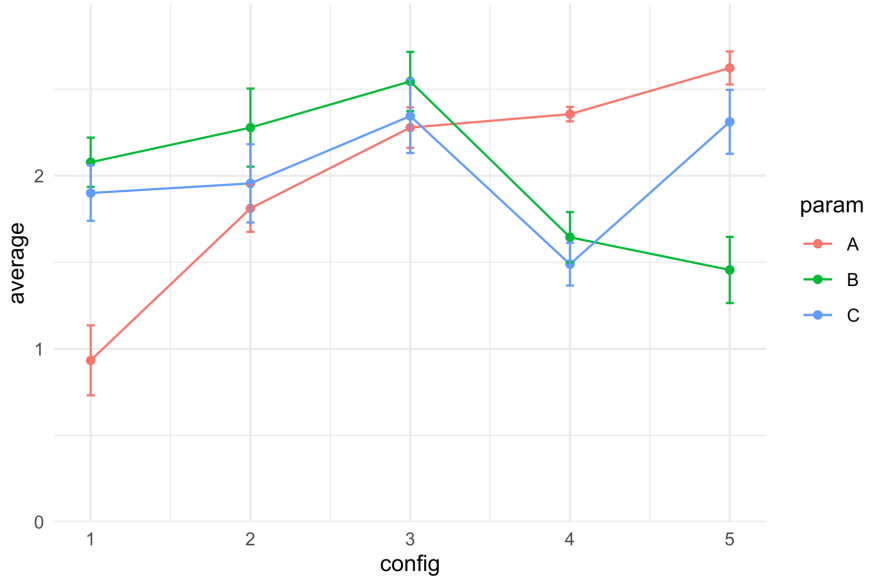

Figure 7: Average wins per configuration per parameter
A (number of iterations) configurations: 1 = 100, 2 = 200, 3 = 300, 4 = 400, 5 = 500;
B (learning rate) configurations: 1 = 0.5e1, 2 = 1e1, 3 = 2e1, 4 = 4e1, 5 = 6e1;
C (content-to-style-weight ratio) configurations: 1 = 10:100, 2 = 50:100, 3 = 100:100, 4 = 200:100, 5 = 300:100

Results

The findings are displayed in Figure 7 and show an aggregated view over 2700 pairwise comparisons yielded by 18 participants, represented by the means (dots) and standard errors (error bars) depicted in Figure 7.

The interpretation for parameter A (number of iterations) is straightforward by visual inspection: People prefer neural style transfer generated images with more iterations. Noteworthy is that the shape of the preference curve is linear only between 100 and 300 iterations, with a strong decline in the gradient thereafter. This suggests a convergence to a climax point near 500 iterations.

The pattern looks different for parameter B (learning rate) that marks an inverted U-shape with the peak in the middle configuration corresponding to a learning rate of 2e1. Mentionable is the right skewness of the preference curve – this indicates that, between small (0.5e1 or 1e1) and large (4e1, 6e1) learning rates, people prefer neural style images created with smaller learning rates.

The preference curve for parameter C (content-to-style-weight ratio) shows an ambiguous picture with two peaks at the middle (100:100) and high configuration (300:100). This pattern cannot be interpreted as a U-shape with an anomaly peak because the latter levels with the middle configuration at par. This pattern rather reveals two different potential preference reasons – whereas the middle configuration allows to easily identify the person but in a strongly stylized form, the high configuration allows recognizing the person's face subtleties better as the image appears closer to a photograph.

Discussion

The findings of this study allow a clear recommendation for configuring the neural style framework by Johnson (2016): Aesthetic results can be expected with 500 iterations, a learning rate of 2e1, and a content-to-style weight ratio of either 100:100 or 300:100.

Interestingly, the preference curve for number of iterations reveals a distinct aesthetic preference (continuously positive and with small standard errors) even though the difference between the configurations was consciously hardly recognizable, as some participants remarked (see Figure 4).

Both optimal configurations for number of iterations and learning rate reveal a general preference for more realistic rendering because these configurations let the viewer best recognize the structure of the portrait image independent from the artistic stylization.

There are several limitations to the findings of this study. The survey was tested by a small sample size. Even though this is consistent with other studies using the 2afc task, a higher sample size could allow detecting more differentiated preference patterns, e.g. differences between demographic groups. Moreover, the double-peak preference curve for the content-to-style-weight ratio could simply reflect a binary preference pattern between abstract versus realistic art. The survey did not control for this explanation. These aspects could be interesting focal points for future research.

The main contribution of this work is threefold.

First, the 2afc methodology removes the measurement problems encountered with other measurement methods. Among them, the most common is the rating with Likert scales

which is prone to acquiescence bias (Friborg, Martinussen, and Rosenvinge 2006). Rank ordering as an alternative method is known to overwhelm the observer with a strong cognitive demand to identify a relative ordering of all images (Palmer, Schloss, and Sammartino 2013).

Second, the findings provide designers with straightforward recommendations on how they can use neural style transfer as an effective design tool for creating artistic visual portrait images. This may be a purpose of its own, or a means to overcome their design fixation by producing artistic variations for inspiration.

Third, it reveals the existence of distinctly different aesthetic preferences. It seems plausible to conjecture that the high configuration for content-to-style weight ratio might be preferred by people with a focus on realistic depictions, whereas the middle configuration might be preferred by people who like abstract arts. This is consistent with other research showing that people differ in their aesthetic preferences (Palmer, Schloss, and Sammartino 2013).

Taken together, under which conditions people favor the proximity of the expected, i.e. a less artistic impression, or a more stylized and thus abstract image rendition, is subject to further research. The present work might have made a crucial step towards opening this new research avenue.

Acknowledgment

This research was supported by the Yonsei University Faculty Research Fund of 2019-22-0199. The author would like to thank Kyuha Jung for his support with the survey setup.

References

Champandard, Alex J. 2016. "Semantic Style Transfer and Turning Two-Bit Doodles into Fine Artworks." *Arxiv*. http://arxiv.org/abs/1603.01768.

Chen, Tian Qi, and Mark Schmidt. 2016. "Fast Patch-Based Style Transfer of Arbitrary Style." *Arxiv*. http://arxiv.org/abs/1612.04337.

Choi, Suk Kyoung. 2018. "Guess, Check and Fix: A Phenomenology of Improvisation in 'Neural' Painting." *Digital Creativity* 29 (1): 96–114.

Friborg, Oddgeir, Monica Martinussen, and Jan H. Rosenvinge. 2006. "Likert-Based vs. Semantic Differential-Based Scorings of Positive Psychological Constructs: A Psychometric Comparison of Two Versions of a Scale Measuring Resilience." *Personality and Individual Differences* 40 (5): 873–84. https://doi.org/10.1016/j.paid.2005.08.015.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. 2015. "A Neural Algorithm of Artistic Style." *Arxiv*, 3–7. https://doi.org/10.1167/16.12.326.

Gatys, Leon A, Alexander S Ecker, and Matthias Bethge. 2016. "Image Style Transfer Using Convolutional Neural Networks." *The IEEE Conference on Computer Vision and Pattern Recognition*, 2414–23. https://doi.org/10.1109/CVPR.2016.265.

Huang, Haozhi, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. 2017. "Real-Time Neural Style Transfer for Videos." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7044–52. https://doi.org/10.1109/CVPR.2017.745.

Jiang, Shuhui, and Yun Fu. 2017. "Fashion Style Generator." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 3721–27.

Johnson, Justin. 2016. "Fast-Neural-Style." *GitHub Repository*. GitHub. https://github.com/jcjohnson/fast-neural-style.

Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. 2016. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9906 LNCS: 694–711. https://doi.org/10.1007/978-3-319-46475-6_43.

Leyssen, Mieke H.R., Sarah Linsen, Jonathan Sammartino, and Stephen E. Palmer. 2012. "Aesthetic Preference for Spatial Composition in Multiobject Pictures." *I-Perception* 3 (1): 25–49. https://doi.org/10.1068/i0458aap.

Linsen, Sarah, Mieke H.R. Leyssen, Jonathan Sammartino, and Stephen E. Palmer. 2011. "Aesthetic Preferences in the Size of Images of Real-World Objects." *Perception* 40 (3): 291–98. https://doi.org/10.1068/p6835.

Macmillan, Neil A, and C Douglas Creelman. 2004. *Detection Theory: A User's Guide*. Psychology press.

Palmer, Stephen E., Karen B. Schloss, and Jonathan Sammartino. 2013. "Visual Aesthetics and Human Preference." *Annual Review of Psychology* 64 (1): 77–107. https://doi.org/10.1146/annurev-psych-120710-100504.

Pelowski, Matthew, Gernot Gerger, Yasmine Chetouani, Patrick S. Markey, and Helmut Leder. 2017. "But Is It Really Art? The Classification of Images as 'Art'/"Not Art" and Correlation with Appraisal and Viewer Interpersonal Differences." *Frontiers in Psychology* 8 (OCT). https://doi.org/10.3389/fpsyg.2017.01729.

Pugach, Cameron, Helmut Leder, and Daniel J. Graham. 2017. "How Stable Are Human Aesthetic Preferences across the Lifespan?" *Frontiers in Human Neuroscience* 11 (May): 1–11. https://doi.org/10.3389/fnhum.2017.00289.

Ruder, Sebastian. 2016. "An Overview of Gradient Descent Optimization Algorithms." *Arxiv*, 1–14. https://doi.org/10.1111/j.0006-341X.1999.00591.x.

Thurstone, L.L. 1927. "A Law of Comparative Judgement." *Psychological Review* 34 (4d): 266–70.

Yang, Qian, and May L Ng. 2017. "Paired Comparison/Directional Difference Test/2-Alternative Forced Choice (2-AFC) Test, Simple Difference Test/Same-Different Test." In *Discrimination Testing in Sensory Science*, 109–34. Elsevier.

Zhu, Shizhan, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change. 2017. "Be Your Own Prada : Fashion Synthesis with Structural Coherence." *ArXiv*, 1680–88.