# 参考文献


[1] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[J]. arXiv preprint arXiv:2108.07258, 2021.

[2] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. [S.l.]: PMLR, 2021: 8748-8763.

[3] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8821-8831.

[4] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. [S.l.]: PMLR, 2020: 1597-1607.

[5] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 9729-9738.

[6] BAO H, DONG L, PIAO S, et al. Beit: Bert pre-training of image transformers[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.

[7] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 16000-16009.

[8] RAZAVI A, VAN DEN OORD A, VINYALS O. Generating diverse high-fidelity images with vq-vae-2[J]. Advances in neural information processing systems, 2019, 32.

[9] DHARIWAL P, NICHOL A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.

[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[12] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.

[13] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[14] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[15] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[16] OPENAI R. Gpt-4v(ision) system card[J]. arXiv, 2023.

[17] YANG Z, LI L, LIN K, et al. The dawn of lmms: Preliminary explorations with gpt-4v (ision)[J]. arXiv preprint arXiv:2309.17421, 2023.

[18] ZENG A, LIU X, DU Z, et al. Glm-130b: An open bilingual pre-trained model[C]//International Conference on Learning Representations. [S.l.: s.n.], 2022.

## 参 考 文 献


[19] THOPPILAN R, DE FREITAS D, HALL J, et al. Lamda: Language models for dialog applications[J]. arXiv preprint arXiv:2201.08239, 2022.

[20] RAE J W, BORGEAUD S, CAI T, et al. Scaling language models: Methods, analysis & insights from training gopher[J]. arXiv preprint arXiv:2112.11446, 2021.

[21] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.

[22] WANG B, KOMATSUZAKI A. Gpt-j-6b: A 6 billion parameter autoregressive language model[Z]. [S.l.: s.n.], 2021.

[23] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J]. arXiv preprint arXiv:2206.07682, 2022.

[24] DU Z, QIAN Y, LIU X, et al. Glm: General language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 320-335.

[25] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[26] DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation[J]. Advances in neural information processing systems, 2019, 32.

[27] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[J]. arXiv preprint arXiv:2203.15556, 2022.

[28] YANG A, XIAO B, WANG B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.

[29] OPENAI R. Gpt-4 technical report[J]. arXiv, 2023: 2303-08774.

[30] ANIL R, DAI A M, FIRAT O, et al. Palm 2 technical report[J]. arXiv preprint arXiv:2305.10403, 2023.

[31] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[32] ZHANG S, ROLLER S, GOYAL N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.

[33] SCAO T L, FAN A, AKIKI C, et al. Bloom: A 176b-parameter open-access multilingual language model[J]. arXiv preprint arXiv:2211.05100, 2022.

[34] TEAM M N, et al. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023[J]. URL www. mosaicml. com/blog/mpt-7b. Accessed, 2023: 05-05.

[35] PENEDO G, MALARTIC Q, HESSLOW D, et al. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only[J]. arXiv preprint arXiv:2306.01116, 2023.

[36] TAORI R, GULRAJANI I, ZHANG T, et al. Alpaca: A strong, replicable instruction-following model[J]. Stanford Center for Research on Foundation Models. 2023, 3(6): 7.

[37] CHIANG W L, LI Z, LIN Z, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality[J]. 2023.

[38] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: Aligning language model with self generated instructions[J]. arXiv preprint arXiv:2212.10560, 2022.

[39] ZHU D, CHEN J, SHEN X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.

[40] ANAND Y, NUSSBAUM Z, DUDERSTADT B, et al. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo[J]. GitHub, 2023.

[41] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[J]. arXiv preprint arXiv:2009.03300, 2020.

[42] LI H, ZHANG Y, KOTO F, et al. Cmmlu: Measuring massive multitask language understanding in chinese[J]. arXiv preprint arXiv:2306.09212, 2023.

[43] HUANG Y, BAI Y, ZHU Z, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models[J]. arXiv preprint arXiv:2305.08322, 2023.

[44] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[J]. arXiv preprint arXiv:2110.14168, 2021.

[45] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[J]. arXiv preprint arXiv:2107.03374, 2021.

[46] JIN D, PAN E, OUFATTOLE N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams[J]. Applied Sciences, 2021, 11(14): 6421.

[47] ZHONG H, XIAO C, TU C, et al. Jec-qa: a legal-domain question answering dataset[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. [S.l.: s.n.], 2020: 9701-9708.

[48] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[49] KUDO T, RICHARDSON J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. [S.l.: s.n.], 2018: 66-71.

[50] SHIBATA Y, KIDA T, FUKAMACHI S, et al. Byte pair encoding: A text compression scheme that accelerates pattern matching[J]. 1999.

[51] TAYLOR R, KARDAS M, CUCURULL G, et al. Galactica: A large language model for science[J]. arXiv preprint arXiv:2211.09085, 2022.

[52] SU J, LU Y, PAN S, et al. Roformer: Enhanced transformer with rotary position embedding[J]. arXiv preprint arXiv:2104.09864, 2021.

[53] PRESS O, SMITH N A, LEWIS M. Train short, test long: Attention with linear biases enables input length extrapolation[J]. arXiv preprint arXiv:2108.12409, 2021.

[54] DAO T, FU D, ERMON S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness[J]. Advances in Neural Information Processing Systems, 2022, 35: 16344-16359.

[55] DAO T. Flashattention-2: Faster attention with better parallelism and work partitioning[J]. arXiv preprint arXiv:2307.08691, 2023.

[56] SHAZEER N. Glu variants improve transformer[J]. arXiv preprint arXiv:2002.05202, 2020.

[57] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[C]//International conference on machine learning. [S.l.]: PMLR, 2017: 933-941.

[58] RABE M N, STAATS C. Self-attention does not need $o(n2)$ memory[J]. arXiv preprint arXiv:2112.05682, 2021.

[59] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[60] XIONG R, YANG Y, HE D, et al. On layer normalization in the transformer architecture[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2020: 10524-10533.

[61] ZHANG B, SENNRICH R. Root mean square layer normalization[J]. Advances in Neural Information Processing Systems, 2019, 32.

[62] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

[63] JIANG Z, GU J, PAN D Z. Normsoftmax: Normalizing the input of softmax to accelerate and stabilize training[C]//2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS). [S.l.]: IEEE, 2023: 1-6.

[64] HENIGHAN T, KAPLAN J, KATZ M, et al. Scaling laws for autoregressive generative modeling[J]. arXiv preprint arXiv:2010.14701, 2020.

[65] NIE X, MIAO X, YANG Z, et al. Tsplit: Fine-grained gpu memory management for efficient dnn training via tensor splitting[C]//2022 IEEE 38th International Conference on Data Engineering (ICDE). [S.l.]: IEEE, 2022: 2615-2628.

[66] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[67] HE P, LIU X, GAO J, et al. Deberta: Decoding-enhanced bert with disentangled attention[C]//International Conference on Learning Representations. [S.l.: s.n.], 2020.

[68] ZHUANG F, QI Z, DUAN K, et al. A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2020, 109(1): 43-76.

[69] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. Advances in neural information processing systems, 2000, 13.

[70] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// ICLR. [S.l.: s.n.], 2013.

[71] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. [S.l.: s.n.], 2019: 4171-4186.

[72] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.

[73] WANG P, YANG A, MEN R, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2022: 23318-23340.

[74] LU J, CLARK C, ZELLERS R, et al. Unified-io: A unified model for vision, language, and multi-modal tasks[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[75] SINGH A, HU R, GOSWAMI V, et al. Flava: A foundational language and vision alignment model[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 15638-15650.

[76] WANG W, BAO H, DONG L, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks[J]. arXiv preprint arXiv:2208.10442, 2022.

[77] NEELAKANTAN A, XU T, PURI R, et al. Text and code embeddings by contrastive pre-training[J]. arXiv preprint arXiv:2201.10005, 2022.

[78] CHRISTIANO P F, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences[J]. Advances in neural information processing systems, 2017, 30.

[79] STIENNON N, OUYANG L, WU J, et al. Learning to summarize with human feedback[J]. Advances in Neural Information Processing Systems, 2020, 33: 3008-3021.

[80] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021: 3045-3059.

[81] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. [S.l.: s.n.], 2021: 255-269.

[82] ZHANG Z, ZHANG A, LI M, et al. Automatic chain of thought prompting in large language models[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[83] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[84] DEHGHANI M, DJOLONGA J, MUSTAFA B, et al. Scaling vision transformers to 22 billion parameters[C]// International Conference on Machine Learning. [S.l.]: PMLR, 2023: 7480-7512.

[85] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational visual media, 2022, 8(3): 331-368.

[86] JOSHI M, CHEN D, LIU Y, et al. Spanbert: Improving pre-training by representing and predicting spans[J]. Transactions of the association for computational linguistics, 2020, 8: 64-77.

[87] LEWIS M, LIU Y, GOYAL N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 7871-7880.

[88] CLARK K, LUONG M T, LE Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.

[89] LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.

[90] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep clustering for unsupervised learning of visual features[C]//Proceedings of the European conference on computer vision (ECCV). [S.l.: s.n.], 2018: 132-149.

[91] CARUANA R. Multitask learning[J]. Machine learning, 1997, 28: 41-75.

[92] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

[93] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.

[94] SONG K, TAN X, QIN T, et al. Mpnet: Masked and permuted pre-training for language understanding[J]. Advances in Neural Information Processing Systems, 2020, 33: 16857-16867.

[95] LI Q, PENG H, LI J, et al. A survey on text classification: From traditional to deep learning[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(2): 1-41.

[96] SONG K, TAN X, QIN T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[97] SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[98] SUN Y, WANG S, LI Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the AAAI conference on artificial intelligence: volume 34. [S.l.: s.n.], 2020: 8968-8975.

[99] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[100] DIAO S, BAI J, SONG Y, et al. Zen: Pre-training chinese text encoder enhanced by n-gram representations[J]. Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, 2020.

[101] TSAI H, RIESA J, JOHNSON M, et al. Small and practical bert models for sequence labeling[J]. arXiv preprint arXiv:1909.00100, 2019.

[102] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[103] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv preprint arXiv:2109.01652, 2021.

[104] HONOVICH O, SCIALOM T, LEVY O, et al. Unnatural instructions: Tuning language models with (almost) no human labor[J]. arXiv preprint arXiv:2212.09689, 2022.

[105] WANG Y, MISHRA S, ALIPOORMOLABASHI P, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2022: 5085-5109.

[106] MISHRA S, KHASHABI D, BARAL C, et al. Cross-task generalization via natural language crowdsourcing instructions[C]//60th Annual Meeting of the Association for Computational Linguistics, ACL 2022. [S.l.]: Association for Computational Linguistics (ACL), 2022: 3470-3487.

[107] WEIDINGER L, MELLOR J, RAUH M, et al. Ethical and social risks of harm from language models[J]. arXiv preprint arXiv:2112.04359, 2021.

[108] KIEGELAND S, KREUTZER J. Revisiting the weaknesses of reinforcement learning for neural machine translation[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2021: 1673-1681.

[109] JAQUES N, SHEN J H, GHANDEHARIOUN A, et al. Human-centric dialog training via offline reinforcement learning[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 3985-4003.

[110] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 7008-7024.

[111] PANG R Y, HE H. Text generation by learning from demonstrations[C]//International Conference on Learning Representations. [S.l.: s.n.], 2020.

[112] HAUSKNECHT M, AMMANABROLU P, CÔTÉ M A, et al. Interactive fiction games: A colossal adventure[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 34. [S.l.: s.n.], 2020: 7903-7910.

[113] RAMAMURTHY R, AMMANABROLU P, BRANTLEY K, et al. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization[J]. arXiv preprint arXiv:2210.01241, 2022.

[114] WU J, OUYANG L, ZIEGLER D M, et al. Recursively summarizing books with human feedback[J]. arXiv preprint arXiv:2109.10862, 2021.

[115] NAKANO R, HILTON J, BALAJI S, et al. Webgpt: Browser-assisted question-answering with human feedback[J]. arXiv preprint arXiv:2112.09332, 2021.

[116] GLAESE A, MCALEESE N, TRĘBACZ M, et al. Improving alignment of dialogue agents via targeted human judgements[J]. arXiv preprint arXiv:2209.14375, 2022.

[117] BAI Y, KADAVATH S, KUNDU S, et al. Constitutional ai: Harmlessness from ai feedback[J]. arXiv preprint arXiv:2212.08073, 2022.

[118] CHUNG H W, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models[J]. arXiv preprint arXiv:2210.11416, 2022.

[119] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[J]. Advances in neural information processing systems, 2022, 35: 22199-22213.

[120] DOSOVITSKIY A, SPRINGENBERG J T, RIEDMILLER M, et al. Discriminative unsupervised feature learning with convolutional neural networks[J]. Advances in neural information processing systems, 2014, 27.

[121] ALEXEY D, FISCHER P, TOBIAS J, et al. Discriminative unsupervised feature learning with exemplar convolutional neural networks[J]. IEEE TPAMI, 2016, 38(9): 1734-1747.

[122] DOERSCH C, GUPTA A, EFROS A A. Unsupervised visual representation learning by context prediction[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2015: 1422-1430.

[123] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 2536-2544.

[124] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. [S.l.]: Springer, 2016: 649-666.

[125] ZHANG R, ISOLA P, EFROS A A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 1058-1067.

[126] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles[C]//European conference on computer vision. [S.l.]: Springer, 2016: 69-84.

[127] KIM D, CHO D, YOO D, et al. Learning image representations by completing damaged jigsaw puzzles[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l.]: IEEE, 2018: 793-802.

[128] NOROOZI M, PIRSIAVASH H, FAVARO P. Representation learning by learning to count[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 5898-5906.

[129] BOJANOWSKI P, JOULIN A. Unsupervised learning by predicting noise[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2017: 517-526.

[130] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised representation learning by predicting image rotations[J]. arXiv preprint arXiv:1803.07728, 2018.

[131] OORD A V D, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

[132] HENAFF O. Data-efficient image recognition with contrastive predictive coding[C]//International conference on machine learning. [S.l.]: PMLR, 2020: 4182-4192.

[133] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[J]. arXiv preprint arXiv:1605.09782, 2016.

[134] DONAHUE J, SIMONYAN K. Large scale adversarial representation learning[J]. Advances in neural information processing systems, 2019, 32.

[135] DUMOULIN V, BELGHAZI I, POOLE B, et al. Adversarially learned inference[J]. arXiv preprint arXiv:1606.00704, 2016.

[136] CHEN M, RADFORD A, CHILD R, et al. Generative pretraining from pixels[C]//International conference on machine learning. [S.l.]: PMLR, 2020: 1691-1703.

[137] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]// Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 9650-9660.

[138] XIE Z, ZHANG Z, CAO Y, et al. Simmim: A simple framework for masked image modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 9653-9663.

[139] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[140] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 10012-10022.

[141] LI X, WANG W, YANG L, et al. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality[J]. arXiv preprint arXiv:2205.10063, 2022.

[142] CHEN J, HU M, LI B, et al. Efficient self-supervised vision pretraining with local masked reconstruction[J]. arXiv preprint arXiv:2206.00790, 2022.

[143] WU Z, XIONG Y, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 3733-3742.

[144] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[145] ZHUANG C, ZHAI A L, YAMINS D. Local aggregation for unsupervised learning of visual embeddings[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2019: 6002-6012.

[146] MISRA I, MAATEN L V D. Self-supervised learning of pretext-invariant representations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 6707-6717.

[147] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in neural information processing systems, 2020, 33: 21271-21284.

[148] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. [S.l.]: MIT press, 2018.

[149] CARON M, MISRA I, MAIRAL J, et al. Unsupervised learning of visual features by contrasting cluster assignments[J]. Advances in neural information processing systems, 2020, 33: 9912-9924.

[150] GOYAL P, CARON M, LEFAUDEUX B, et al. Self-supervised pretraining of visual features in the wild[J]. arXiv preprint arXiv:2103.01988, 2021.
[151] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 10428-10436.
[152] CHEN X, HE K. Exploring simple siamese representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 15750-15758.
[153] LI J, ZHOU P, XIONG C, et al. Prototypical contrastive learning of unsupervised representations[C]// International Conference on Learning Representations. [S.l.: s.n.], 2020.
[154] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.
[155] CUI L, WU Y, LIU J, et al. Template-based named entity recognition using bart[J]. arXiv preprint arXiv:2106.01760, 2021.
[156] PETRONI F, ROCKTÄSCHEL T, RIEDEL S, et al. Language models as knowledge bases?[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 2463-2473.
[157] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S.l.: s.n.], 2021: 4582-4597.
[158] SCHICK T, SCHÜTZE H. Few-shot text generation with natural language instructions[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021: 390-402.
[159] SCHICK T, SCHÜTZE H. It's not just size that matters: Small language models are also few-shot learners[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2021: 2339-2352.
[160] JIANG Z, XU F F, ARAKI J, et al. How can we know what language models know?[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423-438.
[161] YUAN W, NEUBIG G, LIU P. Bartscore: Evaluating generated text as text generation[J]. Advances in Neural Information Processing Systems, 2021, 34: 27263-27277.
[162] HAVIV A, BERANT J, GLOBERSON A. Bertese: Learning to speak to bert[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. [S.l.: s.n.], 2021: 3618-3623.
[163] WALLACE E, FENG S, KANDPAL N, et al. Universal adversarial triggers for attacking and analyzing nlp[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019.
[164] SHIN T, RAZEGHI Y, LOGAN IV R L, et al. Autoprompt: Eliciting knowledge from language models with automatically generated prompts[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 4222-4235.
[165] GAO T, FISCH A, CHEN D. Making pre-trained language models better few-shot learners[C]//Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021. [S.l.]: Association for Computational Linguistics (ACL), 2021: 3816-3830.
[166] GUO H, TAN B, LIU Z, et al. Text generation with efficient (soft) $q$-learning[J]. 2021.
[167] BEN-DAVID E, OVED N, REICHART R. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 414-433.
[168] DAVISON J, FELDMAN J, RUSH A M. Commonsense knowledge mining from pretrained models[C]// Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 1173-1178.

[169] TSIMPOUKELLI M, MENICK J L, CABI S, et al. Multimodal few-shot learning with frozen language models[J]. Advances in Neural Information Processing Systems, 2021, 34: 200-212.

[170] ZHONG Z, FRIEDMAN D, CHEN D. Factual probing is [mask]: Learning vs. learning to recall[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2021: 5017-5033.

[171] QIN G, EISNER J. Learning how to ask: Querying lms with mixtures of soft prompts[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). [S.l.: s.n.], 2021.

[172] HAMBARDZUMYAN K, KHACHATRIAN H, MAY J. Warp: Word-level adversarial reprogramming[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S.l.: s.n.], 2021: 4921-4933.

[173] LIU X, ZHENG Y, DU Z, et al. Gpt understands, too[J]. arXiv preprint arXiv:2103.10385, 2021.

[174] HAN X, ZHAO W, DING N, et al. Ptr: Prompt tuning with rules for text classification[J]. AI Open, 2022, 3: 182-192.

[175] YIN W, HAY J, ROTH D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 3914-3923.

[176] KHASHABI D, MIN S, KHOT T, et al. Unifiedqa: Crossing format boundaries with a single qa system[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. [S.l.: s.n.], 2020: 1896-1907.

[177] JIANG Z, ANASTASOPOULOS A, ARAKI J, et al. X-factr: Multilingual factual knowledge retrieval from pretrained language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 5943-5959.

[178] ZWEIG G, PLATT J C, MEEK C, et al. Computational approaches to sentence completion[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2012: 601-610.

[179] SCHICK T, SCHMID H, SCHÜTZE H. Automatically identifying words that can serve as labels for few-shot text classification[C]//Proceedings of the 28th International Conference on Computational Linguistics. [S.l.: s.n.], 2020: 5569-5578.

[180] CHEN X, ZHANG N, XIE X, et al. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction[C]//Proceedings of the ACM Web conference 2022. [S.l.: s.n.], 2022: 2778-2788.

[181] ALLEN-ZHU Z, LI Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[182] LU Y, BARTOLO M, MOORE A, et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 8086-8098.

[183] LIU J, SHEN D, ZHANG Y, et al. What makes good in-context examples for gpt-3?[C]//Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. [S.l.: s.n.], 2022: 100-114.

[184] MISHRA S, KHASHABI D, BARAL C, et al. Natural instructions: Benchmarking generalization to new tasks from natural language instructions[J]. arXiv preprint arXiv:2104.08773, 2021: 839-849.

[185] KUMAR S, TALUKDAR P. Reordering examples helps during priming-based few-shot learning[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. [S.l.: s.n.], 2021: 4507-4518.

[186] YOO K M, PARK D, KANG J, et al. Gpt3mix: Leveraging large-scale language models for text augmentation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. [S.l.: s.n.], 2021: 2225-2239.

[187] GUU K, HASHIMOTO T B, OREN Y, et al. Generating sentences by editing prototypes[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 437-450.

[188] PETRONI F, LEWIS P, PIKTUS A, et al. How context affects language models' factual predictions[C]// Automated Knowledge Base Construction. [S.l.: s.n.], 2020.

[189] DONG Q, LI L, DAI D, et al. A survey for in-context learning[J]. arXiv preprint arXiv:2301.00234, 2022.

[190] CHEN M, DU J, PASUNURU R, et al. Improving in-context few-shot learning via self-supervised training[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2022: 3558-3573.

[191] MIN S, LEWIS M, ZETTLEMOYER L, et al. Metaicl: Learning to learn in context[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2022: 2791-2809.

[192] WEI J, HOU L, LAMPINEN A, et al. Symbol tuning improves in-context learning in language models[J]. arXiv preprint arXiv:2305.08298, 2023.

[193] GU Y, DONG L, WEI F, et al. Pre-training to learn in context[J]. arXiv preprint arXiv:2305.09137, 2023.

[194] ZHAO Z, WALLACE E, FENG S, et al. Calibrate before use: Improving few-shot performance of language models[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2021: 12697-12706.

[195] SORENSEN T, ROBINSON J, RYTTING C, et al. An information-theoretic approach to prompt engineering without ground truth labels[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 819-862.

[196] TANWAR E, BORTHAKUR M, DUTTA S, et al. Multilingual llms are better cross-lingual in-context learners with alignment[J]. arXiv preprint arXiv:2305.05940, 2023.

[197] GONEN H, IYER S, BLEVINS T, et al. Demystifying prompts in language models via perplexity estimation[J]. arXiv preprint arXiv:2212.04037, 2022.

[198] LEVY I, BOGIN B, BERANT J. Diverse demonstrations improve in-context compositional generalization[J]. arXiv preprint arXiv:2212.06800, 2022.

[199] KIM H J, CHO H, KIM J, et al. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator[J]. arXiv preprint arXiv:2206.08082, 2022.

[200] WU Z, WANG Y, YE J, et al. Self-adaptive in-context learning[J]. arXiv preprint arXiv:2212.10375, 2022.

[201] NGUYEN T, WONG E. In-context example selection with influences[J]. arXiv preprint arXiv:2302.11042, 2023.

[202] LI X, QIU X. Finding supporting examples for in-context learning[J]. arXiv preprint arXiv:2302.13539, 2023.

[203] RUBIN O, HERZIG J, BERANT J. Learning to retrieve prompts for in-context learning[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2022: 2655-2671.

[204] LI X, LV K, YAN H, et al. Unified demonstration retriever for in-context learning[J]. arXiv preprint arXiv:2305.04320, 2023.

[205] YE J, WU Z, FENG J, et al. Compositional exemplars for in-context learning[J]. arXiv preprint arXiv:2302.05698, 2023.

[206] WANG X, ZHU W, WANG W Y. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning[J]. arXiv preprint arXiv:2301.11916, 2023.

[207] ZHANG Y, FENG S, TAN C. Active example selection for in-context learning[J]. arXiv preprint arXiv:2211.04486, 2022.

[208] BELLMAN R. A markovian decision process[J]. Journal of mathematics and mechanics, 1957: 679-684.

[209] HONOVICH O, SHAHAM U, BOWMAN S R, et al. Instruction induction: From few examples to natural language task descriptions[J]. arXiv preprint arXiv:2205.10782, 2022.

[210] ZHOU Y, MURESANU A I, HAN Z, et al. Large language models are human-level prompt engineers[J]. arXiv preprint arXiv:2211.01910, 2022.

[211] QIAO S, OU Y, ZHANG N, et al. Reasoning with language model prompting: A survey[J]. arXiv preprint arXiv:2212.09597, 2022.

[212] FU Y, PENG H, SABHARWAL A, et al. Complexity-based prompting for multi-step reasoning[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[213] PRESS O, ZHANG M, MIN S, et al. Measuring and narrowing the compositionality gap in language models[J]. 2022.

[214] WANG B, DENG X, SUN H. Iteratively prompt pre-trained language models for chain of thought[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2022: 2714-2730.

[215] ZHOU D, SCHÄRLI N, HOU L, et al. Least-to-most prompting enables complex reasoning in large language models[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[216] XU C, XU Y, WANG S, et al. Small models are valuable plug-ins for large language models[J]. arXiv preprint arXiv:2305.08848, 2023.

[217] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[C]//International Conference on Learning Representations. [S.l.: s.n.], 2019.

[218] MIN S, LEWIS M, HAJISHIRZI H, et al. Noisy channel language model prompting for few-shot text classification[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 5316-5330.

[219] HAO Y, SUN Y, DONG L, et al. Structured prompting: Scaling in-context learning to 1,000 examples[J]. arXiv preprint arXiv:2212.06713, 2022.

[220] XU B, WANG Q, MAO Z, et al. $k$ nn prompting: Beyond-context learning with calibration-free nearest neighbor inference[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[221] DAI W, LI J, LI D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning[J]. 2023.

[222] MINDERER M, GRITSENKO A, STONE A, et al. Simple open-vocabulary object detection with vision transformers[J]. 2022. arXiv preprint arXiv:2205.06230.

[223] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for nlp[C]// International Conference on Machine Learning. [S.l.]: PMLR, 2019: 2790-2799.

[224] REBUFFI S A, BILEN H, VEDALDI A. Learning multiple visual domains with residual adapters[J]. Advances in neural information processing systems, 2017, 30.

[225] ZHANG R, ZHENG Y, MAO X, et al. Unsupervised domain adaptation with adapter[J]. arXiv preprint arXiv:2111.00667, 2021.

[226] MALIK B, KASHYAP A R, KAN M Y, et al. Udapter-efficient domain adaptation using adapters[C]// Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.: s.n.], 2023: 2241-2255.

[227] PFEIFFER J, KAMATH A, RÜCKLÉ A, et al. Adapterfusion: Non-destructive task composition for transfer learning[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. [S.l.: s.n.], 2021: 487-503.

[228] CHRONOPOULOU A, PETERS M E, FRASER A, et al. Adaptersoup: Weight averaging to improve generalization of pretrained language models[C]//Findings of the Association for Computational Linguistics: EACL 2023. [S.l.: s.n.], 2023: 2009-2018.

[229] ZHANG R, HAN J, ZHOU A, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention[J]. arXiv preprint arXiv:2303.16199, 2023.

[230] HU Z, LAN Y, WANG L, et al. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models[J]. arXiv preprint arXiv:2304.01933, 2023.

[231] PFEIFFER J, RÜCKLÉ A, POTH C, et al. Adapterhub: A framework for adapting transformers[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. [S.l.: s.n.], 2020: 46-54.

[232] HENDRYCKS D, GIMPEL K. Gaussian error linear units (gelus)[J]. arXiv preprint arXiv:1606.08415, 2016.

[233] BAPNA A, FIRAT O. Simple, scalable adaptation for neural machine translation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.]: Association for Computational Linguistics, 2019.

[234] PFEIFFER J, VULIĆ I, GUREVYCH I, et al. Mad-x: An adapter-based framework for multi-task cross-lingual transfer[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 7654-7673.

[235] ZHAO H, TAN H, MEI H. Tiny-attention adapter: Contexts are more important than the number of parameters[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2022: 6626-6638.

[236] KARIMI MAHABADI R, HENDERSON J, RUDER S. Compacter: Efficient low-rank hypercomplex adapter layers[J]. Advances in Neural Information Processing Systems, 2021, 34: 1022-1035.

[237] ZHANG A, TAY Y, ZHANG S, et al. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters[C]//International Conference on Learning Representations. [S.l.: s.n.], 2020.

[238] HE S, DING L, DONG D, et al. Sparseadapter: An easy approach for improving the parameter-efficiency of adapters[C]//Findings of the Association for Computational Linguistics: EMNLP 2022. [S.l.: s.n.], 2022: 2184-2190.

[239] WANG R, TANG D, DUAN N, et al. K-adapter: Infusing knowledge into pre-trained models with adapters[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. [S.l.: s.n.], 2021: 1405-1418.

[240] SUNG Y L, CHO J, BANSAL M. Lst: Ladder side-tuning for parameter and memory efficient transfer learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 12991-13005.

[241] HU E J, WALLIS P, ALLEN-ZHU Z, et al. Lora: Low-rank adaptation of large language models[C]// International Conference on Learning Representations. [S.l.: s.n.], 2021.

[242] AGHAJANYAN A, GUPTA S, ZETTLEMOYER L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S.l.: s.n.], 2021: 7319-7328.

[243] VALIPOUR M, REZAGHOLIZADEH M, KOBYZEV I, et al. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. [S.l.: s.n.], 2023: 3266-3279.

[244] EDALATI A, TAHAEI M, KOBYZEV I, et al. Krona: Parameter efficient tuning with kronecker adapter[J]. arXiv preprint arXiv:2212.10650, 2022.

[245] HE J, ZHOU C, MA X, et al. Towards a unified view of parameter-efficient transfer learning[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.

[246] MAO Y, MATHIAS L, HOU R, et al. Unipelt: A unified framework for parameter-efficient language model tuning[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 6253-6264.

[247] WANG Y, MUKHERJEE S, LIU X, et al. Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models[J]. arXiv preprint arXiv:2205.12410, 2022, 1(2): 4.

[248] ZAKEN E B, GOLDBERG Y, RAVFOGEL S. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). [S.l.: s.n.], 2022: 1-9.

[249] PONTI E M, SORDONI A, BENGIO Y, et al. Combining modular skills in multitask learning[J]. arXiv preprint arXiv:2202.13914, 2022.

[250] CACCIA L, PONTI E, LIU L, et al. Multi-head adapter routing for data-efficient fine-tuning[J]. arXiv preprint arXiv:2211.03831, 2022.

[251] YANG Z, YI X, LI P, et al. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[252] WANG Y, SI S, LI D, et al. Preserving in-context learning ability in large language model fine-tuning[J]. arXiv preprint arXiv:2211.00635, 2022.

[253] HUANG S, DONG L, WANG W, et al. Language is not all you need: Aligning perception with language models[J]. arXiv preprint arXiv:2302.14045, 2023.

[254] MENICK J, TREBACZ M, MIKULIK V, et al. Teaching language models to support answers with verified quotes[J]. arXiv preprint arXiv:2203.11147, 2022.

[255] PENG B, LI C, HE P, et al. Instruction tuning with gpt-4[J]. arXiv preprint arXiv:2304.03277, 2023.

[256] WANG L, LYU C, JI T, et al. Document-level machine translation with large language models[J]. arXiv preprint arXiv:2304.02210, 2023.

[257] HUANG J, GU S S, HOU L, et al. Large language models can self-improve[J]. arXiv preprint arXiv:2210.11610, 2022.

[258] SCIALOM T, CHAKRABARTY T, MURESAN S. Fine-tuned language models are continual learners[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2022: 6107-6122.

[259] SHIN H, LEE J K, KIM J, et al. Continual learning with deep generative replay[J]. Advances in neural information processing systems, 2017, 30.

[260] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2018: 328-339.

[261] ROBERTS A, RAFFEL C, SHAZEER N. How much knowledge can you pack into the parameters of a language model?[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 5418-5426.

[262] ZIEGLER D M, STIENNON N, WU J, et al. Fine-tuning language models from human preferences[J]. arXiv preprint arXiv:1909.08593, 2019.

[263] DAI D, DONG L, HAO Y, et al. Knowledge neurons in pretrained transformers[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 8493-8502.

[264] DE CAO N, AZIZ W, TITOV I. Editing factual knowledge in language models[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021: 6491-6506.

[265] HERNANDEZ E, LI B Z, ANDREAS J. Measuring and manipulating knowledge representations in language models[J]. arXiv preprint arXiv:2304.00740, 2023.

[266] MENG K, BAU D, ANDONIAN A, et al. Locating and editing factual associations in gpt[J]. Advances in Neural Information Processing Systems, 2022, 35: 17359-17372.

[267] MENG K, SHARMA A S, ANDONIAN A J, et al. Mass-editing memory in a transformer[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[268] MITCHELL E, LIN C, BOSSELUT A, et al. Fast model editing at scale[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.

[269] MITCHELL E, LIN C, BOSSELUT A, et al. Memory-based model editing at scale[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2022: 15817-15831.

[270] JIANG H, HE P, CHEN W, et al. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 2177-2190.

[271] XU R, LUO F, ZHANG Z, et al. Raise a child in large language model: Towards effective and generalizable fine-tuning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021: 9514-9528.

[272] ZHANG H, LI G, LI J, et al. Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively[J]. Advances in Neural Information Processing Systems, 2022, 35: 21442-21454.

[273] MUHAMED A, KEIVANLOO I, PERERA S, et al. Ctr-bert: Cost-effective knowledge distillation for billion-parameter teacher models[C]//NeurIPS Efficient Natural Language and Speech Processing Workshop. [S.l.: s.n.], 2021.

[274] AZERBAYEV Z, NI A, SCHOELKOPF H, et al. Explicit knowledge transfer for weakly-supervised code generation[J]. arXiv preprint arXiv:2211.16740, 2022.

[275] MARJIEH R, SUCHOLUTSKY I, VAN RIJN P, et al. What language reveals about perception: Distilling psychophysical knowledge from large language models[J]. arXiv preprint arXiv:2302.01308, 2023.

[276] VUCETIC D, TAYARANIAN M, ZIAEEFARD M, et al. Efficient fine-tuning of compressed language models with learners[J]. arXiv preprint arXiv:2208.02070, 2022.

[277] HSIEH C Y, LI C L, YEH C K, et al. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes[J]. arXiv preprint arXiv:2305.02301, 2023.

[278] SHRIDHAR K, STOLFO A, SACHAN M. Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions[J]. arXiv preprint arXiv:2212.00193, 2022.

[279] LING W, YOGATAMA D, DYER C, et al. Program induction by rationale generation: Learning to solve and explain algebraic word problems[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2017: 158-167.

[280] ROY S, ROTH D. Solving general arithmetic word problems[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2015: 1743-1752.

[281] CHIANG T R, CHEN Y N. Semantically-aligned equation generation for solving and reasoning math word problems[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). [S.l.: s.n.], 2019: 2656-2668.

[282] AMINI A, GABRIEL S, LIN S, et al. Mathqa: Towards interpretable math word problem solving with operation-based formalisms[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). [S.l.: s.n.], 2019: 2357-2367.

[283] YAO S, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models[J]. arXiv preprint arXiv:2305.10601, 2023.

[284] NEWELL A, SHAW J C, SIMON H A. Report on a general problem solving program[C]//IFIP congress: volume 256. [S.l.]: Pittsburgh, PA, 1959: 64.

[285] NEWELL A, SIMON H A, et al. Human problem solving: volume 104[M]. [S.l.]: Prentice-hall Englewood Cliffs, NJ, 1972.

[286] CAMPBELL M, HOANE JR A J, HSU F H. Deep blue[J]. Artificial intelligence, 2002, 134(1-2): 57-83.

[287] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. nature, 2017, 550(7676): 354-359.

[288] HART P E, NILSSON N J, RAPHAEL B. A formal basis for the heuristic determination of minimum cost paths[J]. IEEE transactions on Systems Science and Cybernetics, 1968, 4(2): 100-107.

[289] BROWNE C B, POWLEY E, WHITEHOUSE D, et al. A survey of monte carlo tree search methods[J]. IEEE Transactions on Computational Intelligence and AI in games, 2012, 4(1): 1-43.

[290] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: Solving elaborate problems with large language models[J]. arXiv preprint arXiv:2308.09687, 2023.

[291] FRISTON K. Hierarchical models in the brain[J]. PLoS computational biology, 2008, 4(11): e1000211.

[292] LEE H, PHATALE S, MANSOOR H, et al. Rlaif: Scaling reinforcement learning from human feedback with ai feedback[J]. arXiv preprint arXiv:2309.00267, 2023.

[293] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]// International conference on machine learning. [S.l.]: PMLR, 2016: 1928-1937.

[294] LIU H, LI C, WU Q, et al. Visual instruction tuning[J]. arXiv preprint arXiv:2304.08485, 2023.

[295] MAAZ M, RASHEED H, KHAN S, et al. Video-chatgpt: Towards detailed video understanding via large vision and language models[J]. arXiv preprint arXiv:2306.05424, 2023.

[296] YE Q, XU H, XU G, et al. mplug-owl: Modularization empowers large language models with multimodality[J]. arXiv preprint arXiv:2304.14178, 2023.

[297] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[298] TIAN Y, KRISHNAN D, ISOLA P. Contrastive multiview coding[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. [S.l.]: Springer, 2020: 776-794.

[299] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[300] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. [S.l.]: PMLR, 2019: 6105-6114.

[301] TOUVRON H, VEDALDI A, DOUZE M, et al. Fixing the train-test resolution discrepancy[J]. Advances in neural information processing systems, 2019, 32.

[302] LI J, LI D, XIONG C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2022: 12888-12900.

[303] ZHANG L, RAO A, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2023: 3836-3847.

[304] CHEN Y C, LI L, YU L, et al. Uniter: Universal image-text representation learning[C]//European conference on computer vision. [S.l.]: Springer, 2020: 104-120.

[305] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in neural information processing systems, 2021, 34: 9694-9705.

[306] KIM W, SON B, KIM I. Vilt: Vision-and-language transformer without convolution or region supervision[C]// International Conference on Machine Learning. [S.l.]: PMLR, 2021: 5583-5594.

[307] WANG Z, YU J, YU A W, et al. Simvlm: Simple visual language model pretraining with weak supervision[C]// International Conference on Learning Representations. [S.l.: s.n.], 2021.

[308] LI J, LI D, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. arXiv preprint arXiv:2301.12597, 2023.

[309] FANG Y, WANG W, XIE B, et al. Eva: Exploring the limits of masked visual representation learning at scale[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 19358-19369.

[310] WENZEK G, LACHAUX M A, CONNEAU A, et al. Ccnet: Extracting high quality monolingual datasets from web crawl data[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. [S.l.: s.n.], 2020: 4003-4012.

[311] GAO L, BIDERMAN S, BLACK S, et al. The pile: An 800gb dataset of diverse text for language modeling[J]. arXiv preprint arXiv:2101.00027, 2020.

[312] LEWKOWYCZ A, ANDREASSEN A, DOHAN D, et al. Solving quantitative reasoning problems with language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 3843-3857.

[313] GAO P, HAN J, ZHANG R, et al. Llama-adapter v2: Parameter-efficient visual instruction model[J]. arXiv preprint arXiv:2304.15010, 2023.

[314] TAORI R, GULRAJANI I, ZHANG T, et al. Stanford alpaca: An instruction-following llama model[Z]. [S.l.: s.n.], 2023.

[315] CHEN X, FANG H, LIN T Y, et al. Microsoft coco captions: Data collection and evaluation server[J]. arXiv preprint arXiv:1504.00325, 2015.

[316] ZAKEN E B, RAVFOGEL S, GOLDBERG Y. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models[J]. arXiv preprint arXiv:2106.10199, 2021.

[317] LIAN D, ZHOU D, FENG J, et al. Scaling & shifting your features: A new baseline for efficient model tuning[J]. Advances in Neural Information Processing Systems, 2022, 35: 109-123.

[318] JIA M, TANG L, CHEN B C, et al. Visual prompt tuning[C]//European Conference on Computer Vision. [S.l.]: Springer, 2022: 709-727.

[319] LI T, WANG L. Learning spatiotemporal features via video and text pair discrimination[J]. arXiv preprint arXiv:2001.05691, 2020.

[320] LI K, HE Y, WANG Y, et al. Videochat: Chat-centric video understanding[J]. arXiv preprint arXiv:2305.06355, 2023.

[321] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]// International Conference on Machine Learning. [S.l.]: PMLR, 2023: 28492-28518.

[322] YUAN L, CHEN D, CHEN Y L, et al. Florence: A new foundation model for computer vision[J]. arXiv preprint arXiv:2111.11432, 2021.

[323] XU H, GHOSH G, HUANG P Y, et al. Videoclip: Contrastive pre-training for zero-shot video-text understanding[J]. arXiv preprint arXiv:2109.14084, 2021.

[324] WANG Y, LI K, LI Y, et al. Internvideo: General video foundation models via generative and discriminative learning[J]. arXiv preprint arXiv:2212.03191, 2022.

[325] LI K, WANG Y, LI Y, et al. Unmasked teacher: Towards training-efficient video foundation models[J]. arXiv preprint arXiv:2303.16058, 2023.

[326] SUN Q, FANG Y, WU L, et al. Eva-clip: Improved training techniques for clip at scale[J]. arXiv preprint arXiv:2303.15389, 2023.

[327] LI K, WANG Y, HE Y, et al. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer[J]. arXiv preprint arXiv:2211.09552, 2022.

[328] BAIN M, NAGRANI A, VAROL G, et al. Frozen in time: A joint video and image encoder for end-to-end retrieval[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2021: 1728-1738.

[329] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 2017, 123: 32-73.

[330] ORDONEZ V, KULKARNI G, BERG T. Im2text: Describing images using 1 million captioned photographs[J]. Advances in neural information processing systems, 2011, 24.

[331] SHARMA P, DING N, GOODMAN S, et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2018: 2556-2565.

[332] CHANGPINYO S, SHARMA P, DING N, et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2021: 3558-3568.

[333] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv preprint arXiv:2001.08361, 2020.

[334] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//International conference on machine learning. [S.l.]: PMLR, 2021: 4904-4916.

[335] GUPTA A, DOLLAR P, GIRSHICK R. Lvis: A dataset for large vocabulary instance segmentation[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 5356-5364.

[336] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[J]. International Journal of Computer Vision, 2020, 128(7): 1956-1981.

[337] ZHOU B, ZHAO H, PUIG X, et al. Semantic understanding of scenes through the ade20k dataset[J]. International Journal of Computer Vision, 2019, 127: 302-321.

[338] MAHADEVAN S, VOIGTLAENDER P, LEIBE B. Iteratively trained interactive segmentation[J]. arXiv preprint arXiv:1805.04398, 2018.

[339] XU N, PRICE B, COHEN S, et al. Deep interactive object selection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 373-381.

[340] KASS M, WITKIN A, TERZOPOULOS D. Snakes: Active contour models[J]. International journal of computer vision, 1988, 1(4): 321-331.

[341] ARBELAEZ P, MAIRE M, FOWLKES C, et al. Contour detection and hierarchical image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(5): 898-916.

[342] REN, MALIK. Learning a classification model for segmentation[C]//Proceedings ninth IEEE international conference on computer vision. [S.l.]: IEEE, 2003: 10-17.

[343] ALEXE B, DESELAERS T, FERRARI V. What is an object?[C]//2010 IEEE computer society conference on computer vision and pattern recognition. [S.l.]: IEEE, 2010: 73-80.

[344] STAUFFER C, GRIMSON W E L. Adaptive background mixture models for real-time tracking[C]//Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149): volume 2. [S.l.]: IEEE, 1999: 246-252.

[345] SHOTTON J, WINN J, ROTHER C, et al. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation[C]//Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. [S.l.]: Springer, 2006: 1-15.

[346] KIRILLOV A, HE K, GIRSHICK R, et al. Panoptic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 9404-9413.

[347] DA SILVA B C, KONIDARIS G, BARTO A G. Learning parameterized skills[C]//Proceedings of the 29th International Coference on International Conference on Machine Learning. [S.l.: s.n.], 2012: 1443-1450.

[348] LI Y, MAO H, GIRSHICK R, et al. Exploring plain vision transformer backbones for object detection[C]// European Conference on Computer Vision. [S.l.]: Springer, 2022: 280-296.

[349] TANCIK M, SRINIVASAN P, MILDENHALL B, et al. Fourier features let networks learn high frequency functions in low dimensional domains[J]. Advances in Neural Information Processing Systems, 2020, 33: 7537-7547.

[350] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. [S.l.]: Springer, 2020: 213-229.

[351] CHENG B, SCHWING A, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation[J]. Advances in Neural Information Processing Systems, 2021, 34: 17864-17875.

[352] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 2980-2988.

[353] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). [S.l.]: Ieee, 2016: 565-571.

[354] SOFIIUK K, PETROV I A, KONUSHIN A. Reviving iterative training with mask guidance for interactive segmentation[C]//2022 IEEE International Conference on Image Processing (ICIP). [S.l.]: IEEE, 2022: 3141-3145.

[355] FORTE M, PRICE B, COHEN S, et al. Getting to 99% accuracy in interactive segmentation[J]. arXiv preprint arXiv:2003.07932, 2020.

[356] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[357] POLU S, HAN J M, ZHENG K, et al. Formal mathematics statement curriculum learning[J]. arXiv preprint arXiv:2202.01344, 2022.

[358] TELLEX S, GOPALAN N, KRESS-GAZIT H, et al. Robots that use language[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2020, 3: 25-55.

[359] AHN M, BROHAN A, BROWN N, et al. Do as i can, not as i say: Grounding language in robotic affordances[J]. arXiv preprint arXiv:2204.01691, 2022.

[360] DRIESS D, XIA F, SAJJADI M S, et al. Palm-e: An embodied multimodal language model[J]. arXiv preprint arXiv:2303.03378, 2023.

[361] ZENG A, ATTARIAN M, ICHTER B, et al. Socratic models: Composing zero-shot multimodal reasoning with language[J]. arXiv preprint arXiv:2204.00598, 2022.

[362] LYNCH C, SERMANET P. Language conditioned imitation learning over unstructured data[J]. arXiv preprint arXiv:2005.07648, 2020.

[363] BROHAN A, BROWN N, CARBAJAL J, et al. Rt-1: Robotics transformer for real-world control at scale[J]. arXiv preprint arXiv:2212.06817, 2022.

[364] CHEN X, WANG X, CHANGPINYO S, et al. Pali: A jointly-scaled multilingual language-image model[J]. arXiv preprint arXiv:2209.06794, 2022.

[365] RYOO M S, PIERGIOVANNI A, ARNAB A, et al. Tokenlearner: What can 8 learned tokens do for images and videos?[J]. arXiv preprint arXiv:2106.11297, 2021.

[366] SAJJADI M S, DUCKWORTH D, MAHENDRAN A, et al. Object scene representation transformer[J]. Advances in Neural Information Processing Systems, 2022, 35: 9512-9524.

[367] LOCATELLO F, WEISSENBORN D, UNTERTHINER T, et al. Object-centric learning with slot attention[J]. Advances in Neural Information Processing Systems, 2020, 33: 11525-11538.

[368] SAJJADI M S, MEYER H, POT E, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 6229-6238.

[369] ANTOL S, AGRAWAL A, LU J, et al. Vqa: Visual question answering[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2015: 2425-2433.

[370] WU Q, WANG P, WANG X, et al. Visual question answering: From theory to application[M]. [S.l.]: Springer Nature, 2022.

[371] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 6904-6913.

[372] REN M, KIROS R, ZEMEL R. Image question answering: A visual semantic embedding model and a new dataset[J]. Proc. Advances in Neural Inf. Process. Syst, 2015, 1(2): 5.

[373] ZHU Y, GROTH O, BERNSTEIN M, et al. Visual7w: Grounded question answering in images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 4995-5004.

[374] HUDSON D A, MANNING C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 6700-6709.

[375] ZHANG P, GOYAL Y, SUMMERS-STAY D, et al. Yin and yang: Balancing and answering binary visual questions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 5014-5022.

[376] JOHNSON J, HARIHARAN B, VAN DER MAATEN L, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 2901-2910.

[377] BEN ABACHA A, HASAN S A, DATLA V V, et al. Vqa-med: Overview of the medical visual question answering task at imageclef 2019[C]//Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes. [S.l.]: 9-12 September 2019, 2019.

[378] SINGH A, NATARAJAN V, SHAH M, et al. Towards vqa models that can read[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 8317-8326.

[379] JANG Y, SONG Y, YU Y, et al. Tgif-qa: Toward spatio-temporal reasoning in visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 2758-2766.

[380] LEI J, YU L, BANSAL M, et al. Tvqa: Localized, compositional video question answering[J]. arXiv preprint arXiv:1809.01696, 2018.

[381] TAPASWI M, ZHU Y, STIEFELHAGEN R, et al. Movieqa: Understanding stories in movies through question-answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 4631-4640.

[382] WANG P, WU Q, SHEN C, et al. Explicit knowledge-based reasoning for visual question answering[J]. arXiv preprint arXiv:1511.02570, 2015.

[383] WANG P, WU Q, SHEN C, et al. Fvqa: Fact-based visual question answering[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(10): 2413-2427.

[384] MARINO K, RASTEGARI M, FARHADI A, et al. Ok-vqa: A visual question answering benchmark requiring external knowledge[C]//Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 3195-3204.

[385] DAS A, DATTA S, GKIOXARI G, et al. Embodied question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 1-10.

[386] MALINOWSKI M, FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input[J]. Advances in neural information processing systems, 2014, 27.

[387] REN M, KIROS R, ZEMEL R. Exploring models and data for image question answering[J]. Advances in neural information processing systems, 2015, 28.

[388] GAO H, MAO J, ZHOU J, et al. Are you talking to a machine? dataset and methods for multilingual image question[J]. Advances in neural information processing systems, 2015, 28.

[389] YU L, PARK E, BERG A C, et al. Visual madlibs: Fill in the blank description generation and question answering[C]//Proceedings of the ieee international conference on computer vision. [S.l.: s.n.], 2015: 2461-2469.

[390] MALINOWSKI M, ROHRBACH M, FRITZ M. Ask your neurons: A neural-based approach to answering questions about images[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2015: 1-9.

[391] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2016.

[392] CHARIKAR M, CHEN K, FARACH-COLTON M. Finding frequent items in data streams[C]//International Colloquium on Automata, Languages, and Programming. [S.l.]: Springer, 2002: 693-703.

[393] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[C]//International Conference on Learning Representations. [S.l.: s.n.], 2016.

[394] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 21-29.

[395] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[J]. Advances in neural information processing systems, 2016, 29.

[396] XIONG C, MERITY S, SOCHER R. Dynamic memory networks for visual and textual question answering[C]// International conference on machine learning. [S.l.]: PMLR, 2016: 2397-2406.

[397] MA C, SHEN C, DICK A, et al. Visual question answering with memory-augmented networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 6975-6984.

[398] ANDREAS J, ROHRBACH M, DARRELL T, et al. Neural module networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 39-48.

[399] DE MARNEFFE M C, MANNING C D. The stanford typed dependencies representation[C]//Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation. [S.l.: s.n.], 2008: 1-8.

[400] ANDREAS J, ROHRBACH M, DARRELL T, et al. Learning to compose neural networks for question answering[C]//Proceedings of NAACL-HLT. [S.l.: s.n.], 2016: 1545-1554.

[401] LEI J, YU L, BERG T L, et al. Tvqa+: Spatio-temporal grounding for video question answering[J]. arXiv preprint arXiv:1904.11574, 2019.

[402] SONG X, SHI Y, CHEN X, et al. Explore multi-step reasoning in video question answering[C]//Proceedings of the 26th ACM international conference on Multimedia. [S.l.: s.n.], 2018: 239-247.

[403] YI K, GAN C, LI Y, et al. Clevrer: Collision events for video representation and reasoning[C]//International Conference on Learning Representations. [S.l.: s.n.], 2020.

[404] GRUNDE-MCLAUGHLIN M, KRISHNA R, AGRAWALA M. Agqa: A benchmark for compositional spatio-temporal reasoning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2021: 11287-11297.

[405] XU L, HUANG H, LIU J. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2021: 9878-9888.

[406] WANG B, XU Y, HAN Y, et al. Movie question answering: Remembering the textual cues for layered visual contents[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. [S.l.: s.n.], 2018.

[407] YU Z, XU D, YU J, et al. Activitynet-qa: A dataset for understanding complex web videos via question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. [S.l.: s.n.], 2019: 9127-9134.

[408] MUN J, HONGSUCK SEO P, JUNG I, et al. Marioqa: Answering questions by watching gameplay videos[C]// Proceedings of the IEEE International Conference on Computer Vision. [S.l.: s.n.], 2017: 2867-2875.

[409] KIM K M, HEO M O, CHOI S H, et al. Deepstory: video story qa by deep embedded memory networks[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. [S.l.: s.n.], 2017: 2016-2022.
[410] ZHU L, XU Z, YANG Y, et al. Uncovering the temporal context for video question answering[J]. International Journal of Computer Vision, 2017, 124: 409-421.
[411] XUE H, ZHAO Z, CAI D. Unifying the video and question attentions for open-ended video question answering[J]. IEEE Transactions on Image Processing, 2017, 26(12): 5656-5666.
[412] ZHAO Z, YANG Q, CAI D, et al. Video question answering via hierarchical spatio-temporal attention networks. [C]//IJCAI: volume 2. [S.l.: s.n.], 2017: 8.
[413] ZHAO Z, ZHANG Z, XIAO S, et al. Open-ended long-form video question answering via adaptive hierarchical reinforced networks.[C]//IJCAI: volume 2. [S.l.: s.n.], 2018: 8.
[414] LIU Y, LI G, LIN L. Cross-modal causal relational reasoning for event-level visual question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
[415] ZHANG C, ZHANG C, ZHENG S, et al. A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need?[J]. arXiv preprint arXiv:2303.11717, 2023.
[416] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
[417] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
[418] CAO H, TAN C, GAO Z, et al. A survey on generative diffusion model[J]. arXiv preprint arXiv:2209.02646, 2022.
[419] CROITORU F A, HONDRU V, IONESCU R T, et al. Diffusion models in vision: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
[420] LI X, THICKSTUN J, GULRAJANI I, et al. Diffusion-lm improves controllable text generation[J]. Advances in Neural Information Processing Systems, 2022, 35: 4328-4343.
[421] DERIU J, RODRIGO A, OTEGI A, et al. Survey on evaluation methods for dialogue systems[J]. Artificial Intelligence Review, 2021, 54: 755-810.
[422] NI J, YOUNG T, PANDELEA V, et al. Recent advances in deep learning based dialogue systems: A systematic survey[J]. Artificial intelligence review, 2023, 56(4): 3055-3155.
[423] PENG B, ZHU C, LI C, et al. Few-shot natural language generation for task-oriented dialog[J]. arXiv preprint arXiv:2002.12328, 2020.
[424] YANG Y, LI Y, QUAN X. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2[C]// Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. [S.l.: s.n.], 2021: 14230-14238.
[425] ZHANG Z, TAKANOBU R, ZHU Q, et al. Recent advances and challenges in task-oriented dialog systems[J]. Science China Technological Sciences, 2020, 63(10): 2011-2027.
[426] ADIWARDANA D, LUONG M T, SO D R, et al. Towards a human-like open-domain chatbot[J]. arXiv preprint arXiv:2001.09977, 2020.
[427] ZHANG Y, SUN S, GALLEY M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation[J]. arXiv preprint arXiv:1911.00536, 2019.
[428] ZHOU L, GAO J, LI D, et al. The design and implementation of xiaoice, an empathetic social chatbot[J]. Computational Linguistics, 2020, 46(1): 53-93.
[429] SINGLA K, CHEN Z, ATKINS D C, et al. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations[C]//Proceedings of the conference. Association for Computational Linguistics. Meeting: volume 2020. [S.l.]: NIH Public Access, 2020: 3797.
[430] SU S Y, HUANG C W, CHEN Y N. Dual supervised learning for natural language understanding and generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2019: 5472-5477.

[431] SHAN Y, LI Z, ZHANG J, et al. A contextual hierarchical attention network with adaptive objective for dialogue state tracking[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 6322-6333.

[432] WANG Y, GUO Y, ZHU S. Slot attention with value normalization for multi-domain dialogue state tracking[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 3019-3028.

[433] HUANG X, QI J, SUN Y, et al. Semi-supervised dialogue policy learning via stochastic reward estimation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 660-670.

[434] XU J, WANG H, NIU Z Y, et al. Conversational graph grounded policy learning for open-domain conversation generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 1835-1845.

[435] BAHETI A, RITTER A, SMALL K. Fluent response generation for conversational question answering[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 191-207.

[436] ELDER H, O' CONNOR A, FOSTER J. How to make neural natural language generation as reliable as templates in task-oriented dialogue[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 2877-2888.

[437] RITTER A, CHERRY C, DOLAN B. Data-driven response generation in social media[C]//Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2011.

[438] SERBAN I V, SANKAR C, GERMAIN M, et al. A deep reinforcement learning chatbot[J]. arXiv preprint arXiv:1709.02349, 2017.

[439] ZHU Q, CUI L, ZHANG W, et al. Retrieval-enhanced adversarial training for neural response generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2019: 3763-3773.

[440] FENG S, REN X, CHEN H, et al. Regularizing dialogue generation by imitating implicit scenarios[J]. arXiv preprint arXiv:2010.01893, 2020.

[441] JIA Q, LIU Y, REN S, et al. Multi-turn response selection using dialogue dependency relations[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 1911-1920.

[442] LIN Z, CAI D, WANG Y, et al. The world is not binary: Learning to rank with grayscale data for dialogue response selection[J]. arXiv preprint arXiv:2004.02421, 2020.

[443] MEHRI S, RAZUMOVSKAIA E, ZHAO T, et al. Pretraining methods for dialog context representation learning[J]. arXiv preprint arXiv:1906.00414, 2019.

[444] HUTCHINS W J. Machine translation: past, present, future[M]. [S.l.]: Ellis Horwood Chichester, 1986.

[445] YANG S, WANG Y, CHU X. A survey of deep learning techniques for neural machine translation[J]. arXiv preprint arXiv:2002.07526, 2020.

[446] FORCADA M L, GINESTÍ-ROSELL M, NORDFALK J, et al. Apertium: a free/open-source platform for rule-based machine translation[J]. Machine translation, 2011, 25: 127-144.

[447] KOEHN P, HOANG H, BIRCH A, et al. Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. [S.l.: s.n.], 2007: 177-180.

[448] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]//Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. [S.l.: s.n.], 2003: 127-133.

[449] SONG F, CROFT W B. A general language model for information retrieval[C]//Proceedings of the eighth international conference on Information and knowledge management. [S.l.: s.n.], 1999: 316-321.

[450] WALLACH H M. Topic modeling: beyond bag-of-words[C]//Proceedings of the 23rd international conference on Machine learning. [S.l.: s.n.], 2006: 977-984.

[451] CHO K, VAN MERRIËNBOER B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.

[452] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[453] KALCHBRENNER N, BLUNSOM P. Recurrent continuous translation models[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. [S.l.: s.n.], 2013: 1700-1709.

[454] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[455] KAISER Ł, BENGIO S. Can active memory replace attention?[J]. Advances in Neural Information Processing Systems, 2016, 29.

[456] KALCHBRENNER N, ESPEHOLT L, SIMONYAN K, et al. Neural machine translation in linear time[J]. arXiv preprint arXiv:1610.10099, 2016.

[457] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//International conference on machine learning. [S.l.]: PMLR, 2017: 1243-1252.

[458] WANG R, TAN X, LUO R, et al. A survey on low-resource neural machine translation[J]. arXiv preprint arXiv:2107.04239, 2021.

[459] JOHNSON M, SCHUSTER M, LE Q V, et al. Google' s multilingual neural machine translation system: Enabling zero-shot translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.

[460] SHATZ I. Native language influence during second language acquisition: A large-scale learner corpus analysis[C]//Proceedings of the Pacific Second Language Research Forum (PacSLRF 2016). [S.l.]: Japan Second Language Association Hiroshima, Japan, 2017: 175-180.

[461] ZOPH B, KNIGHT K. Multi-source neural translation[J]. arXiv preprint arXiv:1601.00710, 2016.

[462] CHENG Y, CHENG Y. Joint training for pivot-based neural machine translation[J]. Joint training for neural machine translation, 2019: 41-54.

[463] REN S, CHEN W, LIU S, et al. Triangular architecture for rare language translation[J]. arXiv preprint arXiv:1805.04813, 2018.

[464] ROTHE S, NARAYAN S, SEVERYN A. Leveraging pre-trained checkpoints for sequence generation tasks[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 264-280.

[465] JIAO W, WANG W, HUANG J T, et al. Is chatgpt a good translator? a preliminary study[J]. arXiv preprint arXiv:2301.08745, 2023.

[466] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2015: 3156-3164.

[467] STEFANINI M, CORNIA M, BARALDI L, et al. From show to tell: A survey on deep learning-based image captioning[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 539-559.

[468] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2015: 1-9.

[469] KARPATHY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2015: 3128-3137.

[470] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[471] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2015: 2625-2634.

[472] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014.

[473] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[474] CHEN X, MA L, JIANG W, et al. Regularizing rnns for caption generation by reconstructing the past with the present[C]//Proceedings of the IEEE Conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 7995-8003.

[475] LU J, XIONG C, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 375-383.

[476] WANG Y, LIN Z, SHEN X, et al. Skeleton key: Image captioning by skeleton-attribute decomposition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 7272-7281.

[477] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// International conference on machine learning. [S.l.]: PMLR, 2015: 2048-2057.

[478] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 6077-6086.

[479] KE L, PEI W, LI R, et al. Reflective decoding network for image captioning[C]//Proceedings of the IEEE/ CVF international conference on computer vision. [S.l.: s.n.], 2019: 8888-8897.

[480] ZHA Z J, LIU D, ZHANG H, et al. Context-aware visual policy network for fine-grained image captioning[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 44(2): 710-722.

[481] YANG X, TANG K, ZHANG H, et al. Auto-encoding scene graphs for image captioning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 10685-10694.

[482] ZHAO W, WU X. Boosting entity-aware image captioning with multi-modal knowledge graph[J]. IEEE Transactions on Multimedia, 2023.

[483] LI G, ZHU L, LIU P, et al. Entangled transformer for image captioning[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2019: 8928-8937.

[484] YANG X, ZHANG H, CAI J. Learning to collocate neural modules for image captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2019: 4250-4260.

[485] ZHANG X, SUN X, LUO Y, et al. Rstnet: Captioning with adaptive attention on visual and non-visual words[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 15465-15474.

[486] LIU W, CHEN S, GUO L, et al. Cptr: Full transformer network for image captioning[J]. arXiv preprint arXiv:2101.10804, 2021.

[487] WANG L, BAI Z, ZHANG Y, et al. Show, recall, and tell: Image captioning with recall mechanism[C]// Proceedings of the AAAI conference on artificial intelligence: volume 34. [S.l.: s.n.], 2020: 12176-12183.

[488] GUO L, LIU J, ZHU X, et al. Normalized and geometry-aware self-attention network for image captioning[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 10327-10336.

[489] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: Transforming objects into words[J]. Advances in neural information processing systems, 2019, 32.

[490] LI X, YIN X, LI C, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. [S.l.]: Springer, 2020: 121-137.

[491] ZHANG P, LI X, HU X, et al. Vinvl: Revisiting visual representations in vision-language models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 5579-5588.

[492] ZHOU L, PALANGI H, ZHANG L, et al. Unified vision-language pre-training for image captioning and vqa[C]//Proceedings of the AAAI conference on artificial intelligence: volume 34. [S.l.: s.n.], 2020: 13041-13049.

[493] INDURKHYA N, DAMERAU F J. Handbook of natural language processing: volume 2[M]. [S.l.]: CRC Press, 2010.

[494] REDDY D R. Speech recognition by machine: A review[J]. Proceedings of the IEEE, 1976, 64(4): 501-531.

[495] KARPAGAVALLI S, CHANDRA E. A review on automatic speech recognition architecture and approaches[J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2016, 9(4): 393-404.

[496] MALIK M, MALIK M K, MEHMOOD K, et al. Automatic speech recognition: a survey[J]. Multimedia Tools and Applications, 2021, 80: 9411-9457.

[497] RAUT P C, DEOGHARE S U. Automatic speech recognition and its applications[J]. International Research Journal of Engineering and Technology, 2016, 3(5): 2368-2371.

[498] JUANG B H, RABINER L R. Hidden markov models for speech recognition[J]. Technometrics, 1991, 33(3): 251-272.

[499] LEVINSON S E, RABINER L R, SONDHI M M. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition[J]. Bell System Technical Journal, 1983, 62(4): 1035-1074.

[500] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on audio, speech, and language processing, 2011, 20(1): 30-42.

[501] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. [S.l.]: Ieee, 2013: 6645-6649.

[502] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[503] CHIU C C, SAINATH T N, WU Y, et al. State-of-the-art speech recognition with sequence-to-sequence models[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). [S.l.]: IEEE, 2018: 4774-4778.

[504] COLLOBERT R, PUHRSCH C, SYNNAEVE G. Wav2letter: an end-to-end convnet-based speech recognition system[J]. arXiv preprint arXiv:1609.03193, 2016.

[505] TÓTH L. A hierarchical, context-dependent neural network architecture for improved phone recognition[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2011: 5040-5043.

[506] ROGER V, FARINAS J, PINQUIER J. Deep neural networks for automatic speech processing: a survey from large corpora to limited data[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2022, 2022(1): 19.

[507] LÜSCHER C, BECK E, IRIE K, et al. Rwth asr systems for librispeech: Hybrid vs attention–w/o data augmentation[J]. arXiv preprint arXiv:1905.03072, 2019.

[508] HORI T, CHO J, WATANABE S. End-to-end speech recognition with word-based rnn language models[C]//2018 IEEE Spoken Language Technology Workshop (SLT). [S.l.]: IEEE, 2018: 389-396.

[509] LI J, LAVRUKHIN V, GINSBURG B, et al. Jasper: An end-to-end convolutional neural acoustic model[J]. arXiv preprint arXiv:1904.03288, 2019.

[510] LI J, WU Y, GAUR Y, et al. On the comparison of popular end-to-end models for large scale speech recognition[J]. arXiv preprint arXiv:2005.14327, 2020.

[511] KARITA S, SOPLIN N E Y, WATANABE S, et al. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration[C]//Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH: volume 2019. [S.l.: s.n.], 2019: 1408-1412.

[512] WANG C, WU Y, DU Y, et al. Semantic mask for transformer based end-to-end speech recognition[J]. arXiv preprint arXiv:1912.03010, 2019.

[513] FENDJI J L K E, TALA D C, YENKE B O, et al. Automatic speech recognition using limited vocabulary: A survey[J]. Applied Artificial Intelligence, 2022, 36(1): 2095039.

[514] BENZEGHIBA M, DE MORI R, DEROO O, et al. Automatic speech recognition and speech variability: A review[J]. Speech communication, 2007, 49(10-11): 763-786.

[515] PASCUAL S, RAVANELLI M, SERRA J, et al. Learning problem-agnostic speech representations from multiple self-supervised tasks[J]. arXiv preprint arXiv:1904.03416, 2019.

[516] RAVANELLI M, ZHONG J, PASCUAL S, et al. Multi-task self-supervised learning for robust speech recognition[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2020: 6989-6993.

[517] BAEVSKI A, AULI M, MOHAMED A. Effectiveness of self-supervised pre-training for speech recognition[J]. arXiv preprint arXiv:1911.03912, 2019.

[518] CONNEAU A, BAEVSKI A, COLLOBERT R, et al. Unsupervised cross-lingual representation learning for speech recognition[J]. arXiv preprint arXiv:2006.13979, 2020.

[519] LIU S, MALLOL-RAGOLTA A, PARADA-CABALEIRO E, et al. Audio self-supervised learning: A survey[J]. Patterns, 2022, 3(12).

[520] YADAV H, SITARAM S. A survey of multilingual models for automatic speech recognition[J]. arXiv preprint arXiv:2202.12576, 2022.

[521] CASTLEMAN K R. Digital image processing[M]. [S.l.]: Prentice Hall Press, 1996.

[522] SHRESTHA S. Image denoising using new adaptive based median filters[J]. arXiv preprint arXiv:1410.2175, 2014.

[523] ZHANG X, FENG X, WANG W, et al. Gradient-based wiener filter for image denoising[J]. Computers & Electrical Engineering, 2013, 39(3): 934-944.

[524] XU L, JIA J. Two-phase kernel estimation for robust motion deblurring[C]//Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11. [S.l.]: Springer, 2010: 157-170.

[525] XU X, LIU H, LI Y, et al. Image deblurring with blur kernel estimation in rgb channels[C]//2016 IEEE international conference on digital signal processing (DSP). [S.l.]: IEEE, 2016: 681-684.

[526] CHENG Z, YANG Q, SHENG B. Deep colorization[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2015: 415-423.

[527] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. [S.l.]: Springer, 2014: 184-199.

[528] JAIN V, SEUNG S. Natural image denoising with convolutional networks[J]. Advances in neural information processing systems, 2008, 21.

[529] LIANG J, LIU R. Stacked denoising autoencoder and dropout together to prevent overfitting in deep neural network[C]//2015 8th international congress on image and signal processing (CISP). [S.l.]: IEEE, 2015: 697-701.

[530] LIU H, WAN Z, HUANG W, et al. Pd-gan: Probabilistic diverse gan for image inpainting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 9371-9381.

[531] XIE J, XU L, CHEN E. Image denoising and inpainting with deep neural networks[J]. Advances in neural information processing systems, 2012, 25.

[532] XU L, REN J S, LIU C, et al. Deep convolutional neural network for image deconvolution[J]. Advances in neural information processing systems, 2014, 27.

[533] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[534] SUN J, CAO W, XU Z, et al. Learning a convolutional neural network for non-uniform motion blur removal[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2015: 769-777.

[535] VARGA D, SZIRÁNYI T. Fully automatic image colorization based on convolutional neural network[C]//2016 23rd International Conference on Pattern Recognition (ICPR). [S.l.]: IEEE, 2016: 3691-3696.

[536] WANG X, TAO Q, WANG L, et al. Deep convolutional architecture for natural image denoising[C]//2015 International Conference on Wireless Communications & Signal Processing (WCSP). [S.l.]: IEEE, 2015: 1-4.

[537] LIANG J, CAO J, SUN G, et al. Swinir: Image restoration using swin transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 1833-1844.

[538] KUMAR M, WEISSENBORN D, KALCHBRENNER N. Colorization transformer[J]. arXiv preprint arXiv:2102.04432, 2021.

[539] LI W, LIN Z, ZHOU K, et al. Mat: Mask-aware transformer for large hole image inpainting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 10758-10768.

[540] FANG J, LIN H, CHEN X, et al. A hybrid network of cnn and transformer for lightweight image super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 1103-1112.

[541] ZHAO M, CAO G, HUANG X, et al. Hybrid transformer-cnn for real image denoising[J]. IEEE Signal Processing Letters, 2022, 29: 1252-1256.

[542] ZHAO Q, YANG H, ZHOU D, et al. Rethinking image deblurring via cnn-transformer multiscale hybrid architecture[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 72: 1-15.

[543] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 4681-4690.

[544] WANG X, YU K, WU S, et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]// Proceedings of the European conference on computer vision (ECCV) workshops. [S.l.: s.n.], 2018: 0-0.

[545] ZHANG W, LIU Y, DONG C, et al. Ranksrgan: Generative adversarial networks with ranker for image super-resolution[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2019: 3096-3105.

[546] NAZERI K, NG E, JOSEPH T, et al. Edgeconnect: Generative image inpainting with adversarial edge learning[J]. arXiv preprint arXiv:1901.00212, 2019.

[547] KAWAR B, ELAD M, ERMON S, et al. Denoising diffusion restoration models[J]. Advances in Neural Information Processing Systems, 2022, 35: 23593-23606.

[548] LI H, YANG Y, CHANG M, et al. Srdiff: Single image super-resolution with diffusion probabilistic models[J]. Neurocomputing, 2022, 479: 47-59.

[549] LUGMAYR A, DANELLJAN M, ROMERO A, et al. Repaint: Inpainting using denoising diffusion probabilistic models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 11461-11471.

[550] REN M, DELBRACIO M, TALEBI H, et al. Image deblurring with domain generalizable diffusion models[J]. arXiv preprint arXiv:2212.01789, 2022.

[551] SAHARIA C, HO J, CHAN W, et al. Image super-resolution via iterative refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4713-4726.

[552] AHN N, KANG B, SOHN K A. Image distortion detection using convolutional neural network[C]//2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). [S.l.]: IEEE, 2017: 220-225.

[553] KIM S, AHN N, SOHN K A. Restoring spatially-heterogeneous distortions using mixture of experts network[C]// Proceedings of the Asian Conference on Computer Vision. [S.l.: s.n.], 2020.

[554] SHIN W, AHN N, MOON J H, et al. Exploiting distortion information for multi-degraded image restoration[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 537-546.

[555] ZHOU J, LEONG C, LIN M, et al. Task adaptive network for image restoration with combined degradation factors[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l.: s.n.], 2022: 1-8.

[556] LIU X, SUGANUMA M, LUO X, et al. Restoring images with unknown degradation factors by recurrent use of a multi-branch network[J]. arXiv preprint arXiv:1907.04508, 2019.

[557] YU K, DONG C, LIN L, et al. Crafting a toolchain for image restoration by deep reinforcement learning[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 2443-2452.

[558] YUAN Y, LIU S, ZHANG J, et al. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. [S.l.: s.n.], 2018: 701-710.

[559] LI B, LIU X, HU P, et al. All-in-one image restoration for unknown corruption[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 17452-17462.

[560] SUGANUMA M, LIU X, OKATANI T. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2019: 9039-9048.

[561] ZHANG K, HUANG D, ZHANG D. An optimized palmprint recognition approach based on image sharpness[J]. Pattern Recognition Letters, 2017, 85: 65-71.

[562] ZHANG M, ZHANG L, SUN Y, et al. Auto cropping for digital photographs[C]//2005 IEEE international conference on multimedia and expo. [S.l.]: IEEE, 2005: 4-pp.

[563] SMOLKA B, CZUBIN K, HARDEBERG J Y, et al. Towards automatic redeye effect removal[J]. Pattern Recognition Letters, 2003, 24(11): 1767-1785.

[564] LI M, ZUO W, ZHANG D. Convolutional network for attribute-driven and identity-preserving human face generation[J]. arXiv preprint arXiv:1608.06434, 2016.

[565] UPCHURCH P, GARDNER J, PLEISS G, et al. Deep feature interpolation for image content changes[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 7064-7073.

[566] LI M, ZUO W, ZHANG D. Deep identity-aware transfer of facial attributes[J]. arXiv preprint arXiv:1610.05586, 2016.

[567] SHEN W, LIU R. Learning residual images for face attribute manipulation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 4030-4038.

[568] HE Z, ZUO W, KAN M, et al. Attgan: Facial attribute editing by only changing what you want[J]. IEEE transactions on image processing, 2019, 28(11): 5464-5478.

[569] KIM T, KIM B, CHA M, et al. Unsupervised visual attribute transfer with reconfigurable generative adversarial networks[J]. arXiv preprint arXiv:1707.09798, 2017.

[570] XIAO T, HONG J, MA J. Dna-gan: Learning disentangled representations from multi-attribute images[J]. arXiv preprint arXiv:1711.05415, 2017.

[571] CHEREPKOV A, VOYNOV A, BABENKO A. Navigating the gan parameter space for semantic image editing[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 3671-3680.

[572] SHEN Y, ZHOU B. Closed-form factorization of latent semantics in gans[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 1532-1540.

[573] JING Y, YANG Y, FENG Z, et al. Neural style transfer: A review[J]. IEEE transactions on visualization and computer graphics, 2019, 26(11): 3365-3385.

[574] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 2414-2423.

[575] XIA W, ZHANG Y, YANG Y, et al. Gan inversion: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3121-3138.

[576] ABDAL R, QIN Y, WONKA P. Image2stylegan: How to embed images into the stylegan latent space?[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2019: 4432-4441.

[577] XU Y, DU Y, XIAO W, et al. From continuity to editability: Inverting gans with consecutive images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2021: 13910-13918.

[578] ZHU J, SHEN Y, ZHAO D, et al. In-domain gan inversion for real image editing[C]//European conference on computer vision. [S.l.]: Springer, 2020: 592-608.

[579] ZHU J Y, KRÄHENBÜHL P, SHECHTMAN E, et al. Generative visual manipulation on the natural image manifold[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. [S.l.]: Springer, 2016: 597-613.

[580] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 4401-4410.

[581] GUAN S, TAI Y, NI B, et al. Collaborative learning for faster stylegan embedding[J]. arXiv preprint arXiv:2007.01758, 2020.

[582] VIAZOVETSKYI Y, IVASHKIN V, KASHIN E. Stylegan2 distillation for feed-forward image manipulation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. [S.l.]: Springer, 2020: 170-186.

[583] WEI T, CHEN D, ZHOU W, et al. E2style: Improve the efficiency and effectiveness of stylegan inversion[J]. IEEE Transactions on Image Processing, 2022, 31: 3267-3280.

[584] CHANDRAMOULI P, GANDIKOTA K V. Ldedit: Towards generalized text guided image manipulation via latent diffusion models[J]. arXiv preprint arXiv:2210.02249, 2022, 3.

[585] HERTZ A, MOKADY R, TENENBAUM J, et al. Prompt-to-prompt image editing with cross attention control[J]. arXiv preprint arXiv:2208.01626, 2022.

[586] WALLACE B, GOKUL A, NAIK N. Edict: Exact diffusion inversion via coupled transformations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 22532-22541.

[587] KIM G, YE J C. Diffusionclip: Text-guided image manipulation using diffusion models[J]. 2021.

[588] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 10684-10695.

[589] ACKERMANN J, LI M. High-resolution image editing via multi-stage blended diffusion[J]. arXiv preprint arXiv:2210.12965, 2022.

[590] AVRAHAMI O, FRIED O, LISCHINSKI D. Blended latent diffusion[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-11.

[591] AVRAHAMI O, LISCHINSKI D, FRIED O. Blended diffusion for text-driven editing of natural images[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 18208-18218.

[592] COUAIRON G, VERBEEK J, SCHWENK H, et al. Diffedit: Diffusion-based semantic image editing with mask guidance[C]//ICLR 2023 (Eleventh International Conference on Learning Representations). [S.l.: s.n.], 2023.

[593] CHAN E R, LIN C Z, CHAN M A, et al. Efficient geometry-aware 3d generative adversarial networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 16123-16133.

[594] KIM G, CHUN S Y. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 14203-14213.

[595] LI G, ZHENG H, WANG C, et al. 3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models[J]. arXiv preprint arXiv:2211.14108, 2022.

[596] SRIVASTAVA N, SALAKHUTDINOV R R. Multimodal learning with deep boltzmann machines[J]. Advances in neural information processing systems, 2012, 25.

[597] YAN X, YANG J, SOHN K, et al. Attribute2image: Conditional image generation from visual attributes[C]// Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. [S.l.]: Springer, 2016: 776-791.

[598] MANSIMOV E, PARISOTTO E, BA J L, et al. Generating images from captions with attention[J]. arXiv preprint arXiv:1511.02793, 2015.

[599] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//International conference on machine learning. [S.l.]: PMLR, 2016: 1060-1069.

[600] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[601] ZHANG H, XU T, LI H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 5907-5915.

[602] ZHANG H, XU T, LI H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962.

[603] XU T, ZHANG P, HUANG Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 1316-1324.

[604] LI B, QI X, LUKASIEWICZ T, et al. Controllable text-to-image generation[J]. Advances in Neural Information Processing Systems, 2019, 32.

[605] DING M, YANG Z, HONG W, et al. Cogview: Mastering text-to-image generation via transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 19822-19835.

[606] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[607] DING M, ZHENG W, HONG W, et al. Cogview2: Faster and better text-to-image generation via hierarchical transformers[J]. Advances in Neural Information Processing Systems, 2022, 35: 16890-16902.

[608] YU J, XU Y, KOH J Y, et al. Scaling autoregressive models for content-rich text-to-image generation[J]. arXiv preprint arXiv:2206.10789, 2022, 2(3): 5.

[609] NICHOL A Q, DHARIWAL P, RAMESH A, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2022: 16784-16804.

[610] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.

[611] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022.

[612] AVRAHAMI O, HAYES T, GAFNI O, et al. Spatext: Spatio-textual representation for controllable image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 18370-18380.

[613] VOYNOV A, ABERMAN K, COHEN-OR D. Sketch-guided text-to-image diffusion models[C]//ACM SIGGRAPH 2023 Conference Proceedings. [S.l.: s.n.], 2023: 1-11.

[614] BLATTMANN A, ROMBACH R, OKTAY K, et al. Retrieval-augmented diffusion models[J]. Advances in Neural Information Processing Systems, 2022, 35: 15309-15324.

[615] SHEYNIN S, ASHUAL O, POLYAK A, et al. Knn-diffusion: Image generation via large-scale retrieval[J]. arXiv preprint arXiv:2204.02849, 2022.

[616] ZHEN R, SONG W, HE Q, et al. Human-computer interaction system: A survey of talking-head generation[J]. Electronics, 2023, 12(1): 218.

[617] CHUNG J S, JAMALUDIN A, ZISSERMAN A. You said that?[J]. arXiv preprint arXiv:1705.02966, 2017.

[618] JAMALUDIN A, CHUNG J S, ZISSERMAN A. You said that?: Synthesising talking faces from audio[J]. International Journal of Computer Vision, 2019, 127: 1767-1779.

[619] SONG Y, ZHU J, LI D, et al. Talking face generation by conditional recurrent adversarial network[J]. arXiv preprint arXiv:1804.04786, 2018.

[620] ZHOU H, LIU Y, LIU Z, et al. Talking face generation by adversarially disentangled audio-visual representation[C]//Proceedings of the AAAI conference on artificial intelligence: volume 33. [S.l.: s.n.], 2019: 9299-9306.

[621] KUMAR R, SOTELO J, KUMAR K, et al. Obamanet: Photo-realistic lip-sync from text[J]. arXiv preprint arXiv:1801.01442, 2017.

[622] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing obama: learning lip sync from audio[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-13.

[623] CUDEIRO D, BOLKART T, LAIDLAW C, et al. Capture, learning, and synthesis of 3d speaking styles[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2019: 10101-10111.

[624] FRIED O, TEWARI A, ZOLLHÖFER M, et al. Text-based editing of talking-head video[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-14.

[625] KARRAS T, AILA T, LAINE S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-12.

[626] THIES J, ELGHARIB M, TEWARI A, et al. Neural voice puppetry: Audio-driven facial reenactment[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. [S.l.]: Springer, 2020: 716-731.

[627] LI T, BOLKART T, BLACK M J, et al. Learning a model of facial shape and expression from 4d scans.[J]. ACM Trans. Graph., 2017, 36(6): 194-1.

[628] RICHARD A, ZOLLHÖFER M, WEN Y, et al. Meshtalk: 3d face animation from speech using cross-modality disentanglement[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2021: 1173-1182.

[629] LI R, TANCIK M, KANAZAWA A. Nerfacc: A general nerf acceleration toolbox[J]. arXiv preprint arXiv:2210.04847, 2022.

[630] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[631] MÜLLER T, EVANS A, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics (ToG), 2022, 41(4): 1-15.

[632] KR P, MUKHOPADHYAY R, PHILIP J, et al. Towards automatic face-to-face translation[C]//Proceedings of the 27th ACM international conference on multimedia. [S.l.: s.n.], 2019: 1428-1436.

[633] PRAJWAL K, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM international conference on multimedia. [S.l.: s.n.], 2020: 484-492.

[634] DORETTO G, CHIUSO A, WU Y N, et al. Dynamic textures[J]. International journal of computer vision, 2003, 51: 91-109.

[635] WEI L Y, LEVOY M. Fast texture synthesis using tree-structured vector quantization[C]//Proceedings of the 27th annual conference on Computer graphics and interactive techniques. [S.l.: s.n.], 2000: 479-488.

[636] ACHARYA D, HUANG Z, PAUDEL D P, et al. Towards high resolution video generation with progressive growing of sliced wasserstein gans[J]. arXiv preprint arXiv:1810.02419, 2018.

[637] CLARK A, DONAHUE J, SIMONYAN K. Efficient video generation on complex datasets[J]. arXiv preprint arXiv:1907.06571, 2019, 2(3): 4.

[638] OHNISHI K, YAMAMOTO S, USHIKU Y, et al. Hierarchical video generation from orthogonal information: Optical flow and texture[C]//Proceedings of the AAAI conference on artificial intelligence: volume 32. [S.l.: s.n.], 2018.

[639] SAITO M, MATSUMOTO E, SAITO S. Temporal generative adversarial nets with singular value clipping[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 2830-2839.

[640] TULYAKOV S, LIU M Y, YANG X, et al. Mocogan: Decomposing motion and content for video generation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 1526-1535.

[641] VONDRICK C, PIRSIAVASH H, TORRALBA A. Generating videos with scene dynamics[J]. Advances in neural information processing systems, 2016, 29.

[642] YUSHCHENKO V, ARASLANOV N, ROTH S. Markov decision process for video generation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. [S.l.: s.n.], 2019: 0-0.

[643] CLARK A, DONAHUE J, SIMONYAN K. Adversarial video generation on complex datasets[J]. arXiv preprint arXiv:1907.06571, 2019.

[644] HO J, SALIMANS T, GRITSENKO A, et al. Video diffusion models, 2022[J]. URL https://arxiv. org/abs/2204.03458.

[645] SAITO M, SAITO S, KOYAMA M, et al. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan[J]. International Journal of Computer Vision, 2020, 128(10-11): 2586-2606.

[646] TIAN Y, REN J, CHAI M, et al. A good image generator is what you need for high-resolution video synthesis[J]. arXiv preprint arXiv:2104.15069, 2021.

[647] GUPTA T, SCHWENK D, FARHADI A, et al. Imagine this! scripts to compositions to videos[C]//Proceedings of the European conference on computer vision (ECCV). [S.l.: s.n.], 2018: 598-613.

[648] LI Y, MIN M, SHEN D, et al. Video generation from text[C]//Proceedings of the AAAI conference on artificial intelligence: volume 32. [S.l.: s.n.], 2018.

[649] LIU Y, WANG X, YUAN Y, et al. Cross-modal dual learning for sentence-to-video generation[C]//Proceedings of the 27th ACM international conference on multimedia. [S.l.: s.n.], 2019: 1239-1247.

[650] MARWAH T, MITTAL G, BALASUBRAMANIAN V N. Attentive semantic video generation using captions[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 1426-1434.

[651] MITTAL G, MARWAH T, BALASUBRAMANIAN V N. Sync-draw: Automatic video generation using deep recurrent attentive architectures[C]//Proceedings of the 25th ACM international conference on Multimedia. [S.l.: s.n.], 2017: 1096-1104.

[652] PAN Y, QIU Z, YAO T, et al. To create what you tell: Generating videos from captions[C]//Proceedings of the 25th ACM international conference on Multimedia. [S.l.: s.n.], 2017: 1789-1798.
[653] HONG W, DING M, ZHENG W, et al. Cogvideo: Large-scale pretraining for text-to-video generation via transformers[J]. arXiv preprint arXiv:2205.15868, 2022.
[654] WU C, HUANG L, ZHANG Q, et al. Godiva: Generating open-domain videos from natural descriptions[J]. arXiv preprint arXiv:2104.14806, 2021.
[655] SINGER U, POLYAK A, HAYES T, et al. Make-a-video: Text-to-video generation without text-video data[J]. arXiv preprint arXiv:2209.14792, 2022.
[656] HO J, CHAN W, SAHARIA C, et al. Imagen video: High definition video generation with diffusion models[J]. arXiv preprint arXiv:2210.02303, 2022.
[657] WU J Z, GE Y, WANG X, et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation[J]. arXiv preprint arXiv:2212.11565, 2022.
[658] WU Z, SONG S, KHOSLA A, et al. 3d shapenets: A deep representation for volumetric shapes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2015: 1912-1920.
[659] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 652-660.
[660] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems, 2017, 30.
[661] HANOCKA R, HERTZ A, FISH N, et al. Meshcnn: a network with an edge[J]. ACM Transactions on Graphics (ToG), 2019, 38(4): 1-12.
[662] MESCHEDER L, OECHSLE M, NIEMEYER M, et al. Occupancy networks: Learning 3d reconstruction in function space[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 4460-4470.
[663] FU R, ZHAN X, CHEN Y, et al. Shapecrafter: A recursive text-conditioned 3d shape generation model[J]. Advances in Neural Information Processing Systems, 2022, 35: 8882-8895.
[664] JAHAN T, GUAN Y, VAN KAICK O. Semantics-guided latent space exploration for shape generation[C]//Computer Graphics Forum: volume 40. [S.l.]: Wiley Online Library, 2021: 115-126.
[665] LIU Z, WANG Y, QI X, et al. Towards implicit text-guided 3d shape generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 17896-17906.
[666] POOLE B, JAIN A, BARRON J T, et al. Dreamfusion: Text-to-3d using 2d diffusion[J]. arXiv preprint arXiv:2209.14988, 2022.
[667] BEDNARIK J, FUA P, SALZMANN M. Learning to reconstruct texture-less deformable surfaces from a single view[C]//2018 international conference on 3d vision (3DV). [S.l.]: IEEE, 2018: 606-615.
[668] GOLYANIK V, SHIMADA S, VARANASI K, et al. Hdm-net: Monocular non-rigid 3d reconstruction with learned deformation model[C]//Virtual Reality and Augmented Reality: 15th EuroVR International Conference, EuroVR 2018, London, UK, October 22–23, 2018, Proceedings 15. [S.l.]: Springer, 2018: 51-72.
[669] LI X, KUANG P. 3d-vrvt: 3d voxel reconstruction from a single image with vision transformer[C]//2021 International Conference on Culture-oriented Science & Technology (ICCST). [S.l.]: IEEE, 2021: 343-348.
[670] TSOLI A, ARGYROS A, et al. Patch-based reconstruction of a textureless deformable 3d surface from a single rgb image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. [S.l.: s.n.], 2019: 0-0.
[671] WANG N, ZHANG Y, LI Z, et al. Pixel2mesh: Generating 3d mesh models from single rgb images[C]//Proceedings of the European conference on computer vision (ECCV). [S.l.: s.n.], 2018: 52-67.
[672] YUAN Y, TANG J, ZOU Z. Vanet: a view attention guided network for 3d reconstruction from single and multi-view images[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). [S.l.]: IEEE, 2021: 1-6.

[673] CHOY C B, XU D, GWAK J, et al. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. [S.l.]: Springer, 2016: 628-644.

[674] HUANG P H, MATZEN K, KOPF J, et al. Deepmvs: Learning multi-view stereopsis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018: 2821-2830.

[675] WANG D, CUI X, CHEN X, et al. Multi-view 3d reconstruction with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2021: 5722-5731.

[676] XIE H, YAO H, SUN X, et al. Pix2vox: Context-aware 3d reconstruction from single and multi-view images[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2019: 2690-2698.

[677] WANG J, CUI Y, GUO D, et al. Pointattn: You only need attention for point cloud completion[J]. arXiv preprint arXiv:2203.08485, 2022.

[678] BAI S, BAI X, LIU W, et al. Neural shape codes for 3d model retrieval[J]. Pattern Recognition Letters, 2015, 65: 15-21.

[679] HOWARD D, EIBEN A E, KENNEDY D F, et al. Evolving embodied intelligence from materials to machines[J]. Nature Machine Intelligence, 2019, 1(1): 12-19.

[680] DUAN J, YU S, TAN H L, et al. A survey of embodied ai: From simulators to research tasks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(2): 230-244.

[681] DEITKE M, BATRA D, BISK Y, et al. Retrospectives on the embodied ai workshop[J]. arXiv preprint arXiv:2210.06849, 2022.

[682] DUAN J, YU S, TAN H L, et al. Actionet: An interactive end-to-end platform for task-based data collection and augmentation in 3d environment[C]//2020 IEEE International Conference on Image Processing (ICIP). [S.l.]: IEEE, 2020: 1566-1570.

[683] SHRIDHAR M, THOMASON J, GORDON D, et al. Alfred: A benchmark for interpreting grounded instructions for everyday tasks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 10740-10749.

[684] ANDERSON P, CHANG A, CHAPLOT D S, et al. On evaluation of embodied navigation agents[J]. arXiv preprint arXiv:1807.06757, 2018.

[685] RAMAKRISHNAN S K, JAYARAMAN D, GRAUMAN K. An exploration of embodied visual exploration[J]. International Journal of Computer Vision, 2021, 129: 1616-1649.

[686] YE X, YANG Y. From seeing to moving: A survey on learning for visual indoor navigation (vin)[J]. arXiv preprint arXiv:2002.11310, 2020.

[687] CHEN T, GUPTA S, GUPTA A. Learning exploration policies for navigation[J]. arXiv preprint arXiv:1903.01959, 2019.

[688] SAVINOV N, DOSOVITSKIY A, KOLTUN V. Semi-parametric topological memory for navigation[J]. arXiv preprint arXiv:1803.00653, 2018.

[689] BEECHING E, DIBANGOYE J, SIMONIN O, et al. Learning to plan with uncertain topological maps[C]//European Conference on Computer Vision. [S.l.]: Springer, 2020: 473-490.

[690] BEATTIE C, LEIBO J Z, TEPLYASHIN D, et al. Deepmind lab[J]. arXiv preprint arXiv:1612.03801, 2016.

[691] KOLVE E, MOTTAGHI R, HAN W, et al. Ai2-thor: An interactive 3d environment for visual ai[J]. arXiv preprint arXiv:1712.05474, 2017.

[692] YAN C, MISRA D, BENNNETT A, et al. Chalet: Cornell house agent learning environment[J]. arXiv preprint arXiv:1801.07357, 2018.

[693] PUIG X, RA K, BOBEN M, et al. Virtualhome: Simulating household activities via programs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018: 8494-8502.

[694] GAO X, GONG R, SHU T, et al. Vrkitchen: an interactive 3d virtual environment for task-oriented learning[J]. arXiv preprint arXiv:1903.05757, 2019.

[695] SAVVA M, KADIAN A, MAKSYMETS O, et al. Habitat: A platform for embodied ai research[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2019: 9339-9347.

[696] XIA F, SHEN W B, LI C, et al. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 713-720.

[697] XIANG F, QIN Y, MO K, et al. Sapien: A simulated part-based interactive environment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2020: 11097-11107.

[698] GAN C, SCHWARTZ J, ALTER S, et al. Threedworld: A platform for interactive multi-modal physical simulation[J]. arXiv preprint arXiv:2007.04954, 2020.

[699] BELLEMARE M G, NADDAF Y, VENESS J, et al. The arcade learning environment: An evaluation platform for general agents[J]. Journal of Artificial Intelligence Research, 2013, 47: 253-279.

[700] PFEIFER R, IIDA F. Embodied artificial intelligence: Trends and challenges[J]. Lecture notes in computer science, 2004: 1-26.

[701] PARTSEY R, WIJMANS E, YOKOYAMA N, et al. Is mapping necessary for realistic pointgoal navigation? [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 17232-17241.

[702] TAN J, ZHANG T, COUMANS E, et al. Sim-to-real: Learning agile locomotion for quadruped robots[J]. arXiv preprint arXiv:1804.10332, 2018.

[703] YU W, KUMAR V C, TURK G, et al. Sim-to-real transfer for biped locomotion[C]//2019 ieee/rsj international conference on intelligent robots and systems (iros). [S.l.]: IEEE, 2019: 3503-3510.

[704] FU Z, KUMAR A, AGARWAL A, et al. Coupling vision and proprioception for navigation of legged robots[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 17273-17283.

[705] SHRIDHAR M, MANUELLI L, FOX D. Cliport: What and where pathways for robotic manipulation[C]// Conference on Robot Learning. [S.l.]: PMLR, 2022: 894-906.

[706] CHEN C, JAIN U, SCHISSLER C, et al. Soundspaces: Audio-visual navigation in 3d environments[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. [S.l.]: Springer, 2020: 17-36.

[707] KADIAN A, TRUONG J, GOKASLAN A, et al. Sim2real predictivity: Does evaluation in simulation predict real-world performance?[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6670-6677.

[708] SHEN B, XIA F, LI C, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [S.l.]: IEEE, 2021: 7520-7527.

[709] MO K, ZHU S, CHANG A X, et al. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 909-918.

[710] SONG S, YU F, ZENG A, et al. Semantic scene completion from a single depth image[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 1746-1754.

[711] CHANG A, DAI A, FUNKHOUSER T, et al. Matterport3d: Learning from rgb-d data in indoor environments[C]//2017 International Conference on 3D Vision (3DV). [S.l.]: IEEE Computer Society, 2017: 667-676.

[712] XIA F, ZAMIR A R, HE Z, et al. Gibson env: Real-world perception for embodied agents[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 9068-9079.

[713] KADIAN A, TRUONG J, GOKASLAN A, et al. Are we making real progress in simulated environments? measuring the sim2real gap in embodied visual navigation[J]. 2019.

[714] DEITKE M, HAN W, HERRASTI A, et al. Robothor: An open simulation-to-real embodied ai platform[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 3164-3174.

[715] BATRA D, GOKASLAN A, KEMBHAVI A, et al. Objectnav revisited: On evaluation of embodied agents navigating to objects[J]. arXiv preprint arXiv:2006.13171, 2020.

[716] JAIN U, WEIHS L, KOLVE E, et al. A cordial sync: Going beyond marginal policies for multi-agent embodied tasks[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. [S.l.]: Springer, 2020: 471-490.

[717] JAIN U, WEIHS L, KOLVE E, et al. Two body problem: Collaborative visual task completion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2019: 6689-6699.

[718] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. Openai gym[J]. arXiv preprint arXiv:1606.01540, 2016.

[719] WEIHS L, SALVADOR J, KOTAR K, et al. Allenact: A framework for embodied ai research[J]. arXiv preprint arXiv:2008.12760, 2020.

[720] SMITH L, GASSER M. The development of embodied cognition: Six lessons from babies[J]. Artificial life, 2005, 11(1-2): 13-29.

[721] NGUYEN P D, GEORGIE Y K, KAYHAN E, et al. Sensorimotor representation learning for an "active self" in robots: a model survey[J]. KI-Künstliche Intelligenz, 2021, 35: 9-35.

[722] CHAPLOT D S, GANDHI D, GUPTA S, et al. Learning to explore using active neural slam[C]//International Conference on Learning Representations. [S.l.: s.n.], 2019.

[723] ZHU F, ZHU Y, CHANG X, et al. Vision-language navigation with self-supervised auxiliary reasoning tasks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2020: 10012-10022.

[724] PATHAK D, AGRAWAL P, EFROS A A, et al. Curiosity-driven exploration by self-supervised prediction[C]// International conference on machine learning. [S.l.]: PMLR, 2017: 2778-2787.

[725] GUPTA S, FOUHEY D, LEVINE S, et al. Unifying map and landmark based representations for visual navigation[J]. arXiv preprint arXiv:1712.08125, 2017.

[726] NARASIMHAN M, WIJMANS E, CHEN X, et al. Seeing the un-scene: Learning amodal semantic maps for room navigation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. [S.l.]: Springer, 2020: 513-529.

[727] RAMAKRISHNAN S K, AL-HALAH Z, GRAUMAN K. Occupancy anticipation for efficient exploration and navigation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. [S.l.]: Springer, 2020: 400-418.

[728] GUPTA S, DAVIDSON J, LEVINE S, et al. Cognitive mapping and planning for visual navigation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 2616-2625.

[729] HENRIQUES J F, VEDALDI A. Mapnet: An allocentric spatial memory for mapping environments[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018: 8476-8484.

[730] FANG K, TOSHEV A, FEI-FEI L, et al. Scene memory transformer for embodied agents in long-horizon tasks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 538-547.

[731] GEORGAKIS G, LI Y, KOSECKA J. Simultaneous mapping and target driven navigation[J]. arXiv preprint arXiv:1911.07980, 2019.

[732] MEZGHANI L, SUKHBAATAR S, SZLAM A, et al. Learning to visually navigate in photorealistic environments without any supervision[J]. arXiv preprint arXiv:2004.04954, 2020.

[733] MISHKIN D, DOSOVITSKIY A, KOLTUN V. Benchmarking classic and learned navigation in complex 3d environments[J]. arXiv preprint arXiv:1901.10915, 2019.

[734] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age[J]. IEEE Transactions on robotics, 2016, 32(6): 1309-1332.

[735] RAMAKRISHNAN S K, JAYARAMAN D, GRAUMAN K. Emergence of exploratory look-around behaviors through active observation completion[J]. Science Robotics, 2019, 4(30): eaaw6326.

[736] LOVEJOY W S. A survey of algorithmic methods for partially observed markov decision processes[J]. Annals of Operations Research, 1991, 28(1): 47-65.

[737] YAMAUCHI B. A frontier-based approach for autonomous exploration[C]//Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'. [S.l.]: IEEE, 1997: 146-151.

[738] BURDA Y, EDWARDS H, PATHAK D, et al. Large-scale study of curiosity-driven learning[J]. arXiv preprint arXiv:1808.04355, 2018.

[739] HOUTHOOFT R, CHEN X, DUAN Y, et al. Vime: Variational information maximizing exploration[J]. Advances in neural information processing systems, 2016, 29.

[740] PATHAK D, GANDHI D, GUPTA A. Self-supervised exploration via disagreement[C]//International conference on machine learning. [S.l.]: PMLR, 2019: 5062-5071.

[741] CHAPLOT D S, JIANG H, GUPTA S, et al. Semantic curiosity for active visual learning[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. [S.l.]: Springer, 2020: 309-326.

[742] BURDA Y, EDWARDS H, STORKEY A, et al. Exploration by random network distillation[J]. arXiv preprint arXiv:1810.12894, 2018.

[743] RAMAKRISHNAN S K, GRAUMAN K. Sidekick policy learning for active visual exploration[C]//Proceedings of the European conference on computer vision (ECCV). [S.l.: s.n.], 2018: 413-430.

[744] SONG S, ZENG A, CHANG A X, et al. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2018: 3847-3856.

[745] DU H, YU X, ZHENG L. Learning object relation graph and tentative policy for visual navigation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. [S.l.]: Springer, 2020: 19-34.

[746] YE J, BATRA D, WIJMANS E, et al. Auxiliary tasks speed up learning point goal navigation[C]//Conference on Robot Learning. [S.l.]: PMLR, 2021: 498-516.

[747] WIJMANS E, KADIAN A, MORCOS A, et al. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames[J]. arXiv preprint arXiv:1911.00357, 2019.

[748] DOSOVITSKIY A, KOLTUN V. Learning to act by predicting the future[J]. arXiv preprint arXiv:1611.01779, 2016.

[749] DAYAN P. Improving generalization for temporal difference learning: The successor representation[J]. Neural computation, 1993, 5(4): 613-624.

[750] ZHU Y, GORDON D, KOLVE E, et al. Visual semantic planning using deep successor representations[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 483-492.

[751] BARRETO A, DABNEY W, MUNOS R, et al. Successor features for transfer in reinforcement learning[J]. Advances in neural information processing systems, 2017, 30.

[752] GORDON D, KADIAN A, PARIKH D, et al. Splitnet: Sim2sim and task2task transfer for embodied visual navigation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2019: 1022-1031.

[753] MOUSAVIAN A, TOSHEV A, FIŠER M, et al. Visual representations for semantic target driven navigation[C]//2019 International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2019: 8846-8852.

[754] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation[J]. arXiv preprint arXiv:1506.02438, 2015.

[755] GUO Z D, AZAR M G, PIOT B, et al. Neural predictive belief representations[J]. arXiv preprint arXiv:1811.06407, 2018.

[756] YANG W, WANG X, FARHADI A, et al. Visual semantic navigation using scene priors[J]. arXiv preprint arXiv:1810.06543, 2018.

[757] GAN C, ZHANG Y, WU J, et al. Look, listen, and act: Towards audio-visual embodied navigation[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2020: 9701-9707.

[758] ANDERSON P, WU Q, TENEY D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 3674-3683.

[759] CAMPARI T, ECCHER P, SERAFINI L, et al. Exploiting scene-specific features for object goal navigation[C]// European Conference on Computer Vision. [S.l.]: Springer, 2020: 406-421.

[760] CHAPLOT D S, GANDHI D P, GUPTA A, et al. Object goal navigation using goal-oriented semantic exploration[J]. Advances in Neural Information Processing Systems, 2020, 33: 4247-4258.

[761] THOMASON J, MURRAY M, CAKMAK M, et al. Vision-and-dialog navigation[C]//Conference on Robot Learning. [S.l.]: PMLR, 2020: 394-406.

[762] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8: 229-256.

[763] DAS A, GKIOXARI G, LEE S, et al. Neural modular control for embodied question answering[C]//Conference on Robot Learning. [S.l.]: PMLR, 2018: 53-62.

[764] YU L, CHEN X, GKIOXARI G, et al. Multi-target embodied question answering[C]//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2019: 6309-6318.

[765] GORDON D, KEMBHAVI A, RASTEGARI M, et al. Iqa: Visual question answering in interactive environments[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 4089-4098.

[766] TAN S, XIANG W, LIU H, et al. Multi-agent embodied question answering in interactive environments[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. [S.l.]: Springer, 2020: 663-678.

[767] STRAUB J, WHELAN T, MA L, et al. The replica dataset: A digital replica of indoor spaces[J]. arXiv preprint arXiv:1906.05797, 2019.

[768] BROHAN A, BROWN N, CARBAJAL J, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[J]. arXiv preprint arXiv:2307.15818, 2023.

[769] HUANG W, WANG C, ZHANG R, et al. Voxposer: Composable 3d value maps for robotic manipulation with language models[J]. arXiv preprint arXiv:2307.05973, 2023.

[770] BHARADHWAJ H, VAKIL J, SHARMA M, et al. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking[J]. arXiv preprint arXiv:2309.01918, 2023.

[771] SU D, XU Y, WINATA G I, et al. Generalizing question answering system with pre-trained language model fine-tuning[C]//Proceedings of the 2nd Workshop on Machine Reading for Question Answering. [S.l.: s.n.], 2019: 203-211.

[772] LEWIS M, LIU Y, GOYAL N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.

[773] LI J, TANG T, ZHAO W X, et al. Pretrained language models for text generation: A survey[J]. arXiv preprint arXiv:2201.05273, 2022.

[774] LI Z, WANG C, LIU Z, et al. Cctest: Testing and repairing code completion systems[C]//2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE). [S.l.]: IEEE, 2023: 1238-1250.

[775] MALINKA K, PERESÍNI M, FIRC A, et al. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?[C]//Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1. [S.l.: s.n.], 2023: 47-53.

[776] LIU J, LIU C, LV R, et al. Is chatgpt a good recommender? a preliminary study[J]. arXiv preprint arXiv:2304.10149, 2023.

[777] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[778] BANG Y, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity[J]. arXiv preprint arXiv:2302.04023, 2023.

[779] ZHANG H, SONG H, LI S, et al. A survey of controllable text generation using transformer-based pre-trained language models[J]. ACM Computing Surveys, 2022.

[780] DANILEVSKY M, QIAN K, AHARONOV R, et al. A survey of the state of explainable ai for natural language processing[J]. arXiv preprint arXiv:2010.00711, 2020.

[781] GOLOVNEVA O, CHEN M P, POFF S, et al. Roscoe: A suite of metrics for scoring step-by-step reasoning[C]// The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2022.

[782] WANG J, HU X, HOU W, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective[J]. arXiv preprint arXiv:2302.12095, 2023.

[783] ZHENG L, CHIANG W L, SHENG Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena[J]. arXiv preprint arXiv:2306.05685, 2023.

[784] GIRDHAR R, EL-NOUBY A, LIU Z, et al. Imagebind: One embedding space to bind them all[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 15180-15190.

[785] CHEN S, HOU Y, CUI Y, et al. Recall and learn: Fine-tuning deep pretrained language models with less forgetting[J]. arXiv preprint arXiv:2004.12651, 2020.

[786] LU P, PENG B, CHENG H, et al. Chameleon: Plug-and-play compositional reasoning with large language models[J]. arXiv preprint arXiv:2304.09842, 2023.

[787] ZHANG Z, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models[J]. arXiv preprint arXiv:2302.00923, 2023.

[788] MÜNDLER N, HE J, JENKO S, et al. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation[J]. arXiv preprint arXiv:2305.15852, 2023.

[789] LI Y, DU Y, ZHOU K, et al. Evaluating object hallucination in large vision-language models[J]. arXiv preprint arXiv:2305.10355, 2023.

[790] RANJIT M, GANAPATHY G, MANUEL R, et al. Retrieval augmented chest x-ray report generation using openai gpt models[J]. arXiv preprint arXiv:2305.03660, 2023.

[791] LIU J, JIN J, WANG Z, et al. Reta-llm: A retrieval-augmented large language model toolkit[J]. arXiv preprint arXiv:2306.05212, 2023.

[792] PAN J, LIN Z, GE Y, et al. Retrieving-to-answer: Zero-shot video question answering with frozen large language models[J]. arXiv preprint arXiv:2306.11732, 2023.

[793] CHOI J H, HICKMAN K E, MONAHAN A, et al. Chatgpt goes to law school[J]. Available at SSRN, 2023.

[794] SHEN Y, HEACOCK L, ELIAS J, et al. Chatgpt and other large language models are double-edged swords[J]. Radiology, 2023, 307(2): e230163.

[795] KHALIL M, ER E. Will chatgpt get you caught? rethinking of plagiarism detection[J]. arXiv preprint arXiv:2302.04335, 2023.

[796] BENGIO Y, LECUN Y, HINTON G. Deep learning for ai[J]. Communications of the ACM, 2021, 64(7): 58-65.

[797] WANG J, LAN C, LIU C, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2022.

[798] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[799] ZHANG Y, KANG B, HOOI B, et al. Deep long-tailed learning: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[800] NATARAJAN N, DHILLON I S, RAVIKUMAR P K, et al. Learning with noisy labels[J]. Advances in neural information processing systems, 2013, 26.

[801] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[J]. Advances in neural information processing systems, 2019, 32.

[802] ZHANG Z, LI Y, WANG J, et al. Remos: reducing defect inheritance in transfer learning via relevant model slicing[C]//Proceedings of the 44th International Conference on Software Engineering. [S.l.: s.n.], 2022: 1856-1868.

[803] CHIN T W, ZHANG C, MARCULESCU D. Renofeation: A simple transfer learning method for improved adversarial robustness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2021: 3243-3252.

[804] MAUS N, CHAO P, WONG E, et al. Adversarial prompting for black box foundation models[J]. arXiv preprint arXiv:2302.04237, 2023.

[805] SUN C, SHRIVASTAVA A, SINGH S, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 843-852.

[806] MILLER J P, TAORI R, RAGHUNATHAN A, et al. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2021: 7721-7735.

[807] TENEY D, LIN Y, OH S J, et al. Id and ood performance are sometimes inversely correlated on real-world datasets[J]. arXiv preprint arXiv:2209.00613, 2022.

[808] ZHU K, WANG J, ZHOU J, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts[J]. arXiv preprint arXiv:2306.04528, 2023.

[809] LENAT D, MARCUS G. Getting from generative ai to trustworthy ai: What llms might learn from cyc[J]. arXiv preprint arXiv:2308.04445, 2023.

[810] LIU Y, YAO Y, TON J F, et al. Trustworthy llms: a survey and guideline for evaluating large language models' alignment[J]. arXiv preprint arXiv:2308.05374, 2023.

[811] PEARL J. Causality[M]. [S.l.]: Cambridge university press, 2009.

[812] KICIMAN E, NESS R, SHARMA A, et al. Causal reasoning and large language models: Opening a new frontier for causality[J]. arXiv preprint arXiv:2305.00050, 2023.

[813] COVERT I C, LUNDBERG S, LEE S I. Explaining by removing: A unified framework for model explanation[J]. The Journal of Machine Learning Research, 2021, 22(1): 9477-9566.

[814] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30.

[815] WACHTER S, MITTELSTADT B, RUSSELL C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr[J]. Harv. JL & Tech., 2017, 31: 841.

[816] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]//International conference on machine learning. [S.l.]: PMLR, 2018: 2668-2677.

[817] ADEBAYO J, GILMER J, MUELLY M, et al. Sanity checks for saliency maps[J]. Advances in neural information processing systems, 2018, 31.

[818] TJOA E, GUAN C. A survey on explainable artificial intelligence (xai): Toward medical xai[J]. IEEE transactions on neural networks and learning systems, 2020, 32(11): 4793-4813.

[819] SAEED W, OMLIN C. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities[J]. Knowledge-Based Systems, 2023, 263: 110273.

[820] DOŠILOVIĆ F K, BRČIĆ M, HLUPIĆ N. Explainable artificial intelligence: A survey[C]//2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). [S.l.]: IEEE, 2018: 0210-0215.

[821] MADSEN A, REDDY S, CHANDAR S. Post-hoc interpretability for neural nlp: A survey[J]. ACM Computing Surveys, 2022, 55(8): 1-42.

[822] SARTI G, FELDHUS N, SICKERT L, et al. Inseq: An interpretability toolkit for sequence generation models[J]. arXiv preprint arXiv:2302.13942, 2023.

[823] SOONG D, SRIDHAR S, SI H, et al. Improving accuracy of gpt-3/4 results on biomedical data using a retrieval-augmented language model[J]. arXiv preprint arXiv:2305.17116, 2023.

[824] IZACARD G, LEWIS P, LOMELI M, et al. Few-shot learning with retrieval augmented language models[J]. arXiv preprint arXiv:2208.03299, 2022.

[825] KHANDELWAL U, LEVY O, JURAFSKY D, et al. Generalization through memorization: Nearest neighbor language models[J]. arXiv preprint arXiv:1911.00172, 2019.

[826] GUU K, LEE K, TUNG Z, et al. Retrieval augmented language model pre-training[C]//International conference on machine learning. [S.l.]: PMLR, 2020: 3929-3938.

[827] BILLS S, CAMMARATA N, MOSSING D, et al. Language models can explain neurons in language models[J]. 2023.

[828] ZHANG H, LI L H, MENG T, et al. On the paradox of learning to reason from data[J]. arXiv preprint arXiv:2205.11502, 2022.

[829] TURPIN M, MICHAEL J, PEREZ E, et al. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting[J]. arXiv preprint arXiv:2305.04388, 2023.

[830] FRIEDER S, PINCHETTI L, GRIFFITHS R R, et al. Mathematical capabilities of chatgpt[J]. arXiv preprint arXiv:2301.13867, 2023.

[831] LIU H, NING R, TENG Z, et al. Evaluating the logical reasoning ability of chatgpt and gpt-4[J]. arXiv preprint arXiv:2304.03439, 2023.

[832] SI C, FRIEDMAN D, JOSHI N, et al. Measuring inductive biases of in-context learning with underspecified demonstrations[J]. arXiv preprint arXiv:2305.13299, 2023.

[833] DEL M, FISHEL M. True detective: A deep abductive reasoning benchmark undoable for gpt-3 and challenging for gpt-4[C]//Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023). [S.l.: s.n.], 2023: 314-322.

[834] LI R, ALLAL L B, ZI Y, et al. Starcoder: may the source be with you![J]. arXiv preprint arXiv:2305.06161, 2023.

[835] FU Y, PENG H, OU L, et al. Specializing smaller language models towards multi-step reasoning[J]. arXiv preprint arXiv:2301.12726, 2023.

[836] LI Y, CHOI D, CHUNG J, et al. Competition-level code generation with alphacode[J]. Science, 2022, 378 (6624): 1092-1097.

[837] UESATO J, KUSHMAN N, KUMAR R, et al. Solving math word problems with process-and outcome-based feedback[J]. arXiv preprint arXiv:2211.14275, 2022.

[838] LE H, WANG Y, GOTMARE A D, et al. Coderl: Mastering code generation through pretrained models and deep reinforcement learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 21314-21328.

[839] TU R, MA C, ZHANG C. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis[J]. arXiv preprint arXiv:2301.13819, 2023.
[840] JIN Z, LIU J, LYU Z, et al. Can large language models infer causation from correlation?[J]. arXiv preprint arXiv:2306.05836, 2023.
[841] PEARL J. Probabilities of causation: three counterfactual interpretations and their identification[M]// Probabilistic and Causal Inference: The Works of Judea Pearl. [S.l.: s.n.], 2022: 317-372.
[842] LIU Y, WEI Y S, YAN H, et al. Causal reasoning meets visual representation learning: A prospective study[J]. Machine Intelligence Research, 2022, 19(6): 485-511.
[843] IMBENS G W, RUBIN D B. Causal inference in statistics, social, and biomedical sciences[M]. [S.l.]: Cambridge University Press, 2015.
[844] PEARL J. Causal inference[J]. Causality: objectives and assessment, 2010: 39-58.
[845] PEARL J. Causal inference in statistics: An overview[J]. 2009.
[846] HELLNER J. Causality and causation in law[J]. Scandinavian studies in law, 2000, 40: 111-134.
[847] KNOBE J, SHAPIRO S. Proximate cause explained[J]. The University of Chicago Law Review, 2021, 88(1): 165-236.
[848] PETERS J, JANZING D, SCHÖLKOPF B. Elements of causal inference: foundations and learning algorithms[M]. [S.l.]: The MIT Press, 2017.
[849] HALPERN J Y. Actual causality[M]. [S.l.]: MiT Press, 2016.
[850] HEGARTY M. Mechanical reasoning by mental simulation[J]. Trends in cognitive sciences, 2004, 8(6): 280-285.
[851] JEANNEROD M. Neural simulation of action: a unifying mechanism for motor cognition[J]. Neuroimage, 2001, 14(1): S103-S109.
[852] GLYMOUR C, ZHANG K, SPIRTES P. Review of causal discovery methods based on graphical models[J]. Frontiers in genetics, 2019, 10: 524.
[853] SHARMA A, SYRGKANIS V, ZHANG C, et al. Dowhy: Addressing challenges in expressing and validating causal assumptions[J]. arXiv preprint arXiv:2108.13518, 2021.
[854] LIPSITCH M, TCHETGEN E T, COHEN T. Negative controls: a tool for detecting confounding and bias in observational studies[J]. Epidemiology (Cambridge, Mass.), 2010, 21(3): 383.
[855] SHARMA A, LI H, JIAO J. The counterfactual-shapley value: Attributing change in system metrics[J]. arXiv preprint arXiv:2208.08399, 2022.
[856] GERSTENBERG T, GOODMAN N, LAGNADO D, et al. From counterfactual simulation to causal judgment[C]//Proceedings of the annual meeting of the cognitive science society: volume 36. [S.l.: s.n.], 2014.
[857] SLOMAN S A, LAGNADO D. Causality in thought[J]. Annual review of psychology, 2015, 66: 223-247.
[858] LONG S, SCHUSTER T, PICHÉ A, et al. Can large language models build causal graphs?[J]. arXiv preprint arXiv:2303.05279, 2023.
[859] WILLIG M, ZEČEVIĆ M, DHAMI D S, et al. Can foundation models talk causality?[J]. arXiv preprint arXiv:2206.10591, 2022.
[860] SHIMIZU S, HOYER P O, HYVÄRINEN A, et al. A linear non-gaussian acyclic model for causal discovery.[J]. Journal of Machine Learning Research, 2006, 7(10).
[861] ZHANG K, CHAN L W. Extensions of ica for causality discovery in the hong kong stock market[C]// International Conference on Neural Information Processing. [S.l.]: Springer, 2006: 400-409.
[862] ZHANG K, HYVÄRINEN A. Causality discovery with additive disturbances: An information-theoretical perspective[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20. [S.l.]: Springer, 2009: 570-585.
[863] KAISER M, SIPOS M. Unsuitability of notears for causal graph discovery when dealing with dimensional quantities[J]. Neural Processing Letters, 2022, 54(3): 1587-1595.

[864] HUANG Y, KLEINDESSNER M, MUNISHKIN A, et al. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere[J]. Frontiers in big Data, 2021, 4: 642182.
[865] LIU Y, CHEN W, LI G, et al. Causalvlr: A toolbox and benchmark for visual-linguistic causal reasoning[J]. arXiv preprint arXiv:2306.17462, 2023.
[866] LECUN Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27[J]. Open Review, 2022, 62.
[867] ACHINSTEIN P. The nature of explanation[M]. [S.l.]: Oxford University Press, USA, 1983.
[868] BRYSON A, HO Y C. Applied optimal control. 1969[J]. Blaisdell, Waltham, Mass, 1969, 8(72): 14.
[869] LEVINE S. Understanding the world through action[Z]. [S.l.: s.n.], 2021.
[870] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2016: 1400-1409.
[871] ASSRAN M, DUVAL Q, MISRA I, et al. Self-supervised learning from images with a joint-embedding predictive architecture[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 15619-15629.
[872] CHEN X, DING M, WANG X, et al. Context autoencoder for self-supervised representation learning[J]. International Journal of Computer Vision, 2023: 1-16.
[873] ZHOU J, WEI C, WANG H, et al. Image bert pre-training with online tokenizer[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.
[874] BARDES A, PONCE J, LECUN Y. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features[J]. arXiv preprint arXiv:2307.12698, 2023.
[875] SUN D, YANG X, LIU M Y, et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 8934-8943.
[876] LYNCH C, WAHID A, TOMPSON J, et al. Interactive language: Talking to robots in real time[J]. IEEE Robotics and Automation Letters, 2023.
[877] LIN J, DU Y, WATKINS O, et al. Learning to model the world with language[Z]. [S.l.: s.n.], 2023.
[878] HAFNER D, PASUKONIS J, BA J, et al. Mastering diverse domains through world models[J]. arXiv preprint arXiv:2301.04104, 2023.
[879] HU A, RUSSELL L, YEO H, et al. Gaia-1: A generative world model for autonomous driving[J]. arXiv preprint arXiv:2309.17080, 2023.
[880] YANG M, DU Y, GHASEMIPOUR K, et al. Learning interactive real-world simulators[Z]. [S.l.: s.n.], 2023.
[881] LIU Y, ZHANG K, LI Y, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models[J]. arXiv preprint arXiv:2402.17177, 2024.
[882] GARRIDO Q, ASSRAN M, BALLAS N, et al. Learning and leveraging world models in visual representation learning[J]. arXiv preprint arXiv:2403.00504, 2024.
[883] RICHARDS T B. Auto-gpt: An autonomous gpt-4 experiment[M]. [S.l.]: May, 2023.
[884] LIU X, YU H, ZHANG H, et al. Agentbench: Evaluating llms as agents[Z]. [S.l.: s.n.], 2023.
[885] TEAM X. Xagent: An autonomous agent for complex task solving[Z]. [S.l.: s.n.], 2023.
[886] YANG J, DING R, BROWN E, et al. V-irl: Grounding virtual intelligence in real life[J]. arXiv preprint arXiv:2402.03310, 2024.
[887] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: A survey[J]. arXiv preprint arXiv:2309.07864, 2023.
[888] REED S, ZOLNA K, PARISOTTO E, et al. A generalist agent[J]. arXiv preprint arXiv:2205.06175, 2022.
[889] LIU H, SON K, YANG J, et al. Learning customized visual models with retrieval-augmented knowledge[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 15148-15158.

[890] WANG Z, MAO S, WU W, et al. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration[Z]. [S.l.: s.n.], 2023.

[891] HONG S, ZHENG X, CHEN J, et al. Metagpt: Meta programming for multi-agent collaborative framework[J]. arXiv preprint arXiv:2308.00352, 2023.

[892] LI G, HAMMOUD H A A K, ITANI H, et al. Camel: Communicative agents for" mind" exploration of large scale language model society[J]. arXiv preprint arXiv:2303.17760, 2023.

[893] CHEN L, ZHANG Y, REN S, et al. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond[J]. arXiv e-prints, 2023: arXiv-2310.

[894] XIANG J, TAO T, GU Y, et al. Language models meet world models: Embodied experiences enhance language models[J]. arXiv preprint arXiv:2305.10626, 2023.

[895] SONG C H, WU J, WASHINGTON C, et al. Llm-planner: Few-shot grounded planning for embodied agents with large language models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2023: 2998-3009.

[896] WU Z, WANG Z, XU X, et al. Embodied task planning with large language models[J]. arXiv preprint arXiv:2307.01848, 2023.

[897] RANA K, HAVILAND J, GARG S, et al. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning[J]. arXiv preprint arXiv:2307.06135, 2023.

[898] SCHICK T, DWIVEDI-YU J, DESSÌ R, et al. Toolformer: Language models can teach themselves to use tools[J]. arXiv preprint arXiv:2302.04761, 2023.

[899] SHEN Y, SONG K, TAN X, et al. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface[J]. arXiv preprint arXiv:2303.17580, 2023.

[900] SURÍS D, MENON S, VONDRICK C. Vipergpt: Visual inference via python execution for reasoning[J]. arXiv preprint arXiv:2303.08128, 2023.

[901] GUPTA T, KEMBHAVI A. Visual programming: Compositional visual reasoning without training[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 14953-14962.

[902] WANG G, XIE Y, JIANG Y, et al. Voyager: An open-ended embodied agent with large language models[J]. arXiv preprint arXiv:2305.16291, 2023.

[903] PARK J S, O'BRIEN J C, CAI C J, et al. Generative agents: Interactive simulacra of human behavior[J]. arXiv preprint arXiv:2304.03442, 2023.

[904] MU Y, ZHANG Q, HU M, et al. Embodiedgpt: Vision-language pre-training via embodied chain of thought[J]. arXiv preprint arXiv:2305.15021, 2023.

[905] YANG J, DONG Y, LIU S, et al. Octopus: Embodied vision-language programmer from environmental feedback[J]. arXiv preprint arXiv:2310.08588, 2023.

[906] BELKHALE S, DING T, XIAO T, et al. Rt-h: Action hierarchies using language[J]. arXiv preprint arXiv:2403.01823, 2024.

[907] PADALKAR A, POOLEY A, JAIN A, et al. Open x-embodiment: Robotic learning datasets and rt-x models[J]. arXiv preprint arXiv:2310.08864, 2023.

[908] FU Z, ZHAO T Z, FINN C. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation[J]. arXiv preprint arXiv:2401.02117, 2024.