Geometry matters: insights from Ollivier Ricci Curvature and Ricci Flow into representational alignment

Nahid Torbati TORBATI@CBS.MPG.DE
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Michael Gaebler GAEBLER@CBS.MPG.DE
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany

Simon M. Hofmann* SIMON.HOFMANN@CBS.MPG.DE
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Nico Scherf* NSCHERF@CBS.MPG.DE
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany

Editors: List of editors' names

Abstract

Representational similarity analysis (RSA) is widely used to analyze the alignment between humans and neural networks; however, conclusions based on this approach can be misleading without considering the underlying representational geometry. Our work introduces a framework using Ollivier Ricci Curvature and Ricci Flow to analyze the fine-grained local structure of representations. This approach is agnostic to the source of the representational space, enabling a direct geometric comparison between human behavioral judgments and a model's vector embeddings. We apply it to compare human similarity judgments for 2D and 3D face stimuli with a baseline 2D-native network (VGG-Face) and a variant of it aligned to human behavior. Our results suggest that geometry-aware analysis provides a more sensitive characterization of discrepancies and geometric dissimilarities in the underlying representations that remain only partially captured by RSA. Notably, we reveal geometric inconsistencies in the alignment when moving from 2D to 3D viewing conditions. This highlights how incorporating geometric information can expose alignment differences missed by traditional metrics, offering deeper insight into representational organization.

Keywords: Ollivier Ricci Curvature, Ricci Flow, Representational Alignment, Convolutional Neural Networks

1. Introduction

Understanding how biological and artificial neural networks (ANNs) solve common tasks such as face recognition or object classification in natural scenes remains a central challenge at the intersection of cognitive science and machine learning. A key question is whether these systems, despite their architectural and physical differences among others, develop similar internal representations when faced with the same computational demands. This has given rise to the study of *representational alignment*, which seeks to uncover not only how much

. * Equal contributions.

representations across systems correspond, but also *how* and *why* such (dis)similarities emerge (Sucholutsky et al., 2023), ultimately exploring both universal and idiosyncratic aspects of representations across systems. Common methods to behaviorally probe the representational space in humans and computational models use perceptual downstream tasks such as similarity judgments (e.g., in *odd-one-out* designs) between objects (Mahner et al., 2025) or faces (Hofmann et al., 2025). *Representational Similarity Analysis* (RSA) is a multivariate method for comparing representational structures across modalities, especially between biological and artificial systems (Kriegeskorte et al., 2008). By correlating Representational Dissimilarity Matrices (RDMs) or similarity kernels, RSA captures second-order isomorphisms, comparing the relational structure of stimulus representations, rather than directly mapping activity patterns. Despite its popularity, RSA is limited by its reliance on Euclidean geometry, which can distort the intrinsic structure of representational spaces. Empirical work further shows that human similarity judgments often deviate from Euclidean assumptions (Rodriguez and Granger, 2017), restricting RSA's ability to capture alignment grounded in the true geometry of the data.

While RSA is widely used, multiple studies highlight critical limitations. Geometrical similarity does not guarantee comparable representational content, as different coding schemes can yield similar RDM structures (Laakso and Cottrell, 2000). Dataset and feature dependencies can inflate similarity, such as between texture-based DNNs and shape-based human vision (Dujmović et al., 2022). Diagonal entries in RDMs can artificially boost correlations (Ritchie et al., 2017), and dissimilarity estimates suffer from bias, covariance, and noise sensitivity, prompting methods such as multivariate noise normalization, cross-validation, and unbiased estimators (Walther et al., 2016; Diedrichsen et al., 2020). Moreover, human similarity judgments often deviate from Euclidean assumptions, aligning better with Riemannian geometry (Rodriguez and Granger, 2017) and violating symmetry and triangle inequality constraints (Tversky, 1977; Tversky and Gati, 1982). Finally, estimation biases tied to task design and within-run correlations further undermine standard RSA, motivating Bayesian approaches that model raw data directly (Cai et al., 2019).

To address the limitations of RSA and capture the intrinsic representational geometry, we model each space as a graph, a discrete analog of a manifold, and apply Ollivier–Ricci curvature (ORC) (Ollivier, 2007) to quantify local connectivity via optimal transport between neighborhood probability measures. These analyses enabled us to quantify both local and global (dis)similarities between representations. In the human dataset, we further compared 2D and 3D viewing conditions, examining their representational congruence and how it is embedded in the underlying geometry.

Main contributions

1. **A novel geometry-aware framework**: We introduce an embedding-agnostic framework for representational alignment using Ollivier Ricci Curvature and Ricci Flow. This approach moves beyond the Euclidean assumptions of traditional Representational Similarity Analysis (RSA) to capture and quantify the intrinsic local geometry and community structure of representational spaces.

2. **Revealing the limits of behavioral alignment**: We demonstrate that successfully aligning a network to human choices does not guarantee an alignment of the underlying representational geometry, especially under a modality mismatch: We train

a 2D-native network (VGG-Face) on human similarity judgments derived from 3D stimuli and find that while task performance is high, the network fails to reconstruct the geometric complexity of human 3D perception. Our method reveals this failure, whereas traditional metrics suggest a moderate alignment.

3. **Identifying alignment deficits to form relevant semantic structures**: Using our framework, we show the network's inability to form key semantic structures, here, male and female clusters, suggesting limitations of the 2D-native model to align its representational space with the human behavioral data from the richer dynamic 3D experience.

2. Related work

Representational alignment Sucholutsky et al. (2023) explore the challenging concept of representational alignment, which assesses the degree to which internal representations between different information processing systems - biological and artificial - correspond. A central approach to bridge neuroscientific domains is Representational Similarity Analysis (RSA), which allows for the comparison of representational structures across measurement modalities Kriegeskorte et al. (2008). RSA contrasts how brain regions encode information using representational differences matrices (RDM) to capture differences in neural responses between stimuli. RSA has been widely used to compare artificial and biological systems, evaluating how well computational models replicate the representational geometry of processes across brain regions. Studies have employed RSA to contrast supervised and unsupervised learning models and to assess the impact of model scale, architecture, training data, and objective functions on alignment with neural representations Khaligh-Razavi and Kriegeskorte (2014); Yamins and DiCarlo (2016); Muttenthaler et al. (2022); Sucholutsky and Griffiths (2024). Beyond RSA, a variety of complementary techniques have been developed to investigate neural network representations. These range from global similarity metrics based on kernel alignment (Kornblith et al., 2019), to finer-grained methods that assess the preservation of local neighborhood structure (Kolling et al., 2023) Huh et al. (2024). Other approaches leverage geometric principles, such as anchor-based encodings (Moschella et al., 2022) (Cannistraci et al., 2023) or spectral maps (Fumero et al., 2024), to enable invariant comparisons between representational spaces. This body of work reflects a trend towards capturing more nuanced structural properties beyond what is revealed by a single global similarity score. Huh et al. (2024) use a mutual nearest-neighbor (mNN) metric to measure how often two models share the same top k neighbors per sample, capturing local similarity better than global metrics like Centered Kernel Alignment (CKA; Kornblith et al. (2019)). Applying mNN across models and modalities, it shows converging representations driven by scale, task diversity, and simplicity bias, supporting the idea of a shared "Platonic" representation in AI.

Ollivier Ricci Curvature and Ricci Flow: Tools for structural analysis The classical notion of Ricci curvature from Riemannian geometry was extended to discrete settings by Ollivier through an optimal transport framework (Ollivier, 2007), giving rise to Ollivier–Ricci curvature (ORC) on graphs and networks. Hamilton's Ricci flow (Hamilton, 1982), originally formulated as the geometric evolution of curvature, analogous to heat

diffusion, was later discretized and adapted to graphs and networks by Ni et al. (2019). Together, these concepts provide a geometric lens for analyzing complex network structures. The discretized Ricci flow has been employed as a principled approach for community detection and network alignment (Ni et al., 2019, 2018). ORC itself has been widely adopted across diverse domains: for instance, Topping et al. (2021) introduced Balanced Forman curvature, inspired by ORC, to characterize the "over-squashing" phenomenon. Extensions of ORC further by incorporating dynamical similarities, revealing how curvature distributions evolve and identify bottleneck edges (Topping et al., 2021). In weighted graphs, it has been used to quantify the local "bending" or geometric distortion at individual edges (Sandhu et al., 2015). In finance, it provides a novel lens to evaluate systemic risk and market fragility (Sandhu et al., 2016). In neuroscience, ORC has proven effective in tracking age-related changes in brain networks as well as connectivity alterations associated with autism spectrum disorder, uncovering both region-specific and global patterns of structural and functional differences (Farooq et al., 2019; Elumalai et al., 2022; Yadav et al., 2023).

3. Methods

Ricci curvature, Ollivier Ricci curvature In Riemannian geometry, curvature measures how a manifold deviates from being locally Euclidean, with Ricci curvature specifically quantifying this deviation in tangent directions Riemann (2016); Samal et al. (2018). It influences the average spread of geodesics in those directions and the rate at which the volume of distance balls and spheres expands.

Ricci curvature affects how the volume of a ball grows with its radius and the overlap between two balls, which depends on their radii and the distance between centers. Greater overlap implies lower transport cost, linking Ricci curvature to optimal transport. Building on this idea, Ollivier introduced a generalized Ricci curvature for metric measure spaces based on optimal transport Ollivier (2007, 2009). For a metric space (X, d) equipped with a probability measure m_x for each $x \in X$, the Ollivier Ricci curvature (ORC) κ_{xy} along a path xy is defined as follows:

$$\kappa_{xy} = 1 - \frac{W(m_x, m_y)}{d(x, y)} \tag{1}$$

where $W(m_x, m_y)$ is the Wasserstein distance. ORC provides an edge-based measure of local connectivity in graphs. Positive curvature indicates cohesive neighborhoods with low transport cost, while negative curvature highlights structural bottlenecks or sparse connections. Intuitively, ORC captures how easily "mass" flows across edges, distinguishing well-connected regions from areas of separation.

Analysis of curvature distributions. To quantitatively characterize the shape of the Ollivier Ricci Curvature distributions from each graph, we modeled them using Gaussian Mixture Models (GMMs). This allows us to identify whether a distribution is simple and unimodal or complex and multimodal. The models were implemented using the scikit-learn (version 1.3.2) library Pedregosa et al. (2011) in Python. In addition, we applied the Wasserstein distance to measure the difference between the curvature distributions, more precisely, Wasserstein 1-distance-1 (WS_1). For the computation, the library SciPy (version 1.10.1) in Python has been used Virtanen et al. (2020).

Ricci flow and Ricci flow-metric The Ricci flow method, based on the geometric concept of curvature introduced by F. Gauss and B. Riemann, describes how the space bends at each point (Perelman, 2002; Gauss, 1828; Riemann, 2016). Areas with high positive curvature are denser, while regions with negative curvature are less so. Hamilton developed the Ricci flow, a curvature-driven diffusion process, which deforms space similarly to heat diffusion (Hamilton, 1982). regions with large positive curvature contract, while those with strong negative curvature expand.

Ni et al. (2019) adapted Ricci flow from Riemannian geometry to discrete networks, using it to detect community structures within graphs. The discrete Ricci flow algorithm on a network is an evolving process; in each iteration, all edge weights are updated simultaneously by the following flow process:

$$w_{xy}^{i+1} = d^i(x,y) - \kappa^i(x,y) \cdot d^i(x,y) \tag{2}$$

where w_{xy}^i is the weight of the edge xy in the i-th iteration, and κ_{xy}^i is the curvature value at the edge xy in the i-th iteration, and $d_{(x,y)}^i$ is the shortest path distance in the graph induced by the weights w_{xy}^i.

Ricci flow-metric, the Ricci flow-induced distance is defined based on the definition of Ricci flow (Ni et al., 2018). It is calculated between nodes using the weighted shortest path, where the edge weights are derived from the final output of the flow process (see equation 5). More details and explanations of computation in Appendix B.

Heat diffusion distance To compare graphs based on edge curvature, we use the graph diffusion distance (Hammond et al., 2013), which quantifies the average similarity of heat diffusion patterns and enables comparison between weighted graphs. Diffusion is simulated by initiating a delta impulse at a vertex and allowing it to evolve over time t; differences in adjacency matrices produce distinct diffusion patterns. The distance between two graphs G_1 and G_2 and for time t can be calculated with:

$$d(G_1, G_2; t) = \|exp(-tL_1) - exp(-tL_2)\|_F^2 \tag{3}$$

where $\|.\|_F$ is the Frobenius norm and L_1 and L_2 are Laplacian matrices of G_1 and G_2, respectively.

Human dataset The human behavioral data (**Human Judgment**) were obtained from a large-scale online experiment (Hofmann et al., 2025). 2,710 participants judged similarities among 100 generated faces (3D head models; $n_{\text{female}} = n_{\text{male}} = 50$; 161,700 trials) in a triplet odd-one-out task under static 2D and dynamic 3D viewing conditions. In the 2D condition, 3D head models (faces) were shown statically from a frontal perspective; in the 3D viewing condition (or 2.5D condition; (Marr, 1982)) the head-models were continuously rotated such that participants could integrate various perspectives in their similarity judgments.

General approach **VGG-Face** is relatively simple by modern standards, yet achieves strong performance in face recognition (Parkhi et al., 2015). Human judgments were modeled with a custom-aligned VGG-Face, which we call **Aligned-VGG**, trained to predict human odd-one-out choices in triplets of face images. Similarity judgments are expressed as similarity matrices. From these, we construct a unified graph-based representation using

adaptive-KNN graphs for both synthetically generated datasets (Figure 6, Appendix B) and model representational spaces, with construction details given in Appendix B. We then compute Ollivier–Ricci curvature (ORC) for each representation (Human Judgment, Aligned-VGG, and VGG-Face) and apply the discrete Ricci flow algorithm. The resulting flow-metric is used to analyze the representations and their underlying geometric properties.

4. Results

Human-like geometry emerges in 2D viewing conditions but generalizes weakly to 3D. To investigate how behavioral alignment shapes representational geometry, we compared human judgments with the aligned and baseline networks across both 2D and 3D viewing conditions (Figure 1). While a human-like geometry is successfully induced in the 2D condition, our analysis shows that this representational alignment is much weaker under the 3D input modality, despite behavioral alignment being comparably strong across conditions (performance scores of Aligned-VGG 2D: 0.72, and 3D: 0.69; noise-ceiling corrected; for details, see Hofmann et al. (2025)). In the **2D condition** (Figure 1a), the RDMs for both Human Judgment (left) and the Aligned-VGG network (middle) show a clear block-diagonal structure when using our flow-metric, indicating a shared geometric organization based on gender. This visual similarity is confirmed by the high quantitative correlation between them (r=0.63). This result establishes that, under matched conditions, behavioral alignment can indeed induce a human-like representational geometry.

However, this alignment does not readily generalize when the network is aligned to behavior from the **3D viewing condition** (Figure 1b). The clear geometric structure in the Aligned-VGG model disappears; its flow-metric RDM (middle column, third row) now lacks the distinct block structure seen in both the 2D case and in the 3D Human Judgment RDM.

Critically, this reduced matching of geometric structure is largely missed by traditional metrics. While the standard Euclidean and Cosine correlations between Human Judgment and Aligned-VGG remain moderate (r=0.54 and r=0.43, respectively), the flow-metric correlation reduces to 0.33. This demonstrates that while the network can still match human choices at a surface level, it doesn't fully reconstruct the underlying geometry of human perception, a misalignment that our geometry-aware method reveals. This change in alignment is specifically tied to the semantic structure of the faces, as the network loses the clear separation of the female face cluster (first 50 images) that is prominent in the human data.

Curvature distributions reveal geometric misalignment. To understand the geometric source of the reduced alignment described above, we analyzed the Ollivier Ricci Curvature (ORC) distributions for each representation using three complementary tools: Gaussian Mixture Models (to identify structural complexity in the distribution), Wasserstein distance (to measure the (dis)similarity of distributions), and heat diffusion distance (to measure global graph (dis)similarity including geometrical information; see Methods for details). This analysis confirms the 2D/3D divergence: the geometry of the Aligned-VGG network mirrors human judgments in the 2D condition but regresses toward the simpler geometry of the baseline VGG-Face in the 3D condition (see Methods, Table 1 and Table 2 in Appendix B).

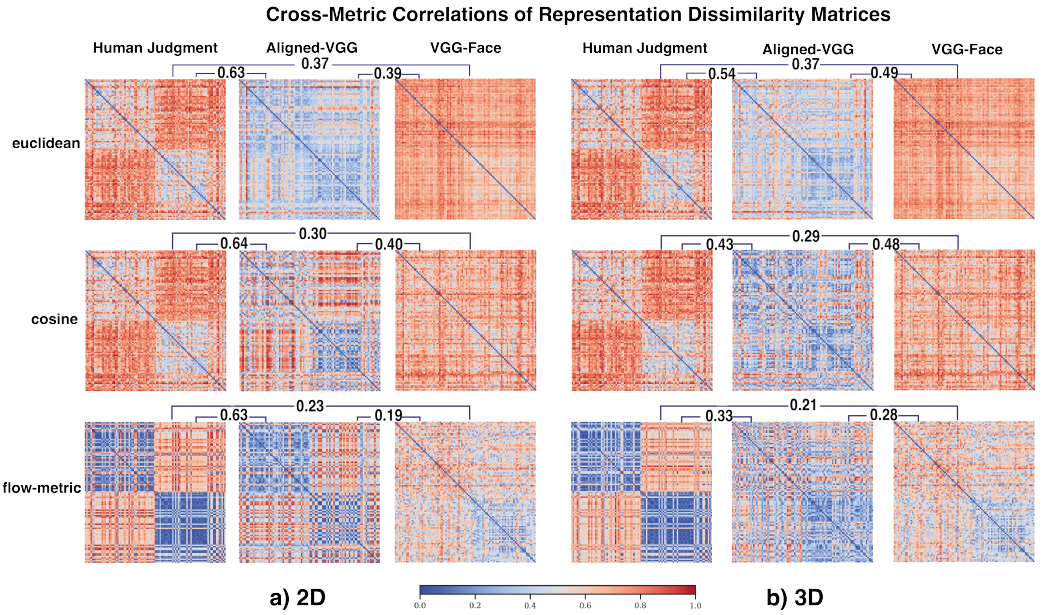


Figure 1: Representational dissimilarity matrices (RDMs) and corresponding alignment scores for (a) 2D and (b) 3D conditions. To compare traditional methods with our geometric approach, we compute RDMs using three distance metrics: *Euclidean* (a standard baseline), *cosine* (a strong empirical baseline Diedrichsen et al. (2020)), and our novel *flow-metric*, which captures intrinsic geometry. Alignment scores (top of each panel) are the Pearson correlation between respective RDMs. The first 50 images (e.g., rows) correspond to female faces and the next 50 images to male faces, revealing a gender-based block structure.

In the **2D condition** (Table 1) the curvature profiles of Human Judgment and Aligned-VGG are highly similar, reflected by a low Wasserstein distance between their distributions (0.04) and a strong correspondence in their heat diffusion patterns (1.6). Both representations exhibit a complex, multi-modal curvature distribution, suggesting a nuanced geometric structure (see Figure 7 and Table 2 in Appendix B).

In contrast, in the **3D condition**, the geometry of the Aligned-VGG model does not show human-like complexity and remains statistically closer to the baseline VGG-Face. Both networks fit a simple, single-Gaussian curvature distribution (Table 2), and the Wasserstein and heat diffusion distances between them are smaller than their distances to Human Judgment(Table 1). In contrast, the Human Judgment representation is geometrically more complex, with two distinct components in its curvature distribution (Table 2).

This indicates that the behavioral signal from the dynamic 3D viewing condition, when fed to the 2D-native model, was insufficient to impose the complex geometric structure of human perception. This critical geometric separation is robust across a range of graph construction parameters. Increasing connectivity parameters modestly improves the alignment score only between Aligned-VGG and VGG-Face, while Human Judgment remains consistently separate (Figure 9 and Figure 11 in Appendix B).

Table 1: Representational comparison across models in 2D and 3D conditions for HJ (Human Judgment), A-VGG (Aligned-VGG), and VGG-Face. Values are the mean and standard deviation of the Wasserstein (WS) and heat diffusion distances, computed over 100 iterations on random subsets of 70 data points to ensure statistical robustness.

Model	**HJ vs. A-VGG**	**HJ vs. VGG-Face**	**A-VGG vs. VGG-Face**
WS-distance 2D	0.04 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
WS-distance 3D	0.07 ± 0.01	0.13 ± 0.01	0.07 ± 0.01
Heat-distance 2D	1.6 ± 0.04	2.05 ± 0.03	1.98 ± 0.03
Heat-distance 3D	1.69 ± 0.04	2.08 ± 0.04	1.79 ± 0.03

Community structure reveals divergent alignment Finally, to understand how these local geometric differences give rise to meso-scale organization, we used discrete Ricci flow (Ni et al., 2019). Ricci flow is a curvature-driven process that contracts well-connected regions and expands sparse ones, thereby it provides a principled way to detect emerging community structure in the graphs (Figure 2). The results confirm the pattern of 2D alignment and the reduced alignment in 3D. In the 2D condition (Figure 2a), the Human Judgment and Aligned-VGG graphs exhibit similar community patterns, both showing a clear tendency to cluster faces by gender. In contrast, the baseline VGG-Face model lacks a coherent structure.

This structure is less pronounced in the model variant trained on data of the 3D condition (Figure 2b). Although the community structure in human judgments also shifts between the 2D and 3D conditions (e.g., by the number of clusters), the Aligned-VGG model fails to capture this organization. Instead, its graph more closely resembles the baseline VGG-Face, a finding supported by quantitative community metrics (see Appendix B and Table 3).

By tracking representative faces across the graphs, we can get visual insights into semantic structures that do not generalize (Figure 2), indicating that emergence of distinct clusters in the 3D Aligned-VGG is more limited compared to 2D. This is exemplified in clusters of female face images, which are successfully formed during 2D alignment, however, appear incoherent in the 3D condition, highlighting the network's failure to learn the geometry of human perception under a modality mismatch (see Appendix B for a fine-grained Representational Profile Analysis, which confirms that this geometric divergence is driven by the incomplete emergence of the female face cluster; see Figure 13 and Figure 14).

5. Discussion and Conclusions

In summary, our study yields three key findings: First, by employing a novel geometry-aware framework that leverages curvature to incorporate local geometric information when measuring alignment between representations, we show that traditional metrics can miss important discrepancies in the underlying representational geometry. Thereby, our approach reveals inherent limitations of behavioral alignment, that is, high task performance does not guarantee a human-like internal organization. Second, it demonstrates that the geo-

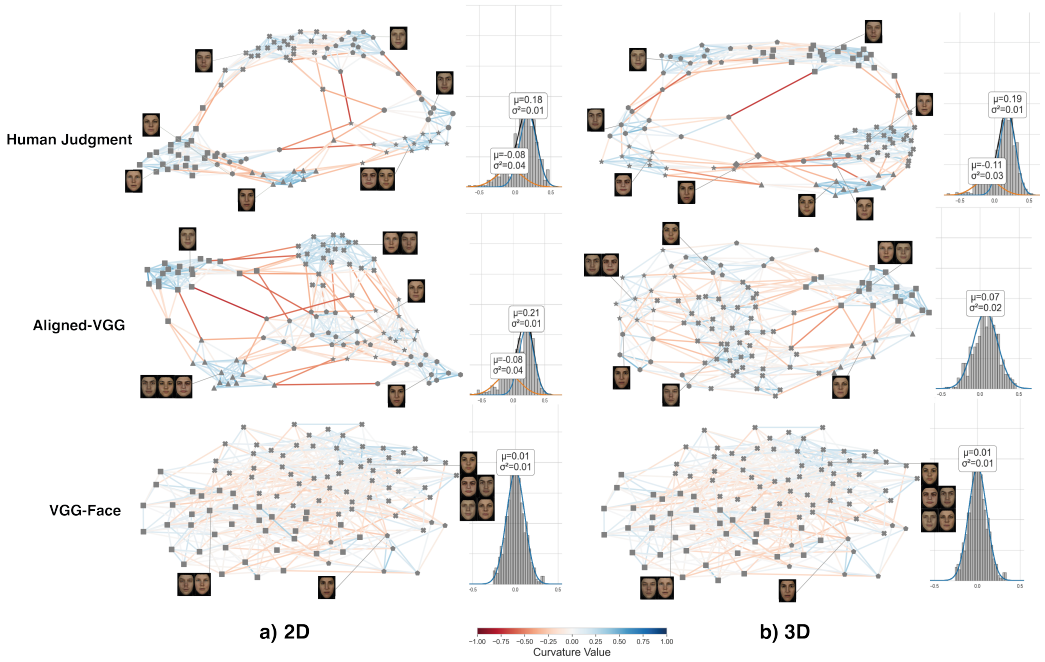


Figure 2: Communities (different symbols) detected by Ricci flow in 2D and 3D viewing conditions. To visualize cluster coherence, representative faces were selected from the communities in the 3D Human Judgment graph and are shown across all panels. Edge colors indicate curvature (red=negative, blue=positive), where red edges highlight intra-community relations; their absence in the 3D Aligned-VGG graph reveals a lack of coherent cluster structure. The GMM component values of the curvature distributions are also shown in the distribution of curvature values.

metric consequences of changing alignment conditions (here, from 2D to 3D)–while task-performance remains relatively constant–are not fully captured by the standard distant metrics. Third, our framework uncovers a model's incapacity to form a coherent semantic structure (here, for female face images in the 3D viewing condition). More specifically, our results indicate that behavioral feedback in the dynamic 3D condition imposes weaker constraints on the model's representational geometry compared to the static 2D case. That is, even when guided by human judgments, the 2D-native network (VGG-Face) fails to converge to a geometrically similar representation. This suggests that successful alignment requires the model's representational capacities (e.g., processing 3D information) to match the complexity of the human experience it is intended to capture.

Future work should test whether these geometric alignment effects generalize across architectures (e.g. transformers (Vaswani et al., 2017)), stimuli (naturalistic inputs), and geometric formulations, and explore geometrically guided alignment methods that incorporate elements of our proposed framework.

To conclude, challenging the limits of classic similarity metrics, our work highlights the potential of discrete geometric tools to provide a more rigorous and insightful lens for comparing perception and representation across biological and artificial systems.

Acknowledgments

N.S., N.T, and S.H. are supported by BMBF (Federal Ministry of Education and Research) through ACONITE (01IS22065) and the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI.) Dresden/Leipzig. N.T. is also supported by the Max Planck IMPRS CoNI Doctoral Program. M.G. and S.H. were funded by a cooperation between the Max Planck Society and the Fraunhofer-Gesellschaft (project NEUROHUM).

References

Ming Bo Cai, Nicolas W Schuck, Jonathan W Pillow, and Yael Niv. Representational structure or task structure? bias in neural representational similarity analysis and a bayesian method for reducing bias. *PLoS computational biology*, 15(5):e1006299, 2019.

Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà. From bricks to bridges: Product of invariances to enhance latent space communication. *arXiv preprint arXiv:2310.01211*, 2023.

Jörn Diedrichsen, Eva Berlot, Marieke Mur, Heiko H Schütt, Mahdiyar Shahbazi, and Nikolaus Kriegeskorte. Comparing representational geometries using whitened unbiased-distance-matrix similarity. *arXiv preprint arXiv:2007.02789*, 2020.

Marin Dujmović, Jeffrey S Bowers, Federico Adolfi, and Gaurav Malhotra. Some pitfalls of measuring representational similarity using representational similarity analysis. *bioRxiv*, pages 2022–04, 2022.

Pavithra Elumalai, Yasharth Yadav, Nitin Williams, Emil Saucan, Jürgen Jost, and Areejit Samal. Graph ricci curvatures reveal atypical functional connectivity in autism spectrum disorder. *Scientific reports*, 12(1):8295, 2022.

Hamza Farooq, Yongxin Chen, Tryphon T Georgiou, Allen Tannenbaum, and Christophe Lenglet. Network curvature as a hallmark of brain structural connectivity. *Nature communications*, 10(1):4937, 2019.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. *arXiv preprint arXiv:2406.14183*, 2024.

Carl Friedrich Gauss. *Disquisitiones generales circa superficies curvas*. Dieterich, 1828.

Richard S Hamilton. Three-manifolds with positive ricci curvature. *Journal of Differential geometry*, 17(2):255–306, 1982.

David K Hammond, Yaniv Gur, and Chris R Johnson. Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel. pages 419–422, 2013.

Simon M Hofmann, Anthony Ciston, Abhay Koushik, Felix Klotzsche, Martin N Hebart, Klaus-Robert Müller, Arno Villringer, Nico Scherf, Anna Hilsmann, Vadim V Nikulin, et al. Dynamic presentation in 3d modulates face similarity judgments–a human-aligned encoding model approach. 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10 (11):e1003915, 2014.

Camila Kolling, Till Speicher, Vedant Nanda, Mariya Toneva, and Krishna P Gummadi. Pointwise representational similarity. *arXiv preprint arXiv:2305.19294*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Aarre Laakso and Garrison Cottrell. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76, 2000.

Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135, 2015. doi: 10.3758/s13428-014-0532-5.

Florian P Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, volume 2. MIT Press, Cambridge, Mass., 1982. ISBN 978-0-262-51462-0.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*, 2022.

Chien-Chun Ni, Yu-Yao Lin, Jie Gao, and Xianfeng Gu. Network alignment by discrete ollivier-ricci flow. In *International symposium on graph drawing and network visualization*, pages 447–462. Springer, 2018.

Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with ricci flow. *Scientific reports*, 9(1):9984, 2019.

Yann Ollivier. Ricci curvature of metric spaces. *Comptes Rendus Mathematique*, 345(11): 643–646, 2007.

Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Grisha Perelman. The entropy formula for the ricci flow and its geometric applications. *arXiv preprint math/0211159*, 2002.

Bernhard Riemann. *On the hypotheses which lie at the bases of geometry*. Birkhäuser, 2016.

J Brendan Ritchie, Stefania Bracci, and Hans Op de Beeck. Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, 148:197–200, 2017.

Antonio M Rodriguez and Richard Granger. The differential geometry of perceptual similarity. *arXiv preprint arXiv:1708.00138*, 2017.

Areejit Samal, RP Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jürgen Jost. Comparative analysis of two discretizations of ricci curvature for complex networks. *Scientific reports*, 8(1):8650, 2018.

Romeil Sandhu, Tryphon Georgiou, Ed Reznik, Liangjia Zhu, Ivan Kolesov, Yasin Senbabaoglu, and Allen Tannenbaum. Graph curvature for differentiating cancer networks. *Scientific reports*, 5(1):12323, 2015.

Romeil S Sandhu, Tryphon T Georgiou, and Allen R Tannenbaum. Ricci curvature: An economic indicator for market fragility and systemic risk. *Science advances*, 2(5):e1501495, 2016.

Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*, 2021.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Amos Tversky and Itamar Gati. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123, 1982.

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.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137:188–200, 2016.

Yasharth Yadav, Pavithra Elumalai, Nitin Williams, Jürgen Jost, and Areejit Samal. Discrete ricci curvatures capture age-related changes in human brain functional connectivity networks. *Frontiers in Aging Neuroscience*, 15:1120846, 2023.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Appendix A. Human experiment data

Face stimuli and human similarity judgments Stimulus images were computed using the 3D reconstruction model DECA (Feng et al., 2021) applied to 2D portraits of the Chicago Face Dataset (Ma et al., 2015). Human face similarity judgments (n = 194,261) were acquired in the form of a triplet-odd-one-out task from 2,710 participants (age range 18 - 65, mean age = 31.9 ± 11.2 years) in an online experiment. For more details on the stimulus set and experimental design, we refer to Hofmann et al. (2025).

Human-aligned VGG-Face The pre-trained VGG-Face architecture (Parkhi et al., 2015) was adopted to predict human face similarity judgments in the experiment (Figure 3). First, all layers up to the fully connected layer FC7 were frozen, making their weights non-trainable (*VGG core*). Second, subsequent layers were replaced with one FC layer (*VGG bridge*), which converts a 4,096-dimensional input to a 300-dimensional vector. A *decision block* was added, consisting of convolutional layers. This block receives stacked activation maps from the bridge for each input image in a triplet (x_i, x_j, x_k), resulting in a 6x300 matrix $[a_i, a_j, a_i, a_k, a_j, a_k]$. The first convolutional layer in the decision block has 2

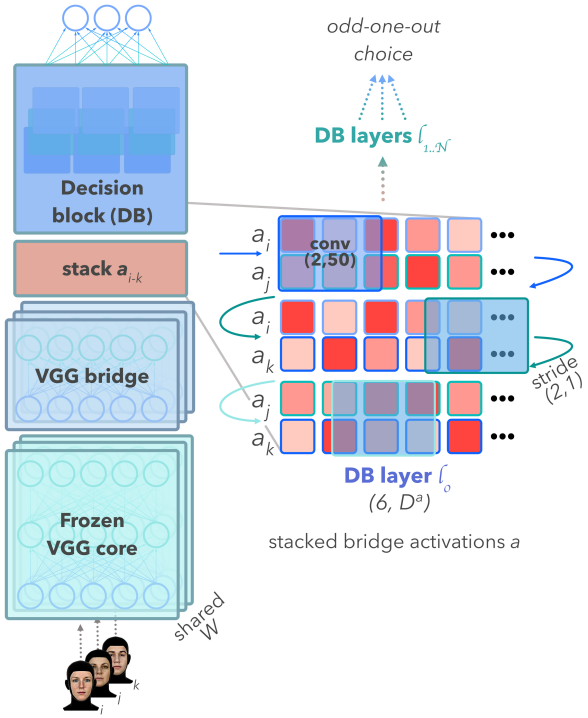


Figure 3: **Human-aligned VGG-Face**. The network is trained to predict human judgments in a face similarity task.

filters of size (2, 50) and stride (2, 1), producing an output of size (batch size, 2, 3, 251). After applying a ReLU activation, another convolutional layer with one filter of size (3, 100) and stride (1, 1) is applied, followed by another ReLU. This results in a (batch size, 2, 1, 152) output. Then, the signal was down-sampled to (batch size, 1, 1, 3) using two more convolutional layers (one filter each, kernel sizes: (1, 100) and (1, 51)) with an intermediate ReLU. The resulting 3-length output vector indicates the model's choice, where the highest value identifies the odd-one-out. The architecture was trained using cross-entropy loss with the Adam optimizer, a learning rate of $5e^{-4}$, and a batch size of 16. The data (X: triplet images, Y: human choices) was split into training (70%; $n_{train} = 135,982$), validation (15%; $n_{val} = 29,139$), and test sets (15%; $n_{test} = 29,140$). For more details, see Hofmann et al. (2025).

Appendix B.

B.1. Graph construction

To construct the graph from vector embeddings, we employ an adaptive nearest-neighbor method. First, we calculate the density as the inverse of the distance using a k-nearest neighbors (KNN) density kernel, incorporating the parameters k_{min} and k_{max}, which specify

the minimum and maximum number of neighbors, respectively. Next, we normalize the local density by scaling it between the minimum and maximum density values, thereby defining the density at each data point. Based on this normalized density, we determine the number of neighbors for each data point and proceed to construct the graph. The same approach is applied to the similarity matrix derived from human similarity judgments. First, a distance matrix is constructed from the similarity matrix provided. Using this distance matrix, the method estimates an appropriate k-value for each data point (image) based on the distance distribution from that point. This results in an **adaptive KNN** structure for the dataset, which in turn forms the foundation for an adaptive graph construction. We initialized the parameters with $k_{min} = 5$ and $k_{max} = 10$, where $k_{min} = 5$ was chosen as the smallest value that ensured connected graphs across all representations. To assess the robustness of our findings, we further evaluated the results over additional parameter pairs. The results show a stable and consistent pattern over a range of parameter choices. Because of limited space, we include only representative results for selected values of $k_{\min}$ and $k_{\max}$.

B.2. Ricci Flow Metric Analysis

Ricci flow, originally introduced by Richard S. Hamilton in the context of Riemannian geometry, is a geometric evolution process analogous to heat diffusion. It iteratively updates the metric tensor to smooth out irregularities in the underlying geometry. In this process, regions of positive curvature contract while regions of negative curvature expand. This concept was notably employed by Perelman in his proof of the Poincaré conjecture, where it facilitated the geometric decomposition of 3-manifolds.

$$\frac{\delta}{\delta t} g_t = -2Ric(g_t) \tag{4}$$

The notion of Ricci flow has since been extended to discrete settings, such as graphs. In particular, Ni et al. (2019) proposed a discrete Ricci flow framework in which edge weights evolve based on their corresponding curvature values, specifically, the Ollivier-Ricci curvature. At each iteration, all edge weights are updated simultaneously. Analogous to the continuous case, the discrete Ricci flow shrinks edges with positive curvature and stretches those with negative curvature. Building on this framework, we apply the discrete Ricci flow process to our graph data, iterating the flow for a fixed number of steps (30 iterations). We initialize all edge weights to one, $w^0_{xy} = w_{xy}$ and $d^0_{xy} = d_{xy}$ in formula 2, allowing the final weights to reflect the extent to which each edge must be contracted or expanded based on its curvature. This transformation reveals intrinsic geometric properties of the graph and facilitates geometry-aware comparisons across graphs.
Let $G = (V, E)$ be a graph. For the fixed source $s \in V$ and target $t \in V$ the shortest path (distance) between s and t is a sequence of vertices, $v_0 = s, v_1, ..., v_k = t$, with $(v_{i-1}, v_i) \in E$. Let $W : E \to \mathbb{R}$ be the final weights, resulted from the Ricci flow process, assigned to each edge $(i, j) \in E$. The **flow-metric** can be defined as follows:

$$d_{fm} = \sum_{i=1}^{k} w_{i-1,i}(v_{i-1}, v_i) \tag{5}$$

for each $w_{i-1,i} \in W$.

B.3. RDMs comparison

To have a better overview of the comparison between the two viewing conditions, 2D and 3D, across different metrics, we provided the visualisation of the RDM matrices in Figure 4. In addition, the result of the pair-wise correlation between RDMs, in Figure 5, is also provided.

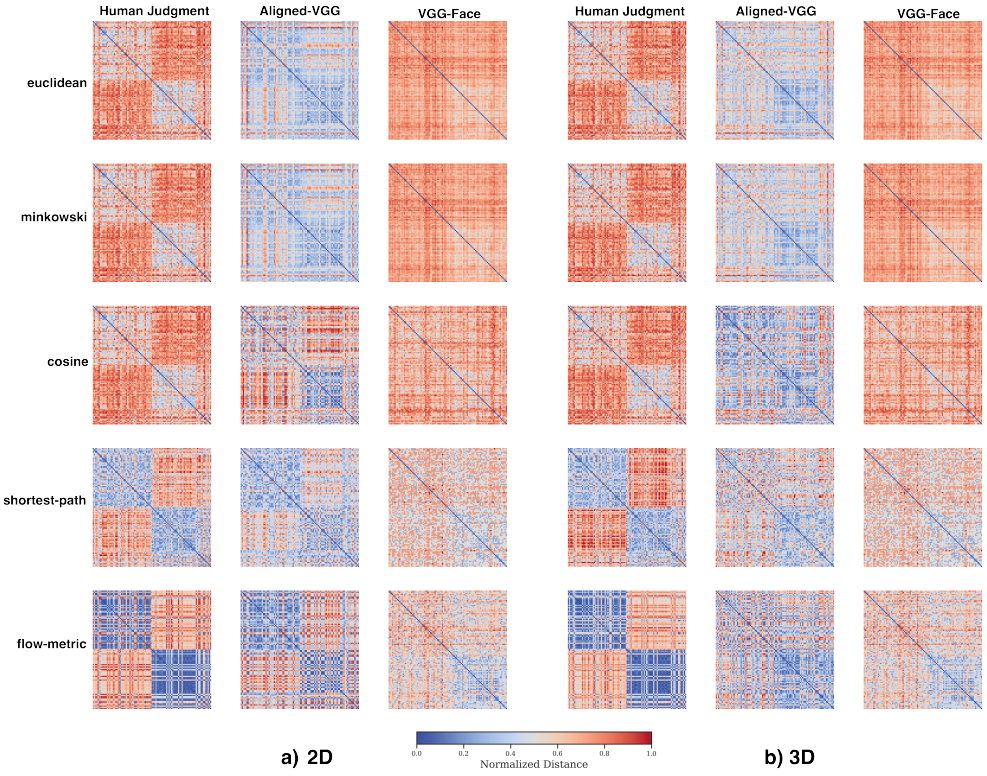


Figure 4: Representational Dissimilarity Matrices (RDMs) of all representations using different metrics for $(k_{\min}, k_{\max}) = (5, 10)$ for a) 2D and b) 3D conditions.

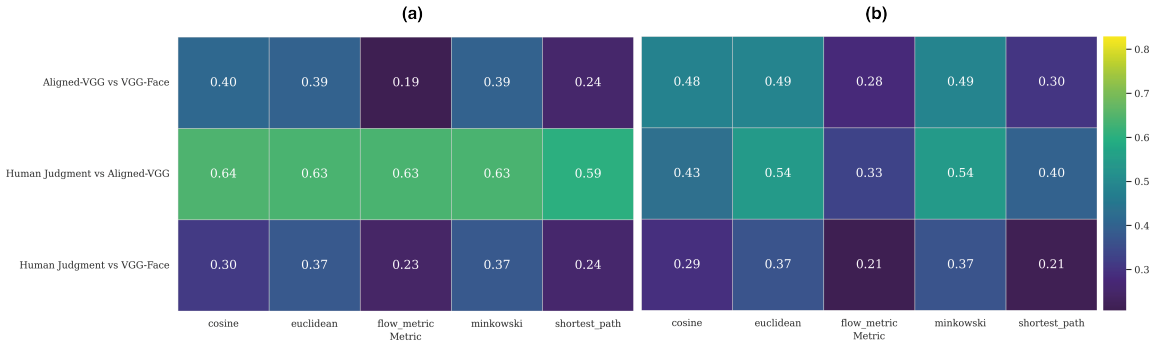


Figure 5: Correlation of RDMs of all pairs of representations using different metrics with $(k_{\min}, k_{\max}) = (5, 10)$ for a) 2D and b) 3D conditions.

B.4. Synthetic data

To evaluate whether our representational alignment approach, based on a graph framework, can capture local geometric properties while comparing different representations with varying underlying geometries, we generated several synthetic datasets. Specifically, we created a 2D torus dataset and a dataset resembling a 2D version of a Swiss roll, as illustrated in Figure6. To examine the extent to which variation in the underlying geometry can be captured by the heat diffusion distance, we transformed the original data using the sigmoid function $S(x) = \frac{1}{1+e^{-x}}$. The data sets are shown in Figure 6. We then constructed a graph representation of each data set as outlined above and computed the pairwise heat diffusion distance. Figure 6 shows the discrepancy captured by the heat diffusion distance, indicating the underlying structural differences between the datasets and their sigmoid compressions. Comparing the distance values within and across data sets shows that the heat diffusion distance could distinguish underlying structures that are more similar within a dataset and its transformation than between datasets or their transforms.

B.5. Gaussian Mixture Models

The results of the GMMs for representations by increasing the $k_{\min}$ and $k_{\max}$ for 2D viewing condition are provided in Figure 7 and for 3D condition in Figure 8. We fix the number of components at 5, thereby enabling the model to infer the optimal number of GMM components.

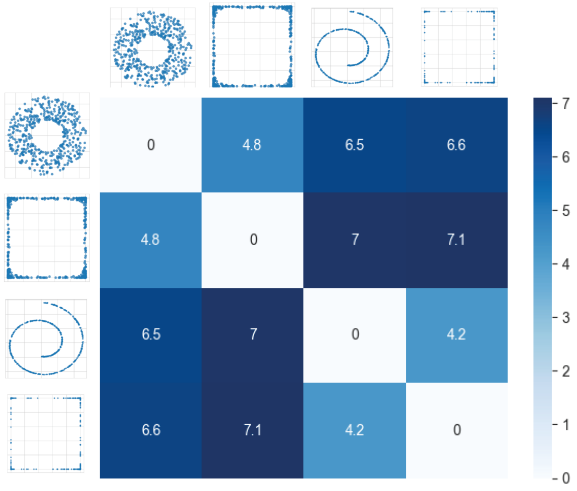


Figure 6: Pairwise heat diffusion distances between the different synthetic datasets. From top to bottom (left to right): 2D torus, transformed 2D torus, 2D swiss roll, transformed 2D swiss roll. Colors indicate heat diffusion distance.

Table 2: Results of the GMMs in 2D and 3D conditions.

Model	**n-components**	**BIC**	**Mean**	**Covariances**
Human Judgment 2D	2	-277	$(0.18, -0.08)$	$(0.01, 0.04)$
Human Judgment 3D	2	-309	$(0.19, -0.10)$	$(0.01, 0.03)$
Aligned-VGG 2D	2	-213	$(0.21, -0.08)$	$(0.01, 0.04)$
Aligned-VGG 3D	1	-461	0.07	0.02
VGG-Face	1	-901	0.01	0.01

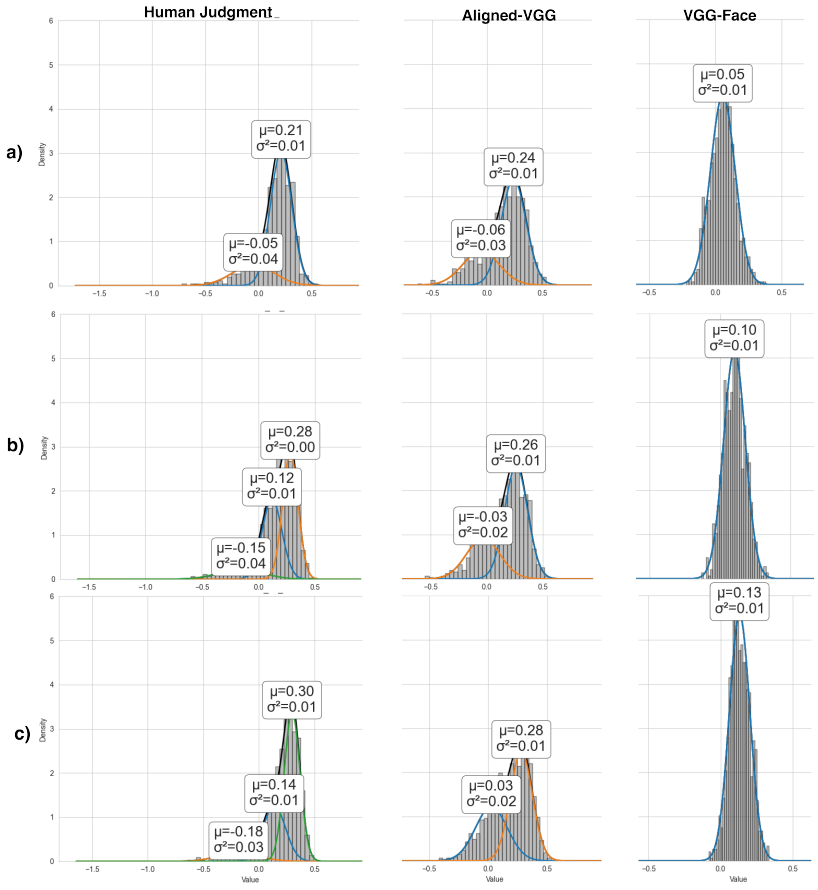


Figure 7: GMMs of curvature distributions of all representations in 2D viewing conditions for (k_{min}, k_{max})= a) (5,15), b) (10,15), and c) (10,20).

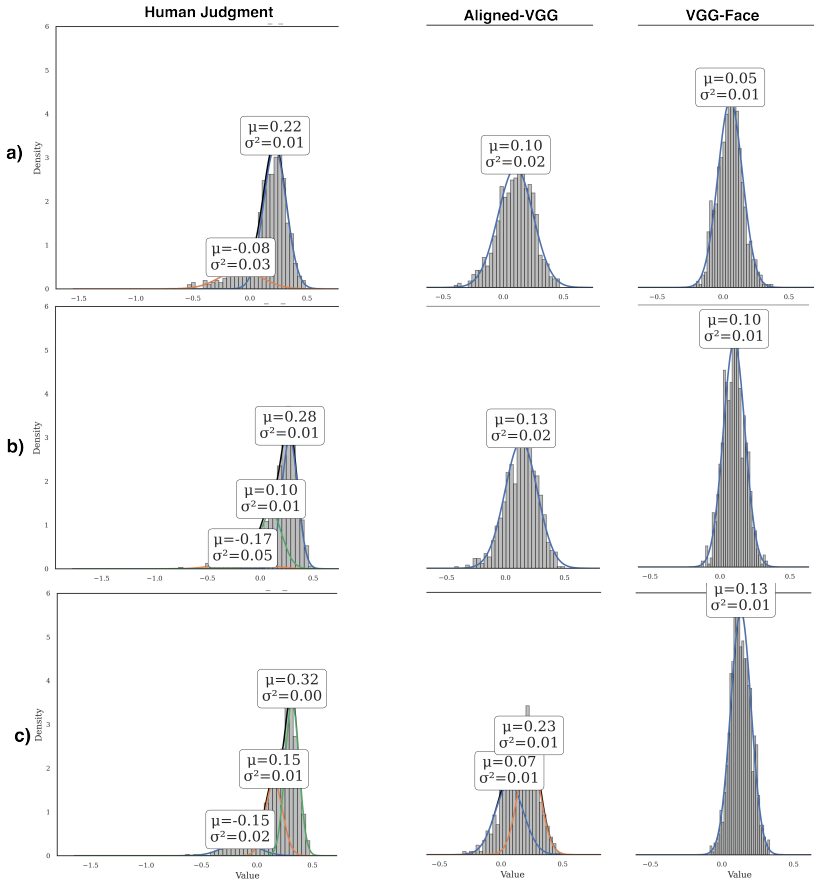
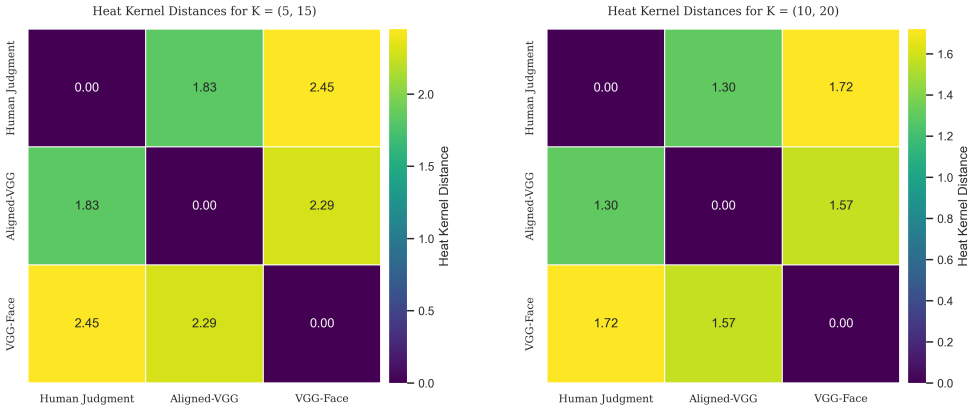


Figure 8: GMMs of curvature distributions of all representations in 3D viewing condition for (k_{min}, k_{max})= a) (5,15), b) (10,15), and c) (10,20).

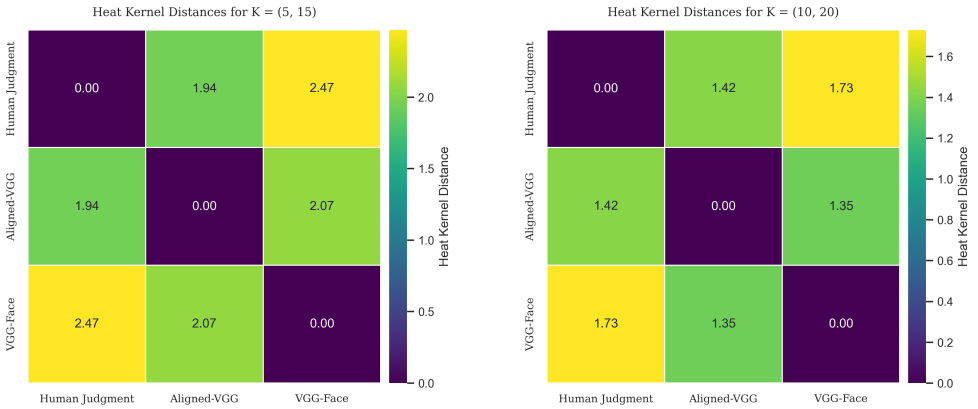
B.6. Heat diffusion distance

The heatmaps of the Heat diffusion distance for $(k_{\min}, k_{\max}) = (5, 15), (10, 20)$ for 2D condition is provided in Figure 9 and for 3D is provided in Figure 10.



(a) $k_{\min}, k_{\max} = (5, 15)$ (b) $k_{\min}, k_{\max} = (10, 20)$

Figure 9: Heat diffusion distance results for 2D conditions.

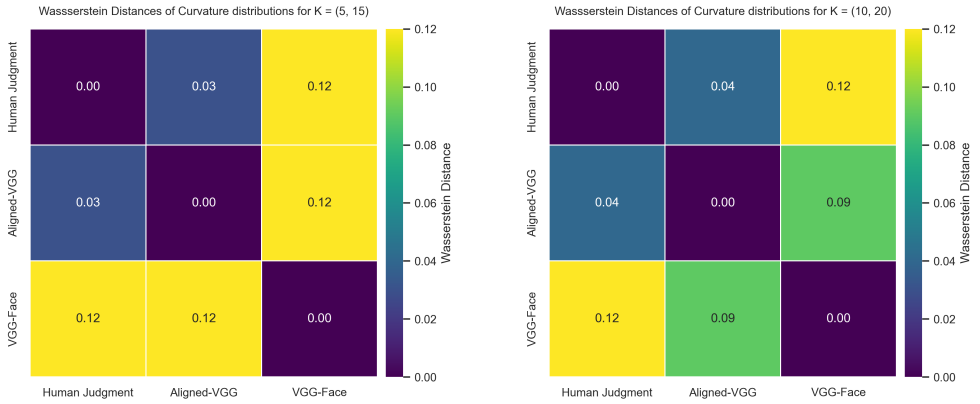


(a) $k_{\min}, k_{\max} = (5, 15)$ (b) $k_{\min}, k_{\max} = (10, 20)$

Figure 10: Heat diffusion distance results for 3D conditions.

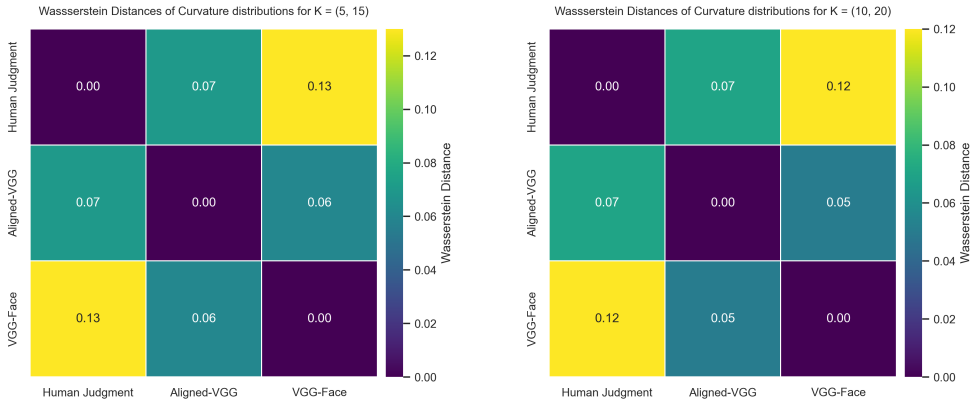
B.7. Wasserstein Distance

The heatmap results of the Wasserstein distance between the curvature distributions of representations for $(k_{\min}, k_{\max}) = (5, 15), (10, 20)$ are provided in Figure 11 for 2D viewing condition and in Figure 12 for 3D viewing condition.



(a) $k_{\min}, k_{\max} = (5, 15)$ (b) $k_{\min}, k_{\max} = (10, 20)$

Figure 11: Wasserstein distance results for 2D condition.



(a) $k_{\min}, k_{\max} = (5, 15)$ (b) $k_{\min}, k_{\max} = (10, 20)$

Figure 12: Wasserstein distance results for 3D condition.

B.8. Representational Profile Analysis

To gain a more fine-grained understanding of how image representations vary and relate across representational spaces, we conducted an additional analysis. For each image, in a given representational space, we extracted its distance profile, any point in the figures, corresponding to the relevant row of its RDM, and correlated it with the same row from

the RDMs of other representations, using multiple distance metrics. In **3D condition** Figure 13, Alignment becomes heterogeneous, splitting into two clusters by gender: male images maintain higher correlations, while female images show weaker alignment, especially between Human Judgment and VGG-based representations under the Euclidean metric. Geometry-aware metrics (shortest path, flow-metric) further reduce and scatter correlations, indicating that increased spatial complexity disrupts geometric preservation. In **2D condition**, Figure 14, Human Judgment and Aligned-VGG representations show consistently high correlations across metrics.

This pattern reinforces our earlier findings: in **2D**, alignment faithfully projects human perceptual geometry into representational space, yielding an arrangement of data points that closely mirrors human face-perception averages. VGG-Face, by comparison, adopts a distinct geometric layout. In **3D**, however, human judgment cannot be mapped onto the underlying geometry of either aligned or unaligned network representations. These results suggest that alignment effectively preserves task-relevant geometric structure only in two dimensions.

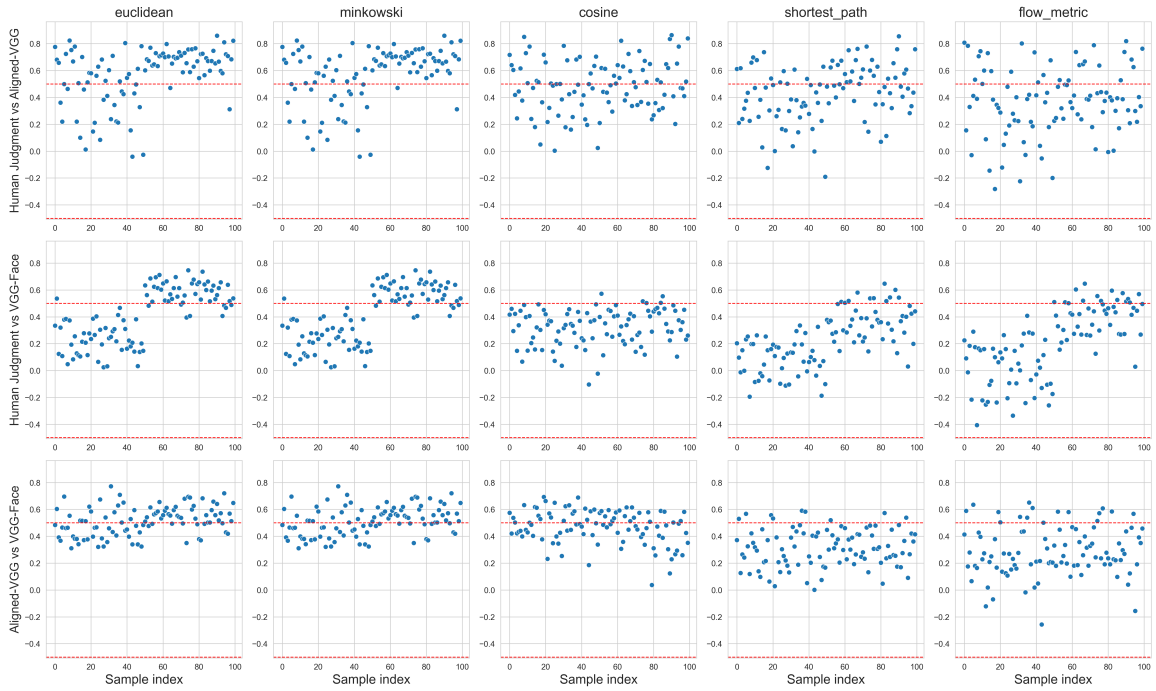


Figure 13: Point-level correlation of the data points between pairs of RDM matrices computed by different metrics (euclidean, shortest-path, flow-metric,minkowski, and cosine from the left, respectively) for $(k_{min}, k_{max}) = (5, 10)$ for **3D** condition. Red dashed lines show r=0.5 (up), and r=-0.5 (down).

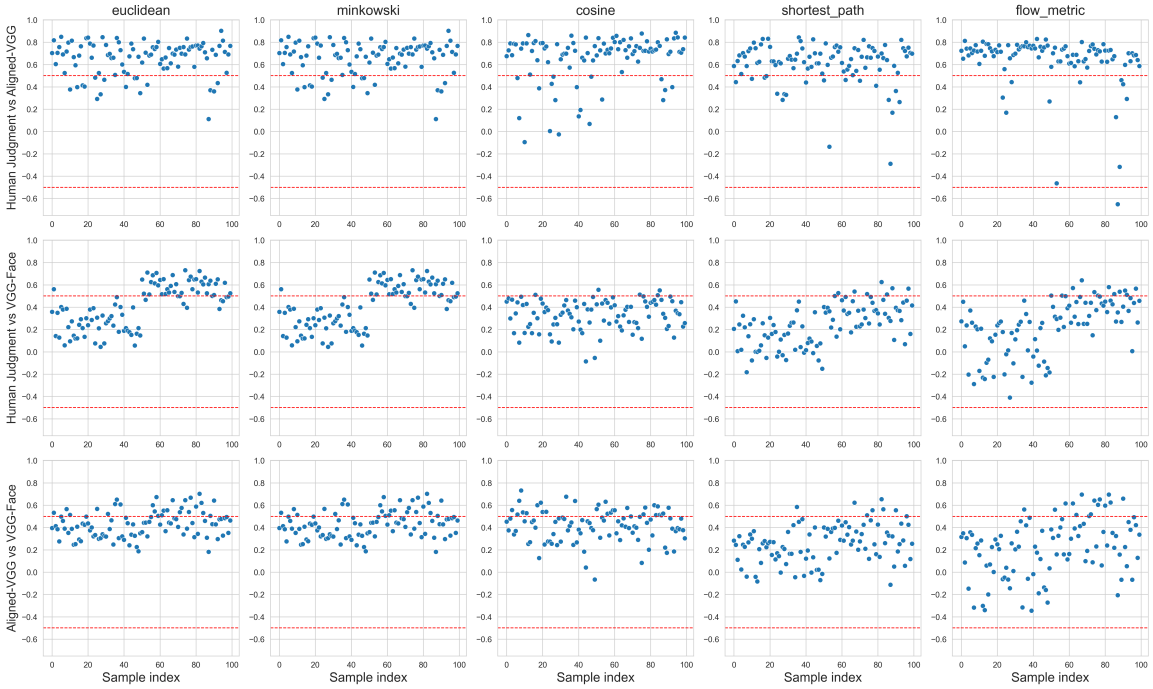


Figure 14: Point-level correlation of the data points between pairs of RDM matrices computed by different distance metrics (euclidean, shortest-path, flow-metric,minkowski, and cosine from the left, respectively) for $(k_{min}, k_{max}) = (5, 10)$ for **2D condition**. Red dashed lines show r=0.5 (up), and r=-0.5 (down).

B.9. Community metrics results

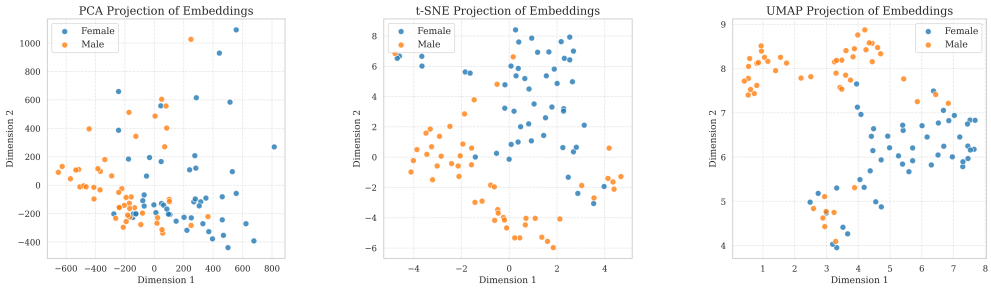
We analyzed the structural properties of the processed graphs, as summarized in Table 3. In particular, we calculated *conductance* (community separation by the between-and-within edge ratio), *internal edge density* (edges within the community vs. maximum possible), *modularity* (community strength by comparing edge densities with a random graph), and *average embeddedness* (number of shared neighbors for pairs of nodes within a community). These results clearly show that the Human Judgment graph and the graph representation of the Aligned-VGG have very similar properties in terms of community structure. In contrast, the Original VGG-Face, despite being trained on face images, does not show a similarly high degree of structure. Going from a 2D viewing condition to 3D, congruence to Fig 2, it is clear that Aligned-VGG has properties more similar to VGG-Face rather than Human Judgment, in contrast to the 2D condition.

Table 3: Comparison of graph structure based on KL-divergence (KLD) between edge curvature distributions and community metrics derived from Ricc flow: *Conductance*, *Internal edge density* (IED), *Modularity*, and *Average embeddedness* (AE).

Graph	Communities	Conductance	IED	Modularity	AE
Human Judgment 2D	7	0.23	0.60	34.73	0.78
Human Judgment 3D	8	0.31	0.68	31.54	0.69
Aligned-VGG 2D	6	0.20	0.52	35.98	0.80
Aligned-VGG 3D	6	0.34	0.57	21.57	0.68
VGG-Face	3	0.33	0.32	11.75	0.69

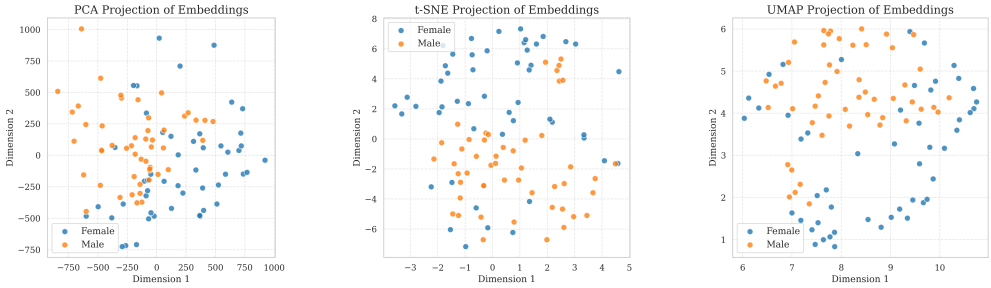
B.10. Dimensionality Reduction Visualisation of Embeddings

To visualize the embedding spaces of both native and behaviorally aligned networks under 2D and 3D viewing conditions, we applied three dimensionality reduction techniques: PCA, t-SNE, and UMAP. Figure 15 presents the Human-Aligned network in the 2D condition, Figure 16 shows the same network under the 3D condition, and Figure 17 depicts the baseline VGG-Face model. In each visualization, male and female image embeddings are distinguished by color.



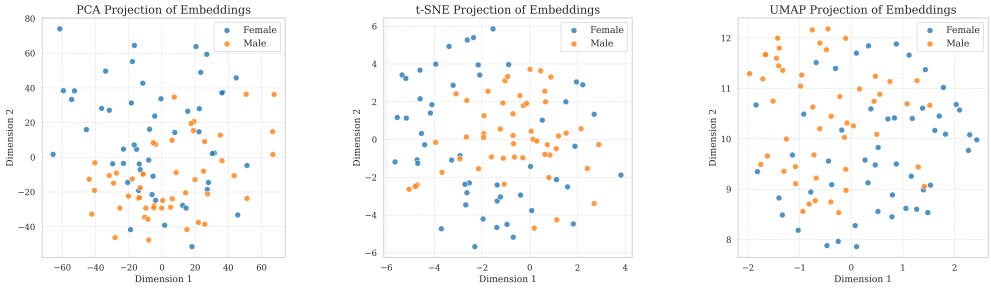
(*a*) PCA (*b*) t-SNE (*c*) UMAP

Figure 15: 2D Visualisation of the embedding vectors for Human-Aligned network in 2D condition. Colors show female and male images.



(*a*) PCA (*b*) t-SNE (*c*) UMAP

Figure 16: 2D Visualisation of the embedding vectors for Human-Aligned network in 3D condition. Colors show female and male images.



(*a*) PCA (*b*) t-SNE (*c*) UMAP

Figure 17: 2D Visualisation of the embedding vectors for the VGG-Face. Colors show female and male images.