ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20240603185810

Original scientific paper
Received: 2024-06-03, Accepted: 2024-07-12

Computing the Deep Semantics of Visual Communications

Trpimir Jeronim Ježić, Marko Maričević, Ivana Pavlović, Miroslav Mikota*

Abstract: The research field of computational aesthetics gives crucial contributions to the development of mechanisms for filtering and/or generating value-laden informational content. This paper acknowledges a recognized escalating problem in the development of contemporary informational technologies and presents a practical solution for communicational quality management by employing an innovative approach to the computational aesthetic evaluation (CAE). After discussing the problem and attempted approaches to its alleviation, the paper offers a novel expert solution by presenting an original research approach and its resulting open-sourced model which outperforms its current state-of-the-art competition in semantic and stylistic classification, at the same time providing an idiomatic measure for objective aesthetic evaluation and demonstrating semantically rich and professionally recognized explanatory power which can serve as the solid basis for development of reliable and user friendly content retrieval, generative or auxiliary design applications. Presented model is resource- and privacy-wise utmost conservative. Its use evades all ethical, legal or security concerns that beset all currently prominent models. Its developmental and operational costs are practically nil.

Keywords: computational aesthetic evaluation; convolutional neural networks; feature engineering; graphic engineering; interpretable machine learning; semantic embeddings

1 INTRODUCTION

One of the first definitions for the term *computational aesthetics* as a name for the research field in informational studies appeared in Hoenig's paper in 2005. He defined it as "the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can" [1, 2]. In that way, he emphasized its two major aspects: the use of computational methods (i.e. it provides measurable output) and the enhancement of applicability. Computational aesthetic evaluation, as any aesthetic judgement, is generally considered to be a highly idiosyncratic act [3]. As Galanter noted, notions of computational aesthetic evaluation led to "deep philosophical waters regarding phenomenology and consciousness" [4]. The area of aesthetic research thus calls for a strongly interdisciplinary approach [5-8]. Nevertheless, computational aesthetics, considered as a field at the intersection of science and art [9, 10], has seen significant progress in recent years. It became focused on aesthetic measurement, generative art, and design generation [7, 11, 12]. Among many applications of computational aesthetics established in last few decades, we focus our attention here on one of its most beneficial aspects: integrating contributions of computational aesthetic evaluation (CAE) and computer vision (CV) algorithms for the sake of advancement of human information accumulation and retrieval capacities. The research question posed in this paper is whether there is a way to computationally detect salient visual features which provide semantic context for decoding the meaning of visual communications.

2 RESEARCH METHODOLOGY

A necessary condition for success of any machine learning (ML) project is the appropriateness and quality of its training data. Our research question imposes on samples a few criteria that would make them suitable contributors to its answer. Firstly and obviously, the sample instances need to be recognized as successful communicators of their intended meaning. Secondly, to prove the universality of their status, they need to be acknowledged as role models for other successful communication attempts time and again since their first occurrence. Thirdly, there are a couple of bonus features of visual communication instances that would propound them as exceptionally suitable for enlightening our proposed research question. It would be preferable if the syntactic form of those instances were as arbitrary as possible, measurable by some standardized laboratory equipment, under the authors' control, not dictated by available resources or other means of production and not constrained by any predefined linguistic or other cultural mandate in adding or removing features of shapes to attune the entire composition for transmitting the intended semantic meaning, and lastly explainable by some design theory. In accordance with the research question and its qualifiers, we concluded that the narrowed category of pictorial high art would best serve as a representative of successful unhindered visual communications [13-15]. For that reason, we chose the community driven WikiArt site as the source of our sample.

2.1 Data Accumulation

First, we gathered data by scraping the whole of WikiArt website. In this way we gathered information on over 215,000 instances of artworks made by 5300 artists featured on the site at the time. The data featured information about exhibits such as artwork's title, author, style and genre and unique ID given to the artwork in the WikiArt site's database. We gathered additional data about artists such as their lists of names, birthdays, deathdays, and URLs of their Wikipedia pages. WikiArt organizes artworks into 220 styles and 68 genres. Artworks could be labelled with multiple style and genre categories, culminating in total 1552 unique style and 726 unique genre combinations. In the next step we cleaned up the dataset so it would contain only research relevant instances. Our focus was on achieving high precision in this selection without any pressure to achieve a high recall rate due to the abundant cardinality of the sample. Beside

pictorial art, WikiArt presents many exhibits of architectural, product design, VR and AR, calligraphic, ornamental, and other types of art. All instances belonging to those genre categories were not relevant to our research. We also applied an ensemble of custom filters on the scraped data to ensure that the remaining instances represent high-quality works, exemplars in communication design, authored by artists confident in their style, affirmed by their influence on culture and other artists and recognized by experts in the field. We based the construction of those filters on some proxy features. Some of these filters connote the requirements: that the instance's author has a substantial body of appreciated work; has a Wikipedia page dedicated to his influence as a visual artist; the style in which the image was categorized has substantial following and a body of authors who adhere to it; that the style has its dedicated Wikipedia page; that the work managed to stay relevant through at least couple of major cultural shifts (which eliminated a plethora of recent artworks); and so on.

Second, we needed to decide on the semantically relevant features for whose detection we will train our ML model. In the beginnings, CAE researchers attempted to construct expert systems for detecting hand-crafted and statistically generic aesthetic features, accomplishing modest but promising results [16, 17]. In 2014 [18], motivated by the advances in CV technologies [19], researchers turned their focus on convolutional deep neural network (CNN) architectures which immediately yielded better results in blind prediction of past ratings of images, but forced the researchers to abandon their search for aesthetic features that could clarify the problem of understanding the appreciation of visual communications [20]. Many black-box approaches have been tried over the next decade with increasing success rates due to the addition of more precise search criteria: by discriminating between portions of images that feature subjects from those of backgrounds [21]; or by discriminating portions by the compositional relevance of their positions in an image format [22, 23]; or by giving higher priority to features prominent in art and design theory [24]; or by adding descriptive "semantic" features that give context to aesthetic evaluation [25-31], which at times proved to be alone better predictors of aesthetic judgements than the pixel data.

Current research in CAE recognizes the need and benefits of focusing on semantically relevant features of images even when aiming solely to create a naive aesthetic approval prediction machine. However, all up to date and state-of-the-art research approaches try to add semantic features to their datasets by appending columns featuring linguistic labels gathered through some survey on dataset images or through semantic-web scraping. The idea is that these inputted features are universally relevant for instances of the dataset because they represent human judgement of those images, and that those features are automatically semantic because they are linguistically labelled. None of those stances are exculpatory. There are no guarantees that the surveys and their multiple choices present a valid spectrum or measure of semantically potent attributes of images, nor are there any guarantees that the surveys' subjects are representative of any model's future user. But crucially, there is no justification for claiming that the added feature columns store any semantic information at all. Those appended columns bring to the dataset only additional syntax, foreign to the original samples and absent during the original sample selection which qualified its instances as valid. Without guarantees for semantic stability of added syntax, additional columns only enhance the problem complexity of information extraction and cast doubt on the representative validity of altered samples. To avoid those common pitfalls, we abstained from conducting any *ad hoc* surveys on the collected dataset instances. For this we decided to base our decision mechanism exclusively on the well-established art-historical classification of styles already present in the given dataset. Those style labels served as a meaningful (semantically fixed) way of separating the instances into syntactically distinct buckets. Many of those buckets appeared to be uninformative for our purpose. For example, we found that sample sets labelled with style names prefixed with terms such as "post-", "neo-" or "new" more often than not presented aggregates of images with no internal syntactic or semantic coherence. Those labels rather functioned as negative signifiers, encompassing artworks associated only by the signal they are responding to, and not by the response communicated through the artworks. Similarly incoherent aggregations appeared in sets based on style labels that contained vague qualifying terms such as "classical", "analytical", "period" or "school". Using the criteria described in this and previous paragraphs alone we managed to reduce the sample size by 25%, purifying it down to 160,000 relevant instances.

2.2 Feature Selection

For the relevant features to help semantically categorize visual communications we decided to use two dimensions, well established and basic to any classification of possible thoughts and their expressions. The first dimension is the one on which the scientific method with its task of extracting generalizations from samples is founded; which is to say, the dimension relied upon when distinguishing concrete objects via abstractions. The second dimension we chose is the one on which all theories of applied communications, whether linguistic or visual, base their discussion on the structure of the relationships between the signifier and the signified; which is to say, when mapping the syntax to semantics of a message. Specifically, we base our second line of distinction on Charles S. Peirce's original theory of *semiotics* [32], which in practice underlies all design theories. Our first dimension determines the nature of the subjects in communication, and the second one determines the way the subjects are referred to. Clearly, these two dimensions describe the necessary basic semantic context which is a prerequisite for any further semantic decoding. We will refer to these dimensions as the *breadth* and the *depth* of a communication, in the same spirit in which the terms are used in common speech. It is likewise important to note that these two dimensions are by definition orthogonal to each other which makes them very convenient when mapping the

communicative space. Humans can talk about concrete objects or abstract concepts via examples or through symbolic means. There is no predetermined correlation at hand.

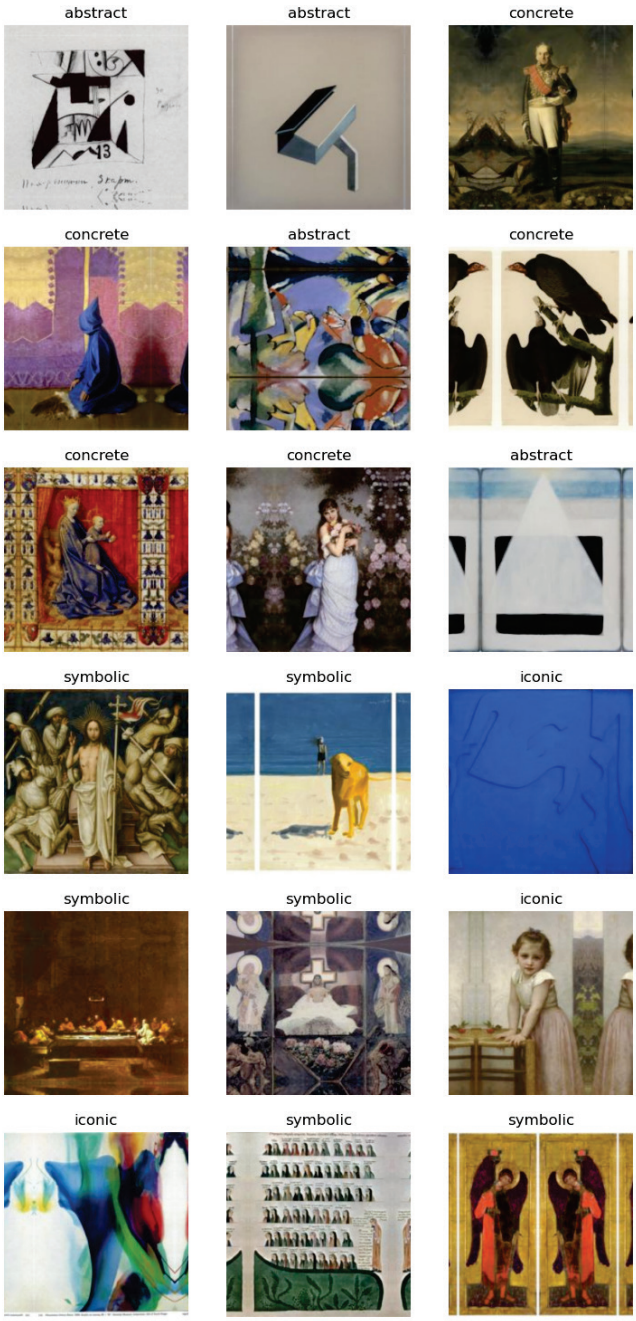

Figure 1 Random subsets of representative instances collected from WikiArt

To select the representative instances for the extrema of these dimensions, we created buckets of styles that are manifestly and universally recognized as belonging by impressions and communicative intentions to one of our four extrema. The concrete-abstract (breadth) dimension is represented on one end by figurative art styles such as Renaissance, Academicism, Naturalism, Realism and Hyper-Realism; and on the other end by abstract styles such as Concretism, Suprematism, Abstract Expressionism and Action painting. The iconic-symbolic (depth) dimension is represented on one end with naturalistic, visceral art styles such as Rococo, Academicism, Pointillism, Lyrical Abstraction and Color Field Painting; and on the other end by the intellectual styles, rich in symbolism, such as Byzantine, International Gothic, Classicism, Romanesque and Pop Art. While the distinction between the styles on the breadth dimension is visually obvious, there is no such obviousness in the depth dimension. The styles were handpicked based on consensus among the art historians and critics regarding the tone and intended meaning of the artworks belonging to these styles [33-35]. When it has been decided which styles are representative, further selection of specific artworks belonging to those stylistic categories was left to random choice. An image dataset containing about 500 images per extrema was obtained by downloading digital replicas of a stylistically curated, but otherwise randomly chosen samples from the WikiArt's website (Fig. 1).

2.3 Architecture Design and Training

To guarantee objectivity, training was conducted on pixel-data alone, supervised only by meta-stylistic labels for four extrema of two semantic dimensions defined above. By recognizing the fact that the two chosen dimensions are mutually orthogonal and that there could be very little or none instances that could serve as good representatives of extrema on both dimensions simultaneously, and that forcing the neural networks to distinguish between intermediate representations would motivate the model to overfit to any given dataset, we decided to train two separate neural networks, each for implementing one semantic dimension, and thereafter to pair the two unidimensional models into one proficient parallelized system. Model development was conducted with the PyTorch library for ML in Python, more specifically, using Jeremy Howards fastai wrapper-framework. By following the advice of Yosinski et al. [36], Dong et al. [37], and Howard [38], we base our training approach on fine-tuning an existing state-of-the-art CV model. This approach is conventional in CAE [18], [39], [28], [40] for its noted benefits. Fine-tuning on a trained CNN has been proven to be a resource effective initialization approach, but the benefits of fine-tuning do not stop at conservation. As has been shown in Deng et al. [20], fine-tuning for aesthetic evaluation from the vanilla AlexNet yields better performance than simply training the base net from scratch [18]. One possible explanation is the claim that multitask model training, or in case of fine-tuning, task accumulation, forces ML models to construct more realistic embedding space for inputted data, and thus build a better understanding of real-world relations between data points [41, 42, 30]. Another, more reasonable explanation is that the task overload serves as a realistic regularization mechanism which prevents models from overfitting to sampled data.

Papers referenced in previous paragraph used the AlexNet [19], a 13-layered convolutional neural network (CNN) that brought neural networks into the spotlight back in 2012 when it outperformed all other ML architectures

competing on ILSVRC, and by doing so, kick-started the ongoing AI revolution. We found its successor, the winner of ILSVRC 2014, a 22-layered CNN named GoogleNet [43], most efficient for our task. All foundational models were accessed through PyTorch's Timm library of open-sourced legacy CV models. In our training, the GoogleNet model outperformed AlexNet and all other up to date and moderately sized pretrained CNNs. It was deemed appropriate because of its minimalist design architecture and extremely small number of layers, considering today's trends. The interpolation between semantic extrema is expected in reality to be somewhat linear and we were cautious in preventing the possibility of overfitting a model to the data by giving it too many degrees of freedom. The only significant alteration we conducted on the model was that of severing GoogleNet's original classification head, for it was modeled for ImageNet's one thousand classes, and replacing it with a binary classification head.

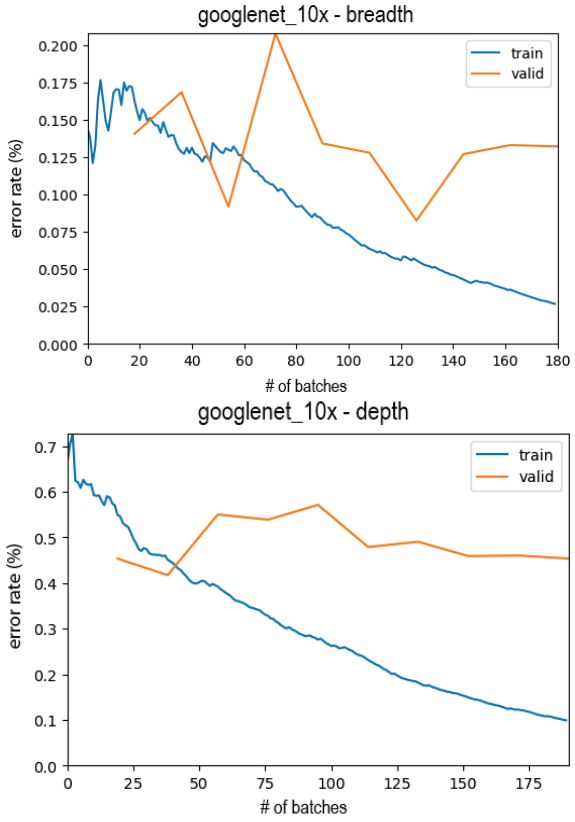

Figure 2 Plot of training and validation losses over 10 epochs

Training was conducted on two twin models in parallel, each on its respective sample-set of about 1000 images, driving Nvidia RTX 3070Ti graphics processor. The models were trained for 10 epochs each using *error rate* as the dominant metric. Each epoch took 4 seconds to compute. Results were similar for all tried models, but here we focus on the training of GoogleNet. For the breadth dimension, best results were achieved around seventh epoch, while for the depth dimension error rate kept declining through all ten epochs, although with diminishing returns. The results of best epochs were 2% error rate on the breadth dimension, and 14% error rate on the depth dimension. A surprising insight during training was, as visible in Fig. 2, the fact that the initial states of the CV models were inversely well fitted for the two tasks of semantic classification. Models were initially guessing classes on the breadth dimension correctly well beyond the binary chance probability, and on the depth dimension, missing the mark in the same manner. The other interesting aspect of those results, as visible on plots below, is the grade to which the weights of a CV model pretrained for object recognition needed to be altered to accommodate for a new task of semantic feature detection, especially for the breadth (abstract-concrete) dimension - the one that was initially good at blind guessing. Less surprising was the fact that the training for the depth (iconic-symbolic) dimension was more confounding to the algorithms, but the resulting model is still surprisingly efficient, given that there is no obvious visual cue that distinguishes those two semantic extrema. The training pipeline and detailed results have been made available online [44].

3 RESULTS AND DISCUSSION

To ensure the robustness of results, models were additionally tested on a completely new and unseen stylistically conditioned random set of images downloaded from the WikiArt's website. The respective *F1* scores are 0.978 for the breadth dimension and 0.882 for the depth dimension (Fig. 3). Those results are impressive for any practical feature detection, but state-of-the-art in general and universal semantic feature detection.

To further evaluate potential applicability of models, we downloaded entirely new dataset of 1000 images from the WikiArt's website, unconditioned on style classification, and plotted the models' predictions on a joint scatter plot. This shows how the models handle hundreds of styles unseen during training. The plot shown in Fig. 4a provides us with a couple of insights. Firstly, we can see that the breadth and depth dimensions are not strongly correlated, but there seems to be a slight imbalance in randomly downloaded sample which suggests weak artistic tendency to produce more figurative than abstract artworks when meaning to communicate a symbolic message. Likewise, there is a greater aspiration of visual artists for production of iconic rather than symbolic images (which explains model's iconic bias visible in Fig. 3).

The same plot displays how the CNN models tend to sharply distinct between classes, especially on the breadth dimension. All of those observations make real-world sense. High art is expected to be more iconic in contrast with applied art's need for heavy symbolic transmission. It is reasonable to expect that the artworks recognized as highly successful would feature clear indications of their intended subject matter and wouldn't be visually ambiguous about position of their content along the breadth dimension. Greater ambiguity on the depth dimension can be explained in couple of ways. There could be a greater tendency in visual communications to mix symbolic and iconic aspects when communicating layered messages common to high art. Moreover, Peirce's semiotic theory itself foresees that there should be a

discernible space for indexical messages amid the iconic-symbolic extrema. On the other hand, the moderate spread of distribution of artworks along the depth dimension could be the consequence of the loose method of sample collection, where the style labels by themselves could not demarcate the boundaries between the two limits of the category or that the WikiArt's labelling method wasn't stringent enough in this respect to begin with. Nevertheless, both dimensions provide a significant leverage for predicting the cultural period from which the artworks originated. By using the year of origin as a semi-reliable proxy indicator of cultural background, both dimensions show a noticeable measure of correlation with their semantic context (see Fig. 4b). Since the boundaries of visual styles are lax and cyclic, predictive power is measured through degree of mutual information.

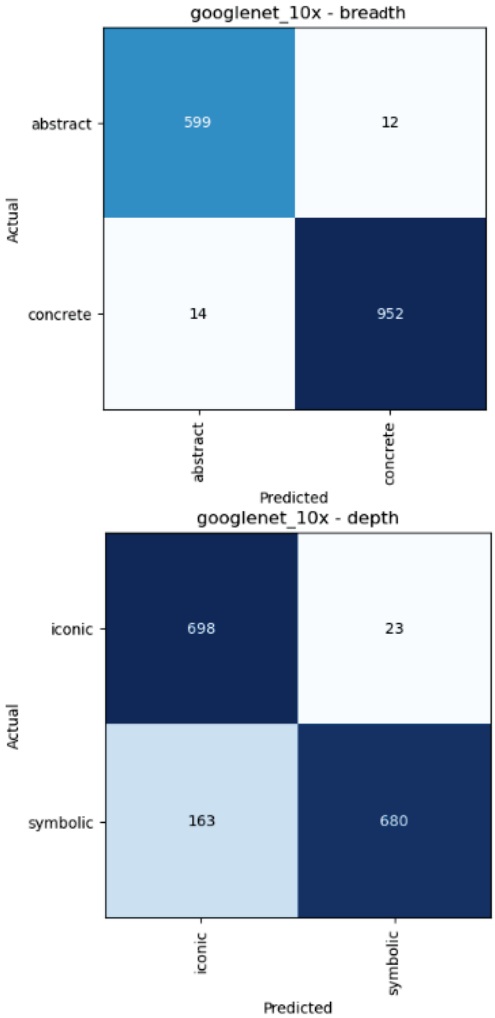

Figure 3 Confusion matrices of two fine-tuned GoogleNets, evaluated on a completely new validation set.

To get a clearer understanding of art styles distribution in this semantic space, Fig. 5 shows a scatter plot of joint distribution of random collection of artworks accompanied with style labels for random subset of 50 artworks. The plot clearly shows how the networks were able to cluster together artworks from diverse cultural periods, featuring distinct styles but associated by common semantic themes. Predominantly, symbolic and concrete artworks belong to periods of great narratives; concrete and iconic artworks convey subjective impressions and concerns; iconic and abstract works deal with avant-garde and conceptual expressions; while the abstract and symbolic art focuses on imaginative, intellectual and decorative trends. All art styles that were not presented in the training dataset found their reasonable placement within the proposed semantic space placing their artworks in a clear and explicatory relation to other, more unequivocal and less ambiguous works of art.

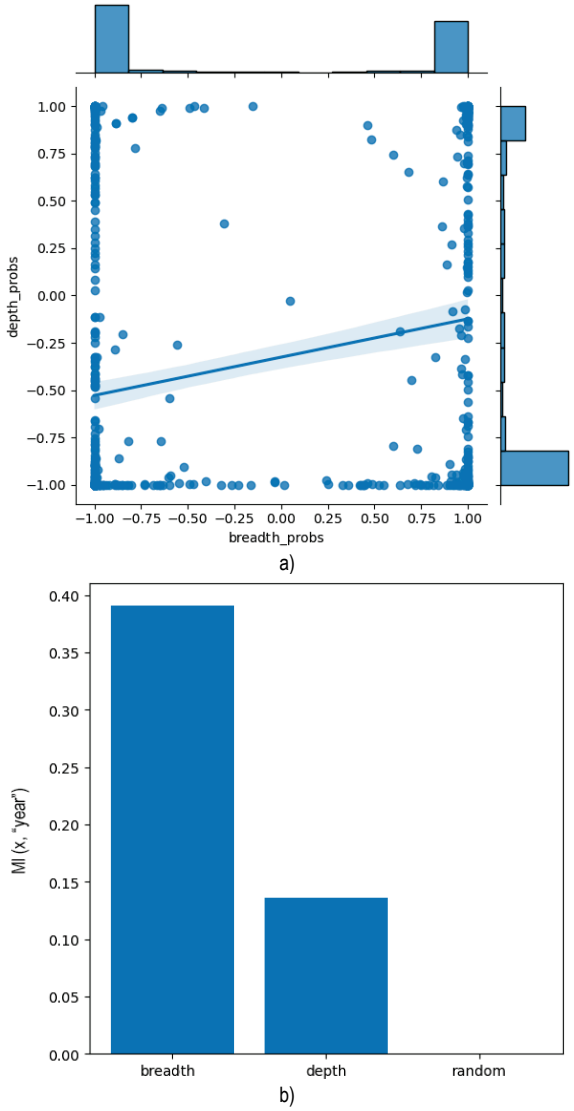

Figure 4 Diagrams of (a) joint plot of semantic feature predictions, (b) graph of mutual information between semantic features and the year of artwork's creation

Alongside aiding semantic interpretation, the model provides an idiomatic aesthetic evaluation criterion. Since all of the instances used in the sample present successful visual communications, and since the sample was picked at random from the entire body of historically and globally collected artworks, then the emptiness of large areas in the constructed space suggests the impossibility of creating successful visual communications that would feature compositional attributes congruent to those areas. If the model places the evaluated

artwork in a stranded area of semantic space, it is highly likely that the artwork wouldn't satisfy many peoples' aesthetic tastes. Likewise, visible in the same manner, if the model places the evaluated artwork in an area distant from its thematic kin, it is highly unlikely that it will efficiently communicate its intended meaning or be regarded as aesthetically significant.

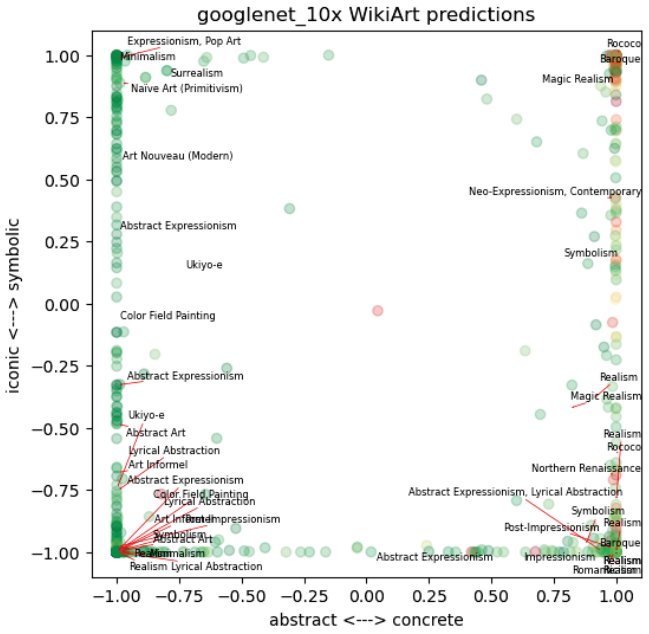

Figure 5 Scatter plot of semantic embeddings with style labels shown for a subset of sample's instances.

4 CONCLUSION

In conclusion, this study presents a novel approach to semantic feature detection using convolutional neural networks fine-tuned on stylistically conditioned images of artworks from WikiArt. The models achieved high F1 scores and provided robust results in distinguishing between abstract-concrete (breadth) and iconic-symbolic (depth) dimensions of artistic expression. The findings suggest that the proposed method can be used for universal semantic feature detection, aiding in the interpretation of visual communications across different, past and future cultural periods. Additionally, the models provide an idiomatic aesthetic evaluation criterion, allowing for the assessment of artworks' success in visual communication based on their positioning within the semantic space. Further research could explore the model's applicability expanding the scope of its application to include more diverse and historically representative samples of artworks, as well as investigating the potential applications of this method in other domains such as graphic or multimedia design.

5 REFERENCES

[1] Hoenig, F. (2005). Defining Computational Aesthetics. *Computational Aesthetics in Graphics, Visualization and Imaging*, p. 6.

[2] Greenfield, G. (2005). On the origins of the term "Computational aesthetics". *Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 9-12.

[3] Debnath, S. & Changder, S. (2020). Computational Approaches to Aesthetic Quality Assessment of Digital Photographs: State of the Art and Future Research Directives. *Pattern Recognit. Image Anal., 30*(4), 593-606. https:/doi.org/10.1134/S1054661820040082

[4] Galanter, P. (2012). Computational Aesthetic Evaluation: Past and Future. *Computers and Creativity*, McCormack, J. & d'Inverno, M., Eds. Springer Berlin Heidelberg, 255-293. https://doi.org/10.1007/978-3-642-31727-9_10

[5] Graham, D. J. & Redies, C. (2010). Statistical regularities in art: Relations with visual coding and perception. *Vision Research, 50*(16), 1503-1509. https:/doi.org/10.1016/j.visres.2010.05.002

[6] Shimamura, A. P. & Palmer, S. E. (Eds.) (2011). Aesthetic science: Connecting minds, brains, and experience. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199732142.001.0001

[7] Brachmann, A. & Redies, C. (2017). Computational and Experimental Approaches to Visual Aesthetics. *Front. Comput. Neurosci., 11*, p. 102. https:/doi.org/10.3389/fncom.2017.00102

[8] Li, R. & Zhang, J. (2020). Review of computational neuroaesthetics: Bridging the gap between neuroaesthetics and computer science. *Brain Inform, 7*(1), p. 16, https:/doi.org/10.1186/s40708-020-00118-w

[9] Spratt, E. L. & Elgammal, A. (2014). Computational Beauty: Aesthetic Judgment at the Intersection of Art and Science. https://arxiv.org/abs/1410.2488. https://doi.org/10.48550/arXiv.1410.2488

[10] Bo, Y., Yu, J. & Zhang, K. (2018). Computational aesthetics and applications. *Vis Comput Ind Biomed Art, 1*, p. 6. https:/doi.org/10.1186/s42492-018-0006-1

[11] Zhang, J., Miao, Y. & Yu, J. (2021). A Comprehensive Survey on Computational Aesthetic Evaluation of Visual Art Images: Metrics and Challenges. *IEEE Access, 9*, 77164-77187. https:/doi.org/10.1109/ACCESS.2021.3083075

[12] Valenzise, G., Kang, C. & Dufaux, F. (2022). Advances and challenges in computational image aesthetics. *Human Perception of Visual Information: Psychological and Computational Perspectives*, Springer, 133-181. https://doi.org/10.1007/978-3-030-81465-6_6

[13] Elgammal, A., Mazzone, M., Liu, B., Kim, D. & Elhoseiny, M. (2018). The Shape of Art History in the Eyes of the Machine. http://arxiv.org/abs/1801.07729. https://doi.org/10.48550/arXiv.1801.07729

[14] Mohammad, S. & Kiritchenko, S. (2018). WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art,. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[15] Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M. & Guibas, L. (2021). ArtEmis: Affective Language for Visual Art. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 11564-11574. https://doi.org/10.1109/CVPR46437.2021.01140

[16] Datta, R., Joshi, D., Li, J. & Wang, J. Z. (2006). Studying Aesthetics in Photographic Images Using a Computational Approach. *Computer Vision – ECCV 2006, vol. 3953*, Leonardis, A., Bischof, H. & Pinz, A., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 288-301. https://doi.org/10.1007/11744078_23

[17] Ke, Y., Tang, X. & Jing, F. (2006). The Design of High-Level Features for Photo Quality Assessment. *The IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06), 419-426. https://doi.org/10.1109/CVPR.2006.303

[18] Lu, X., Lin, Z., Jin, H., Yang, J. & Wang, J. Z. (2014). RAPID: Rating Pictorial Aesthetics using Deep Learning. *Proceedings of the 22nd ACM international conference on Multimedia*, 457-466. https://doi.org/10.1145/2647868.2654927

[19] Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM, 60*(6), 84-90. https://doi.org/10.1145/3065386

[20] Deng, Y., Loy, C. C. & Tang, X. (2017). Image Aesthetic Assessment: An experimental survey. *The IEEE Signal Processing Magazine, 34*(4), 80-106. https://doi.org/10.1109/MSP.2017.2696576

[21] Luo, Y. & Tang, X. (2008). Photo and Video Quality Evaluation: Focusing on the Subject. *Computer Vision – ECCV 2008, vol. 5304*, Forsyth, D., Torr, P. & Zisserman, A. Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 386-399. https://doi.org/10.1007/978-3-540-88690-7_29

[22] Mai, L., Jin, H. & Liu, F. (2016). Composition-Preserving Deep Photo Aesthetics Assessment. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016)*, 497-506. https://doi.org/10.1109/CVPR.2016.60

[23] Ma, S., Liu, J. & Chen, C. W. (2017). A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, 722-731. https://doi.org/10.1109/CVPR.2017.84

[24] Aydın, T. O., Smolic, A. & Gross, M. (2015). Automated Aesthetic Analysis of Photographic Images. *The IEEE Transactions on Visualization and Computer Graphics, 21*(1), 31-42. https://doi.org/10.1109/TVCG.2014.2325047

[25] Dhar, S., Ordonez, V. & Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 1657-1664. https://doi.org/10.1109/CVPR.2011.5995467

[26] San Pedro, J., Yeh, T. & Oliver, N. (2012). Leveraging user comments for aesthetic aware image search reranking. *Proceedings of the 21st international conference on World Wide Web*, 439-448. https://doi.org/10.1145/2187836.2187896

[27] Tang, X., Luo, W. & Wang, X. (2013). Content-Based Photo Quality Assessment. *The IEEE Transactions on Multimedia, 15*(8), 1930-1943. https://doi.org/10.1109/TMM.2013.2269899

[28] Kao, Y., He, R. & Huang, K. (2017). Deep Aesthetic Quality Assessment with Semantic Information. *The IEEE Trans. on Image Process., 26*(3), 1482-1495. https://doi.org/10.1109/TIP.2017.2651399

[29] Liu, X., Li, N. & Xia, Y. (2018). Affective Image Classification by Jointly Using Interpretable Art Features and Semantic Annotations. *Journal of Visual Communication and Image Representation, 58*. https://doi.org/10.1016/j.jvcir.2018.12.032

[30] Tian, X. (2021). Using multi-task residual network to evaluate image aesthetic quality. *The IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC2021)*, 171-174. https://doi.org/10.1109/IAEAC50856.2021.9391005.

[31] Duan, J., Chen, P., Li, L., Wu, J. & Shi, G. (2022). Semantic Attribute Guided Image Aesthetics Assessment. *The IEEE International Conference on Visual Communications and Image Processing (VCIP2022)*, 1-5. https://doi.org/10.1109/VCIP56404.2022.10008896

[32] Jappy, T. (2013). *Introduction to Peircean visual semiotics*. London: Bloomsbury.

[33] Clark, K. (1969). Civilisation: *A personal view*, 1st publ. London: British Broadcasting Corp.

[34] Hauser, A. (1999). *The social history of art. 1: From prehistoric times to the Middle Ages*. 3rd ed. London: Routledge.

[35] Davies, P. J. E., Denny, W. B., Hofrichter, F. F., Jacobs, J., Roberts, A. M. & Simon, D. L. (2016). *Janson's History of art: The Western tradition*. Reissued eighth edition. Boston: Pearson.

[36] Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014). How transferable are features in deep neural networks? https://arxiv.org/abs/1411.1792. https://doi.org/10.48550/arXiv.1411.1792

[37] Dong, C., Deng, Y., Loy, C. C. & Tang, X. (2015). Compression Artifacts Reduction by a Deep Convolutional Network. *The IEEE International Conference on Computer Vision (ICCV2015)*, 576-584. https://doi.org/10.1109/ICCV.2015.73

[38] Howard, J., Gugger, S. & Chintala, S. (2020). *Deep learning for coders with fastai and PyTorch: AI applications without a PhD*. First edition. Sebastopol, California: O'Reilly Media, Inc.

[39] Denzler, J., Rodner, E. & Simon, M. (2016). Convolutional Neural Networks as a Computational Model for the Underlying Processes of Aesthetics Perception. *Computer Vision – ECCV 2016 Workshops*, 871-887. https://doi.org/10.1007/978-3-319-46604-0_60

[40] Anwar, A., Kanwal, S., Tahir, M., Saqib, M., Uzair, M., Rahmani, M. K. I. & Ullah, H. (2022). A Survey on Image Aesthetic Assessment. https://arxiv.org/abs/2103.11616. https://doi.org/10.48550/arXiv.2103.11616

[41] Cavanagh, P. (2005). The artist as neuroscientist. *Nature, 434*(7031), 301-307. https://doi.org/10.1038/434301a

[42] Peng, K.-C. & Chen, T. (2016). Toward correlating and solving abstract tasks using convolutional neural networks. *The IEEE Winter Conference on Applications of Computer Vision (WACV2016)*, 1-9. https://doi.org/10.1109/WACV.2016.7477616

[43] Szegedy C. et al. (2015). Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), Boston, MA, USA, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[44] Ježić, T. J. (2024). Trpquo/art_critic: Repository for WikiArt model training. https://github.com/Trpquo/art_critic. (Accessed: 26-Apr-2024).

Authors' contacts:

Trpimir Jeronim Ježić, B.Sc. Eng.
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
trpimir.jeronim.jezic@grf.unizg.hr

Marko Maričević, Assist. Prof., PhD
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
marko.maricevic@grf.unizg.hr

Ivana Pavlović, Teach. Assist., PhD
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
ivana.pavlovic@grf.unizg.hr

Miroslav Mikota, Assoc. Prof., PhD
(Corresponding author)
University of Zagreb, Faculty of Graphic Arts,
Getaldićeva 2, 10 000 Zagreb, Croatia
miroslav.mikota@grf.unizg.hr