



UvA-DARE (Digital Academic Repository)

Visual understanding of dynamic scenes using object relationships and open vocabularies

Ülger, O.

Publication date
2026

[Link to publication](#)

Citation for published version (APA):

Ülger, O. (2026). *Visual understanding of dynamic scenes using object relationships and open vocabularies*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Bibliography

- [1] K. Rajaei, Y. Mohsenzadeh, R. Ebrahimpour, and S.-M. Khaligh-Razavi, "Beyond core object recognition: Recurrent processes account for object recognition under occlusion," *PLOS Computational Biology*, vol. 15, no. 5, pp. 1–30, 2019.
- [2] S. Ullman, L. Assif, E. Fetaya, and D. Harari, "Atoms of recognition in human and computer vision," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2744–2749, 2016.
- [3] J. S. Johnson and B. A. Olshausen, "The recognition of partially visible natural objects in the presence and absence of their occluders," *Vision Research*, vol. 45, no. 25, pp. 3262–3276, 2005.
- [4] I. Biederman and M. Bar, "One-shot viewpoint invariance in matching novel objects," *Vision Research*, vol. 39, no. 17, pp. 2885–2899, 1999.
- [5] V. Ayzenberg and S. F. Lourenco, "Skeletal descriptions of shape provide unique perceptual information for object recognition," *Scientific Reports*, vol. 9, no. 1, p. 9359, 2019.
- [6] P. Spröte, F. Schmidt, and R. W. Fleming, "Visual perception of shape altered by inferred causal history," *Scientific Reports*, vol. 6, no. 1, p. 36245, 2016.
- [7] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *ICLR*, 2019.
- [8] B. S. Tjan, W. L. Braje, G. E. Legge, and D. Kersten, "Human efficiency for recognizing 3-d objects in luminance noise," *Vision Research*, vol. 35, no. 21, pp. 3053–3069, 1995.
- [9] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.
- [10] J. E. Hummel and I. Biederman, "Dynamic binding in a neural network for shape recognition," *Psychol Rev*, vol. 99, no. 3, pp. 480–517, 1992.
- [11] M. J. Tarr, P. Williams, W. G. Hayward, and I. Gauthier, "Three-dimensional object recognition is viewpoint dependent," *Nature Neuroscience*, vol. 1, no. 4, pp. 275–277, 1998.
- [12] H. H. Bülthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 1, pp. 60–64, 1992.
- [13] T. Poggio and S. Edelman, "A network that learns to recognize three-dimensional objects," *Nature*, vol. 343, no. 6255, pp. 263–266, 1990.
- [14] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and T. Poggio, "View-dependent object recognition by monkeys," *Current Biology*, vol. 4, no. 5, pp. 401–414, 1994.
- [15] M. J. Tarr and H. H. Bülthoff, "Image-based object recognition in man, monkey and machine," *Cognition*, vol. 67, no. 1, pp. 1–20, 1998.
- [16] J. E. Hummel, "Object recognition," in *The Oxford Handbook of Cognitive Psychology*. Oxford University Press, 2013.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.
- [19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.

- [20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR*, 2021.
- [21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *ECCV*, 2020.
- [22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [23] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *ICML*, 2022.
- [24] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *ECCV*, 2014.
- [25] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *IJCV*, vol. 123, pp. 32–73, 2017.
- [26] B. Li, X. Li, Z. Zhang, and F. Wu, “Spatio-temporal graph routing for skeleton-based action recognition,” in *AAAI*, 2019.
- [27] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *AAAI*, 2018.
- [28] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, “Learning human-object interactions by graph parsing neural networks,” in *ECCV*, 2018.
- [29] P. Zhou and M. Chi, “Relation parsing neural network for human-object interaction detection,” in *ICCV*, 2019.
- [30] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, “Situation recognition with graph neural networks,” in *ICCV*, 2017.
- [31] X. Wang and A. Gupta, “Videos as space-time region graphs,” in *ECCV*, 2018.
- [32] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, “Graph convolutional networks for temporal action localization,” in *ICCV*, 2019.
- [33] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, “Semantic object parsing with graph lstm,” in *ECCV*, 2016.
- [34] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, “3d graph neural networks for rgb-d semantic segmentation,” in *ICCV*, 2017.
- [35] L. Mi and Z. Chen, “Hierarchical graph attention network for visual relationship detection,” in *CVPR*, 2020.
- [36] Z. Wu, M. Wang, J. Wang, W. Zhang, M. Fang, and T. Xu, “Deepword: A gcn-based approach for owner-member relationship detection in autonomous driving,” *ICME*, 2021.
- [37] G. Singh, S. Akrigg, M. D. Maio, V. Fontana, R. J. Alitappeh, S. Saha, K. J. Saravi, F. Yousefi, J. Cully, T. Nicholson, J. Omokeowa, S. Khan, S. Grazioso, A. Bradley, G. D. Gironimo, and F. Cuzzolin, “ROAD: the road event awareness dataset for autonomous driving,” *TPAMI*, 2023.
- [38] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” in *ICLR*, 2018.
- [39] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, “Clevrer: Collision events for video representation and reasoning,” in *ICLR*, 2019.
- [40] J. Ji, R. Krishna, L. Fei-Fei, and J. Niebles, “Action genome: Actions as compositions of spatio-temporal scene graphs,” in *CVPR*, 2020.

- [41] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, 2008.
- [42] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.
- [43] L. Gong and Q. Cheng, "Exploiting edge features for graph neural networks," in *CVPR*, 2019.
- [44] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *CVPR*, 2019.
- [45] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *KDD*, 2019.
- [46] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *NeurIPS*, 2013.
- [47] Y. Hu, S. Chen, X. Chen, Y. Zhang, and X. Gu, "Neural message passing for visual relationship detection," in *ICML workshops*, 2019.
- [48] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.
- [49] M. Zhang and Y. Chen, "Weisfeiler-lehman neural machine for link prediction," in *KDD*, 2017.
- [50] —, "Link prediction based on graph neural networks," in *NeurIPS*, 2018.
- [51] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR*, 2019.
- [52] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *ICML*, 2017.
- [53] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *ICML*, 2019.
- [54] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, 2003.
- [55] S. Kim, S. Nowozin, P. Kohli, and C. Yoo, "Higher-order correlation clustering for image segmentation," in *NeurIPS*, 2011.
- [56] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, 2016.
- [57] R. van den Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," in *KDD*, 2017.
- [58] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *The World Wide Web Conference*, 2019.
- [59] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, 2009.
- [60] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "Kgat: Knowledge graph attention network for recommendation," in *ACM SIGKDD*, 2019.
- [61] Y. Li, Y. Luo, and Z. Huang, "Fashion recommendation with multi-relational representation learning," in *PAKDD*, 2020.
- [62] T. Kipf, E. Fetaya, K. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *ICML*, 2018.
- [63] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *CVPR*, 2019.
- [64] R. Herzig, E. Levi, H. Xu, H. Gao, E. Brosh, X. Wang, A. Globerson, and T. Darrell, "Spatio-temporal action graph networks," in *ICCV workshops*, 2019.

- [65] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, “Multi-range attentive bicomponent graph convolutional network for traffic forecasting,” in *AAAI*, 2020.
- [66] Z. Diao, X. Wang, D. Zhang, Y. Liu, K. Xie, and S. He, “Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting,” in *AAAI*, 2019.
- [67] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional neural network: A deep learning framework for traffic forecasting,” in *IJCAI*, 2018.
- [68] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-rnn: Deep learning on spatio-temporal graphs,” in *CVPR*, 2016.
- [69] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, “Spatial-temporal graph convolutional network for video-based person re-identification,” in *CVPR*, 2020.
- [70] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, and K. Achan, “Inductive representation learning on temporal graphs,” 2020.
- [71] L. Gao, B. Wang, and W. Wang, “Image captioning with scene-graph based semantic concepts,” in *ICMLC*, 2018.
- [72] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” *CVPR*, 2018.
- [73] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, “Dense relational captioning: Triple-stream networks for relationship-based captioning,” in *CVPR*, 2019.
- [74] L. Li, Z. Gan, Y. Cheng, and J. Liu, “Relation-aware graph attention network for visual question answering,” *CVPR*, 2019.
- [75] J. Johnson, A. Gupta, and L. Fei-Fei, “Image generation from scene graphs,” *CVPR*, 2018.
- [76] G. Mittal, S. Agrawal, A. Agarwal, S. Mehta, and T. Marwah, “Interactive image generation using scene graphs,” *ICLR*, 2019.
- [77] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *CVPR*, 2017.
- [78] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *NeurIPS*, 2015.
- [79] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *NeurIPS*, 2019.
- [80] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2014.
- [81] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.
- [82] Z. Lu, W. Lv, Y. Cao, Z. Xie, H. Peng, and B. Du, “Lstm variants meet graph neural networks for road speed prediction,” *Neurocomputing*, 2020.
- [83] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, “Visual relationship detection with language priors,” in *ECCV*, 2016.
- [84] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” *CVPR*, 2018.
- [85] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” *ECCV*, 2018.
- [86] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, “Scene graph generation from objects, phrases and caption regions,” *ICCV*, 2017.
- [87] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” *CVPR*, 2017.
- [88] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, “Graphical contrastive losses for scene graph generation,” *CVPR*, 2019.

- [89] M. F. Bonner and R. A. Epstein, “Object representations in the human brain reflect the co-occurrence statistics of vision and language,” *Nature Communications*, 2021.
- [90] M. Greene, “Statistics of high-level scene context,” *Frontiers in Psychology*, vol. 4, 2013.
- [91] D. Kaiser, G. L. Quek, R. M. Cichy, and M. V. Peelen, “Object vision in a structured world,” *TiCS*, vol. 23, no. 8, pp. 672–685, 2019.
- [92] H. Xu, C. Jiang, X. Liang, L. Lin, and Z. Li, “Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection,” in *CVPR*, 2019.
- [93] C. Jiang, H. Xu, X. Liang, and L. Lin, “Hybrid knowledge routed modules for large-scale object detection,” *NeurIPS*, 2018.
- [94] S. Sharifzadeh, S. M. Baharlou, and V. Tresp, “Classification by attention: Scene graph classification with prior knowledge,” *AAAI*, 2021.
- [95] P. O. Pinheiro, R. Collobert, and P. Dollár, “Learning to segment object candidates,” in *NeurIPS*, 2015.
- [96] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” in *ICCV*, 2019.
- [97] R. Xu, Y. Li, C. Wang, S. Xu, W. Meng, and X. Zhang, “Instance segmentation of biological images using graph convolutional network,” *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104739, 2022.
- [98] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, “Solq: Segmenting objects by learning queries,” in *NeurIPS*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 21 898–21 909.
- [99] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *CVPR*, 2022.
- [100] L. Chen, W. Zhang, Y. Wu, M. Strauch, and D. Merhof, “Semi-supervised instance segmentation with a learned shape prior,” in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, 2020.
- [101] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, “Weakly supervised instance segmentation using the bounding box tightness prior,” in *NeurIPS*, 2019.
- [102] S. Hao, G. Wang, and R. Gu, “Weakly supervised instance segmentation using multi-prior fusion,” *CVIU*, 2021.
- [103] C. d. M. Braz, L. F. D. Santos, and P. A. V. Miranda, “Graph-based image segmentation with shape priors and band constraints,” in *DGMM*, 2022.
- [104] S. Vicente, V. Kolmogorov, and C. Rother, “Graph cut based image segmentation with connectivity priors,” in *CVPR*, 2008.
- [105] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, “Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval,” *GRSS*, 2020.
- [106] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, “Relation distillation networks for video object detection,” *CVPR*, 2019.
- [107] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” *CVPR*, 2018.
- [108] B. Dai, Y. Zhang, and D. Lin, “Detecting visual relationships with deep relational networks,” *CVPR*, 2017.
- [109] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. F. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, Ç. Gülçehre, H. F. Song, A. J. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. R. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, “Relational inductive biases, deep learning, and graph networks,” *arXiv*, 2018.
- [110] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, and H. Hajishirzi, “Text generation from knowledge graphs with graph transformers,” *NAACL*.

- [111] A. Maas, A. Hannun, and A. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *ICML*, 2013.
- [112] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016.
- [113] G. A. Miller, “WordNet: A lexical database for English,” in *Human Language Technology*, 1994.
- [114] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask scoring R-CNN,” *CVPR*, 2019.
- [115] K. Lis, K. K. Nakka, P. Fua, and M. Salzmann, “Detecting the unexpected via image resynthesis,” *ICCV*, 2019.
- [116] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge.” *IJCV*, 2010.
- [117] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ade20k dataset,” in *CVPR*, 2017.
- [118] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using deep learning: A survey,” *TPAMI*, vol. 44, no. 7, pp. 3523–3542, 2021.
- [119] P. Pandey, M. Chasmai, M. Natarajan, and B. Lall, “A language-guided benchmark for weakly supervised open vocabulary semantic segmentation,” *arXiv*, 2023.
- [120] C. Ma, Y. Yang, Y. Wang, Y. Zhang, and W. Xie, “Open-vocabulary semantic segmentation with frozen vision-language models,” in *BMVC*, 2022.
- [121] B. Li, K. Q. Weinberger, S. J. Belongie, V. Koltun, and R. Ranftl, “Language-driven semantic segmentation,” *ICLR*, 2022.
- [122] G. Ghiasi, X. Gu, Y. Cui, and T. Lin, “Scaling open-vocabulary image segmentation with image-level labels,” *ECCV*, 2022.
- [123] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu, “Maskclip: Masked self-distillation advances contrastive language-image pretraining,” *CVPR*, 2023.
- [124] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, “Open-vocabulary semantic segmentation with mask-adapted clip,” *CVPR*, 2023.
- [125] S. Cho, H. Shin, S. Hong, S. An, S. Lee, A. Arnab, P. H. Seo, and S. Kim, “Cat-seg: Cost aggregation for open-vocabulary semantic segmentation,” *CVPR*, 2024.
- [126] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, “Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip,” in *NeurIPS*, 2023.
- [127] X. Chen, S. Li, S.-N. Lim, A. Torralba, and H. Zhao, “Open-vocabulary panoptic segmentation with embedding modulation,” in *ICCV*, 2023.
- [128] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, “Side adapter network for open-vocabulary semantic segmentation,” in *CVPR*, 2023.
- [129] M. Yi, Q. Cui, H. Wu, C. Yang, O. Yoshie, and H. Lu, “A simple framework for text-supervised semantic segmentation,” in *CVPR*, 2023.
- [130] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, “Segment anything,” in *ICCV*, 2023.
- [131] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, “Lisa: Reasoning segmentation via large language model,” *CVPR*, 2024.
- [132] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, N. Peng, L. Wang, Y. J. Lee, and J. Gao, “Generalized decoding for pixel, image, and language,” *CVPR*, 2023.
- [133] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, “Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models,” *ICCV*, 2023.

- [134] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, “Zero-shot semantic segmentation,” in *NeurIPS*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., 2019.
- [135] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, “Semantic projection network for zero- and few-label semantic segmentation,” in *CVPR*, 2019.
- [136] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, , and X. Bai, “A simple baseline for open vocabulary semantic segmentation with pre-trained vision-language model,” *ECCV*, 2022.
- [137] S. Ren, A. Zhang, Y. Zhu, S. Zhang, S. Zheng, M. Li, A. Smola, and X. Sun, “Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition,” in *NeurIPS*, 2023.
- [138] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan, “Going denser with open-vocabulary part segmentation,” in *ICCV*, 2023.
- [139] L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, “Diffusion Models for Zero-Shot Open-Vocabulary Segmentation,” *ECCV*, 2024.
- [140] J. Xu, S. D. Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, “Groupvit: Semantic segmentation emerges from text supervision,” *CVPR*, 2022.
- [141] Q. Liu, Y. Wen, J. Han, C. Xu, H. Xu, and X. Liang, “Open-world semantic segmentation via contrasting and clustering vision-language embedding,” in *ECCV*, 2022.
- [142] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie, “Learning open-vocabulary semantic segmentation models from natural language supervision,” in *CVPR*, 2023.
- [143] P. Rewatbowornwong, N. Chatthee, E. Chuangsuwanich, and S. Suwajanakorn, “Zero-guidance segmentation using zero segment labels,” in *ICCV*, 2023.
- [144] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *ICCV*, 2021.
- [145] R. Mokady, A. Hertz, and A. H. Bermano, “Clipcap: Clip prefix for image captioning,” *arXiv*, 2021.
- [146] J. Cho, S. Yoon, A. Kale, F. Démoncourt, T. Bui, and M. Bansal, “Fine-grained image captioning with clip reward,” *NAACL*, 2022.
- [147] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *NeurIPS*, 2023.
- [148] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, “Grounding of textual phrases in images by reconstruction,” in *ECCV*, 2016.
- [149] T. Shaharabany, Y. Tewel, and L. Wolf, “What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs,” *NeurIPS*, 2022.
- [150] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281–297.
- [151] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [152] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” *NeurIPS*, 2011.
- [153] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *EMNLP*, 2019.
- [154] A. Yadav, A. Patel, and M. Shah, “A comprehensive review on resolving ambiguities in natural language processing,” *AI Open*, 2021.
- [155] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” 2023.

- [156] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “The role of context for object detection and semantic segmentation in the wild,” in *CVPR*, 2014.
- [157] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *CVPR*, 2016.
- [158] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020, <https://github.com/explosion/spaCy>.
- [159] A. Tao, K. Sapra, and B. Catanzaro, “Hierarchical multi-scale attention for semantic segmentation,” *arXiv*, 2020.
- [160] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, “Inverseform: A loss function for structured boundary-aware segmentation,” in *CVPR*, 2021.
- [161] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, “Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers,” in *ECCV*, 2022.
- [162] Y. Yan, Y. Mao, and B. Li, “Second: Sparsely embedded convolutional detection,” *Sensors*, vol. 18, no. 10, p. 3337, 2018.
- [163] T. Yin, X. Zhou, and P. Krähenbühl, “Center-based 3d object detection and tracking,” in *CVPR*, 2021.
- [164] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, “PointPillars: Fast encoders for object detection from point clouds,” in *CVPR*, 2019.
- [165] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception for autonomous driving: Waymo open dataset,” in *CVPR*, 2020.
- [166] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, H. Xu, and C. Xu, “One million scenes for autonomous driving: Once dataset,” in *NeurIPS*, 2021.
- [167] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in *CVPR*, 2020.
- [168] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences,” in *ICCV*, 2019.
- [169] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *ICML*, 2021.
- [170] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, “Videoclip: Contrastive pre-training for zero-shot video-text understanding,” in *EMNLP*, 2021.
- [171] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang, B. Ghanem, and D. Tao, “Towards open vocabulary learning: A survey,” *TPAMI*, 2024.
- [172] X. Zhou, M. Liu, B. L. Zagar, E. Yurtsever, and A. C. Knoll, “Vision language models in autonomous driving and intelligent transportation systems,” *arXiv*, 2023.
- [173] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017.
- [174] D. Rozenberszki, O. Litany, and A. Dai, “Language-grounded indoor 3d semantic segmentation in the wild,” in *ECCV*, 2022.
- [175] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” in *CVPR*, 2023.
- [176] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, “Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding,” in *CVPR*, 2023.

- [177] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, and W. Wang, “Clip2scene: Towards label-efficient 3d scene understanding by clip,” in *CVPR*, 2023.
- [178] Y. Zeng, C. Jiang, J. Mao, J. Han, C. Ye, Q. Huang, D.-Y. Yeung, Z. Yang, X. Liang, and H. Xu, “Clip²: Contrastive language-image-point pretraining from real-world point cloud data,” in *CVPR*, 2023.
- [179] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, and D. Anguelov, “Unsupervised 3d perception with 2d vision-language distillation for autonomous driving,” in *ICCV*, 2023.
- [180] O. Ülger, M. Kulicki, Y. Asano, and M. R. Oswald, “Auto-vocabulary semantic segmentation,” 2025.
- [181] A. Conti, E. Fini, M. Mancini, P. Rota, Y. Wang, and E. Ricci, “Vocabulary-free image classification and semantic segmentation,” *arXiv*, 2024.
- [182] G. Mei, L. Riz, Y. Wang, and F. Poesi, “Vocabulary-free 3d instance segmentation with vision and language assistant,” in *3DV*, 2025.
- [183] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Babytalk: Understanding and generating simple image descriptions,” *TPAMI*, 2013.
- [184] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, “I2t: Image parsing to text description,” *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.
- [185] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” *CVPR*, 2015.
- [186] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *ICML*, 2015.
- [187] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *CVPR*, 2016.
- [188] L. Xue, M. Shu, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S. Ryoo, S. Kendre, J. Zhang, C. Qin, S. Zhang, C.-C. Chen, N. Yu, J. Tan, T. M. Awalgaonkar, S. Heinecke, H. Wang, Y. Choi, L. Schmidt, Z. Chen, S. Savarese, J. C. Niebles, C. Xiong, and R. Xu, “xgen-mm (blip-3): A family of open large multimodal models,” *arXiv*, 2024.
- [189] L. Xue, N. Yu, S. Zhang, J. Li, R. Martin-Martin, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, “Ulip-2: Towards scalable multimodal pre-training for 3d understanding,” in *CVPR*, 2024.
- [190] T. Luo, C. Rockwell, H. Lee, and J. Johnson, “Scalable 3d captioning with pretrained models,” in *NeurIPS*, 2023.
- [191] Z. Guo, R. Zhang, X. Zhu, Y. Tang, X. Ma, J. Han, K. Chen, P. Gao, X. Li, H. Li, and P.-A. Heng, “Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following,” *arXiv*, 2023.
- [192] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, “Pointllm: Empowering large language models to understand point clouds,” in *ECCV*, 2024.
- [193] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, “3d-vista: Pre-trained transformer for 3d vision and text alignment,” in *ICCV*, 2023.
- [194] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, “3d-llm: Injecting the 3d world into large language models,” in *NeurIPS*, 2023.
- [195] D. Z. Chen, R. Hu, X. Chen, M. Nießner, and A. X. Chang, “Unit3d: A unified transformer for 3d dense captioning and visual grounding,” in *ICCV*, 2023.
- [196] Y. Chen, S. Yang, H. Huang, T. Wang, R. Lyu, R. Xu, D. Lin, and J. Pang, “Grounded 3d-llm with referent tokens,” *arXiv*, 2024.
- [197] Z. Wang, H. Huang, Y. Zhao, Z. Zhang, and Z. Zhao, “Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes,” *arxiv preprint arxiv:2308.08769*, 2023.
- [198] H. Huang, Z. Wang, R. Huang, L. Liu, X. Cheng, Y. Zhao, T. Jin, and Z. Zhao, “Chat-3d v2: Bridging 3d scene and large language models with object identifiers,” *NeurIPS*, 2023.

- [199] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, “L13da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning,” in *CVPR*, 2024.
- [200] G. Hess, A. Tonderski, C. Petersson, K. Åström, and L. Svensson, “Lidarclip or: How i learned to talk to point clouds,” in *WACV*, 2024.
- [201] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu *et al.*, “Recognize anything: A strong image tagging model,” *CVPR*, 2024.
- [202] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, “Point transformer v3: Simpler, faster, stronger,” in *CVPR*, 2024.
- [203] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.
- [204] P. Mishra and K. Sarawadekar, “Polynomial learning rate policy with warm restart for deep neural network,” in *TENCON IEEE Region 10 Conference (TENCON)*, 2019, pp. 2087–2092.
- [205] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, A. Maalouf, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. B. Tenenbaum, C. M. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, “Conceptfusion: Open-set multimodal 3d mapping,” in *RSS*, 2023.
- [206] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, “Openmask3d: Open-vocabulary 3d instance segmentation,” in *NeurIPS*, 2023.
- [207] X. Kang, L. Chu, J. Li, X. Chen, and Y. Lu, “Hierarchical intra-modal correlation learning for label-free 3d semantic segmentation,” in *CVPR*, 2024.
- [208] P. Zou, S. Zhao, W. Huang, Q. Xia, C. Wen, W. Li, and C. Wang, “Adaco: Overcoming visual foundation model noise in 3d semantic segmentation via adaptive label correction,” in *AAAI*, 2025.
- [209] R. Chen, Y. Liu, L. Kong, N. Chen, X. Zhu, Y. Ma, T. Liu, and W. Wang, “Towards label-free scene understanding by vision foundation models,” in *NeurIPS*, 2023.
- [210] X. Zhu, H. Zhou, P. Xing, L. Zhao, H. Xu, J. Liang, A. Hauptmann, T. Liu, and A. Gallagher, “Open-vocabulary 3d semantic segmentation with text-to-image diffusion models,” in *ECCV*, 2024.
- [211] H. Wang, C. Yan, S. Wang, X. Jiang, X. Tang, Y. Hu, W. Xie, and E. Gavves, “Towards open-vocabulary video instance segmentation,” in *ICCV*, 2023.
- [212] P. Guo, H. Huang, P. He, X. Liu, T. Xiao, and W. Zhang, “Openvis: Open-vocabulary video instance segmentation,” in *AAAI*, 2025.
- [213] A. Athar, J. Luiten, P. Voigtlaender, T. Khurana, A. Dave, B. Leibe, and D. Ramanan, “Burst: A benchmark for unifying object recognition, segmentation and tracking in video,” in *WACV*, 2023.
- [214] L. Yang, Y. Fan, and N. Xu, “Video instance segmentation,” in *ICCV*, 2019.
- [215] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” in *ICLR*, 2022.
- [216] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, “Detecting twenty-thousand classes using image-level supervision,” in *ECCV*, 2022.
- [217] M. Piccardi, “Background subtraction techniques: a review,” in *IEEE international conference on systems, man and cybernetics*, vol. 4, 2004, pp. 3099–3104.
- [218] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, “Siamese neural networks for one-shot image recognition,” in *ICML Deep Learning Workshop*, 2015.
- [219] W. G. C. Bandara and V. M. Patel, “A transformer-based siamese network for change detection,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 207–210.
- [220] U. Mall, B. Hariharan, and K. Bala, “Change-aware sampling and contrastive learning for satellite images,” in *CVPR*, 2023.

- [221] J. Kim and U. Kim, “Towards generalizable scene change detection,” in *CVPR*, 2025.
- [222] A. Dave, T. Khurana, P. Tokmakov, C. Schmid, and D. Ramanan, “Tao: A large-scale benchmark for tracking any object,” in *ECCV*, 2020.
- [223] Y. Liu, I. E. Zulfikar, J. Luiten, A. Dave, D. Ramanan, B. Leibe, A. Ošep, and L. Leal-Taixé, “Opening up open-world tracking,” in *CVPR*, 2022.
- [224] S. Li, T. Fischer, L. Ke, H. Ding, M. Danelljan, and F. Yu, “Ovtrack: Open-vocabulary multiple object tracking,” in *CVPR*, 2023.
- [225] G. Heigold, M. Minderer, A. Gritsenko, A. Bewley, D. Keysers, M. Lučić, F. Yu, and T. Kipf, “Video owl-vit: Temporally-consistent open-world localization in video,” in *CVPR*, 2023.
- [226] Z. Yang and Y. Yang, “Decoupling features in hierarchical propagation for video object segmentation,” in *NeurIPS*, 2022.
- [227] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, “Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features,” *arXiv*, 2025.
- [228] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, 2018.
- [229] A. Gupta, P. Dollar, and R. Girshick, “LVIS: A dataset for large vocabulary instance segmentation,” in *CVPR*, 2019.
- [230] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, A. Benhaim, M. Bilenko, J. Bjorck, S. Bubeck, M. Cai, Q. Cai, V. Chaudhary, D. Chen, D. Chen, W. Chen, Y.-C. Chen, Y.-L. Chen, H. Cheng, P. Chopra, X. Dai, M. Dixon, R. Eldan, V. Fragoso, J. Gao, M. Gao, M. Gao, A. Garg, A. D. Giorno, A. Goswami, S. Gunasekar, E. Haider, J. Hao, R. J. Hewett, W. Hu, J. Huynh, D. Iter, S. A. Jacobs, M. Javaheripi, X. Jin, N. Karampatziakis, P. Kauffmann, M. Khademi, D. Kim, Y. J. Kim, L. Kurilenko, J. R. Lee, Y. T. Lee, Y. Li, Y. Li, C. Liang, L. Liden, X. Lin, Z. Lin, C. Liu, L. Liu, M. Liu, W. Liu, X. Liu, C. Luo, P. Madan, A. Mahmoudzadeh, D. Majercak, M. Mazzola, C. C. T. Mendes, A. Mitra, H. Modi, A. Nguyen, B. Norick, B. Patra, D. Perez-Becker, T. Portet, R. Pryzant, H. Qin, M. Radmilac, L. Ren, G. de Rosa, C. Rosset, S. Roy, O. Ruwase, O. Saarikivi, A. Saied, A. Salim, M. Santacroce, S. Shah, N. Shang, H. Sharma, Y. Shen, S. Shukla, X. Song, M. Tanaka, A. Tupini, P. Vaddamanu, C. Wang, G. Wang, L. Wang, S. Wang, X. Wang, Y. Wang, R. Ward, W. Wen, P. Witte, H. Wu, X. Wu, M. Wyatt, B. Xiao, C. Xu, J. Xu, W. Xu, J. Xue, S. Yadav, F. Yang, J. Yang, Y. Yang, Z. Yang, D. Yu, L. Yuan, C. Zhang, C. Zhang, J. Zhang, L. L. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, and X. Zhou, “Phi-3 technical report: A highly capable language model locally on your phone,” 2024.
- [231] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *ECCV*, 2024.
- [232] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” 2024.
- [233] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *ICIP*, 2016.
- [234] H. K. Cheng and A. G. Schwing, “XMem: Long-term video object segmentation with an atkinson-shiffrin memory model,” in *ECCV*, 2022.
- [235] L. T. at Meta, “Llama 3: Advancing open-source large language models,” *arXiv*, 2024.
- [236] OpenAI, “Gpt-4o system card,” *arXiv*, 2024.

Samenvatting

Het doel van dit proefschrift was om visueel begrip van dynamische scènes te bestuderen vanuit twee invalshoeken: de structuur en extra informatie die wordt geboden door relaties tussen objecten, evenals het gebruik van “open vocabulaires”, oftewel modellen die bij het maken van voorspellingen niet beperkt zijn tot een vaste lijst object-categorieën. Samen streven deze twee lijnen van onderzoek ernaar om machinale leermodellen dichter te brengen bij de manier waarop mensen redeneren in veranderlijke omgevingen, waarin zowel de verzameling objecten als hun betekenissen in de loop van de tijd kunnen veranderen.

In **hoofdstuk twee** ontwikkelden we MTD-GNN om relaties te voorspellen in grafen waarbij de verzameling entiteiten verandert terwijl een video zich ontvouwt. In plaats van de verzameling knopen statisch te houden en ontbrekende objecten met lege knopen te modelleren, benut dit model verbindingen in ruimte en tijd afzonderlijk waardoor deze dynamisch kan omgaan met data. Tegelijkertijd generaliseert het model ook effectief naar andere relatie-vormen, doordat het ook multi-label relatievoorspelling per objectpaar ondersteunt. Experimenten op de publieke datasets CLEVRER en Action Genome lieten zien dat onze aanpak de integriteit van de informatie-stroom respecteert, zelfs wanneer objecten verschijnen, verdwijnen of op nieuwe manieren met elkaar interacteren. Een praktische les uit dit hoofdstuk is dat relatie-voorspelling gevoelig blijft voor gemiste of foutieve detecties uit de gebruikte *backbone*. Een veelbelovende volgende stap is om detectie en relatiemodellering nauwer te koppelen, zodat relationeel bewijs kan helpen om aanvankelijk zwakke detecties te herstellen.

Hoofdstuk drie verlegde de focus van het voorspellen van relaties naar het benutten ervan om detectie en instance-segmentatie te verbeteren. Ons model, RP-FEM, gebruikt relationele *priors* (voorkennis) uit een kennisbank om de verbindingen in een scènegraaf, opgebouwd uit region proposals als knopen, aan te passen. Het vergelijkt verbindingen uit de kennisbank met verbindingen in de scènegraaf en werkt op basis daarvan iteratief de verbindingen en knoopkenmerken bij. Op die manier worden kenmerken die passen bij een waarschijnlijke context versterkt en worden kenmerken die onwaarschijnlijk samen voorkomen onderdrukt. Deze aanpak verminderde dubbele voorspellingen, gaf betrouwbaardere voorspellingen met betrekking tot het aantal instanties en zorgde voor veel prestatiewinst bij zeldzame objectcategorieën. De aanpak blijft echter

afhankelijk van de dekking en kwaliteit van de externe kennisbank. Toekomstig werk kan die afhankelijkheid verkleinen door priors op grotere schaal te leren uit tekst en video, en door het model toe te staan priors ter discussie te stellen of te verfijnen wanneer de scène sterk tegenbewijs biedt.

Hoofdstukken vier en vijf behandelden een andere vorm van dynamiek: de rijkdom aan verschillende objectcategorieën in de echte wereld, aangeduid met de term *open vocabularies*. **Hoofdstuk vier** introduceerde AutoSeg, dat een scène-specifieke vocabulaire opbouwt door *vision-language features* - representaties die visuele en tekstuele informatie combineren - te groeperen en beschrijvingen op objectniveau te genereren, om vervolgens deze beschrijvingen te gebruiken voor semantische segmentatie. Omdat deze opzet voorspellingen van objectcategorieën oplevert die zelden één-op-één overeenkomen met publiekelijk beschikbare annotaties, en vergelijking met de state-of-the-art daardoor lastig is, introduceerde dit hoofdstuk ook LAVE. LAVE is een mappingprotocol dat voorspelde namen verbindt met door mensen gelabelde namen in publieke datasets om standaardmetrics te kunnen berekenen. De methode behaalde sterke resultaten in de zero-labelsetting, terwijl zij competitief bleef met methoden die zich baseren op een vooraf, door een mens opgegeven vocabulaire. De belangrijkste beperking is de afhankelijkheid van beeldkwaliteit en lichtomstandigheden, die het vermogen van het model om semantische clusters te vormen, essentieel voor het genereren van goede beschrijvingen, kan verzwakken.

Hoofdstuk vijf was erop gericht de beperking uit hoofdstuk vier te verhelpen door gebruik te maken van geometrische informatie uit LiDAR-data (puntenwolken). Onze oplossing, 3D-AVS, draagt kennis over van een 2D vision-languagemodel naar een 3D-backbone en gebruikt een *Sparse Masked Attention Pooling*-module om een vocabulaire te genereren uit specifieke puntenwolken, zelfs wanneer het 2D-beeld niet beschikbaar is. Dit werk introduceerde de *Text-Point-Semantic-Similarity*-maatstaf om op puntniveau te beoordelen hoe goed representaties van puntenwolken aansluiten bij automatisch gegenereerde namen. De resultaten lieten zien dat onze segmentaties in sommige gevallen semantisch preciezer kunnen zijn dan de labels in de door mensen geannoteerde dataset. Dit hoofdstuk onderstreept het voordeel van het combineren van verschillende modaliteiten, waarbij punten stabiele geometrie aanleveren wanneer enkel 2D data onbetrouwbaar is en taal bruikbare labels biedt wanneer 3D-annotaties schaars zijn.

Ten slotte behandelde **hoofdstuk zes** het thema van voorspellen zonder vaste lijst aan categorieën in een online videostroom-setting. Ons model, FOVIS, voert instance-segmentatie uit op binnenkomende frames zonder vooraf gedefinieerde klasselabels. Het doet dit door voor elke regio in het beeld, eerst aangeduid met een masker dat alleen aangeeft *waar* een object zit en nog niet *wat* het is, een conditionele visuele representatie te berekenen. Om de berekeningen efficiënt te houden over lange video's, gebruikt het systeem VLChange om inhoudelijke veranderingen in

de scène te detecteren en alleen dan nieuwe segmentaties uit te voeren, bijvoorbeeld wanneer er nieuwe objecten of andere objectklassen verschijnen. Een extra stap zorgt ervoor dat het model zich expliciet richt op objectregio's in plaats van de achtergrond, zodat deze objectmaskers effectief kunnen worden benut. Het raamwerk behaalt state-of-the-artprestaties op video's met veel eerder onbekende categorieën en laat daarmee het potentieel zien van modellering voor een wereld waarin steeds nieuwe typen objecten kunnen opduiken.

Er zijn twee belangrijke conclusies die over de zes hoofdstukken heen naar voren komen. Ten eerste: structuur doet ertoe. Door relaties tussen objecten te leren en te gebruiken, krijgen AI-systemen *handvatten* om informatie te delen tussen objecten, onduidelijkheid weg te nemen en met minder computatie toch effectief te werken. Ten tweede: het is belangrijk om modellen te ontwikkelen die met een wereld vol onverwachte en nieuwe objecten kunnen omgaan. Automatisch gegenereerde vocabulaires of klasselabels stellen deze systemen in staat objecten te benoemen op basis van wat zij daadwerkelijk waarnemen, in plaats van alleen op wat mensen vooraf hebben vastgelegd. De evaluatie blijft een uitdaging, juist doordat de voorspellingen niet netjes binnen een vaste lijst categorieën passen, maar technieken zoals LAVE en de TPSS-maat helpen om eerlijkere vergelijkingen mogelijk te maken.

Acknowledgements

After almost ten years, my long journey at the University of Amsterdam comes to an end (for now) with this thesis. In that time, I have completed a bachelor's, master's and PhD in Artificial Intelligence. I have always been proud to be a "homegrown" PhD candidate, as it was within this excellent education system that I developed my *love for research*, which made the step into a PhD feel natural. That step, however, would not have been possible without the trust of my promoter, Theo Gevers.

Theo, I remember walking past your office to ask you to be an examiner for my master's thesis, with a dry throat and trembling hands. After reading my thesis, you told me about your newly started Atlas Lab in collaboration with TomTom and invited me to apply. It was special to me that this was coming from the person who made me enthusiastic about computer vision in the first place. Thank you for involving me in the project and for your guidance throughout the entire journey. I appreciated that you stayed calm when I was not, and that when I was *too calm* you woke me up. You consistently sensed what the situation required and offered advice that helped both me and my research move forward. Even when I was stuck with my work like a rusty bolt, or when we disagreed on how to approach something, you remained reflective, supportive and solution-oriented. Your kindness and openness towards colleagues are truly inspiring. You take us seriously and expect us to do our best, yet you act as a mentor rather than a boss. You invest in relationships, and that can be felt. I have learned a great deal from you, not only about writing papers for anonymous reviewers, but also about speaking to real people I may meet in life. Whether we are proofreading my papers, stranded together in Los Angeles at night or shooting paintballs at each other, you always show the same mix of calmness, humour and dedication. That combination of taking the work seriously while still enjoying it with the people around you is something I hope to carry forward in my own career.

The next person I would like to thank is my co-promoter and daily supervisor, Martin Oswald. Martin, you have played a crucial role in the success of my PhD. Your perseverance, creativity and genuine commitment have had a huge impact on my papers, but even more so on me as a professional. We worked through many nights together to submit something we were proud of, and then talked about the latest memes during lunch the very next day. From the moment you

moved to Amsterdam, you opened your home to us and treated us to some of the best cheese and chocolate I have ever had. I associate you with many warm memories and sincerely hope we stay in touch. Thank you for teaching me to believe in myself and in my ideas, and to give them time to reach a wider audience. When others doubted, you doubled down on your confidence in me, and I am deeply grateful for that.

My other co-promoter, Sezer Karaoglu, also deserves special thanks. Sezer, I remember being so impressed by your lecture in Computer Vision 1 that I immediately called my parents to tell them about you. Becoming your colleague was a true honour, but the greatest honor was to be crowned the next BBQ-master by you. Thank you for not only trusting me with the spatula, but also with my ideas. Whenever I was stuck in my research, fifteen minutes with you would clear my head and help me see the way forward. Your ability to reason critically yet respectfully from a high-level perspective, and to simplify a problem without losing its essence, helped me many times.

My gratitude also extends to the committee members, Prof. Dr. Vasileios Belagiannis, Prof. Dr. Ir. Peter de With, Prof. Dr. Ir. Arnold Smeulders, Dr. Dimitris Tzionas and Dr. Efstratios Gavves, for dedicating their valuable time to serve in the defence committee. Likewise, I would like to sincerely thank Prof. Dr. Christof Monz for chairing the defense.

The past four years were both rewarding and fun thanks to my many colleagues from the Atlas Lab and the Computer Vision Group. Weijie, you were a joy to work with and even nicer to hang out with. We share many good memories, but my favourite is still taking a dive in the sea after losing our way in the sand dunes on Gran Canaria. I am happy we published a paper together, a solid symbol of our synergy as colleagues and friends. I am especially grateful that you shared so much of your culture with me. Aritra, thank you for bringing so much energy into the lab with your silly jokes and softly singing along to reggaeton while having your headphones on (we could hear you at all times). I really appreciated that we started our PhDs side by side, which was very welcome in the COVID years. Kien, you are a brilliant researcher with the most relaxed attitude. I cannot count the number of times you answered “it is okay” to a situation, always calm and in control, which helped me not to overthink. Thank you as well for stepping in to help with my projects from time to time. Vlad, from the day you joined our lab, the atmosphere became noticeably more social. Thank you for teasing Aritra every other day so we could all have a laugh. Yue, you joined our lab last, so we did not spend that much time working together, but I have vivid memories of exploring Milan with you, and I am grateful for that. I also want to thank Cees and Virginie for being great colleagues, ready to help whenever issues arose.

My appreciation extends to my other colleagues from the Computer Vision Group, with whom I was lucky to share brainstorming sessions, insights from bi-weekly presentations, conversations

over drinks, barbecues, football matches and more. I felt at home in the group from the first day. Thank you, Leo, Arnoud, Qi Bi, Partha, Xiaoyan, Melis, Dimitris, Dimitrije, George, Qi Zhang, Ronny, Ruihong, Jincheng, Li, Shaodi, Rick, Wei and Yahui, for the many good moments. I am also grateful to my non-CV Group colleagues Mina, Mohammadreza and Dennis for the many valuable brainstorms and fun chats throughout the years.

I would like to dedicate a few words of appreciation to four collaborators with whom I had the honour to work closely: Pascal, Julian, Yuki and Maksymilian. Pascal and Julian, thank you for investing your time and encouraging me to publish the second chapter of this thesis, which grew out of the master's thesis project we did together. What I learned from you formed a strong foundation for the years that followed. Yuki, we approached you to discuss our ideas on Auto-Vocabulary Segmentation during a period when you were very busy. You still made time for us, and the main contributions of Chapter 4 emerged from that meeting. Thank you for sharing your perspective on this project, which ended up shaping a large part of my thesis. Finally, Maksymilian, thank you for your perseverance in bringing the Auto-Vocabulary Segmentation project to a successful conclusion together. Thanks to our hard work, sometimes deep into the night, we not only finished the project but also got to witness a volcanic eruption and swim with turtles before presenting it at the conference in Hawaii. I will cherish our memories forever.

The research in this thesis would not have been possible without the financial support of TomTom. I would like to thank the organisation for investing in local research and providing us with the tools we needed to excel in our projects. Even more importantly, I am grateful to the colleagues with whom I have had the opportunity to work. Mohsen, Yu, Ysbrand, Fatemeh and Georgi, thank you for being such helpful and enjoyable colleagues. I also want to extend my thanks to the colleagues at TomTom with whom I did not directly collaborate, but with whom I shared many lunches and coffee breaks: Rinat, Maxence, Nils, Mick, Vlad, Shirish, Julia, Arent, Attila and Deboleena. Even if you were not involved in my projects, it was always nice chatting with you on or off-topic.

My sincere gratitude also goes out to my many friends who have been part of this journey in one way or another. The creative and sports-related activities we did together often acted as a catalyst, helping me clear my head, which was often the key to progress in my research. I especially want to thank my closest friends Daniël, Kevin, Ruth, Yunus and Betal for always showing interest, offering support, helping out when needed and celebrating milestones with me. A special thanks to my talented friend Guanyan for helping me with the cover of my thesis, you translated my ideas into an amazing design.

Lovely Lucia, thank you for always being so supportive. You entered my life when I was in the *grind* phase of my PhD, and you understood that context very well. You gave me space when

I needed it, and a listening ear and a helping hand when it was important. I have made you look at segmentation outputs of dachshund dogs far too often and yet you still smiled through the pain. Thank you for being part of this.

There is no better way to close this thesis than with words of gratitude for my family. Thank you, mom and dad, for always standing by my side. Your unconditional support is so strong and special that nothing in the world could ever break me, because I have you two in my life. You raised me to be considerate of others, creative and never to be afraid of anything. Even when I chose paths that went against the grain, you always believed in me and supported me with the words “*biz senin arkadayız*” (we are behind you). You sensed it immediately when I was going through difficult times, and you dropped everything you were doing to get me back up. You are the best parents in the world, thank you for everything.

I also happen to have the best younger sister in the world. Elif, I am smiling as I write this. You are the sunshine in my head. Part of me wishes you had stayed the small, giggly girl you once were, but you grew into a powerful woman who truly inspires me. Your strongest trait is your determination. You are a courageous go-getter, and I have learned a lot from you while watching you grow into an adult. Thank you as well to the rest of my family, in particular Erdal, Türkan, Tarkan, Meltem, Gülser, Harry, Melisa, Emir and Hamdi, for being part of my journey. Whether it was with supportive words, advice, a basketball or checkers game to clear my head or Turkish food and tea between deadlines, all these gestures meant a lot to me.

I want to end with a deep and heartfelt thanks to the people I dedicate this thesis to: my grandparents Osman, Asiye, Ömer and Döndü. You had the courage and determination to move thousands of kilometres away from home to provide a better future for your families, without knowing the language or the culture. You worked hard all your lives to give us everything we needed to succeed. I am proud to have completed this thesis, because it symbolises what you have been working so hard for. It fills me with joy that I wrote the final parts of the thesis in both your homes in Türkiye, while being accompanied by your presence. Thank you for being the best grandparents I could wish for in my life.