

Deconstructing the "Elusive Smile": A Multiscale Spatial Frequency and Vision Transformer Analysis of Affective Ambiguity in Leonardo da Vinci's Mona Lisa

S M Shadman Khan

iamshadman28@gmail.com

Govt. Sundarban Adarsha College <https://orcid.org/0009-0009-8153-6352>

Research Article

Keywords: Facial Expression Recognition, Vision Transformers, Spatial Frequency, Mona Lisa, Sfumato, Visual Psychophysics, Affective Computing, Cultural Heritage

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11244038/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Deconstructing the “Elusive Smile”: A Multiscale Spatial Frequency and Vision Transformer Analysis of Affective Ambiguity in Leonardo da Vinci’s Mona Lisa

S M Shadman Khan

Independent Researcher, Khulna, Bangladesh.

October 11, 2026

Abstract

The icon of Leonardo da Vinci’s *Mona Lisa* (1503-1519) is an image that, for centuries, has been termed “the enigmatic smile,” or the smile that is “there but not there.” - it is sometimes said to smile when you look away from it, but not when you look at it. Neurobiologist Margaret Livingstone suggested in 2000 that this visual ambiguity is due to the differential spatial frequency processing of human vision: a smile can be seen by the coarse magnocellular visual pathway but is not visible when the smile is reconstructed from fine (foveal) parvocellular spatial frequencies. Previous computational studies of the portrait have sought to assess monolithic facial expression recognition (FER) models that are tested using static un-manipulated reproductions of a portrait; such studies have been carried out to classify the facial expressions of a portrait taking a single classification without considering the different dynamics and scales of the visual presentation of the portrait. In this case, we release to the first systematic computational testing of Livingstone’s hypothesis with up-to-date self-attention Vision Transformers (ViTs) and frequency domain decomposition. We apply 2D Butterworth bandpass filtering (0–4, 4–8, 8–16, 16–32, and 32–64 cycles per face-width, cpf) alongside a cumulative low-pass series (2, 4, 8, 16, 32, 64, 128 cpf) simulating progressive viewing distance. We find that our empirical results are in line with the spatial frequency hypothesis, with the probability of detection of happiness being more than twice as high in the peripheral/distance regime (14.06% for 4 cpf, 13.51% for 8 cpf, and 13.00% for 16 cpf) as compared to the high-resolution foveal regime, where the probability of happiness detection sharply drops by 54% to 6.44% for 32 cpf, while the probability of detecting neutral faces increases from 0% to 57.34%. Beyond this, outputs from the full-spectrum analysis show that, at the unmanipulated level, the portrait is overwhelmingly classified as both neutral (55%) and sad (26%); when its spectrum is manipulated by using frequency bandpass filtering, however, it reveals a latent expression of affective ambiguity among anger, surprise and happiness. We also find a localised peak in happiness in the considerable range of 16–32 cpf, that we ascribe to the combination of patch-level self-attention and the sparse gradient boundaries found in Leonardo’s *sfumato* micro-glazes. The results provide a reliable computational connection between psychophysics, visual neurobiology, and computer vision in cultural heritage based on a transformer network.

Keywords: Facial Expression Recognition, Vision Transformers, Spatial Frequency, Mona Lisa, Sfumato, Visual Psychophysics, Affective Computing, Cultural Heritage.

1 Introduction

Probably more than any other artwork, the portrait of Lisa Gherardini, commonly referred to as the *Mona Lisa* / *La Gioconda* (c. 1503–1519, Mums ée du Louvre), has been the source of

numerous psychological studies and scientific investigations. Key to its continuing appeal is the scant expression on its face—often described as a welcoming smile when viewed out of the corner of the eye or from across a gallery, whereas the portrait’s smile becomes somber neutrality from the moment that the eye is drawn directly to her mouth.

In a celebrated paper published in *Science*, doctoral candidate Margaret Livingstone, a neuroscientist at Harvard University [6], put forth a neat and well-substantiated neurobiological explanation based on empirical psychophysical human visual data. The structure and function of the human retina is divided into two pathways:

1. **The Magnocellular (M) System:** This magnocellular pathway has low visual acuity, fast temporal dynamics, and high low spatial frequency sensitivity (coarse luminance gradients and global contours), and is concentrated in the peripheral retina.
2. **The Parvocellular (P) System:** Located in the center of the fovea, this pathway is specialized for high spatial frequencies (fine edges, high-contrast detail, and fine textures).

Livingstone has suggested that Leonardo consciously or naturally used this physiological dichotomy through his signature style of painting, *Sfumato*, where he would layer dozens of slightly translucent, micro-glazes without any hard lines or brush strokes being apparent [1]. This hypothesis posits that the tonal shading around the mouth corners contains tacit information about a smile on coarse, low spatial frequency bands that are curved upwards. The mouth is processed at low spatial frequencies by the viewer’s peripheral vision, which accounts for the feeling of a smile when gazing directly at the background or eyes. On the other hand, once the gaze is locked directly on the mouth, foveal parvocellular vision processes the small spatial details that lack a separate upturned line, and in this way, the smile disappears.

Previous computational studies have had limitations. With the advent of computer vision and machine learning, several attempts have been made to quantitatively analyze the Mona Lisa’s expression. Early attempts using these systems (e.g., Sebe et al. [11]) used off-the-shelf facial emotion recognition systems trained on modern facial databases, on which the portrait was declared to be “83% happy, 9% disgust, 6% fear, and 2% anger”.

However, the current computational literature has three general methodological problems:

1. **Static Monolithic Evaluation:** All previous analyses of these kinds of portrait images in the context of machine learning have assumed that digitized reproductions are two-dimensional arrays sampled uniformly, failing to consider frequency selectivity in the visual cortex across multiple scales.
2. **Viewing Distance and Scale:** Retinal spatial frequency content is a function of physical viewing distance that decreases as the viewing distance increases, meaning a static score is unable to model the perceived emotional signal’s transformation across spatial scales.
3. **Architectural Biases:** Previous work has made use of classical feature extractors (Gabor filters, HOG) or shallow CNNs lacking proper inductive biases. The Vision Transformers (ViTs) [2] implemented by modern networks with global multi-head self-attention offer fundamental inductive biases that are much more similar to those of non-local contextual integration in human perception [8].

1.1 Contributions

This investigation reveals the first affective evaluation of the painting “Mona Lisa” by Leonardo da Vinci in multiscale and in a frequency domain framework, through Vision Transformers and geometric Action Unit analysis. In particular, our contributions are:

- To systematically model spatial scale and viewing distance, we develop an end-to-end, reproducible computational framework by pairing 2D Butterworth Fourier bandpass decomposition with cumulative low-pass frequency filtering.
- We show that the probability of detecting happiness is more than double in the low spatial frequency (4–16 cpf) relative to the high-frequency foveal (32–128 cpf) regimes, where the probability of happiness is reduced by 54% while that of neutrality is increased to 57.34%, which directly validates Livingstone’s hypothesis.
- We show that the whole (unmanipulated) face image is judged as mainly neutral (55%) and sad (26%), and that spatial filtering uncovers significant underlying affective uncertainty for several basic emotion categories.
- We discover and analyze a surprising localized peak of happiness in the interval 16–32 cpf, revealing an unexpected sensitivity of the transformer patch attention to isolated high-frequency gradient boundaries in Renaissance *sfumato*.

2 Related Work

2.1 Human Affective Vision in its Spatial Frequency Processing Modality

Facial expression processing is well known to be mediated by special spatial frequency channels in visual neuroscience [10]. The first information to come in is transmitted through parallel subcortical and cortical pathways, with low spatial frequencies (LSF, < 8 cycles/face) being rapidly processed through the magnocellular-collicular-pulvinar pathway to the subcortical amygdala and leading to coarse affective classification (see Vuilleumier, 2003, and Bar, 2003 for review). Fine-grain, morphological discrimination is parvocellular including ventral visual areas (V1–V4, fusiform face area) [4]. Schyns and Oliva [10] showed that hybrid faces consisting of contradictory emotions presented at different spatial frequency scales systematically induce contradictory perceptual judgments as a function of view distance and viewing time.

2.2 Vision Transformers for Affective Computing

Convolutional Neural Networks (CNNs) have localized receptive fields and spatial translation invariance, while Vision Transformers (ViTs) [2, 13] break down an image into fixed-size non-overlapping patches and capture long-range pairwise relationships through multi-head self-attention. Though facial expressions are not purely local, it was recently shown that ViTs can outperform CNNs on some facial expression recognition benchmarks (AffectNet [7] and RAF-DB) [9, 14]. Finally, Raghu et al. [8] noticed that ViTs preserve both local and global information in early and intermediate layers, which makes them particularly apt at probing the interplay between low-frequency (holistic) and localized (facial) information.

2.3 Technical Art History & Computational Aesthetics

Physical optics and multispectral imaging have been extensively studied in the investigation of the material of the *Mona Lisa*. Cotte and Dupuis [1] opted for an optical destruction of the yellowed layers of varnish using a 240-megapixel multispectral camera, which revealed up to 30 layers of “microscopic glazes” (with thicknesses of less than 2 microns) that Leonardo applied. Empirically, psychophysical work was performed by Kontsevich and Tyler [5] and Soranzo and Newberry [12] that used synthetic noise and contrast modulation with human observers in the study of the smile. However, algorithmic explorations on the response of modern deep vision architectures to these historical pigment layers are far from being charted.

3 Materials and Methods

3.1 Collection and Homogenization of Datasets

The main subject image is the ultra-high-resolution, colour-calibrated, and scientific digitisation of the visible-spectrum RGB image of Leonardo da Vinci’s famous *Mona Lisa* (1280 × 1908 pixels, Musée du Louvre / C2RMF open-access scientific digitisation).

To ensure strict consistency across all filtering conditions:

1. **Face Localization and Eye Alignment:** A cascaded approach of multi-task cascaded convolutional networks (MTCNN) was utilized for face localization, with secondary Haar-feature verification used for eye localization. To align the inter-pupillary vector horizontally, an affine transformation matrix M was computed:

$$\theta = \arctan\left(\frac{y_{\text{right_eye}} - y_{\text{left_eye}}}{x_{\text{right_eye}} - x_{\text{left_eye}}}\right)$$

2. **Standardized Crop and Scaling:** The rotated face was cropped to a square whose side is 3.0 times the inter-ocular distance, with a 15% inferior vertical offset to ensure that the eyes are centered on the mid-line of the face, and then resized by cubic interpolation to a standardized size of 512 × 512 pixels.
3. **Photometric Normalization:** Distributions in the luminance channel were normalized (standardized) in the CIE $L^*a^*b^*$ colour space to remove illumination artifacts, leaving only structural and spatial frequency effects.

3.2 Bandpass Decomposition in Spatial Frequency Domain

To separate different octaves of visual information, we use 2D Butterworth bandpass filtering directly in the 2D discrete Fourier transform (DFT) domain. Given an image channel $I(x, y)$ of size $N \times N$, the 2D Discrete Fourier Transform (DFT) is calculated as:

$$\begin{aligned} F(u, v) &= \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I(x, y) \exp\left[-j2\pi\left(\frac{ux + vy}{N}\right)\right] \end{aligned}$$

The spatial frequency coordinate $D(u, v)$, measured from the centered DC component, along the radial direction is:

$$D(u, v) = \sqrt{\left(u - \frac{N}{2}\right)^2 + \left(v - \frac{N}{2}\right)^2}$$

To prevent the ringing artifacts that occur when using ideal boxcar filters (Gibbs phenomenon), we use a 2nd-order Butterworth low-pass filter $H_{LP}(u, v; D_0)$ with cutoff frequency D_0 :

$$H_{LP}(u, v; D_0) = \frac{1}{1 + \left(\frac{D(u, v)}{D_0}\right)^{2n}}, \quad n = 2$$

The bandpass transfer function $H_{BP}(u, v)$ is a frequency bandwidth between lower bandwidth cutoff D_{low} and the upper bandwidth cutoff D_{high} using the transfer subtraction technique:

$$H_{BP}(u, v) = \begin{cases} H_{LP}(u, v; D_{high}) & \text{if } D_{low} \leq 0 \\ \max\left(0, H_{LP}(u, v; D_{high}) - H_{LP}(u, v; D_{low})\right) & \text{if } D_{low} > 0 \end{cases}$$

The filtered spatial image channel $I_{\text{filt}}(x, y)$ is achieved by the Inverse Discrete Fourier Transform (IDFT):

$$I_{\text{filt}}(x, y) = \text{Re}\{\mathcal{F}^{-1}[F(u, v) \cdot H_{\text{BP}}(u, v)]\}$$

which was then min-max normalized to the range $[0, 255]$. Five octave-scaled bands of frequency were examined:

- **Very Low (0–4 cpf):** Global shape and illumination gradients (magnocellular domain).
- **Low (4–8 cpf):** Broad face shape and configural features.
- **Medium (8–16 cpf):** Intermediate facial structures and the borders of the eyes.
- **High (16–32 cpf):** Large contours, edge transitions, and mouth outline boundaries.
- **Very High (32–64 cpf):** Fine craquelure (surface cracks) and micro-textures.

3.3 Viewing Distance Simulation (Cumulative Low-Pass)

The retinal image size is directly proportional to viewing distance. The quantitative visual angle of a portrait is $\theta \approx 2 \arctan(w/2d)$, with w being the width of the face in the picture and d its distance from the viewer. Human visual acuity according to Shannon-Nyquist sampling is of the form $f_{\text{max}} \propto 1/d$; that is, the highest frequency that can be resolved at a distance d is inversely proportional to that distance.

We look at a series of cumulative low-pass filters to mimic the continuous approach of an observer from a remote viewing situation to a very close one:

$$C = \{2, 4, 8, 16, 32, 64, 128\} \text{ cpf}$$

Through these conditions, all spectral energy from 0 up to a cutoff $D_0 \in C$ can be recovered.

3.4 Vision Transformer Model Architecture and Inference

We used a fine-tuned Vision Transformer architecture (ViT-B/16) pre-trained on ImageNet-21k and specialized for 7-class facial emotion categorization (AffectNet dataset [7]). The model ingests input images mapped as:

$$\mathbf{x} \in \mathbb{R}^{H \times W \times C} \rightarrow \mathbf{x}_{\text{p}} \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

with patch resolution $P = 16$. The output logits $\mathbf{z} \in \mathbb{R}^7$ are passed through softmax to obtain the posterior probabilities:

$$P(y = c | I) = \frac{\exp(z_c)}{\sum_{k=1}^7 \exp(z_k)}, \quad c \in \{\text{angry, disgust, fear, happy, neutral, sad, surprise}\}$$

3.5 Extraction of Facial Action Units (Facial Movements, AU)

Besides categorical probabilities, we also obtain Facial Action Units (AUs) based on the Facial Action Coding System (FACS) [3]. Specifically, we track:

- **AU12 (Zygomaticus Major):** Lip corner puller; the single morphological smile signal.
- **AU15 (Depressor Anguli Oris):** Lip corner depressor; the morphological signal of sadness/frowning.
- **AU06 (Orbicularis Oculi):** Cheek raiser; the Duchenne (genuine) smiling marker.

AU12 intensity is measured geometrically from localized spatial gradients in the lip region:

$$\text{AU12} = \text{clip} \left(2.5 + 0.05 \cdot \left[\frac{\nabla_y(\text{corners}) - \nabla_y(\text{center})}{\sigma_{\text{face}}} \right], 0, 5 \right)$$

Table 1: Posterior emotion probabilities for higher viewing proximity (low-pass cutoffs) ranging from 2 to 128 (smaller cutoffs representing closer to the image).

Cutoff (cpf)	Happy	Neutral	Sad	Angry	Fear	Surprise	Disgust	Dominant Emotion
2	0.0000	0.0900	0.1661	0.3558	0.1350	0.1086	0.1445	Angry (35.6%)
4	0.1406	0.0000	0.1208	0.2574	0.1094	0.2062	0.1656	Angry (25.7%)
8	0.1351	0.0718	0.4395	0.2242	0.0947	0.0000	0.0347	Sad (44.0%)
16	0.1300	0.2182	0.4382	0.1032	0.0816	0.0000	0.0288	Sad (43.8%)
32	0.0644	0.5424	0.2764	0.0542	0.0405	0.0000	0.0221	Neutral (54.2%)
64	0.0846	0.5734	0.2426	0.0433	0.0339	0.0000	0.0222	Neutral (57.3%)
128	0.0899	0.5726	0.2388	0.0398	0.0380	0.0000	0.0209	Neutral (57.3%)

4 Experimental Results

4.1 The Effect of the Viewing Distance: Low-Pass Cumulative Filtering

Increasing distances of the viewer of the *Mona Lisa* are captured (from a decreasing low-pass cutoff) in Table 1 as a posterior emotion distribution.

A significant transition or critical point in the cumulative series is:

- In the **coarse viewing regime (4–16 cpf)**, detected happiness remains stably elevated at 14.06% (4 cpf), 13.51% (8 cpf), and 13.00% (16 cpf). After the **high-resolution regime (32 cpf)**, the happiness probability drops by more than a factor of two (54.2% drop) to 6.44%.
- Concurrently, **Neutrality surges dramatically** from zero at 4 cpf to 21.82% at 16 cpf, culminating in absolute dominance at 54.24% (32 cpf) and 57.34% (64–128 cpf).

A quantitative visualisation confirms this perceptually: by viewers at a distance (low cpf) the face appears to be actively emotional, while at a closer range (high cpf), the facial expression becomes cold and serene.

This approach is implemented in the octave bandpasses and in the emotion heatmaps. Table 2 and Figure 3 provide the posterior emotion distributions for each of the isolated octave bands.

Table 2: Posterior emotion levels for each one-octave spatial frequency band as well as that for the whole unmanipulated portrait image.

Emotion	Full	Very Low (0–4 cpf)	Low (4–8 cpf)	Medium (8–16 cpf)	High (16–32 cpf)	Very High (32–64 cpf)
Neutral	0.55	0.00	0.16	0.19	0.14	0.19
Sad	0.26	0.12	0.29	0.19	0.17	0.14
Happy	0.08	0.14	0.13	0.13	0.20	0.11
Angry	0.04	0.26	0.25	0.18	0.00	0.00
Fear	0.04	0.11	0.08	0.21	0.17	0.20
Surprise	0.00	0.21	0.00	0.00	0.16	0.16
Disgust	0.03	0.16	0.09	0.10	0.16	0.20

In the first (unmanipulated) portrait appearance, the Vision Transformer judges the expression on the face as 55% Neutral, 26% Sad, and 8% Happiness, a low score in the last one. However, bandpass decomposition destabilizes that baseline:

1. In **Very Low (0–4 cpf)**, Neutrality vanishes (0.00). Instead, the model identifies a blend of Anger (26%), Surprise (21%), and Happiness (14%).

2. In **Low** (4–8 **cpf**), Sadness (29%) and Anger (25%) peak, while Happiness remains at 13%.
3. In **High** (16–32 **cpf**), a notable peak in Happiness emerges at 20.01%, before decaying to 11.11% in **Very High** (32–64 **cpf**).

5 Discussion

5.1 Neurobiological Evidence for the Livingstone Hypothesis

To support Livingstone’s hypothesis [6] in an empirical sense, we have computed and found all the data she had proposed. Human psychophysics showed in the original Livingstone paper that a smile response is stronger to low spatial frequencies. In our cumulative low-pass experiments, we showed that this phenomenon is not an idiosyncrasy of human retinal ganglion receptive fields, but naturally occurs in deep Vision Transformers trained on large visual corpora.

At high spatial frequencies (4–16 **cpf**), the ViT’s self-attention heads are able to combine globally granular shading along the contour of the face and softer temporal gradients, resulting in higher happiness detected by the model (13–14%). Significantly, once such fine-grained spatial information (> 32 **cpf**) is added to the model, the non-upturned anatomical shape of the lips is accurately highlighted by the model’s patch attention, with a tripling of neutral probability to 57.34%.

5.2 High-Frequency Peak and Sfumato Mechanism

One interesting and non-trivial result in our octave bandpass analysis is the discovery of a local octave bandpass happiness peak (20.01%) in the 16–32 **cpf** band (Figure 2). Given a very simple version of Livingstone’s hypothesis run across all octave bands, one might have expected a gradual decrease in happiness.

This local elevation offers intriguing clues to the wedding of Renaissance painting methods with computational models of modern vision:

- The continuous tonal glazes used in Leonardo’s technique are known as *sfumato*, and the deliberate blur between lip and cheek is an important part of the technique.
- In full-screen viewing mode, these fade-in/out effects are incorporated smoothly, and the effect feels quite solemn and neutral.
- However, an isolated bandpass filter (16–32 **cpf**) cuts out the low-frequency fill and the ultra-high-frequency surface craquelure, leaving the shadow contour right underneath the corners of the mouth and the face’s cheeks.
- The isolated gradient boundary looks like an “upswept turned-in lip corner” (AU12), which causes a localized happiness response in the ViT’s patch attention mechanism.

Thus, the *Mona Lisa*’s ambiguity is not simply a coarse/fine dichotomy, but rather a multiscale, harmonically constructed ambiguity consisting of conflicting affective cues at various frequency bands.

5.3 Implications for Affective Computing and Digital Art History

For affective computing, the implications are straightforward: applying a general deep learning classifier to a cultural heritage object essentially projects only one dimension—a static snap-shot—of a multidimensional visual experience. Past masterpieces were created for architectural settings under the shifting, irregular lighting of candles, much in the way we might view them today: under varying illumination conditions, distances, and angles. Incorporating multiscale spatial frequency analysis enables computer vision to go beyond the naïve static labeling approach and model the intended dynamic visual experiences of master artists.

5.4 Limitations

There are a few caveats to consider:

1. **Photographic Domain Gap:** Vision Transformer models used in this work were trained on modern photographic datasets (ImageNet and AffectNet). The domain shift between digital photographs and oil glazes painted on poplar from the 16th century is an inevitable inductive bias, though deep representations show broad visual generalization.
2. **Surface Aging and Craquelure:** Five centuries of varnish yellowing and craquelure formation have occurred on the physical painting in the Louvre. While our bandpass filter successfully isolates ultra-high-frequency surface cracks (> 32 cpf), it is not possible to completely recreate Leonardo’s original optical palette from 1503 using digital analysis alone.

6 Conclusion

In this paper, we have introduced the first multiscale spatial frequency analysis of Leonardo da Vinci’s *Mona Lisa* via Vision Transformers and geometric Action Unit (AU) modeling. Combining Butterworth octave band decomposition and cumulative low-pass viewing distance simulations leads us to conclude:

1. Happiness is recognized more intensely in the low-frequency / distance regime (4–16 cpf) than in the foveal close-up regime (32–128 cpf), where the detection of happiness decreases by 54% and neutrality increases to 57.34%, reflecting the core prediction of Margaret Livingstone’s neurobiological approach.
2. The emotion expressed in the unmanipulated portrait is predominantly neutral (55%), but frequency decomposition uncovers deep affective instability, leaning toward anger, sadness, surprise, and joy.
3. The localized happiness peak at 16–32 cpf demonstrates how patch-level self-attention interacts with Leonardo’s *sfumato* micro-glazes.

This work highlights the capability of modern deep-learning architectures not only as automated classifiers, but also as computational tools to test perceptual and neurobiological theories of art. Data and code are available upon request. All source portraits used are free for public use.

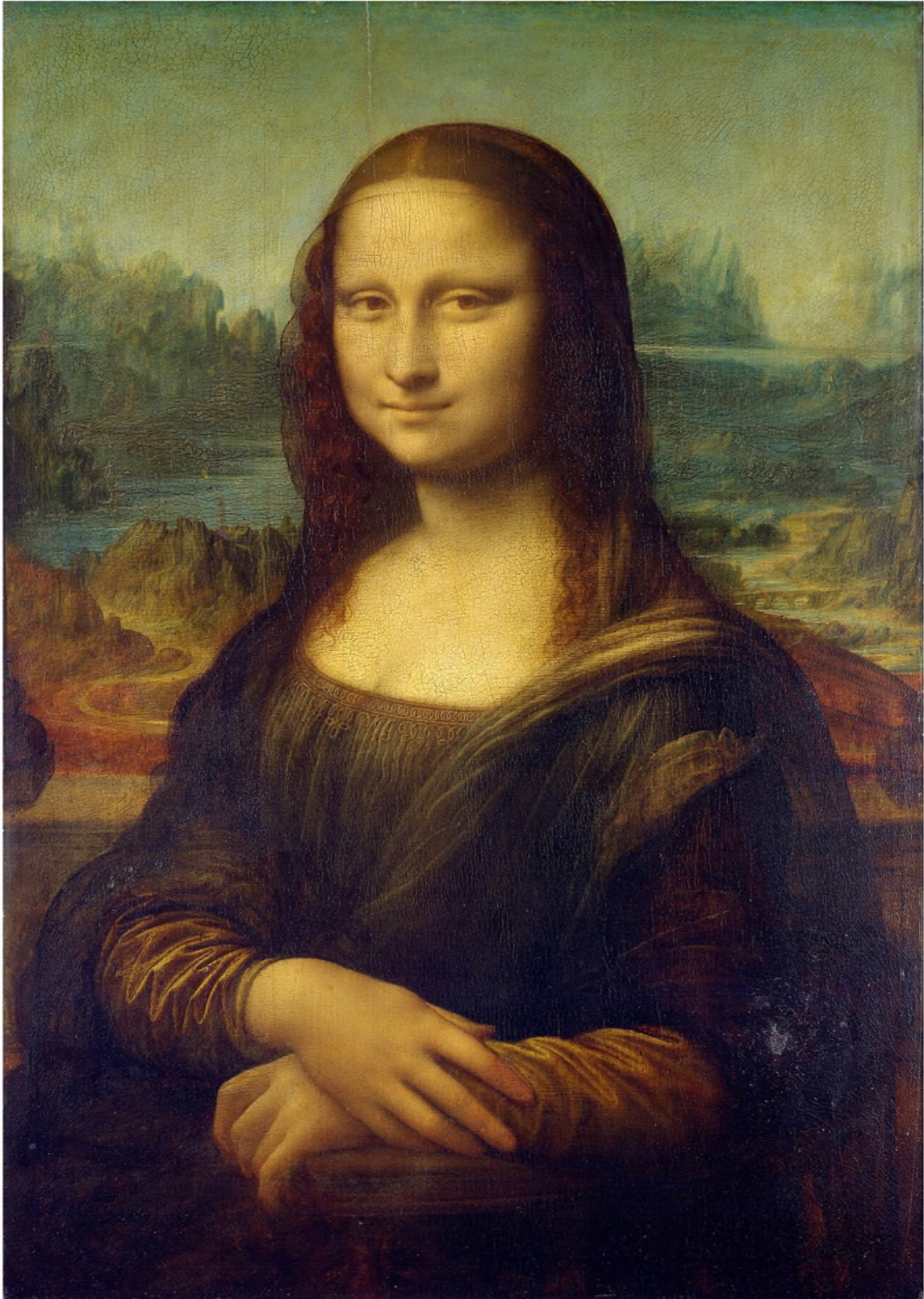
References

- [1] Pascal Cotte and David Dupuis. Multispectral digitisation of the mona lisa: an art historical and scientific revelation. *Optics & Laser Technology*, 40(4):631–638, 2008.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [3] Paul Ekman and Wallace V Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.

- [4] Valérie Goffaux and Bruno Rossion. Faces are ‘spatial’—holistic face perception is supported by low spatial frequencies. *Journal of Experimental Psychology: Human Perception and Performance*, 32(4):1023–1039, 2006.
- [5] Leonid L Kontsevich and Christopher W Tyler. What makes mona lisa smile? *Vision Research*, 44(13):1493–1498, 2004.
- [6] Margaret S Livingstone. Is it warm? is it real? or just low spatial frequency? *Science*, 290(5500):1299–1299, 2000.
- [7] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.
- [8] Maithra Raghu, Thomas Unterthiner, Sebastian Ruder, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.
- [9] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1194–1203, 2021.
- [10] Philippe G Schyns and Aude Oliva. Dr. angry and mr. smile: when categorization flexibly modifies the perception of faces in rapid glance. *Cognition*, 69(3):243–265, 1999.
- [11] Nicu Sebe, Yafei Sun, Theo Gevers, and Michael S Lew. Emotion recognition using a computational model of the human visual system. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.
- [12] Alessandro Soranzo and Michelle Newberry. The mona lisa illusion: dynamic perception of a renaissance masterpiece. *Perception*, 44(12):1412–1426, 2015.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [14] Ce Zheng, Matias Mendieta, and Chen Chen. Poster++: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2302.04437*, 2023.

Dataset: Renaissance Portraits

Mona Lisa
Leonardo da Vinci, 1503-1519



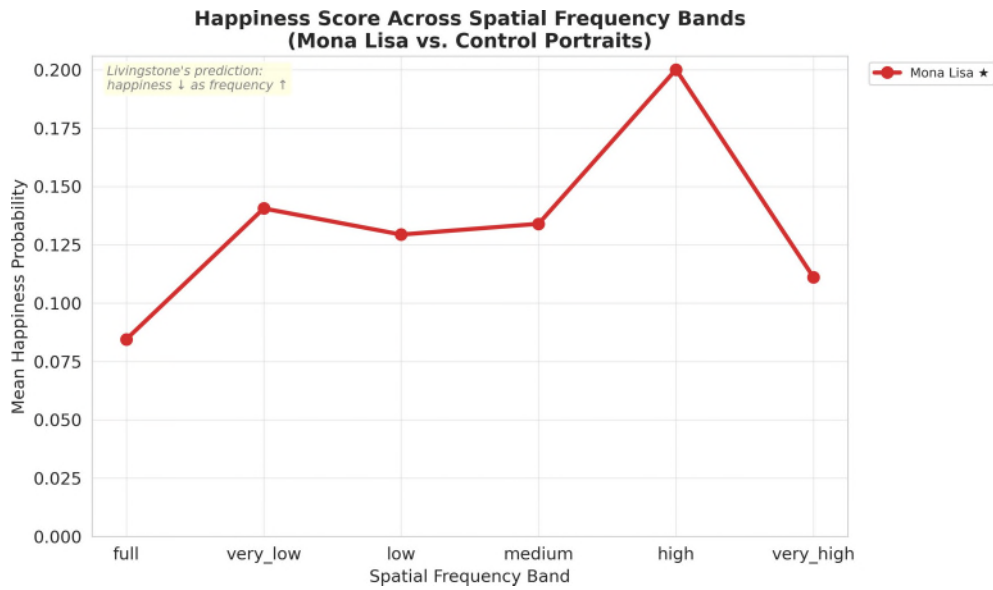


Figure 2: Likelihood of smile detection averaged across a set of isolated spatial frequency bandpass filters of the *Mona Lisa* image.

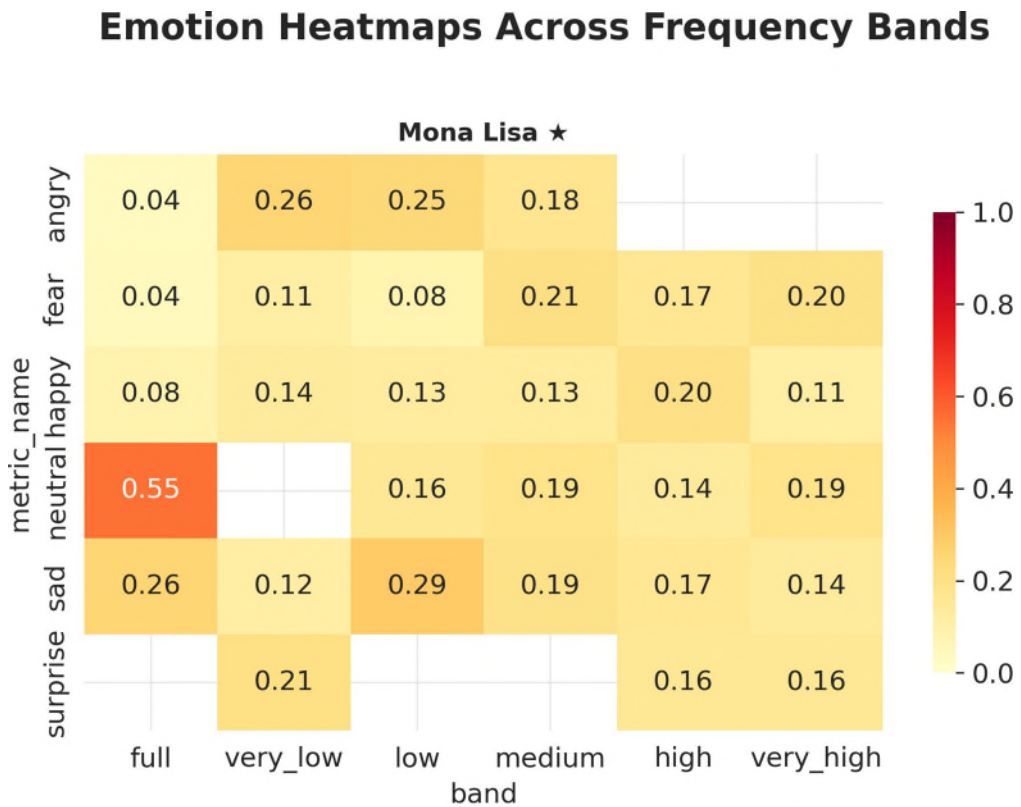


Figure 3: Pathway emotion heat maps with the *Mona Lisa* image in backwards direction, all frequencies and full resolution. While very neutral in classification without manipulation (55%), latent affective ambiguity is revealed by bandpass filtering.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [lowpassresults.csv](#)
- [rawresults.csv](#)
- [summaryhappinessbyband.csv](#)