Gradus Vol 13, No 2 (2026)
ISSN 2064-8014
Article No. 2026.2.CSC.003

Az Eloszláson Kívüli (OOD) Detektálás Módszertani Alapjai és Taxonómiája a Percepciós Modellekben

Methodological Foundations and Taxonomy of Out-of-Distribution (OOD) Detection in Perception Models

Szabó Loránt[1], Dr. Weltsch Zoltán[2], Ádámné Dr. Major Andrea[3]*

[1]GAMF Műszaki és Informatikai Kar, Neumann János Egyetem, Kecskemét, Magyarország, https://orcid.org/0009-0003-9587-4981
[2]Széchenyi István Egyetem, Győr, Magyarország, https://orcid.org/0000-0002-6366-8281
[3]GAMF Műszaki és Informatikai Kar, Neumann János Egyetem, Kecskemét, Magyarország, https://orcid.org/0009-0009-8546-718X

https://doi.org/10.47833/2026.2.CSC.003

Kulcsszavak:
OOD detektálás
Percepciós modellek
Módszertani taxonómia
Anomáliadetektálás
Mélytanulás

Keywords:
OOD Detection
Perception Models
Methodological Taxonomy
Anomaly Detection
Deep Learning

Cikktörténet:
Beérkezett 2026. április 11.
Átdolgozva 2026. augusztus 27.
Elfogadva 2026. szeptember 22.

Összefoglalás

A biztonságkritikus mesterséges intelligencia rendszerek megbízhatóságának előfeltétele az eloszláson kívüli (Out-of-Distribution, OOD) minták helyes kezelése. Jelen cikk célzott narratív módszertani áttekintés és fogalmi taxonómia a percepciós modellek OOD detektálásáról; nem szisztematikus irodalmi áttekintés. A tanításmentes, jellemző-alapú és integrált tanítási módszerek célzott áttekintésén túl részletesen elemezzük a terület legérettebb megközelítéseit – a SAFE érzékenység-tudatos jellemzőket és a NAPTRON neuronaktivációs mintákat – amelyek a kétlépéses detektorokon élvonalbeli eredményeket értek el. Az architektúrális evolúció vizsgálatán keresztül azonosítjuk a kutatási rést: a jellemző-alapú módszerek modern, egylépéses és horgonymentes YOLO-architektúrákon való validálásának korlátozottságát.

Abstract

Reliable handling of Out-of-Distribution (OOD) samples is a prerequisite for safety-critical artificial intelligence systems. This paper is a targeted narrative methodological review and conceptual taxonomy of OOD detection in perception models; it is not a systematic literature review. Beyond a targeted review of post-hoc, feature-based, and integrated training methods, we analyze in detail the most mature approaches – SAFE sensitivity-aware features and NAPTRON neuron activation patterns – that have achieved state-of-the-art results on two-stage detectors. Through examination of architectural evolution, we identify the research gap: the limited validation of feature-based methods on modern one-stage, anchor-free YOLO architectures.

1. Bevezetés

A számítógépes látás (Computer Vision, CV) az elmúlt években paradigmaváltáson ment keresztül, amelyet a konvolúciós neurális hálók (CNN-ek) és a transzformátor-alapú architektúrák dominanciája vezérelt. E modellek teljesítménye számos feladaton – mint az ImageNet képfelismerés vagy

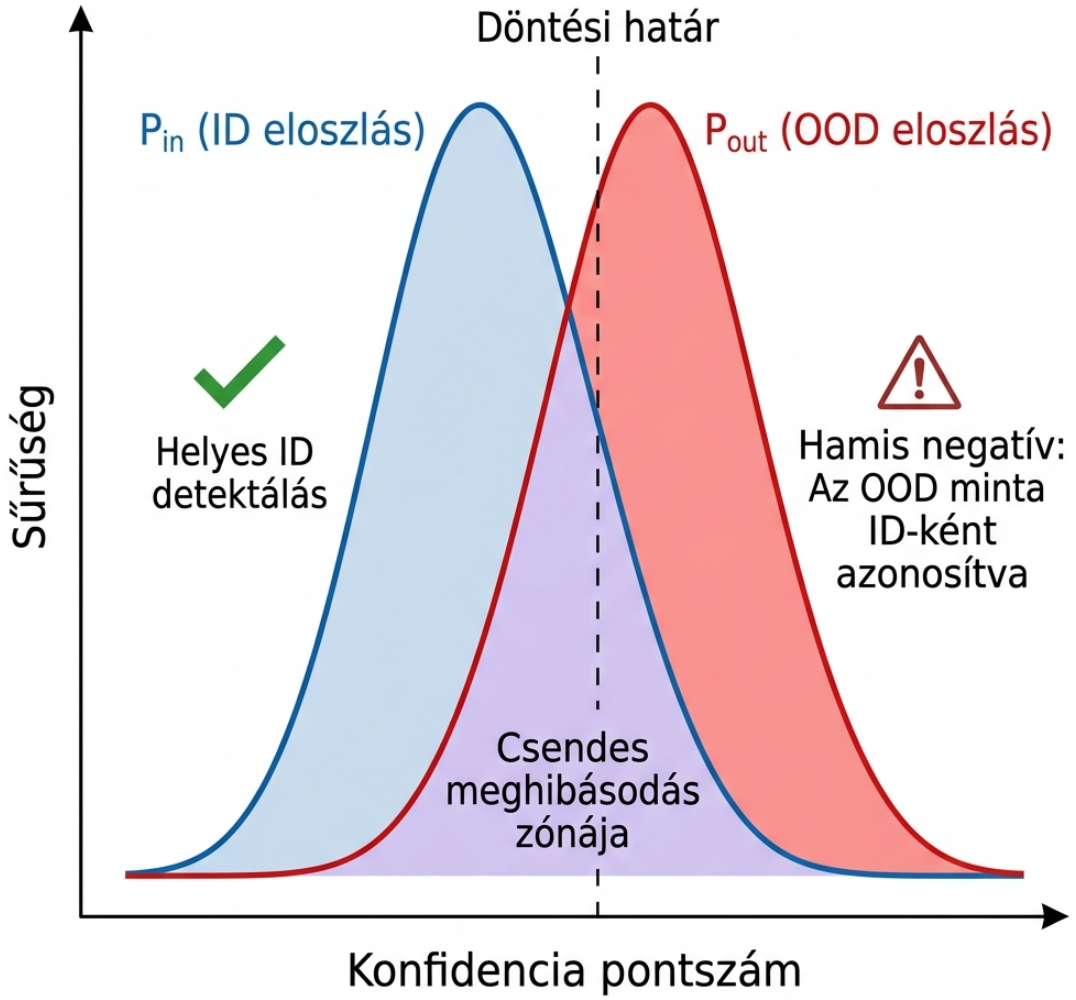
*Kapcsolattartó szerző.
E-mail cím: szabo.lorant@nje.hu

a COCO [1] objektumdetektálás – meghaladja az emberi pontosságot, ám többségük a zárt világ (closed-world) feltételezésre épül [2]: arra, hogy a tesztelés során előforduló adatok eloszlása megegyezik a tanítóadatokéval.

A valóság ugyanakkor természeténél fogva nyílt világú (open-world) [3], és a mélytanuló modellek legkritikusabb biztonsági kockázata a túlzott magabiztosság: ismeretlen bemenetre a modell hajlamos magas valószínűséggel ismert osztályt prediktálni [4]. Ez a csendes meghibásodás (silent failure) közvetlen veszélyt jelent a fizikai világban operáló rendszereknél.

1.1. Az OOD-detektálás definíciós keretei

Az Out-of-Distribution detektálás célja egy döntési mechanizmus integrálása, amely megkülönbözteti az eloszláson belüli (ID) P_{in} eloszlású mintákat az eloszláson kívüli (OOD) P_{out} eloszlású mintáktól, ahol P_{in} jelöli a tanítási, P_{out} pedig a tesztidőben fellépő, ismeretlen eloszlást.



1. ábra. Az ID (P_{in}) és OOD (P_{out}) eloszlások konceptuális ábrázolása. Az átfedő régió a csendes meghibásodás zónáját jelöli.

Az OOD detektálás kapcsolódó, de nem azonos területek:

- **Anomáliadetektálás:** Általánosabb fogalom, a szenzorhibák észlelését is magában foglalja.
- **Novelty Detection:** Új, később integrálható minták felismerése.
- **Open-Set Recognition:** A tanítás során nem definiált osztályok felismerése [3].

Jelen cikk a percepciós modellek OOD detektálásának módszertani alapjait mutatja be, különös tekintettel az objektumdetektálás kontextusára, ahol a modellnek lokális régiók szintjén kell értékelnie a bizonytalanságot [2].

1.2. A tárgyalt módszerek kiválasztása és az áttekintés hatóköre

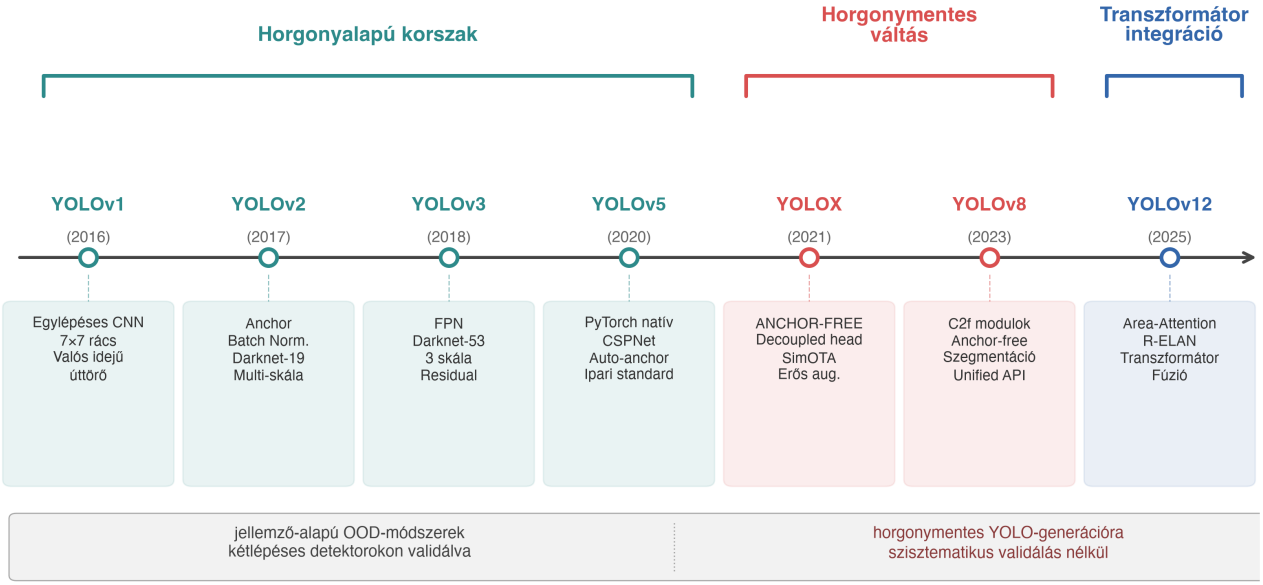
A taxonómia a beavatkozás mélysége és a tanítási igény szerint tagolja a területet: e két dimenzió dönti el, hogy egy eljárás betanított detektorra utólag ráépíthető-e, vagy újratanítást igényel. A részletesen tárgyalt módszereket négy szempont szerint választottuk ki: (i) lektorált publikációban jelentek meg, (ii) objektumdetektálási feladaton kerültek validálásra, (iii) saját kategóriájuk működési elvét reprezentálják, és (iv) relevánsak az egylépéses, valós idejű detektorok szempontjából.

Reprezentativitásuk abban áll, hogy kategóriájukon belül az arra jellemző pontszám-számítási stratégiát valósítják meg: a SAFE [10] és a NAPTRON [11] a rejtett rétegek reprezentációiból, a disszimilaritási pontozás [16] és a HBOS [17] kizárólag a detektor kimenetéből, a VOS [14] és a YolOOD [15] pedig a módosított tanítási célfüggvényen keresztül. A Deep Hybrid Models [12] a (ii) szempont alóli tudatos kivétel: eredményeit képosztályozáson közlik, a módszercsalád elvi bemutatása és hatókörkorlátai miatt tárgyaljuk (4.3).

Nem tárgyaljuk a kizárólag képosztályozásra kidolgozott módszercsaládokat, a bizonytalanságkalibrációs és Bayes-i megközelítéseket, sem a generatív rekonstrukciós eljárásokat. Jelen munka célzott narratív áttekintés: nem alkalmaz előre rögzített keresési protokollt, beválasztási kritériumrendszert, tanulmány-kiválasztási folyamatot vagy formális minőségértékelést, ezért nem törekszik a terület kimerítő lefedésére; célja a fogalmi rendszerezés és a legérettebb eljárások működési elvének bemutatása.

2. A percepciós modellek architektúrális evolúciója és hatása az OOD-detektálásra

Az OOD detektálás hatékonysága elválaszthatatlan az alapmodell architektúrájától. Az objektumdetektálás két fő paradigmát követ: a kétlépéses (two-stage) detektorokat (Faster R-CNN) és az egylépéses (one-stage) detektorokat (YOLO [5]).



2. ábra. A YOLO architektúra evolúciója a YOLOv1-től a YOLOv12-ig. Az alsó sáv sematikusan jelzi a jellemző-alapú OOD-módszerek validációs lefedettségét: ezeket az eljárásokat kétlépéses detektorokon validálták, a horgonymentes YOLO-generációra vonatkozó szisztematikus validálásuk nem áll rendelkezésre. A sáv nem publikációszámot ábrázol.

2.1. Horgonyalapú vs. horgonymentes detektálás

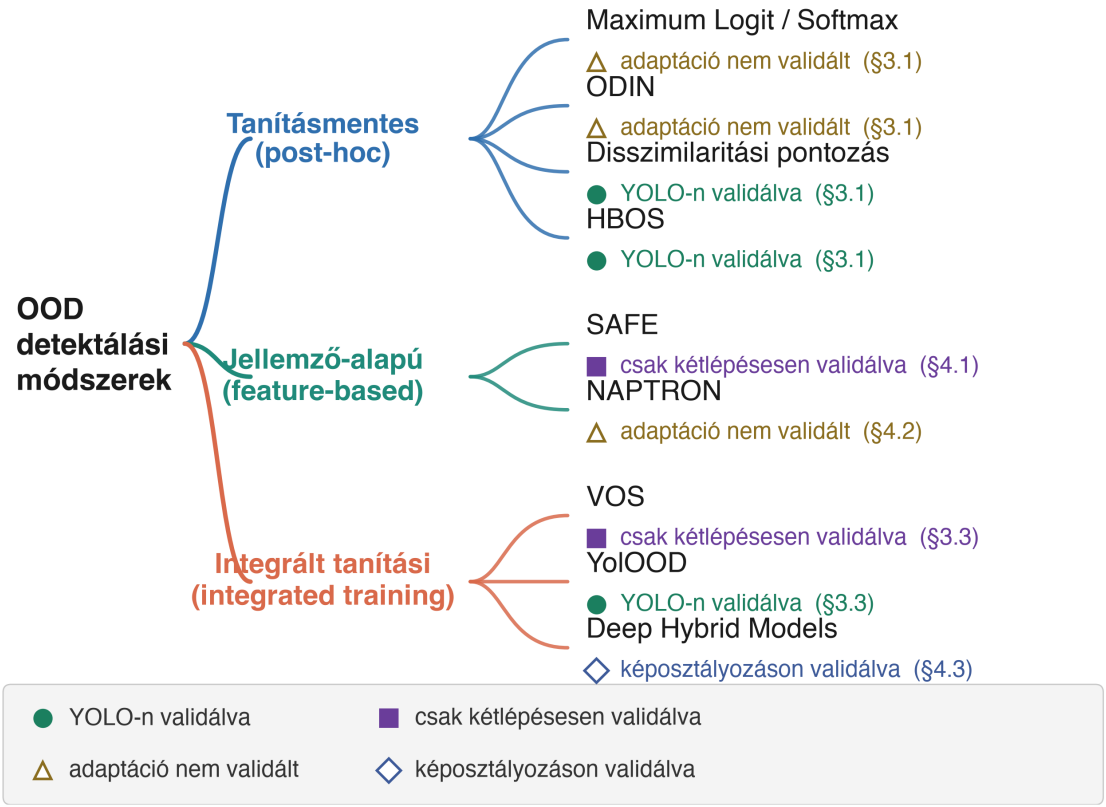
A korai verziók (YOLOv2–v5) horgonyalapú (anchor-based) megközelítést alkalmaztak, ahol előre definiált dobozok szolgáltak a detektálás alapjául; a YOLOX [7] és YOLOv8-tól az iparág a horgonymentes (anchor-free) detektálás felé mozdult [6]. Ez az OOD detektálás szempontjából lényeges: a horgonyokra optimalizált módszerek hatékonysága csökkenhet a prior információ hiánya miatt [6].

2.2. A transzformátorok megjelenése

A YOLOv12 [8] és az RT-DETR [9] figyelem-mechanizmusokat integrálnak. Előnyük a globális kontextus-reprezentáció, amely segíthet kiszűrni a kontextusba nem illő OOD objektumokat. Kihívásuk, hogy a hagyományos, CNN-re tervezett OOD módszerek (pl. a SAFE batch normalizációra építő logikája [10]) adaptációt igényelnek. A jelenlegi OOD-kutatások többsége ezt az architektúrális ugrást eddig csak korlátozott mértékben követte le [2].

3. Az OOD-detektálás módszertani taxonómiája

Az OOD módszertanokat a beavatkozás mélysége és a tanítási igény szerint három kategóriába soroljuk.



3. ábra. Az OOD detektálási módszerek hierarchikus taxonómiája. A módszerek melletti jelölés a validáció állapotát adja meg, négy fokozatban: YOLO-n validálva (kör), csak kétlépésesen validálva (négyzet), adaptáció nem validált – elvben lehetséges, de a szerzők YOLO-eredményt nem közölnek (háromszög), illetve képosztályozáson validálva (rombusz). Minden jelölés mellett szerepel az azt alátámasztó szekció száma.

3.1. Tanításmentes (post-hoc) módszerek

Nem igénylik az alapmodell módosítását:

– **Maximum Logit / Softmax Score:** A legegyszerűbb megközelítés [4]. Az ODIN [13] hőmérséklet-skálázással finomít rajta.

– **Disszimilaritási pontozás:** A kimeneti eloszlások közötti távolságot méri. Honorato és társai YOLOv5 modellen alkalmazták sikeresen [16].

– **HBOS:** Szabó és társai a YOLOv11 kimeneteire integrálták [17]. Lineáris, $\mathcal{O}(N)$ komplexitású (N a minták száma), ezért valós idejű rendszerekben is alkalmazható.

3.2. Jellemző-alapú (feature-based) módszerek

A mélyebb rétegek gazdagabb térbeli és szemantikai információt hordoznak:

– **NAPTRON:** Bináris aktivációs mintázatokat elemez Hamming-távolsággal [11].

– **SAFE:** Batch normalizációs rétegek érzékenységét használja [10].

3.3. Integrált tanítási (integrated training) módszerek

A modellt eleve OOD-képessé tervezik:

– **VOS:** Szintetikus OOD mintákat generál a tanítás során [14].

– **YolOOD:** Többcímkés osztályozóvá alakítja a detektort, az objektumszerűség és osztálypontszámok együttes értékelésével [15].

– **Deep Hybrid Models:** Generatív modellezést kombinálnak diszkriminatív osztályozással [12].

4. Érett OOD-módszerek: részletes elemzés

Az objektumdetektálás OOD módszertanainak jelenlegi technológiai felső korlátját a kétlépéses detektorokon validált módszerek jelölik ki.

Az alább idézett számszerű eredmények eltérő ID–OOD adathalmaz-párokon, különböző backbone-okon és nem egységesített protokollok szerint születtek, ezért *illusztratív* jellegűek: az egyes módszerek működési elvének hatékonyságát szemléltetik, és nem tekinthetők a módszerek közötti közvetlen teljesítmény-összehasonlításnak. Az összemérhetőség előfeltétele egy egységes, objektumdetektálásra szabott értékelési keretrendszer lenne.

4.1. SAFE: érzékenység-tudatos jellemzők

Wilson és társai (2023) SAFE módszere [10] a batch normalizációval ellátott reziduális konvolúciós rétegek érzékenységét használja ki. A módszer ezekből a rétegekből vektorokat von ki, és egy kis méretű neurális hálót (MLP) tanít a normál adatok és az adverzariálisan perturbált adatok megkülönböztetésére.

Eredmények: Az OpenImages adathalmazon a SAFE 30,6 százalékpontos FPR@95 csökkenést ért el (48,3% $\rightarrow$ 17,7%). BDD100K-n ResNet-50 backbone-nal AUROC = 94,64%, FPR@95 = 16,04%; RegNet backbone-nal AUROC = 95,97%, FPR@95 = 13,98%. Ez arra utal, hogy a modellek belső reprezentációi az adott feladatkörnyezetben több információt hordoznak az ismeretlenről, mint a végső döntési kimenet.

4.2. NAPTRON: neuronaktivációs minták

Olber és társai (2023) NAPTRON módszere [11] bináris mintázatokat (on/off állapotok) generál a hálózat rejtett rétegeiből. Ha egy új bemenet aktivációs mintázata túlságosan eltér (Hamming-távolság) a tanítás során látottaktól, anomáliának tekinti. Élvonalbeli FPR@95 eredmények Faster R-CNN, RetinaNet és FCOS architektúrákon. Architektúra-függetlensége miatt elvben adaptálható a YOLO családra is, YOLO-architektúrán való validálását azonban a szerzők nem közlik.

4.3. Deep Hybrid Models

A Deep Hybrid Models [12] a generatív és diszkriminatív modellezés kombinálásával értek el 100%-os AUROC-ot a CIFAR-10 vs. SVHN feladaton. Hangsúlyozni kell, hogy ez képosztályozási benchmark, nem objektumdetektálási: a módszert itt a módszercsalád működési elvének illusztrációjaként tárgyaljuk, eredménye a jelen cikk fókuszába tartozó detektálási feladatokra közvetlenül nem vihető át. Fontos korlát továbbá, hogy ez a teljesítmény csak szemantikailag diszjunkt adatbázisokon érhető el, és nem skálázódik átfedő osztályokat tartalmazó feladatokra.

5. Összefoglalás

Jelen cikkben az eloszláson kívüli (OOD) detektálás módszertani alapjait és hierarchikus taxonómiáját vizsgáltuk a percepciós modellek kontextusában. Áttekintettük a tanításmentes (post-hoc), a jellemző-alapú és az integrált tanítási megközelítések fő stratégiáit, részletesen elemeztük a terület legérettebb objektumdetektálási módszereit (SAFE, NAPTRON) és – hatóköri korlátaival együtt – a Deep Hybrid Models megközelítést, valamint összevetettük az egylépéses és kétlépéses detektorok architektúrális különbségeit.

A vizsgálat fő mondanivalója, hogy bár a kétlépéses detektorokon (Faster R-CNN, RetinaNet) validált jellemző-alapú módszerek élvonalbeli eredményeket nyújtanak, az egylépéses, horgonymentes YOLO architektúrákra [5] való adaptációjuk szisztematikus validálása még nem áll rendelkezésre. Fontos elhatárolás, hogy ez nem a YOLO-alapú OOD-megoldások hiányát jelenti – a tanításmentes kategóriában több ilyen eljárás működik (3.1) –, hanem azt, hogy éppen a legpontosabb, jellemző-alapú módszerek validálása korlátozott a modern detektor-generációra. Ez a rés biztonsági kockázatot jelent a valós idejű percepciós rendszerekben, és a YOLO-specifikus OOD-megoldások célzott fejlesztését, valamint egységes értékelési protokollját igényli. Egy kísérő tanulmányban részletesen elemezzük azokat a YOLO-specifikus megközelítéseket – a YolOOD többcímkés paradigmáját, a disszimilaritási pontozást és a HBOS valós idejű integrációját –, amelyek e rés áthidalására irányulnak.

Hivatkozások

[1] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, „Microsoft COCO: Common Objects in Context," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755. arXiv:1405.0312.

[2] J. Yang, K. Zhou, Y. Li, and Z. Liu, „Generalized Out-of-Distribution Detection: A Survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024. DOI: 10.1007/s11263-024-02117-4

[3] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, „Toward Open Set Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013. DOI: 10.1109/TPAMI.2012.256

[4] D. Hendrycks and K. Gimpel, „A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2017. arXiv:1610.02136.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, „You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91

[6] R. Sapkota, R. Qureshi, M. Flores-Calero, C. Badgujar, U. Nepal, A. Poulose, P. Zeno, U. B. P. Vaddevolu, H. Yan, and M. Karkee, „YOLO Advances to Its Genesis: A Decadal and Comprehensive Review of The You Only Look Once (YOLO) Series," *arXiv preprint* arXiv:2406.19407, 2024.

[7] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, „YOLOX: Exceeding YOLO Series in 2021," *arXiv preprint* arXiv:2107.08430, 2021.

[8] Y. Tian, Q. Ye, and D. Doermann, „YOLOv12: Attention-Centric Real-Time Object Detectors," *arXiv preprint* arXiv:2502.12524, 2025.

[9] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, „DETRs Beat YOLOs on Real-Time Object Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2304.08069.

[10] S. Wilson, T. Fischer, F. Dayoub, D. Miller, and N. Sünderhauf, „SAFE: Sensitivity-Aware Features for Out-of-Distribution Object Detection," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 23565–23576. DOI: 10.1109/ICCV51070.2023.02160

[11] B. Olber, K. Radlak, K. Chachuła, J. Łyskawa, and P. Frątczak, „Detecting Out-of-Distribution Objects Using Neuron Activation Patterns," in *Proc. 26th European Conference on Artificial Intelligence (ECAI)*, 2023. DOI: 10.3233/FAIA230467 (arXiv:2307.16433)

[12] P.-R. Schlumbom and E. Frank, „Revisiting Deep Hybrid Models for Out-of-Distribution Detection," *Transactions on Machine Learning Research (TMLR)*, Apr. 2025. OpenReview: yeITEuhv4Q

[13] S. Liang, Y. Li, and R. Srikant, „Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2018. arXiv:1706.02690.

[14] X. Du, Z. Wang, M. Cai, and Y. Li, „VOS: Learning What You Don't Know by Virtual Outlier Synthesis," in *Proc. International Conference on Learning Representations (ICLR)*, 2022. arXiv:2202.01197.

[15] A. Zolfi, G. Amit, A. Baras, S. Koda, I. Morikawa, Y. Elovici, and A. Shabtai, „YolOOD: Utilizing Object Detection Concepts for Multi-Label Out-of-Distribution Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5788–5797.

[16] E. S. Honorato, M. A. S. Uchida, T. H. S. Silva, and D. F. Wolf, „Out-of-Distribution Object Detection in Autonomous Vehicles With Yolo Model," in *Proc. 21st Latin American Robotics Symposium (LARS)*, 2024, pp. 1–6. DOI: 10.1109/LARS64411.2024.10786436

[17] L. Szabó, Z. Weltsch, and A. Major, „Real-Time Out of Distribution Detection in 2D Object Detection for Autonomous Cars," *Engineering Perspective*, vol. 5, no. Special Issue: 1st Future of Vehicles Conf., pp. 28–36, 2025. DOI: 10.64808/engineeringperspective.1814718