Affectless Visual Machines Explain a Majority of Variance in Human Visual Affect and Aesthetics for Natural Images

Daniel J. Graham (artstats@gmail.com)
Hobart and William Smith Colleges, Department of Psychological Science

Colin Conwell (conwell@g.harvard.edu)
Johns Hopkins University, Department of Cognitive Science

Chelsea Boccagno (cboccagno@g.harvard.edu)
Harvard T.H. Chan School of Public Health, Department of Epidemiology; Massachusetts General Hospital, Department of Psychiatry

Edward A. Vessel (evessel@ccny.cuny.edu)
City College, City University of New York, Department of Psychology

Abstract

Looking at the world involves not just seeing things, but feeling things. Feedforward machine vision systems that learn to perceive the world without physiology, thought, or feedback that resembles human affective experience offer tools to demystify the relationship between seeing and feeling, and to assess how much of affective experiences may be a function of representation learning over natural image statistics. We deploy 180 deep neural networks trained only on canonical computer vision tasks to predict human ratings of arousal, valence, and beauty for images from multiple categories (objects, faces, landscapes, art) across two datasets. We use features of these networks without additional learning, such that we linearly decode human affective responses from network activity just as one decodes information from neural recordings. We find that features of purely perceptual models predict average ratings of arousal, valence, and beauty with high accuracy: On average, models in our survey explain 53% of explainable variance in human responses; the most predictive model explains 72%. These results add to growing evidence for an information-processing account of visually-evoked affect linked to representation learning over natural statistics, and hint at a locus of affective and aesthetic valuation proximate to perception.

Keywords: Keywords: visual aesthetics; machine vision; affect

Introduction

Looking at the world usually means *feeling* the world. Though often studied in isolation, perception and affect are linked in everyday experience Merleau-Ponty & Smith (1962), at both conscious and unconscious levels Barrett & Bar (2009); Barrett & Bliss-Moreau (2009). Exposure to a beautiful, moving, inviting, or aversive visual stimulus evokes processes beyond what is often called vision, but where "seeing" stops and "feeling" begins is unknown. The intimate link between "seeing" and "feeling" in everyday experience makes disentangling the computations that undergird these processes challenging.

Here we use a survey of visual machines – which only see and cannot feel – to predict how humans respond to a diverse set of natural images. Our goal is to better isolate the unique contributions of visual perception to visually-evoked affect.

Past work in vision science Redies et al. (2007); Graham & Field (2007); Graham & Redies (2010); Hughes et al. (2010); Brachmann et al. (2017); Redies et al. (2012); Mallon et al. (2014); Graham et al. (2016) and machine learning Dong et al. (2015); Lu et al. (2014); Kong et al. (2016); Sheng et al. (2018); Goetschalckx et al. (2019); Hosu et al. (2019); Iigaya et al. (2021); AlZayer et al. (2021); Geller et al. (2022); Karjus et al. (2023) has examined relationships between image-computable statistics and aesthetic and artistic image properties. Models such as 'Emo-Net' Kragel et al. (2019), a modified machine vision system, suggest image-computable feature extraction pipelines work for predicting a variety of affective and emotional responses (e.g. fear, surprise).

These efforts provide methodological groundwork for exploring the intersection of perceptual computation and affect. However, no work at present addresses how far we can go in predicting affect with perceptual computations alone, and why such systems are effective. Here we survey 180 machine vision systems to determine the upper limit on affective prediction–specifically, prediction of human arousal, valence, and beauty ratings in response to natural images–from perceptual computations alone. We use variation across models (in terms of architecture, task, and input) to begin to answer why prediction is possible Cao & Yamins (2021a,b); Kanwisher et al. (2023). Using a 'model zoology' approach Conwell et al. (2021, 2023), we find that purely perceptual computations of 'affectless' machines can predict the majority of explainable variance in human arousal and valence and beauty ratings. We find that the ability of these machines to predict arousal, valence, and beauty is a function of representations these machines learn through *experience over many images*, i.e., their hierarchically structured knowledge of the visual world.

Methods

Our approach is to fit *linear* decoding models to features extracted from every layer of the models to assess whether information that is predictive of affect is inherent to these features, despite never being shaped to predict affective experiences.

We use OASIS Kurdi et al. (2017), a set of 900 images spanning 4 categories (people, animals, objects and scenes), with normed ratings of arousal and valence from 822 human subjects. Ratings of beauty (from 751 subjects) were obtained from a separate source Brielmann & Pelli (2019). We use a second dataset consisting of 512 images across 5 distinct categories (art, faces, landscapes, internal & external architecture) Vessel et al. (2018), but for which only ratings of beauty ("aesthetic appeal") are available. This dataset allows us to compare judgments of art to those of natural scenes, and to internally replicate a subset of the results obtained with OASIS.

We calculate two forms of reliability as gauges of the comparative performance of our models: 'mean-minus-one' reliability Vessel et al. (2018), i.e., iteratively removing one subject from the subject pool and correlating that subject's ratings with the average ratings of the subjects remaining((r_{mmo})); and split-half reliability (r_{split}), which involves splitting group-level data in half 10000 times and correlating each half with the other. This latter metric provides an upper bound (a noise ceiling) on how well any predictive model could do in predicting the mean rating.

The 180 models (252 including randomly-initialized versions of a designated subset) are sourced from 6 repositories: Torchvision (PyTorch) model zoo Paszke et al. (2019); Pytorch-image-models (timm) library Wightman (2019); VISSL (self-supervised) model zoo Goyal et al. (2021); Taskonomy (visualpriors) project Zamir et al. (2018); Sax et al. (2018, 2019); The CLIP repository Radford et al. (2021); and the SLIP repository Mu et al. (2021). To predict arousal, valence, and beauty from a given set of deep net features, we use regularized linear regression with cross-validation. Our regression pipeline con-

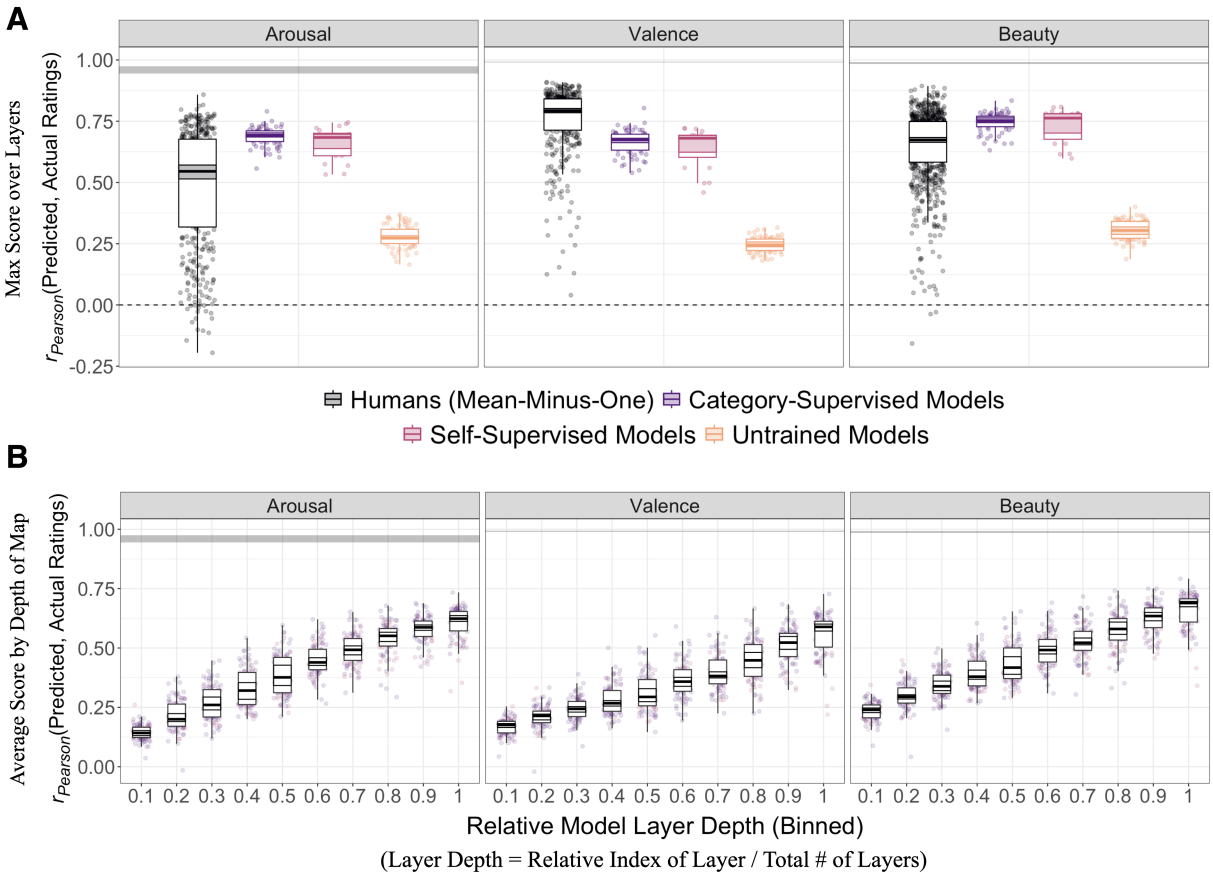

Figure 1: **Accuracy of Model-Predicted Affect Ratings.** Each point is an individual subject or model. Gray horizontal bars are Spearman-Brown split-half reliability noise ceilings for group-average affect ratings, and shaded horizontal cross-bars are 95% bootstrapped confidence interval of the mean across points; **A** shows scores of the *most predictive* layers in each model (points in orange are untrained models). Gray points are mean-minus-one correlations of individual subjects to the group average. **A** shows that the average *trained* model is (for arousal and beauty) about as predictive of group-average affect as the 32.5% most taste-typical subjects, and about 70% accurate overall. Category-supervised models (purple) are no more predictive than self-supervised models (red), but trained models are categorically more predictive than untrained models. **B** shows accuracies across layer depth for category-trained and self-supervised models. X-axis is relative depth of layer in the network (0 = earliest, 1 = deepest), binned into slices of 10. Y-axis is average accuracy in that slice, in units of $r_{Pearson}$. Each point is a model trained on ImageNet, with category-supervised models (trained on 1000-way supervision) in purple, and self-supervised models (trained with category-supervision) in red. The deepest layers are most predictive, with a nearly monotonic increase in predictivity over layers.

sists of 4 phases: feature extraction; dimensionality reduction; ridge regression; cross-validation; and scoring.

Results

We find that the average and highest affective predictive accuracies of the object recognition networks is far above the ceiling of 'shared taste', and over halfway to the noise ceiling. We also find that trained models are categorically more predictive than untrained models and that average prediction scores increase with model layer depth. Models trained with category supervision are as predictive of affect as models trained without. Details are shown and described in Figure 1.

In the Taskonomy set, representations learned for object and scene recognition are best for predicting affect. In all affective categories of OASIS (arousal, valence, and beauty) and the single affective category of the Vessel dataset (beauty), object and scene recognition tasks are the top 2 of the 24 (+1) Taskonomy task weights tested.

In the Vessel dataset, there are large differences in predictivity across image category, with Scenes and Landscapes being more predictable categories, and Person and Art among the least predictable categories.

Summary

Our work suggests that learning over natural image statistics may be central to the ontology of visually-evoked affect. We also go some way to disentangle the roles of task-related representations, hierarchical structure, and model training in predicting human affective responses with affectless perceptual machines.

References

AlZayer, H., Lin, H., & Bala, K. (2021). Autophoto: Aesthetic photo capture using reinforcement learning. In *2021 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 944–951).

Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1325–1334. (Publisher: The Royal Society London) doi: 10.1098/rstb.2008.0312

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, *41*, 167–218. (Publisher: Elsevier) doi: 10.1016/S0065-2601(08)00404-8

Brachmann, A., Barth, E., & Redies, C. (2017). Using cnn features to better understand what makes visual artworks special. *Frontiers in Psychology*, *8*, 830.

Brielmann, A. A., & Pelli, D. G. (2019). Intense beauty requires intense pleasure. *Frontiers in psychology*, *10*, 2420.

Cao, R., & Yamins, D. (2021a). Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490*.

Cao, R., & Yamins, D. (2021b). Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*.

Conwell, C., Mayo, D., Barbu, A., Buice, M., Alvarez, G., & Katz, B. (2021). Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex. *Advances in Neural Information Processing Systems*, *34*, 5590–5607.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

Dong, Z., Shen, X., Li, H., & Tian, X. (2015). Photo quality assessment with dcnn that understands image well. In *International conference on multimedia modeling* (pp. 524–535).

Geller, H. A., Bartho, R., Thömmes, K., & Redies, C. (2022). Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer. *Frontiers in neuroscience*, *16*, 999720.

Goetschalckx, L., Andonian, A., Oliva, A., & Isola, P. (2019). Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5744–5753).

Goyal, P., Duval, Q., Reizenstein, J., Leavitt, M., Xu, M., Lefaudeux, B., ... Misra, I. (2021). *Vissl.* `https://github.com/facebookresearch/vissl`.

Graham, D. J., & Field, D. J. (2007). Statistical regularities of art images and natural scenes: spectra, sparseness and nonlinearities. *Spatial vision*, *21*.

Graham, D. J., & Redies, C. (2010). Statistical regularities in art: Relations with visual coding and perception. *Vision research*, *50*(16), 1503–1509.

Graham, D. J., Schwarz, B., Chatterjee, A., & Leder, H. (2016). Preference for luminance histogram regularities in natural scenes. *Vision research*, *120*, 11–21.

Hosu, V., Goldlucke, B., & Saupe, D. (2019). Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9375–9383).

Hughes, J. M., Graham, D. J., & Rockmore, D. N. (2010). Quantification of artistic style through sparse coding analysis in the drawings of pieter bruegel the elder. *Proceedings of the National Academy of Sciences*, *107*(4), 1279–1283.

Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2021). Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features. *Nature Human Behaviour*, *5*(6), 743–755.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254.

Karjus, A., Solà, M. C., Ohm, T., Ahnert, S. E., & Schich, M. (2023). Compression ensembles quantify aesthetic complexity and the evolution of visual art. *EPJ Data Science*, *12*(1), 21.

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision* (pp. 662–679).

Kragel, P. A., Reddan, M. C., LaBar, K. S., & Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Science advances*, *5*(7), eaaw4358.

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (oasis). *Behavior research methods*, *49*(2), 457–470.

Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2014). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd acm international conference on multimedia* (pp. 457–466).

Mallon, B., Redies, C., & Hayn-Leichsenring, G. U. (2014). Beauty in abstract paintings: perceptual contrast and statistical properties. *Frontiers in human neuroscience*, *8*, 161.

Merleau-Ponty, M., & Smith, C. (1962). *Phenomenology of perception* (Vol. 26). Routledge London.

Mu, N., Kirillov, A., Wagner, D., & Xie, S. (2021). Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Redies, C., Amirshahi, S. A., Koch, M., & Denzler, J. (2012). Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In *Computer vision–eccv 2012. workshops and demonstrations: Florence, italy, october 7-13, 2012, proceedings, part i 12* (pp. 522–531).

Redies, C., Hasenstein, J., & Denzler, J. (2007). Fractal-like image statistics in visual art: similarity to natural scenes. *Spatial vision*, *21*.

Sax, A., Emi, B., Zamir, A. R., Guibas, L., Savarese, S., & Malik, J. (2018). Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. *arXiv preprint arXiv:1812.11971*.

Sax, A., Zhang, J. O., Emi, B., Zamir, A., Savarese, S., Guibas, L., & Malik, J. (2019). Learning to navigate using mid-level visual priors. *arXiv preprint arXiv:1912.11121*.

Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., & Hu, B.-G. (2018). Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th acm international conference on multimedia* (pp. 879–886).

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, *179*, 121–131.

Wightman, R. (2019). *Pytorch image models*. `https://github.com/rwightman/pytorch-image-models`. GitHub. doi: 10.5281/zenodo.4414861

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3712–3722).