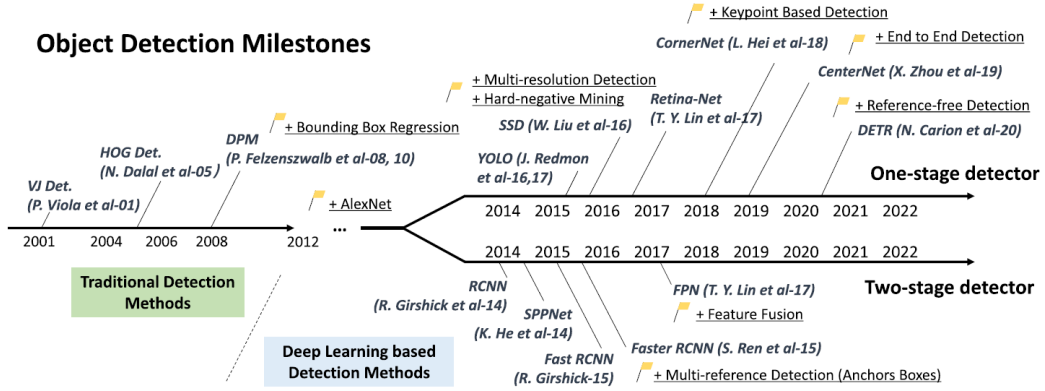
BAB II

LANDASAN TEORI

2.1 Deteksi Objek (*Object Detection*)

Deteksi objek merupakan bagian dari visi komputer yang bertugas untuk mendeteksi visual yang memiliki kelas tertentu pada sebuah gambar digital (Zou dkk., 2023). Tujuan utama dari deteksi objek ini adalah mengembangkan sebuah model dan teknik komputasi seperti konvolusi, segmentasi, dan lain-lain untuk mendeteksi objek beserta letak objek tersebut. Dalam deteksi objek terdapat dua komponen utama metrik yang digunakan untuk menghitung performa dari model deteksi objek di antaranya akurasi dan kecepatan deteksi (Zou dkk., 2023).

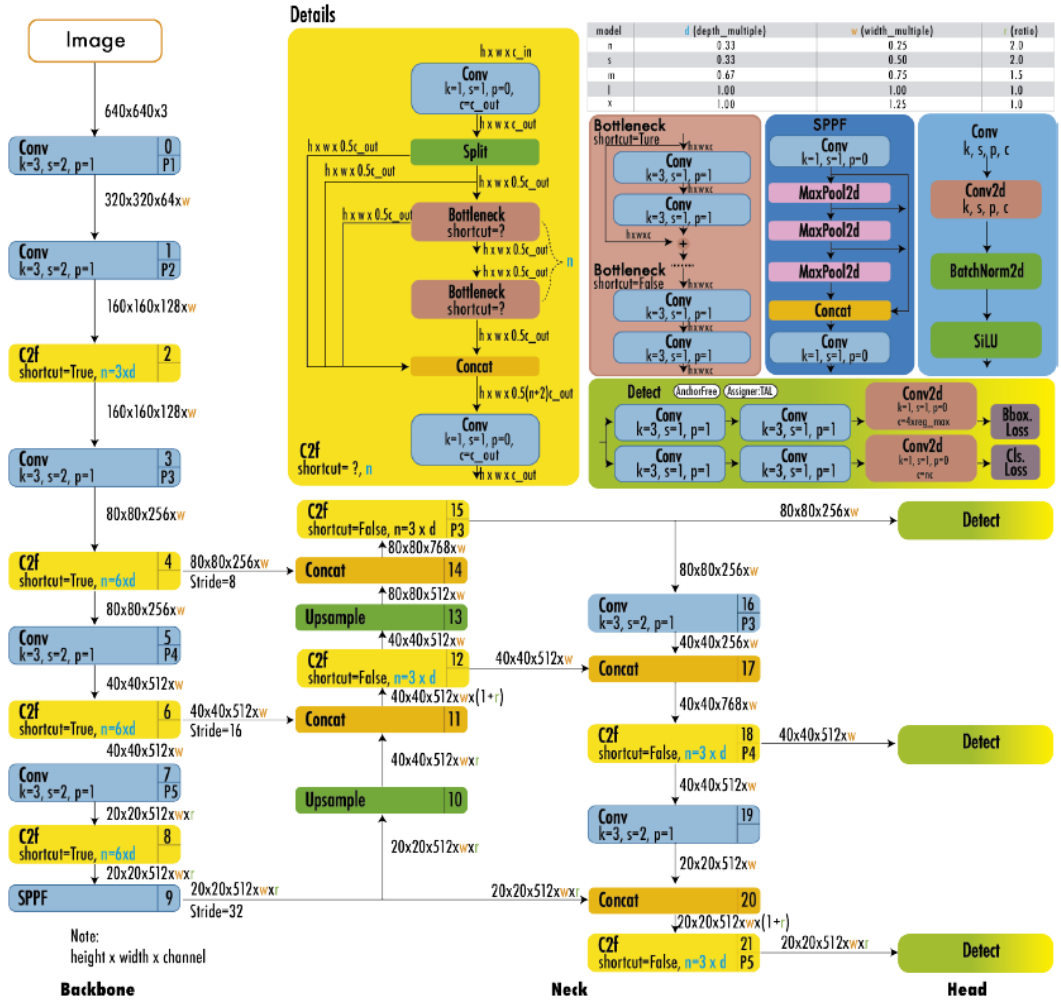
Deteksi objek pertama kali diawali oleh *Traditional Detector Method* yang dikembangkan oleh (Viola, 2001) dengan nama *Viola Jones Detector* (*VJ Detection*). Model yang dikembangkan tersebut menerapkan perhitungan sederhana, yaitu *sliding window* dengan menelusuri semua kemungkinan lokasi dan melakukan penskalaan pada setiap gambar yang ditangkap untuk mengidentifikasi apakah terdapat wajah manusia. Pada masanya, kecepatan deteksi *VJ Detection* paling tinggi jika dibandingkan dengan *detector* lain dengan mengintegrasikan tiga teknik penting, yaitu *integral image, feature selection,* dan *detection cascade* (Viola, 2001). Selanjutnya, pada tahun 2012 tercipta suatu inovasi bernama *Deep Convolutional Neural Network* (CNN) (Krizhevsky dkk., 2012) yang menggantikan *Traditional Detector Method* menjadi *Deep Learning Based Detection Method* dari *You Only Look Once* (YOLO) (Zou dkk., 2023).



Gambar 2.1 Perkembangan deteksi objek dalam 20 tahun terakhir (Zou dkk., 2023)

YOLO pertama kali diperkenalkan oleh (Redmon dkk., 2016) dengan menerapkan *single neural network* untuk memproses seluruh gambar secara penuh. Arsitektur jaringan yang digunakan akan membagi gambar menjadi beberapa wilayah dan selanjutnya melakukan prediksi *bounding box* serta menghitung probabilitas untuk setiap kelas pada setiap wilayah secara bersamaan. Model ini dapat meningkatkan kecepatan deteksi dengan pesat, meskipun hasil akurasi deteksi yang dihasilkan harus dikorbankan (Redmon dkk., 2016). Saat ini, versi YOLO telah mencapai versi 8 dengan segala keunggulannya (J. R. Terven & Cordova-esparza, 2024). YOLOv8 dikembangkan oleh (Jocher dkk., 2023) dengan memperbaiki atau meningkatkan arsitektur yang ada pada YOLOv5 (Jocher, 2020). YOLOv5 menggunakan *Cross Stage Partial* (CSP) *Layer* dibuat untuk meningkatkan kinerja jaringan dengan membagi jalur masukan menjadi dua jalur yang berbeda, kemudian menggabungkan kembali fitur-fitur tersebut (Jocher, 2020). Ini membantu dalam mengatasi masalah degradasi jaringan (*network degradation*) yang bisa terjadi dalam jaringan yang sangat dalam. Pada YOLOv8 *layer* tersebut diganti dengan

Cross Stage Feature Fusion (*C2f*) *module* yang digunakan untuk menyatukan informasi fitur dari berbagai tingkatan resolusi dalam jaringan (Jocher dkk., 2023). Ini memungkinkan jaringan untuk memperoleh pemahaman yang lebih baik tentang objek dalam berbagai skala dan tingkat detail, meningkatkan kemampuan deteksi objek secara keseluruhan. Kebaruan arsitektur dari YOLOv8 dapat dilihat pada Gambar 2.2



Gambar 2.2 Arsitektur YOLOv8 (Jocher dkk., 2023; J. R. Terven & Cordova-esparza, 2024)

2.2 *Speech Recognition*

Speech recognition merupakan salah satu teknologi yang dapat melakukan konversi suara menjadi sebuah teks yang dapat dikenali oleh mesin (Malik dkk., 2021). Secara ideal *speech recognition* harus dapat menerima input suara yang diberikan oleh manusia, mengenali kata-kata yang diucapkan, dan menggunakan kata-kata tersebut sebagai perintah kepada mesin atau komputer untuk melakukan aksi (Y. dkk., 2017). Perkembangan dari *speech recognition* telah dimulai dari tahun 1950an dan masih berkembang sampai sekarang (Malik dkk., 2021). *Audrey* merupakan sistem *speech recognition* pertama yang dikembangkan oleh laboratorium Bell (Davis dkk., 1952). Sistem yang dirancang hanya dapat mengenali dan membedakan digit-digit berbeda yang diucapkan oleh seorangan pengguna tunggal.

Saat ini, penggunaan *Recurrent Neural Network* (RNN) yang merupakan bagian dari *Long Short Term Memory* (LSTM) menjadi model *speech recognition* paling banyak digunakan dan memiliki kemampuan pengenalan suara yang baik (Malik dkk., 2021). Model tersebut terinspirasi dari penggabungan antara *Hidden Markov Model* (HMM) dengan *Artificial Neural Network* (ANN) (Bourlard & Morgan, 1994). Sehingga model *speech recognition* yang menggunakan RNN dapat mengenali lebih banyak ucapan dalam berbagai aksen karena model tersebut menyimpan pengetahuannya dalam memori jangka pendek pada LSTM. Implementasi penggunaan *speech recognition* yang menggunakan RNN dapat dilakukan pada bahasa pemrograman *python* dengan mudah. Terdapat modul *SpeechRecognition* yang dapat diunduh dan modul tersebut telah mendukung

banyak API untuk *speech recognition* seperti CMU Spinx, *Google Speech Recognition*, *Microsoft Bing Voice Recognition*, dan lain-lain.

2.3 *Text-to-Speech* (TTS)

Text-to-Speech (TTS) atau dapat disebut juga dengan *speech synthesis* merupakan suatu metode yang bertujuan untuk melakukan sintesis ucapan yang dapat dipahami dan dikeluarkan secara alami (Tan dkk., 2021). Secara sederhana, TTS akan melakukan konversi terhadap teks yang menjadi input dengan mengubahnya menjadi sebuah gelombang yang dapat memunculkan suara. Penelitian yang tercatat pertama kali mengenai pengembangan dari TTS ini ada pada tahun 1976 dengan nama model *Articulatory Synthesis* (Coker, 1976). *Articulatory Synthesis* adalah pendekatan dalam pembuatan sistem sintesis ucapan yang mencoba meniru bagaimana manusia menghasilkan suara dengan memodelkan gerakan fisik dari organ-organ bicara seperti lidah dan bibir (Coker, 1976). Dengan menggunakan model ini, sistem dapat menghasilkan suara yang lebih realistis yang dapat disesuaikan dengan berbagai gaya bicara dan aksen pada masanya.

Dengan kemajuan dalam *deep learning*, diperkenalkan TTS berbasis jaringan saraf (*neural* TTS), yang menggunakan *neural network* sebagai kerangka model untuk sintesis suara. WaveNet (Oord dkk., 2016) kemudian diusulkan untuk langsung menghasilkan gelombang suara dari fitur linguistik yang dianggap sebagai model TTS berbasis jaringan saraf modern pertama (Tan dkk., 2021). Model lain seperti DeepVoice (Arik dkk., 2017) masih mengikuti komponen-komponen dalam sintesis parametrik statistik, tetapi diperbarui dengan model berbasis jaringan saraf

yang sesuai. Selain itu, beberapa model *end-to-end* seperti Tacotron (Wang dkk., 2017) , Deep Voice 3 (Ping dkk., 2018), dan FastSpeech (Ren dkk., 2019) menyederhanakan modul analisis teks dan langsung menggunakan urutan karakter/fonem sebagai input, serta menyederhanakan fitur akustik dengan *mel-spectrogram*. Kemudian, sistem TTS end-to-end sepenuhnya dikembangkan untuk menghasilkan gelombang suara langsung dari teks, seperti ClariNet (Ping dkk., 2019), FastSpeech 2 (Ren dkk., 2021), dan *End-to-End Adversarial Text-to-Speech* (EATS) (Donahue dkk., 2021). Dibandingkan dengan sistem TTS sebelumnya, kelebihan *neural text-to-speech* meliputi kualitas suara yang tinggi dalam hal kejelasan dan kealamiannya, serta persyaratan yang lebih sedikit dalam pra-pemrosesan manusia dan pengembangan fitur.

2.4 Penelitian Terkait

Saat ini telah banyak pengembangan-pengembangan terhadap perangkat yang digunakan untuk membantu penderita tunanetra dalam menjalani aktivitasnya sehari-sehari. Pengembangan yang dilakukan tidak semata-mata hanya dilakukan tanpa aspek-aspek yang mempengaruhi kualitas perangkat. Terdapat beberapa aspek yang menjadi fitur utama dalam mengukur performa pengembangan perangkat yang diantaranya *capturing device, working hour, response time, coverage area, feedback, working range, weight, robustness,* dan *cost* (Mashiata dkk., 2022). Pengukuran performa tersebut didasarkan pada tantangan-tantangan yang dialami oleh para penderita tunanetra serta berasal dari lingkungan para penderita, kehidupan sosialnya, kesulitan dalam menggunakan teknologi dan lain sebagainya (Manjari dkk., 2020).

Pemanfaatan *object detection* dan *text-to-speech feedback* berbasis *Artificial Intelligence* (AI) semakin banyak digunakan untuk mengatasi setiap tantangan yang ada dan memenuhi setiap aspek yang dibutuhkan (Walle dkk., 2022). Namun, terdapat masalah pada keakuratan model deteksi dan *feedback* yang diberikan oleh model apakah memiliki kesesuaian dengan apa yang dibutuhkan oleh pengguna. Sehingga, untuk menjawab permasalahan tersebut dibuatlah *state of the art* terkait dengan penelitian yang dilakukan serta melihat kesesuaian pemanfaatan dari *object detection* dan TTS pada perangkat pembantu tunanetra.

Tabel 2.1 *State of the art* penelitian terkait

Penulis	Judul	Metode Deteksi Objek	Metode TTS	Keterangan
(Guravaiah dkk., 2022)	*Third Eye: Object Recognition and Speech Generation for Visually Impaired*	YOLOv5	*google Text-to-Speech* (gTTS) dan pyttsx3	**Kelebihan** : Hasil perhitungan matrik mAP menunjukkan nilai mAP:0.5 pada 39% dan mAP:0.95 pada 25%, **Kekurangan** : Hasil mAP yang dihasilkan oleh model YOLOv5 masih kurang baik karena untuk hasil mAP:0.5 saja berada di bawah 50%.
(Setiadi dkk., 2020)	*Navigation and Object Detection for Blind Persons Based on Neural Network*	CNN dan LiDAR (*Light Detection and Ranging*)	*Voice attitudes information*	**Kelebihan** : menunjukkan akurasi hasil pendeteksian jalur pejalan kaki sebesar 89.7% pada lux<15000 dan 87.5% untuk akurasi *object detection* **Kekurangan** : Penelitian yang dilakukan tidak menjelaskan bagaimana arsitektur yang digunakan dan tidak memperlihatkan bagaimana optimasi model bekerja
(Ravindra Karmarkar &	*Object Detection System for The*	YOLO	gTTS	**Hasil** : Menunjukkan bahwa model dapat

Honmane, 2021)	*Blind with Voice Guidance*			melakukan pendeteksian objek dan mengeluarkan suara dari hasil pendeteksian. **Kekurangan** : Penelitian tersebut tidak menunjukkan perbandingan dan akurasi hasil model
(Najm dkk., 2022)	*Assisting Blind People Using Object Detection with Vocal Feedback*	YOLOv3	gTTS	**Hasil** : Untuk gambar 608x608 memiliki akurasi sebesar 90.17%, untuk gambar 512x512 memiliki akurasi sebesar 87.32%, dan untuk gambar 416x416 memiliki akurasi sebesar 97.17%. **Kekurangan** : Model deteksi objek yang digunakan hanya terpaku pada YOLOv3 saja
(Wong dkk., 2019)	*Convolutional Neural Network for Object Detection System for Blind People*	CNN dengan arsitektur SSD Mobilenet	*Voice synthesis*	**Hasil** : Model menunjukkan akurasi sebesar 73.7% pada *dataset* CIFAR-10. **Kekurangan** : Model hanya dapat mencapai FPS sebesar 5 yang membuat pendeteksian secara real time menjadi kurang baik.
(Abdurrasyid dkk., 2019)	*Detection of immovable objects on visually impaired people walking aids*	*Template matching*	-	**Hasil** : Hasil menunjukkan bahwa ketika kamera diletakkan pada 90^{o} dengan objek, hasil deteksi menjadi lebih maksimal karena noise yang didapatkan sedikit. **Kekurangan** : Ketika kamera berada pada 45^{o} dan 135^{o} dengan objek, hasil deteksi menjadi tidak maksimal dan mendapatkan akurasi terendah pada pengujian.
(Konaite dkk., 2021)	*Smart Hat for the blind with Real-Time Object Detection using Raspberry Pi and TensorFlow Lite*	SSD MobileNet v2 320x320	SPEAK	**Hasil** : Deteksi *real-time* FPS sebesar 5.26 frame/s pada raspberry pi 4. Untuk waktu deteksi gambar

				mencapai waktu tercepat, yaitu 19 ms. **Kekurangan :** Model memiliki nilai mAP yang paling kecil jika dibandingkan dengan model lain yang dibandingkan yaitu 20.2%
(Mahendru & Dubey, 2021)	*Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3*	YOLOv3 dengan Darknet53 dan YOLO dengan SSD MobilNet	gTTS	**Kelebihan** : Pendeteksian *multi-class* pada satu gambar dengan baik. rata-rata akurasi hasil deteksi gambar yang dilakukan oleh model ada pada 65-98 % **Kekurangan** : Penelitian ini masih menggunakan YOLOv3 yang original sehingga tidak membangkitkan performa yang maksimal bagi model
(Ramalakshmi dkk., 2020)	*Object Detector for Visually Impaired with Distance Calculation for Humans*	Integrasi YOLO dengan *Haar Classifier*	gTTS	**Kelebihan** : Model dapat melakukan deteksi pada objek yang saling berhimpitan. Selain itu, model dapat melakukan kalkulasi perkiraan jarak antara objek manusia dengan kamera melalui *Haar classifier* **Kekurangan** : Penelitian ini tidak memperlihatkan bagaimana hasil performa model secara menyeluruh
(Alzahrani & Al-Baity, 2023)	*Object Recognition System for the Visually Impaired: A Deep Learning Approach using Arabic Annotation*	Mask R-CNN	gTTS	**Kelebihan** : Model Mask R-CNN memiliki nilai mAP sebesar 83.9% dengan objek gambar yang berhasil dideteksi beserta letak objek tersebut. **Kekurangan** : Terdapat beberapa kelas objek yang memiliki nilai mAP rendah dan model ini masih belum mendukung pendeteksian secara *real time*.

(R. C. Joshi dkk., 2020)	*Efficient Multi-Object Detection and Smart Navigation Using Artificial Intelligence for Visually Impaired People*	YOLOv3 dengan tambahan sensor ultrasonic	*Optical Character Recognizer* (OCR)	**Kelebihan** : 95.19% untuk akurasi deteksi dan 99.69% untuk akurasi pengenalan objek. **Kekurangan** : Dibutuhkan tambahan kelas pada dataset untuk memaksimalkan pendeteksian pada model
(Alahmadi dkk., 2023)	*Enhancing Object Detection for VIPs Using YOLOv4_Resnet101 and Text-to-Speech Conversion Model*	YOLOv4 dengan backbone Resnet101	pyttsx3	**Kelebihan** : Akurasi model sebesar 96.34% dan error rate 0.073%. Kecepatan FPS yang dihasilkan ada pada 60-80 frame/s untuk input gambar 416x416 **Kekurangan** : Memerlukan *resource* yang besar dalam proses pelatihan maupun percobaan.
(Noman dkk., 2020)	*Portable offline indoor object recognition system for the visually impaired*	Faster R-CNN	eSpeakNG	**Kelebihan** : Akurasi model sebesar 73.24% dengan kesalahan pendeteksian jarak sekitar 5%. **Kekurangan** : Akurasi yang dihasilkan masih kurang karena model yang digunakan hanya menggunakan Tensorflow Object Detection API saja
(Lee & Cho, 2022)	*Automatic Object Detection Algorithm-Based Braille Image Generation System for the Recognition of Real-Life Obstacles for Visually Impaired People*	YOLOv3	-	**Kelebihan** : Hasilnya, sistem bantuan hidup yang diusulkan dalam penelitian ini terbukti efisien dan berguna dengan akurasi rata-rata konversi braille 85%, akurasi deteksi 90% atau lebih, dan waktu konversi braille rata-rata 6,6 detik **Kekurangan** : Hasil ekstraksi objek yang diperoleh melalui GrabCut menggunakan koordinat objek yang terdeteksi dengan YOLOv3 tidak sesuai dengan objek sebenarnya.

(Rahman & Sadi, 2021)	*IoT Enabled Automated Object Recognition for the Visually Impaired*	SSD MobilNet	-	**Kelebihan** : Akurasi keseluruhan sistem yang diusulkan dalam deteksi dan pengenalan objek masing-masing adalah 99,31% dan 98,43%. **Kekurangan** : Penelitian tidak menjelaskan metode TTS apa yang digunakan.
(Rocha dkk., 2023)	*Using Object Detection Technology to Identify Defects in Clothing for Blind People*	YOLOv5	*Smartphone Audio Feedback*	**Kelebihan** : Dari perbandingan yang dilakukan model YOLOv5s6 dapat melakukan waktu inferensi yang paling cepat, yaitu 0.0092s, sedangkan untuk nilai mAP terbaik ada pada model YOLOv5l6 dengan mAP0.5 sebesar 76%. **Kekurangan** : Metode optimasinya masih menggunakan SGD yang masih kurang maksimal jika dibandingkan dengan metode lain
(Kadhim & Oleiwi, 2022)	*Blind Assistive System Based on Real Time Object Recognition using Machine Learning*	YOLOv3	-	**Kelebihan** : Hasil menunjukkan bahwa YOLOv3 dapat melakukan deteksi dengan baik pada kelas-kelas yang ada pada dataset. **Kekurangan** : Tidak adanya perbandingan dengan penelitian lain membuat model pada penelitian ini tidak dapat dipastikan performanya
(Yohannes dkk., 2020)	*Robot Eye: Automatic Object Detection And Recognition Using Deep Attention Network to Assist Blind People*	*Deep Attention Network*	-	**Kelebihan** : Metode yang diusulkan mencapai tingkat akurasi sekitar 81%, lebih baik daripada YOLO v3 original. **Kekurangan** : Performa model yang dihasilkan tidak mencapai performa yang maksimal karena tidak adanya proses optimasi model

(Islam dkk., 2023)	*Deep learning based object detection and surrounding environment description for visually impaired people*	TensorFlow Object Detection API dan SSDLite MobileNetV2	*Google text-to-speech*, PyAudio, playsound	**Kelebihan** : Akurasi yang dihasilkan oleh model ada pada 88.89% dan untuk FPS-nya berada pada 2.15 frame/s **Kekurangan** : Ukuran kumpulan data cuaca yang digunakan untuk mengembangkan ambiance mode relatif kecil.
(Khan dkk., 2023)	*Outdoor mobility aid for people with visual impairment: Obstacle detection and responsive framework for the scene perception during the outdoor mobility of people with visual impairment*	YOLOv5 dan Mask R-CNN	Google services	**Kelebihan** : . Hasil pengukuran mAP(0.5) menujukkan untuk YOLOv5s mendapatkan nilai 97% dan untuk Mask R-CNN mendapat nilai 93%. **Kekurangan** : Kelas yang digunakan pada dataset terbilang sedikit karena hanya terdiri dari 8 kelas saja.
(Mukhiddinov & Cho, 2021)	*Smart Glass System Using Deep Learning for the Blind and Visually Impaired*	DETR (*Detection Transformer*)	OCR	**Kelebihan** : Model memiliki nilai mAP:0.5 sebesar 63.5%. Untuk model ekstraksi objek menonjol menghasilkan nilai maxFM, MAE, dan WFM secara berturut-turut adalah 0.814, 0.058, dan 0.725. **Kekurangan** : Dalam beberapa situasi, model deteksi objek mendeteksi lebih dari sepuluh objek, di mana beberapa di antaranya adalah objek kecil atau terdeteksi dengan tidak benar,
(Mandhala dkk., 2020)	*Object Detection Using Machine Learning for Visually Impaired People*	Retina Net, YOLOv3, dan Tiny-YOLOv3	-	**Kelebihan** : Akurasi masing-masing untuk Retina Net 80.5%, YOLOv3 71.1%, dan Tiny-YOLOv3 76.3%. **Kekurangan** : Tidak adanya proses optimasi membuat model tidak menghasilkan performa yang maksimal.
(Pi dkk., 2022)	*Low-Cost Smart Glasses for Blind Individuals using Raspberry Pi 2*	YOLOv3	OCR	**Kelebihan** : . Dengan jarak kurang lebih 5 meter, model dengan

				baik dapat mengenali objek yang ditangkap. **Kekurangan** : Terdapat perbedaan dalam pendeteksian beberapa objek karena jarak yang menyebabkan pendeteksiannya tidak dapat mendeteksi hal yang spesifik pada percobaan tertentu
(Hsieh dkk., 2021)	*A CNN-Based Wearable Assistive System for Visually Impaired People Walking Outdoors*	FAST-SCNN dan YOLOv5s	*Voice prompt manager*	**Kelebihan :** Performa dari model hasil pelatihan Fast-SCNN akurasi pixel 95%. Sedangkan untuk model deteksi rintangan yang menggunakan YOLOv5s menunjukkan performa mAP50:95 sebesar 52.6%, mAP50 sebesar 85.1%, dan mAP75 sebesar 58.5% **Kekurangan :** Kecepatan deteksi yang masih kurang dapat membahayakan penggunaa jika tiba-tiba muncul objek di depannya
(Ganesan dkk., 2022)	*Deep Learning Reader for Visually Impaired*	CNN dan LSTM	Text-to-Speech API	**Kelebihan :** Dengan menggunakan Resnet arsitektur performa model menunjukkan hasil akurasi sebesar 83%. Selain itu, sistem dapat melakukan pemahaman terhadap gambar yang dideteksi dengan menambahkan caption pada setiap hasil deteksi yang membuatnya semakin natural. **Kekurangan :** Tidak ada fitur untuk menghitung jarak objek
(Chou dkk., 2023)	*A Lightweight Robust Distance Estimation Method for Navigation Aiding in Unsupervised Environment Using Monocular Camera*	YOLOv4	-	**Kelebihan :** Dengan baik model dapat menghitung jarak objek pada rentang di bawah 10 meter dengan total kesalahan sebesar 4%. Selain itu, penggunaan YOLOv4 membuat

				proses deteksi objek menjadi lebih baik lagi. **Kekurangan :** Saat ini, model hanya dapat mengenali dua jenis objek publik di luar ruangan, yang berarti model hanya dapat berfungsi di luar ruangan. Selain itu, benda bergerak seperti manusia dan mobil juga belum dapat terdeteksi

Penggunaan YOLO untuk model deteksi objek sebagai komponen dari alat pembantu tunanetra menunjukkan kesuksesan yang sangat luar biasa. Dibuktikan dengan beberapa penelitian yang menggunakan model tersebut sebagai basis deteksi objeknya pada Tabel 2.1. Selain itu, penggunaan gTTS masif digunakan sebagai bentuk implementasi untuk TTS pada keluaran model. Dari *state of the art* yang ada dan pernyataan yang diberikan, terciptalah ide penelitian pada proposal ini dengan mengusung model yang memiliki kemampuan deteksi objek, translasi TTS, *speech recognition*, dan menghitung estimasi jarak antara kamera dengan objek. Tabel 2.2 menunjukkan perbandingan capaian yang didapatkan pada penelitian ini dengan capaian yang telah didapatkan oleh penelitian sebelumnya. Perbandingan tersebut mengacu pada dukungan model versi YOLO, *parameter optimization*, *voice feedback*, *speech recognition*, dan *distance estimation*

Tabel 2.2 Perbandingan target capaian penelitian dengan penelitian terkait

Penelitian	YOLO	*Parameter Optimization*	*Voice Feedback*	*Speech Recognition*	*Distance Estimation*
(Guravaiah dkk., 2022)	YOLOv5	-	√	-	-
(Ravindra Karmarkar &	YOLOv1	-	√	-	-

Honmane, 2021)					
(Najm dkk., 2022)	YOLOv3	-	√	-	-
(Mahendru & Dubey, 2021)	YOLOv3	-	√	-	-
(Ramalakshmi dkk., 2020)	YOLOv1	-	√	-	√
(N. Joshi dkk., 2021)	YOLOv3	-	√	-	-
(Alahmadi dkk., 2023)	YOLOv4	-	√	-	-
(Lee & Cho, 2022)	YOLOv3	-	√	-	-
(Kadhim & Oleiwi, 2022)	YOLOv3	-	√	-	-
(Nugraha, 2024)	YOLOv8	√	√	√	√

Tabel 2.2 menunjukkan perbandingan capaian setiap penelitian terkait dengan penelitian yang dilakukan. Pemilihan penelitian terkait pada tabel tersebut didasari pada penggunaan model YOLO yang digunakan sebagai model objek deteksi untuk membantu penyandang tunanetra. Pada penelitian (Guravaiah dkk., 2022) versi YOLO yang digunakan adalah versi 5 dengan dikombinasikan oleh TTS menggunakan gTTS dan pyttsx3 sehingga menghasilkan suatu sistem bernama Third Eye. Selain itu, penambahan 15 kelas baru pada MS COCO *dataset* membuat model yang dihasilkan dapat lebih banyak mengenali banyak objek. (Ramalakshmi dkk., 2020; Ravindra Karmarkar & Honmane, 2021) menggunakan YOLOv1 sebagai dasar dari model objek deteksinya. Meskipun menggunakan YOLO versi pertama, tetapi model tersebut dapat bekerja dengan baik dalam melakukan deteksi pada kelas-kelas yang terdaftar pada *dataset* COCO. Pada penelitian (N. Joshi dkk., 2021; Kadhim & Oleiwi, 2022; Lee & Cho, 2022; Mahendru & Dubey, 2021; Najm

dkk., 2022) model deteksi yang digunakan adalah YOLOv3 dimana versi ini terkenal dengan fleksibilitas dan ketangguhannya dalam hal deteksi objek. Hal ini membuktikan bahwa penggunaan YOLOv3 saat ini masih relevan untuk digunakan. Penambahan fitur feedback berupa text-to-speech menjadi pelengkap dari semua sistem yang dibangun pada penelitian-penelitian tersebut. (Alahmadi dkk., 2023) mengembangkan YOLOv4 dengan mengganti *backbone*-nya dari Darknet menjadi Resnet101. Sehingga dihasilkan sebuah model deteksi yang memiliki performa yang sangat baik dengan rata-rata akurasi sebesar 96.34% pada 16 kelas yang ada pada MS COCO *dataset*.

Penelitian-penelitian terkait tersebut memiliki kekurangan yang identik, yaitu tidak dilakukannya pemilihan parameter yang baik khususnya parameter untuk masalah optimasi. Selain itu, hanya segelintir penelitian terkait yang menggunakan fitur *distance estimation* bahkan tidak ada sama sekali yang menggunakan fitur *speech recognition*. Padahal kedua fitur tersebut dapat meningkatkan kebergunaan dari sistem yang dihasilkan bagi para penyandang tunanetra. Maka dari itu, penelitian ini menghasilkan kebaruan yang tidak dimiliki oleh penelitian-penelitian terkait pada Tabel 2.2. Penelitian ini menerapkan metode *parameter optimization* menggunakan PSO pada pemilihan nilai *learning rate* dan *beta 1 Adam optimization* yang paling optimal bagi proses pelatihan model deteksi. Selain itu, penambahan fitur TTS, *speech recognition*, dan *distance estimation* menggunakan perhitungan *monocular camera* ditambahkan untuk mendukung kegunaan sistem bagi para penyandang tunanetra.