Preface of UniReps: the First Workshop on Unifying Representations in Neural Models

Marco Fumero
Dept. of Computer Science
Sapienza University of Rome
Rome, IT
fumero@di.uniroma1.it

Clémentine C. J. Dominé
Gatsby Computational Neuroscience Unit
University College London
London, UK
clementine.domine.20@ucl.ac.uk

Emanuele Rodolà
Dept. of Computer Science
Sapienza University of Rome
Rome, IT
e.rodola@di.uniroma1.it

Francesco Locatello
Dept. of Computer Science
IST Austria
Klosterneuburg, Austria
francesco.locatello@ist.ac.at

Gintare Karolina Dziugaite
Google DeepMind
Toronto, Canada
gkdz@google.com

Mathilde Caron
Google DeepMind
Paris, France
mathilde.caron@google.com

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

Discover why, when and how distinct learning processes yield similar representations, and the degree to which these can be unified.

https://unireps.org

Workshop Summary Neural models tend to learn similar representations when subject to similar stimuli; this behavior has been observed both in biological [10, 19] and artificial settings [21, 17, 22]. The word *similar* here plays a fundamental role: under different conditions and assumptions on the observed data and the neural model (for instance, two distinct individuals exposed to the same stimulus [30] or different initializations of the same neural architecture [44]), inner representations of distinct models can be reconnected to one another, e.g. up to a linear transformation [34]. The similarities in the observational space can refer to settings where data are acquired in a multimodal environment, for instance textual and image representations of the same entity [28], or in a multiview setting [41] where observations in a single modality are acquired under different conditions.

The emergence of these similar representations is a ubiquitous phenomenon, which is igniting a growing interest in the fields of Neuroscience, Artificial Intelligence and Cognitive Science. By convening researchers with expertise in these fields, this workshop addressed the following key points:

- (*When*): To explore the specific patterns by which these similarities emerge in different neural models. Modelling the transformations, symmetries and invariances between similar representations is key to measure if these can be unified [17, 16].

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

- (*Why*): To investigate the underlying causes of these similarities in neural representations, with a focus on both artificial and biological models, as well as across them. Promising directions include analyzing the learning dynamics of neural models [1, 3, 36], studying model identifiability in the functional and parameter space [38, 12, 34, 14] and investigating the relations between different local minima reached by the optimization process [8, 7, 20].
- (*What for*): To explore and showcase applications in modular deep learning ranging from model merging [2], reuse [6, 15] and stitching [4, 26] to efficient strategies for fine-tuning and knowledge transfer between models [45] even in out-of-distribution settings [31], or to exploit cross-domain representation similarities (e.g. in *fMRI*-to-image models [40]).

The workshop provided an exciting, timely, and diverse environment for discussing theoretical findings, empirical evidence, and practical applications of the emergence of similar representations across models. It benefited from the cross-pollination of different fields—Machine Learning (ML), Neuroscience, Cognitive Science—to foster the exchange of ideas and encourage collaborations. The suggested *topics* include:

- Model merging, stitching and reuse [2]
- Identifiability in neural models [34]
- Learning dynamics [36]
- Representation similarity analysis [18]
- Similarity based learning [42, 46]
- Representational alignment [23]
- Symmetry and equivariance in NNs [11]
- Synergy of biological & artificial NNs[5]
- Multiview representation learning [41]
- Linear mode connectivity [8]
- Multimodal learning [28]

Workshop Format We proposed a dynamic workshop that fostered discussion among researchers. To this end, we designed a program integrating invited talks with a panel discussion, a mentorship program, and a poster session. In the panel discussion, we gathered renowned experts from the fields of AI, Neuroscience, and Cognitive Sciences for a dynamic roundtable discussion on key topics explored during the workshop. Our aim was to establish a cohesive understanding of the emergence of similar representations in neural models and pave the way for a new interdisciplinary community and research area. By fostering collaboration among diverse fields, we envisioned fruitful cross-pollination of ideas. Additionally, participants had the opportunity to address questions posed by attendees, which were further explored in our mentoring program. This took place during our coffee breaks and lunch, along with casual discussions, serving as an opportunity to conduct research discussions, engage in informal conversations, and introduce a new 1:1 mentoring initiative for junior and senior researchers. Our primary objective was to facilitate networking and foster collaboration opportunities for all workshop attendees, even in the remote format. Finally, a dedicated poster session provided the chance to showcase recent work, share findings, and engage in meaningful discussions among peers. Borrowing ideas from WiML and ICLR 2023, we assigned experienced participants to opted-in posters, ensuring feedback to our most junior participants, seeding conversations, and potentially research relationships.

Schedule	
08.15 AM	Opening Remarks
08.30 AM	Invited Talk: T. Griffiths
09.00 AM	Invited Talk: S. Sanborn
09.30 AM	Invited Talk: A. Saxe
10.00 AM	Coffee Break (Mentorship)
10.30 AM	Contributed talks
11.45 AM	Panel Discussion
12.30 AM	Lunch (Mentorship)
1.45 PM	Invited Talk: S. Kornblith
2.15 PM	Invited Talk: E. Triantafillou
2.45 PM	Invited Talk: A. Lampinen
3:15 PM	Closing Remarks
3.30 PM	Poster Session

Double submission track Submissions to the workshop were organized into two tracks, both requiring novel and unpublished results: an extended abstract track, which addressed early-stage results, insightful negative findings, opinion pieces, and a proceedings track, which focused on complete papers that were published in a dedicated workshop proceedings volume. Both tracks were featured in the workshop poster session, giving authors the opportunity to present their work. Additionally, a subset of the submissions was selected for a spotlight talk session during the workshop. This structure ensured a diverse and engaging presentation of ideas, fostering dialogue and exchange among participants.

Diversity and inclusivity Our workshop upheld diversity and inclusivity as fundamental principles to fostering a balanced and productive environment. To achieve this, we strived for diversity in various aspects, including seniority, gender balance, and nationality. Our organizers and invited speakers ranged from PhD students to junior and senior researchers, reflecting a broad spectrum of experience levels. We made a conscious effort to ensure gender balance among both our organizers and keynote speakers, and included participants from different regions, covering Europe, the United States, and Middle Eastern Asia. To promote an inclusive environment, we actively sought participation from the BlackInAI, Women In Machine Learning (WiML), QueerInAI, and LatinxInAI communities by sending Program Committee calls, spotlight talk invitations, and invitations to attend the workshop through their mailing lists and communication channels. In this regard, with the generous contribution in funding from the Gatsby Foundation for UniReps and Google Deepmind, we were able to establish a travel and registration assistance program for attendees. This program was designed to provide financial aid to researchers, students, or individuals who encountered financial obstacles when trying to attend NeurIPS and UniReps. Thanks to this financial support, we directly offset expenses such as the registration fee, which typically amounts to around $\$500$, making it more feasible for a wider range of participants to attend and contribute to our workshop.

Attendance We surpassed our expectations by drawing in a diverse crowd of 800 attendees in person, along with an additional 50 participants joining virtually. The audience was a rich tapestry of students, researchers, and industry practitioners from a variety of communities and cultures. The welcoming nature of our event was further enhanced by the thoughtful room setup and environment we created, which fostered a sense of inclusion and engagement among all attendees.

Speakers and Panelists

Simon Kornblith *Google DeepMind*
Senior Research Scientist studying similarities across different neural representations. Simon proposed CKA to measure similarity across different neural representations in [17], compared the representations between different networks in [25, 29] and finally investigated the alignment between neural network representations and cognitive representations in [24].

Sophia Sanborn *University of California, Santa Barbara*
Postdoctoral Scholar at UC Santa Barbara, Sophia leverages group theory, differential geometry and topology to understand representations in biological and artificial neural networks, with a focus on studying symmetry-preserving representations [35].

Thomas L. Griffiths *Princeton University*
Director of the Computational Cognitive Science Lab at Princeton University. Among numerous contributions in cognitive science, Thomas is interested in exploring how ideas from artificial intelligence, machine learning, and statistics connect to human cognition, with a focus on representational alignment [39, 13, 27].

Andrew Lampinen *Google DeepMind*
Andrew Lampinen is a Senior Research Scientist at DeepMind, having previously§ completed his PhD in Cognitive Psychology at Stanford University. He has a keen interest in cognitive flexibility and generalization, particularly in how these abilities are enabled by factors such as language, memory, and embodiment. Additionally, he is intrigued by the instances and mechanisms of intelligence failure. His research considers these issues from both human cognition and artificial intelligence perspectives.

Andrew Saxe *University College London*
Associate Professor studying principles of learning in the brain and mind and its connection to theory of deep learning. His work in analyzing the dynamics of deep linear models [36, 37] and in representational similarity analysis [9] can shed light on the reasons why similar representations emerge from neural models (both artificial and biological) when exposed to similar stimuli.

Eleni Triantafillou *Google DeepMind*
Research Scientist studying methods to allow efficient and effective adaptation of deep neural networks to cope with distribution shifts, introduction of new concepts, or removal of outdated or harmful knowledge. Eleni's research falls in the areas of few-shot learning [33, 43], meta-learning [32], domain adaptation and machine unlearning.

Organizers

Emanuele Rodolà *Sapienza University of Rome*
Emanuele is Full Professor of Computer Science at Sapienza University of Rome, where he leads the GLADIA group of learning and applied AI, funded by an ERC Grant and a Google Research Award. Previously, he was Assistant and then Associate Professor at Sapienza (2017-2020), a postdoc at USI Lugano (2016-2017), an Alexander von Humboldt Fellow at TU Munich (2013-2016), and a JSPS Research Fellow at The University of Tokyo (2013). He is a fellow of ELLIS and the Young Academy of Europe, has received a number of research prizes, has been serving in the program and organizing committees of the top rated conferences in computer vision, machine learning and graphics, founded and chaired several successful workshops. His research interests lie at the intersection of representation learning, graph / geometric deep learning, language and learning for audio, and has published more than 120 papers in these areas. Previously, he has organized and lectured at 15 tutorials, and has co-organized and chaired 10 workshops co-located with the major conferences in machine learning, geometry processing and computer vision including the successful Geometry Meets Deep Learning workshop (ECCV 2016, ICCV 2017, ECCV 2018, ICCV 2019).

Gintare Karolina Dziugaite *Google DeepMind*
Gintare Karolina Dziugaite is a Senior Research Scientist at Google DeepMind, an Adjunct Professor in the McGill University School of Computer Science, and an Associate Industry Member of Mila, the Quebec AI Institute. Dr. Dziugaite's research combines theoretical and empirical approaches to understanding deep learning, with a focus on studying deep learning training dynamics, symmetries and linear mode connectivity. She was one of the main organizers of a NeurIPS 2019 workshop on "ML with Guarantees", one of the largest workshops in 2019. She also co-organized the 2022 Eastern European Machine Learning summer school, 2022 and 2023 Mila-Google Brain scientific workshop, and the NeurIPS 2020 Generalization Measure competition. While at ServiceNow, Gintare led the Trustworthy AI team.

Francesco Locatello *Institute of Science and Technology Austria (ISTA)*
Francesco Locatello is an assistant professor at the Institute of Science and Technology Austria (ISTA) leading the Causal Learning and Artificial Intelligence lab. Previously, he was a Senior Applied Scientist at Amazon Web Services (AWS) where he leads the Causal Representation Learning research team. He is interested in the intersection between causal methods and deep learning. He received his Ph.D. in Computer Science from ETH Zurich (2020), where he was awarded the ETH medal for outstanding doctoral dissertation. During his Ph.D. he was supported by a Google Fellowship and was a Fellow at the Max Planck ETH Center for Learning Systems and ELLIS. His research has received awards at several premier conferences and workshops, most notably the best paper award at the International Conference on Machine Learning in 2019 and the award from the Hector foundation in 2023. Francesco Locatello co-organized the first and second international conference on Causal Learning and Reasoning (CLeaR) as sponsorship and general chair, ELLIS, ICLR and UAI workshops and a NeurIPS competition.

Clementine Domine *University College London*
Clementine Domine is a Ph.D. candidate at the Gatsby Computational Neuroscience Unit, supervised by Andrew Saxe. Her research seeks to develop mathematical toolkits suitable for describing complex and flexible learning mechanisms in both artificial and biological agents. Her work has been published in major ML conferences, including NeurIPS. Clémentine's commitment extends beyond academics, as demonstrated by her active involvement in Equality, Diversity, and Inclusion initiatives. She's been an integral part of the Athena Swan Committee at the Sainsbury Wellcome Centre, she is a member of WiML, and at Gatsby Unit she has been responsible for mentoring and teaching multiple students, notably through the In2research program.

Marco Fumero *Institute of Science and Technology Austria (ISTA)*
Marco Fumero is a PostDoc at the institute of science and technology Austria (ISTA) under the supervision of prof. Francesco Locatello. He was an ELLIS PhD candidate (industry track) in computer science at Sapienza University, in the GLADIA lab. His primary research focuses on representation learning and its application in real-world tasks. He has wide expertise in topics such as disentangled representation learning and out-of-distribution generalization. He has published in major conferences and journals (ICML, ICLR, CVPR, TOG, CGF), including works directly aligned with the workshop themes. He has gathered industry experience, helding positions at Amazon AI Research and Autodesk AI.

Mathilde Caron *Google*
Mathilde is currently a Research Scientist at Google, and previously at Facebook AI Research (FAIR), working on large-scale self-supervised representation learning for vision. Previously a Ph.D. student at Inria Grenoble, she graduated from both Ecole Polytechnique and KTH Royal Institute of Technology. She won the annual ELLIS PhD award in 2022, her works appeared in NeurIPS, ECCV, ICCV, and TPAMI. Mathilde is also proposing an unrelated workshop for NeurIPS 2023 on "Self-Supervised Learning: Theory and Practice".

Program Committee & Chairs

We are proud to introduce our esteemed reviewing committee, comprised of 156 dedicated reviewers who have collectively contributed 474 reviews. Their expertise and commitment have been instrumental in ensuring the high quality and rigor of the discussions and findings presented at our workshop. Likewise, we thank our chairs Luca Moschella, Donato Crisostomi and Antonio Norelli for helping in the organization of the event.

1. Aaditya Singh (UCL)
2. Adeniyi Adetola Omolara (Aston University)
3. Aditi Jha (Princeton University)
4. Adwaita Janardhan Jadhav (Apple)
5. Andrea Caciolai (Amazon)
6. Andrea Santilli (Sapienza University of Rome)
7. Danilo Numeroso (University of Pisa)
8. Devon Jarvis (UCL)
9. Ajay Subramanian (New York University)
10. Akash Nagaraj (Brown University)
11. Alessandro Raganato (University of Milan - Bicocca)
12. Alessio Devoto (University of Roma "La Sapienza")
13. Alex H Williams (New York University)
14. Alexander F Spies (Imperial College London)
15. Amirhesam Abedsoltan (University of California, San Diego)
16. Anastasia Borovykh (Imperial College London)
17. Andrew Kyle Lampinen (Google DeepMind)
18. Andrew William Engel (Pacific Northwest National Laboratory)
19. Anna Bair (Carnegie Mellon University)
20. Antonio Longa (University of Trento)
21. Antonio Pio Ricciardi (University of Roma "La Sapienza")
22. Arvind Saraf (Massachusetts Institute of Technology)
23. Berivan Isik (Google)
24. Binxu Wang (Harvard University)
25. Bo Zhao (University of California, San Diego)
26. Brian S Robinson (Johns Hopkins University Applied Physics Laboratory)
27. Brice Ménard (Johns Hopkins University)
28. Celia Cintas (International Business Machines)
29. Changfeng Wang (Boston Data Science)
30. Chanwoo Chun (Cornell University)
31. Ching Fang (Columbia University)
32. Dan Friedman (Princeton University)
33. Daniel Gedon (Uppsala University)
34. Daniel Marczak (IDEAS NCBR)
35. Daniele Baieri (University of Roma "La Sapienza")
36. David A. Klindt (Stanford (SLAC))
37. David Torpey (University of the Witwatersrand)
38. Davide Eynard (Mozilla.ai)
39. Davit Soselia (University of Maryland, College Park)
40. Dean A Pospisil (Princeton University)
41. Debora Caldarola (Computer Science Department, Stanford University)
42. Devon Jarvis (University College London, University of London)
43. Eeshan Hasan (Indiana University)
44. Elom Amematsro (Columbia University)
45. Emanuele Marconato (University of Pisa)
46. Emilian Postolache (University of Roma "La Sapienza")
47. Eric J Bigelow (Harvard University)
48. Filip Szatkowski (IDEAS NCBR)
49. Gabor Lengyel (University of Rochester)
50. Gabriele Dominici (Universita della Svizzera Italiana)
51. Garrison W. Cottrell (University of California, San Diego)

52. George Stoica (Georgia Institute of Technology)
53. Giovanni Ficarra (University of Roma "La Sapienza")
54. Gozde Merve Demirci (City University of New York, City University of New York)
55. Gregor Bachmann (Swiss Federal Institute of Technology)
56. Greta Tuckute (Massachusetts Institute of Technology)
57. Gül Sena Altıntaş (ETHZ - ETH Zurich)
58. Hadi Pouransari (Apple)
59. Haofei Yu (Carnegie Mellon University)
60. Haoli Yin (Vanderbilt University)
61. HyungGoo Kim (SungKyunKwan University)
62. Irene Cannistraci (University of Roma "La Sapienza")
63. Irene Tallini (University of Roma "La Sapienza")
64. Itay Evron (Technion, Technion)
65. Jaeho Lee (Pohang University of Science and Technology)
66. Jannis Born (International Business Machines)
67. Jaweria Amjad (University College London, University of London)
68. Jia Shi (Carnegie Mellon University)
69. Jiawen Xu (Technische Universität Berlin)
70. Jin Hwa Lee (University College London, University of London)
71. Jingtong Su (New York University)
72. Jingyang Zhou (New York University)
73. Jinyung Hong (Arizona State University)
74. Jiwoon Lee (Pohang University of Science and Technology)
75. Juan Miguel Navarro Carranza (Stanford University)
76. Julia Huiming Wang (Cold Spring Harbor Laboratory)
77. Khai Loong Aw (Singapore Management University)
78. Khalid Oublal (École Polytechnique)
79. Kira Michaela Düsterwald (University College London, University of London)
80. Konstantin Hemker (University of Cambridge)
81. Lars Kai Hansen (Technical University of Denmark)
82. Longbiao Cheng (Insititute of Neuroinformatics, University of Zurich and ETH Zurich)
83. Luca Cosmo (University of Venice)
84. Lucie Charlotte Magister (University of Cambridge)
85. Luigi Gresele (Max-Planck-Institute for Intelligent Systems, Max-Planck Institute)
86. Luke Hollingsworth (University College London, University of London)
87. Luke McDermott (University of California, San Diego)
88. Marco Pegoraro (University of Roma "La Sapienza")
89. Martha Gahl (University of California, San Diego)
90. Marvin Schmitt (Universität Stuttgart)
91. Mateusz Pyla (Jagiellonian University Cracow)
92. Mathias Sablé-Meyer (Ecole Normale Supérieure de Cachan)
93. Matteo Boschini (University of Modena and Reggio Emilia)
94. Matteo Ferrante (Università di Roma Tor Vergata)
95. Matthew James Sargent (University College London)
96. Max Klabunde (Universität Passau)
97. Meenakshi Khosla (Massachusetts Institute of Technology)
98. Michael Moeller (University of Siegen)
99. Mohamed Shawky Sabae (Faculty of Engineering Cairo University, Cairo University)
100. Mohammadreza Salehi (Apple)
101. Mycal Tucker (Massachusetts Institute of Technology)
102. Nanda H Krishna (Université de Montréal)
103. Nasik Muhammad Nafi (Kansas State University)
104. Nassim Oufattole (Massachusetts Institute of Technology)
105. Nico Daheim (Technische Universität Darmstadt)
106. Nikolaos Tsilivis (New York University)
107. Nima Dehmamy (International Business Machines)
108. Nishil Patel (University College London, University of London)
109. Nishkrit Desai (University of Toronto)
110. Osamu Hirose (Kanazawa University)
111. Partha Pratim Saha (University of Massachusetts at Amherst)
112. Patrik Reizinger (Eberhard-Karls-Universität Tübingen)
113. Pietro Barbiero (University of Cambridge)
114. Raviteja Vemulapalli (Apple)
115. Riccardo Marin (Eberhard-Karls-Universität Tübingen)

116. Rylan Schaeffer (Computer Science Department, Stanford University)
117. Samuel Lippl (Columbia University)
118. Samy Wu Fung (Colorado School of Mines)
119. Sarah E Harvey (Flatiron Institute)
120. Sebastian Cygert (IDEAS NCBR)
121. Shubhi Asthana (International Business Machines)
122. Simone Calderara (University of Modena and Reggio Emilia)
123. Simone Melzi (University of Milan - Bicocca)
124. Simone Scardapane (Sapienza University of Rome)
125. Srinivasan Sivanandan (Insitro)
126. Stefan Horoi (Université de Montréal)
127. Stefan T. Radev (Rensselaer Polytechnic Institute)
128. Steve Azzolin (University of Trento)
129. Sungjin Ahn (KAIST)
130. Tahereh Toosi (Columbia University)
131. Tala Fakhoury (University of Pennsylvania)
132. Tamlin Love (Universidad Politécnica de Cataluna)
133. Tanya Akumu (Carnegie Mellon University)
134. Tassilo Wald (Deutsches Krebsforschungszentrum)
135. Teresa Scheidt (Technical University of Denmark)
136. Thomas Edward Yerxa (New York University)
137. Thomas Möllenhoff (RIKEN Center for Advanced Intelligence Project (AIP))
138. Till Aczel (ETHZ - ETH Zurich)
139. Tilman Räuker (Universität Hannover)
140. Valentino Maiorca (University of Roma "La Sapienza")
141. Valeria Ruscio (University of Roma "La Sapienza")
142. Will Dorrell (University College London, University of London)
143. Xiuyuan Hu (Tsinghua University)
144. Yang Zhao (Tsinghua University)
145. Yatin Dandi (EPFL - EPF Lausanne)
146. Yedi Zhang (University College London, University of London)
147. Yi-Fu Wu (Google)
148. Yufan Zhuang (University of California, San Diego)
149. Yujia Bao (Insitro)
150. Ziming Liu (Massachusetts Institute of Technology)
151. Ziqian Zhong (Massachusetts Institute of Technology)
152. Zuowen Wang (Institute of Neuroinformatics, University of Zurich and ETH Zurich)

Community

Join us to stay up-to-date with the latest workshop news, connect with a vibrant community, display your latest projects, and remain informed about exciting opportunities, events, and research. Our aim is to foster an engaging and inclusive environment, allowing each participant to contribute, learn, and maintain lasting connections beyond the workshop. Check out the UniReps Website! In addition, you can follow the last updates on the UniReps community on our Twitter profile!

Sponsors

We extend our deepest gratitude to our sponsors, Google DeepMind and The Gatsby Foundation, for their generous support and commitment to advancing research and innovation. Their contributions have been invaluable in making our event a success, enabling us to create a platform for sharing knowledge, fostering collaborations, and promoting the latest advancements in the field. We are truly thankful for their support and look forward to continuing our partnership in the future.

Future directions

We consider it both critical and opportune to establish a research forum and nurturing community that promotes knowledge exchange at the confluence ofmachine learning, and neuroscience on the topic of unified representations. As we progress, we are committed to facilitating opportunities for dialogue and discourse on these subjects at NeurIPS and various other gatherings. In line with our overarching goal of fostering a sense of community, we've also formed an active network of students and researchers. This community is envisioned as a central hub for coordinating related activities, including seminars and hackathons, further enriching the UniReps workshop experience

References

[1] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[2] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries, 2023.

[3] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, and Masato Okada. Dynamics of learning in mlp: Natural gradient and singularity revisited. *Neural computation*, 30(1):1–33, 2017.

[4] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations, 2021.

[5] David G. T. Barrett, Ari S. Morcos, and Jakob H. Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55:55–64, 2018.

[6] Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks, 2021.

[7] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks, 2022.

[8] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis, 2020.

[9] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[10] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[11] Irina Higgins, Sébastien Racanière, and Danilo Rezende. Symmetry-based representations for artificial and biological general intelligence, 2022.

[12] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[13] Aditi Jha, Joshua Peterson, and Thomas L. Griffiths. Extracting low-dimensional psychological representations from convolutional neural networks, 2020.

[14] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[15] Louis Kirsch, Julius Kunze, and David Barber. Modular networks: Learning to decompose neural computation, 2018.

[16] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2023.

[17] 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.

[18] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

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

[20] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning, 2020.

[21] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication, 2023.

[23] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations, 2023.

[24] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. In *SVRHM 2022 Workshop@ NeurIPS*, 2022.

[25] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2020.

[26] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training, 2023.

[27] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Adapting deep network features to capture psychological representations, 2016.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[29] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, 2021.

[30] Rajeev D. S. Raizada and Andrew C. Connolly. What makes different people's representations alike: Neural similarity space solves the problem of across-subject fmri decoding. *Journal of Cognitive Neuroscience*, 24:868–877, 2012.

[31] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization, 2023.

[32] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification, 2018.

[33] Mengye Ren, Eleni Triantafillou, Kuan-Chieh Wang, James Lucas, Jake Snell, Xaq Pitkow, Andreas S Tolias, and Richard Zemel. Flexible few-shot learning with contextual similarity. *arXiv preprint arXiv:2012.05895*, page 1, 2020.

[34] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

[35] Sophia Sanborn, Christian Shewmake, Bruno Olshausen, and Christopher Hillar. Bispectral neural networks. *arXiv preprint arXiv:2209.03416*, 2022.

[36] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[37] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, may 2019.

[38] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

[39] Ilia Sucholutsky and Thomas L. Griffiths. Alignment with human representations supports robust few-shot learning, 2023.

[40] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[43] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *International Conference on Machine Learning*, pages 10424–10433. PMLR, 2021.

[44] Liwei Wang, Lunjia Hu, Jiayuan Gu, Yue Wu, Zhiqiang Hu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation, 2018.

[45] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022.

[46] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.