

**MODIFIKASI ARSITEKTUR *REAL-TIME DETECTION*
TRANSFORMER (RT-DETR) UNTUK *INSTANCE*
SEGMENTATION YANG EFISIEN**

SKRIPSI

Diajukan untuk memenuhi bagian dari syarat memperoleh gelar Sarjana Komputer
Program Studi Ilmu Komputer



Oleh:

Dwiki Fajar Kurniawan

1903761

**PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN
ALAM UNIVERSITAS PENDIDIKAN INDONESIA**

2025

**MODIFIKASI ARSITEKTUR *REAL-TIME DETECTION*
TRANSFORMER (RT-DETR) UNTUK *INSTANCE*
SEGMENTATION YANG EFISIEN**

Oleh
Dwiki Fajar Kurniawan
NIM 1903761

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Komputer
pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

© Dwiki Fajar Kurniawan
Universitas Pendidikan Indonesia
Agustus 2025

Hak cipta dilindungi Undang-Undang
Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian, dengan
dicetak ulang, difotokopi, atau cara lainnya tanpa izin dari penulis.

DWIKI FAJAR KURNIAWAN

MODIFIKASI ARSITEKTUR *REAL-TIME DETECTION TRANSFORMER* (RT-DETR) UNTUK *INSTANCE SEGMENTATION* YANG EFISIEN

Disetujui dan disahkan oleh pembimbing:

Pembimbing I,



Dr. Muhamad Nursalman, M.T.

NIP. 197909292006041002

Pembimbing II,



Yaya Wihardi, S.Kom., M.Kom.

NIP 198903252015041001

Mengetahui,

Ketua Program Studi Ilmu Komputer



Dr. Muhamad Nursalman, M.T.

NIP. 197909292006041002

KATA PENGANTAR

Puji dan syukur kehadiran Allah Subhanahu Wa Ta'ala, yang atas rahmat, berkah, dan karunia-Nya telah memberikan kelancaran dan kemudahan, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Modifikasi Arsitektur *Real-Time Detection Transformer* (RT-DETR) untuk *Instance Segmentation* yang Efisien” dengan baik. Penyusunan skripsi ini diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Ilmu Komputer, Departemen Pendidikan Ilmu Komputer, Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam, Universitas Pendidikan Indonesia.

Dalam penyusunan skripsi ini, penulis telah berusaha semaksimal mungkin untuk menyajikan hasil penelitian yang komprehensif. Namun, penulis menyadari sepenuhnya bahwa skripsi ini masih memiliki banyak kekurangan dan keterbatasan. Oleh karena itu, penulis dengan tulus menerima segala bentuk kritik dan saran yang membangun demi perbaikan di masa yang akan datang. Penulis berharap skripsi ini dapat memberikan kontribusi positif dan manfaat, baik bagi perkembangan ilmu pengetahuan di bidang *computer vision* maupun sebagai referensi bagi penelitian selanjutnya.

Bandung, Agustus 2025

Penulis,



Dwiki Fajar Kurniawan

UCAPAN TERIMA KASIH

Alhamdulillah Rabbil 'Alamin, segala puji dan syukur penulis panjatkan kehadirat Allah Subhanahu Wa Ta'ala, yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini. Selama proses penyusunan skripsi, penulis mendapatkan banyak bimbingan, dukungan, dan bantuan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Muhamad Nursalman, M.T. selaku Ketua Program Studi Ilmu Komputer Universitas Pendidikan Indonesia yang telah memberikan dukungan dan arahan pada penulis.
2. Bapak Yaya Wihardi, S.Kom., M.Kom. selaku Koordinator Sidang Skripsi Program Studi Ilmu Komputer yang telah memberikan pengarahan kepada penulis.
3. Ibu Dr. Rani Megasari, M.T selaku Dosen Pembimbing Akademik yang telah membimbing penulis selama perkuliahan.
4. Bapak Dr. Muhamad Nursalman, M.T. selaku pembimbing I yang telah memberikan penulis pengarahan dan pengetahuan selama proses penulisan skripsi
5. Bapak Yaya Wihardi, S.Kom., M.Kom. selaku pembimbing II atas segala waktu yang diberikan untuk memberikan penulis bimbingan, pengarahan serta pengetahuan selama proses penulisan skripsi.
6. Bapak dan Ibu Dosen Program Studi Ilmu Komputer yang telah membimbing dan memberikan ilmu yang bermanfaat semasa kuliah pada penulis.
7. Keluarga tercinta, Bapak Syamsu Kunaefi dan Ibu Sri Pujiastuti, serta Mba Noverina Rahmانيanti yang senantiasa memberikan doa, dukungan, dan kepercayaan sehingga skripsi ini dapat terselesaikan.
8. Rekan-rekan Mahasiswa Program Studi Ilmu Komputer angkatan 2019, khususnya kelas C1, atas segala kebersamaan, kenangan, dan bantuan yang telah diberikan selama masa perkuliahan.
9. Semua pihak yang telah membantu dalam kelancaran penelitian dan penyusunan skripsi ini yang tidak dapat penulis sebutkan satu per satu.

ABSTRAK

Implementasi *instance segmentation* secara *real-time* menghadapi tantangan fundamental berupa *trade-off* antara akurasi dan kecepatan. Model deteksi objek yang cepat seperti *Real-Time Detection Transformer* (RT-DETR) belum memiliki kemampuan segmentasi, sementara model segmentasi yang akurat seperti *Mask DINO* tidak efisien untuk aplikasi *real-time*. Penelitian ini bertujuan untuk memodifikasi arsitektur *RT-DETR* agar mampu melakukan *instance segmentation* dengan tetap mempertahankan efisiensi inferensinya. Metode yang diusulkan adalah dengan mengintegrasikan tiga komponen kunci dari *Mask DINO*, yaitu mask branch untuk prediksi level piksel, mekanisme hybrid matching, dan unified denoising training ke dalam arsitektur *RT-DETR*. Model hasil modifikasi yang dinamakan *Insta-RT-DETR* ini dilatih dan dievaluasi menggunakan dataset COCO 2017 dengan backbone *ResNet-50*. Hasil eksperimen menunjukkan bahwa *Insta-RT-DETR* berhasil mencapai performa yang sangat kompetitif dengan 42.5% *Mask AP* dan 50.2% *Box AP*. Keunggulan utamanya terletak pada efisiensi, di mana model ini mampu beroperasi pada kecepatan 35.6 FPS (GPU NVIDIA A100), secara signifikan melampaui kecepatan *Mask DINO* (14.2 FPS) dan akurasi model *real-time* lainnya. Kesimpulannya, *Insta-RT-DETR* berhasil mencapai *trade-off* yang optimal antara akurasi dan kecepatan, menjadikannya solusi yang unggul dan seimbang untuk kebutuhan *instance segmentation* secara *real-time*.

Keywords: *Deep Learning, Instance Segmentation, Mask DINO, Object Detection, Real-Time, RT-DETR, Transformer*

ABSTRACT

The implementation of real-time instance segmentation faces a fundamental trade-off between accuracy and speed. Fast object detection models like the Real-Time Detection Transformer (RT-DETR) lack segmentation capabilities, while accurate segmentation models such as Mask DINO are inefficient for real-time applications. This research aims to modify the RT-DETR architecture to enable it to perform instance segmentation while maintaining its inference efficiency. The proposed method involves integrating three key components from Mask DINO, a mask branch for pixel-level prediction, a hybrid matching mechanism, and unified denoising training into the RT-DETR architecture. The resulting model, named Insta-RT-DETR, was trained and evaluated on the COCO 2017 dataset using a ResNet-50 backbone. Experimental results demonstrate that Insta-RT-DETR achieves highly competitive performance with 42.5% Mask AP and 50.2% Box AP. Its primary advantage lies in efficiency, operating at a speed of 35.6 FPS (on an NVIDIA A100 GPU), significantly surpassing the speed of Mask DINO (14.2 FPS) and the accuracy of other real-time models. In conclusion, Insta-RT-DETR successfully achieves an optimal trade-off between accuracy and speed, establishing it as a superior and well-balanced solution for real-time instance segmentation needs.

Keywords: *Deep Learning, Instance Segmentation, Mask DINO, Object Detection, Real-Time, RT-DETR, Transformer*

DAFTAR ISI

KATA PENGANTAR.....	iii
UCAPAN TERIMA KASIH	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI	vii
DAFTAR GAMBAR	x
DAFTAR TABEL	xii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan Penelitian.....	4
1.4 Manfaat Penelitian.....	4
1.5 Batasan Masalah.....	4
1.6 Sistematika Penulisan.....	5
BAB II KAJIAN PUSTAKA	7
2.1 <i>Deep Learning</i>	7
2.2 Objek Deteksi	9
2.2.1 <i>Bounding Box</i>	11
2.2.2 <i>Non-Maximum Suppression (NMS)</i>	13
2.3 Segmentasi Citra.....	15
2.3.1 <i>Semantic Segmentation</i>	17
2.3.2 <i>Instance Segmentation</i>	19
2.3.3 <i>Panoptic Segmentation</i>	20
2.3.4 <i>Loss Function</i> untuk Segmentasi.....	22

2.4	Transformer	23
2.4.1	<i>Detection Transformer</i> (DETR)	27
2.4.2	<i>Real-Time Detection Transformer</i> (RT-DETR)	28
2.4.3	Mask DINO	31
2.4.4	FastInst	34
2.5	Metrik Evaluasi Model	35
2.6	Dataset	37
2.7	Penelitian Sebelumnya	39
BAB III METODE PENELITIAN		41
3.1.	Desain Penelitian	41
3.1.1.	Perumusan Masalah	42
3.1.2.	Studi Literatur	42
3.1.3.	Pengumpulan Data	42
3.1.4.	Perancangan dan Modifikasi Arsitektur Model	42
3.1.5.	Eksperimen	43
3.1.6.	Analisis Hasil dan Pembahasan	43
3.1.7.	Penarikan Kesimpulan	44
3.2.	Alat dan Bahan Penelitian	44
3.2.1.	Alat Penelitian	44
3.2.2.	Bahan Penelitian	45
3.3.	Prosedur Penelitian	46
3.3.1.	<i>Pre-Processing</i> dan <i>Load Data</i>	46
3.3.2.	Skenario Pelatihan Model	46
3.3.3.	Skenario Evaluasi Model	49
BAB IV HASIL DAN PEMBAHASAN PENELITIAN		51

4.1.	Hasil Penelitian.....	51
4.1.1.	Arsitektur Model yang Diusulkan (Insta-RT-DETR)	51
4.1.2.	Hasil Eksperimen Training Model	54
4.1.3.	Hasil Evaluasi Performa Model	58
4.1.4.	Hasil Analisis Trade-off AP vs FPS.....	62
4.2.	Pembahasan Penelitian	63
4.2.1.	Analisis Arsitektur Model	63
4.2.2.	Analisis Performa Model.....	65
4.2.3.	Analisis <i>Trade-off</i> antara Akurasi (AP) dan Kecepatan (FPS).....	66
4.2.4.	Implikasi Hasil Penelitian	68
BAB V KESIMPULAN DAN SARAN		69
5.1.	Kesimpulan.....	69
5.2.	Saran.....	70
DAFTAR PUSTAKA.....		72

DAFTAR GAMBAR

Gambar 2.1 <i>Multilayer Neural Network</i> (Lecun et al., 2015)	8
Gambar 2.2 <i>Backpropagation</i> (Lecun et al., 2015)	9
Gambar 2.3 Contoh objek deteksi	10
Gambar 2.4 Contoh <i>bounding box</i> pada anotasi dataset MS COCO.....	11
Gambar 2.5 Perbandingan pendekatan anchor-based dan anchor-free	13
Gambar 2.6 Ilustrasi proses NMS pada objek deteksi.....	14
Gambar 2.7 Contoh segmentasi gambar (Minaee et al., 2020)	16
Gambar 2.8 Segmentasi gambar semantik (Ghosh et al., 2019)	17
Gambar 2.9 <i>Instance Segmentation</i> (Minaee et al., 2020)	20
Gambar 2.10 <i>Timeline</i> evolusi segmentasi gambar (Elharrouss et al., 2021)	21
Gambar 2.11 Dice Coefficient.....	23
Gambar 2.12 Arsitektur model <i>original Transformer</i> (Vaswani et al., 2017)	25
Gambar 2.13 Arsitektur <i>Detection Transformer</i> (DETR) (Carion et al., 2020)....	28
Gambar 2.14 Arsitektur Model Real-Time Detection Transformer (RT-DETR) .	29
Gambar 2.15 Struktur <i>encoder</i> untuk setiap varian.....	30
Gambar 2.16 Hasil Studi Ablasi dari <i>Hybrid Encoder</i>	30
Gambar 2.17 Arsitektur Model <i>Mask DINO</i>	32
Gambar 2.18 Arsitektur FastInst	34
Gambar 2.19 Contoh dataset PASCAL VOC	38
Gambar 2.20 Contoh dataset COCO 2017	39
Gambar 3.1 Desain Penelitian	41
Gambar 4.1 Diagram Arsitektur Insta-RT-DETR yang Diusulkan (Full).....	52
Gambar 4.2 Diagram Arsitektur <i>Insta-RT-DETR</i> yang Diusulkan (<i>Feature Extraction - Query Selection</i>).....	52

Gambar 4.3 Diagram Arsitektur <i>Insta-RT-DETR</i> yang Diusulkan (<i>Query Selection - Criterion</i>)	53
Gambar 4.4 Grafik Rata-rata Loss per Epoch <i>Training</i>	55
Gambar 4.5 Training Log yang menunjukkan <i>skip update weight</i>	56
Gambar 4.6 <i>Training</i> Log yang menunjukkan penerapan <i>Penalty Loss</i>	57
Gambar 4.7 Grafik evaluasi <i>Box AP (Average Precision)</i> saat <i>Training</i>	57
Gambar 4.8 Grafik evaluasi <i>Mask AP (Average Precision)</i> saat <i>Training</i>	58
Gambar 4.9 Hasil kualitatif prediksi instance segmentation pada 6 citra subset COCO val2017 (score threshold = 0.5).....	61
Gambar 4.10 Grafik Perbandingan <i>Trade-off</i> Akurasi vs. Kecepatan pada Model Umum.....	62
Gambar 4.11 Grafik Perbandingan Trade-off Akurasi vs. Kecepatan pada Model <i>Real-time</i>	63

DAFTAR TABEL

Tabel 3.1 Konfigurasi <i>Hyperparameter</i> yang digunakan pada <i>training</i>	47
Tabel 4.1 Perbandingan <i>Average Precision</i> (AP) Model	59
Tabel 4.2 Perbandingan Performa Model dengan GFLOPS dan FPS.....	59
Tabel 4.3 Perbandingan Performa Model dengan dengan Model-model <i>Real-time Instance Segmentation</i>	60

DAFTAR PUSTAKA

- Ahmed, S., Hasan, Md. B., Ahmed, T., Sony, R. K., & Kabir, Md. H. (2021). Less is More: Lighter and Faster Deep Neural Architecture for Tomato Leaf Disease Classification. <https://doi.org/10.1109/ACCESS.2022.3187203>
- Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). Soft-NMS--improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision* (pp. 5561-5569).
- Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9157-9166).
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). *End-to-End Object Detection with Transformers*. <http://arxiv.org/abs/2005.12872>
- Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., ... & Lin, D. (2019). Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4974-4983).
- Chen, L., Wu, W., Fu, C., Han, X., & Zhang, Y. (n.d.). *Weakly Supervised Semantic Segmentation with Boundary Exploration*.
- Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1290-1299).
- Cheng, T., Wang, X., Chen, S., Zhang, W., Zhang, Q., Huang, C., ... & Liu, W. (2022). Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4433-4442).
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

- Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., & Himeur, Y. (n.d.). *Panoptic Segmentation: A Review*. <https://github.com/elharroussomar/Awesome-Panoptic-Segmentation>
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2), 303-338.
- Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., ... & Liu, W. (2021). Instances as queries. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6910-6919).
- Gao, B.-B., Chen, X., Huang, Z., Nie, C., Liu, J., Lai, J., Jiang, G., Wang, X., & Wang, C. (2025). Decoupling Classifier for Boosting Few-shot Object Detection and Instance Segmentation. <https://arxiv.org/abs/2505.14239>
- Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T. Y., Cubuk, E. D., ... & Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2918-2928).
- Ghosh, S., Das, N., Das, I., & Maulik, U. (2019). Understanding deep learning techniques for image segmentation. *ACM Computing Surveys*, 52(4). <https://doi.org/10.1145/3329784>
- Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).
- Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2), 87–93. <https://doi.org/10.1007/s13735-017-0141-z>
- Gupta, A., Dollar, P., & Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5356-5364).
- Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., & Molchanov, P. (2023, July). Global context vision transformers. In International Conference on Machine Learning (pp. 12633-12646). PMLR.

- Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2020). *A Survey on Visual Transformer*. <https://doi.org/10.1109/TPAMI.2022.3152247>
- Hafiz, A. M., & Bhat, G. M. (2020). A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3), 171-189.
- He, J., Li, P., Geng, Y., & Xie, X. (2023). Fastinst: A simple query-based model for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23663-23672).
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- Hosang, J., Benenson, R., & Schiele, B. (2017). Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4507-4515).
- Jadon, S. (2020, October). A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)* (pp. 1-7). IEEE.
- Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., & Qu, R. (2019). A survey of deep learning-based object detection. *IEEE access*, 7, 128837-128868.
- Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s), 1-41.
- Law, H., & Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 734-750).
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. <https://doi.org/10.1038/nature14539>
- Li, X., Ding, H., Zhang, W., Yuan, H., Pang, J., Cheng, G., Chen, K., Liu, Z., & Loy, C. C. (2023). *Transformer-Based Visual Segmentation: A Survey*. <http://arxiv.org/abs/2304.09854>
- Li, X., Wu, J., Zhao, H., Lin, Z., & Wang, K. (2022). Mask DINO: Towards A Unified Transformer-Based Framework for Object Detection and Segmentation. <https://arxiv.org/abs/2206.02777>

- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2), 261-318.
- LV, W., Zhao, Y., Xu, S., Wei, J., Wang, G., Cui, C., Du, Y., Dang, Q., & Liu, Y. (2023). *DETRs Beat YOLOs on Real-time Object Detection*. <http://arxiv.org/abs/2304.08069>
- Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* (pp. 565-571). Ieee.
- Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). *Image Segmentation Using Deep Learning: A Survey*. <http://arxiv.org/abs/2001.05566>
- Neubeck, A., & Van Gool, L. (2006, August). Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)* (Vol. 3, pp. 850-855). IEEE.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28, 2015
- Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658-666).
- Roberts, D. A., Yaida, S., & Hanin, B. (2021). The principles of deep learning theory. arXiv.
- Singh, V., & Misra, A. K. (2017). Detection of plant leaf diseases using image segmentation and soft computing techniques. *Information Processing in Agriculture*, 4(1), 41–49. <https://doi.org/https://doi.org/10.1016/j.inpa.2016.10.005>
- Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627-9636).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. <http://arxiv.org/abs/1706.03762>

- Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). Ieee.
- Wang, J., Zhang, T., & Cheng, Y. (2021). Deep Learning for Object Detection: A Survey. *Computer Systems Science & Engineering*, 38(2).
- Wu, H., Xu, Y., Zhang, J., Wu, W., Hu, X., & Lv, Y. (2022). A Review of Object Detection Based on Transformer. *Multimedia Tools and Applications*, 81, 22325–22355.
- Yang, Y., Jiao, L., Liu, X., Liu, F., Yang, S., Feng, Z., & Tang, X. (2022). *Transformers Meet Visual Learning Understanding: A Comprehensive Review*. <http://arxiv.org/abs/2203.12944>
- Yu, L., Tang, L., & Mu, L. (2025). A Review of DETection TRansformer: From Basic Architecture to Advanced Developments and Visual Perception Applications. *Sensors*, 25(13), 3952. <https://doi.org/10.3390/s25133952>
- Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., ... & Shum, H. Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.
- Zhao, X., Wang, L., Zhang, Y. *et al.* A review of convolutional neural networks in computer vision. *Artif Intell Rev* 57, 99 (2024). <https://doi.org/10.1007/s10462-024-10721-6>
- Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020, April). Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 12993-13000).
- Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.
- Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3), 257-276.