Zero-Shot Recognition of Utility Meter Readings Using Vision-Language Models and General-Purpose OCRs

Luís Gustavo Werle Tozevich[1], Luan Portella da Silva[1], Gabriel Machado Lunardi[1], Thiago Lopes Trugillo da Silveira[1], Adriano Quilião de Oliveira[1]

[1]Federal University of Santa Maria

lgtozevich@inf.ufsm.br, luan.portella@ufsm.br,

{gabriel.lunardi, thiago.silveira, adriano.q.oliveira}@ufsm.br

Abstract. *Supervised Automatic Meter Reading (AMR) methods achieve high accuracy but require costly domain-specific training. Despite the popularity of Vision-Language Models and general Optical Character Recognition systems offering theoretical zero-shot capabilities, their effectiveness on constrained real-world meters remains unexplored. We evaluated the zero-shot performance of seven models (PyTesseract, Keras-OCR, EasyOCR, PARSeq, GOT-OCR 2.0, DeepSeek-VL2 and Qwen2.5-VL) on the UFPR-AMR and Copel-AMR datasets. Results show supervised recognition rates average 95.87%, outperforming the best zero-shot model, which averaged 64.86% across both datasets. This accuracy gap makes generalist models currently unsuitable for AMR.*

1. Introduction

The emergence of large-scale Vision-Language Models (VLMs) is fundamentally shifting computer vision away from task-specific engineering [Yin et al. 2024]. Unlike traditional paradigms that rely on specialized training for constrained tasks, these networks leverage massive pre-training to generalize to unseen domains in a zero-shot regime [Bommasani et al. 2022, Awais et al. 2025]. In service automation, this capability may reduce dependence on custom-engineered pipelines and the costs associated with extensive local data collection and annotation. This zero-shot capability potentially enables the deployment of reading solutions across different utility providers without the need for data annotation [Zhao et al. 2023].

Automatic Meter Reading (AMR) aims to automatically obtain the consumption values displayed by electricity, water, and gas meters for monitoring and billing [Shu et al. 2007, Vanetti et al. 2013, Spichkova et al. 2019]. In image-based AMR, a photograph of the meter is processed to locate the counter region and convert the displayed sequence of digits into machine-readable data [Shu et al. 2007, Vanetti et al. 2013]. Automating this process can reduce the errors and human effort associated with manually transcribing and subsequently verifying meter readings [Shu et al. 2007, Vanetti et al. 2013]. Reliable meter readings support billing and, more broadly, the monitoring and management of smart energy and water systems [Shu et al. 2007, Wang et al. 2019, Zapata-Sierra et al. 2023]. However, recognition under uncontrolled acquisition conditions remains challenging because of image degradation, illumination

and contrast variations, glare, shadows, and partial obstruction [Vanetti et al. 2013]. Electromechanical counters introduce an additional ambiguity when a rotating wheel exposes portions of two adjacent digits [Rodríguez et al. 2014].

Methods proposed to overcome these issues can achieve high accuracy but rely heavily on fully annotated datasets and domain-specific training [Laroca et al. 2019, Laroca et al. 2021]. This reliance restricts these systems' adaptability when deployed across novel meter layouts or unseen operational environments. Although generalist models may theoretically provide the flexibility needed to handle domain shifts, their zero-shot performance on utility meters captured in the wild remains underexplored. To investigate this feasibility, this work evaluates the zero-shot performance of general-purpose Optical Character Recognition (OCR) engines and VLMs applied to the AMR task. Furthermore, we assess the cost-benefit trade-offs of these architectures, analyzing the computational overhead of VLMs and whether they provide accuracy gains over specialized frameworks.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the experimental methodology, Section 4 discusses the results, and Section 5 concludes the work.

2. Related Work

Image-based AMR has been addressed through supervised, multi-stage pipelines that decompose the problem into subtasks. Laroca et al. [Laroca et al. 2019] proposed a pipeline using the Fast-YOLO detector for counter localization, followed by a Convolutional Neural Network (CNN) based approach to jointly segment and recognize the meter digits. This pipeline was extended [Laroca et al. 2021] with corner detection and geometric rectification, enabling spatial normalization of the counter region and early rejection of low-quality inputs before recognition. Despite their effectiveness, such approaches rely on domain-specific training and intermediate stages. In parallel, alternative strategies have explored adapting generic Scene Text Recognition (STR) architectures to the AMR domain. Gómez et al. [Gómez et al. 2018] proposed an end-to-end convolutional model for transcribing meter readings from cropped images, eliminating explicit segmentation. Similarly, recent industrial solutions [Carvalho et al. 2023] employ a multi-stage deep learning pipeline for mobile devices that uses screen detection, text localization, and a Convolutional Recurrent Neural Network (CRNN) to automate digital meter readings.

Another approach to AMR uses general-purpose OCR and STR models as flexible alternatives that require little to no task-specific training. Classical OCR engines, such as Tesseract [Smith 2007], alongside deep learning-based approaches including CRNN-based pipelines and attention-driven architectures, have demonstrated strong performance on unconstrained text recognition benchmarks [Shi et al. 2016, Baek et al. 2019]. More recently, transformer-based models such as PARSeq [Bautista and Atienza 2022] have advanced the field by leveraging permutation-based autoregressive modeling to improve performance across irregular text layouts. In addition, widely used frameworks such as EasyOCR and Keras-OCR integrate detection and recognition modules pre-trained on large-scale scene text datasets, enabling straightforward deployment in diverse applications. Nevertheless, despite their generalization capabilities, these models are predominantly optimized for natural-scene and document text, where variations in font, orientation, and background differ substantially from the patterns observed in utility meters.

Recent VLMs further extend the scope of general-purpose recognition by combining high-capacity visual encoders with large language models trained on large-scale image-text datasets. Architectures such as GOT-OCR 2.0 [Wei et al. 2024], DeepSeek-VL2 [Wu et al. 2024], and Qwen2.5-VL [Bai et al. 2025] demonstrate strong zero-shot capabilities across a wide range of vision-language tasks, including structured document understanding and OCR scenarios. Their ability to model long-range dependencies and implicit linguistic priors suggests potential applicability to meter reading.

However, specialized domains indicate that zero-shot generalization remains limited under strong domain constraints. In medical imaging, large VLMs evaluated under zero-shot and few-shot settings exhibit lower diagnostic accuracy than supervised baselines trained on domain-specific data [Tong et al. 2026]. Likewise, in industrial anomaly detection, large vision-language models struggle to capture localized details due to limited domain-specific knowledge, limiting their effectiveness when compared to tailored approaches [Gu et al. 2024]. These findings suggest that large-scale pretraining alone does not guarantee robustness in constrained visual domains. Previous evaluations closer to our setting remain limited. Spichkova et al. [Spichkova et al. 2019] compared the off-the-shelf AWS Rekognition and Google Cloud Vision services for recognizing utility meter readings. Their study used 30 original images and generated additional test conditions through scaling, blur, gamma correction, and noise. The study predates modern VLMs and did not use public AMR benchmarks or supervised AMR baselines. More recently, Wang et al. [Wang et al. 2025] reported results for 18 foundation models on RPM-10K and proposed MRLM, a task-specific vision-language architecture for pointer-meter reading. Their experiments concern analog dials, and MRLM is trained end-to-end on RPM-10K. Mueez et al. [Mueez et al. 2026] investigated the direct reading of analog gauges with Qwen2.5-VL-7B-Instruct, comparing zero-shot prompting, in-context learning, and QLoRA-based fine-tuning across public and proprietary datasets. Although their work includes a zero-shot evaluation, they also focus on analog pointer gauges and do not examine digit-counter transcription or compare general-purpose OCR with supervised digit-based AMR methods. To our knowledge, no prior study has evaluated general-purpose OCR/STR systems and modern VLMs together on public digit-based AMR datasets.

3. Methodology

This section details the experimental methodology used to evaluate the zero-shot performance of general-purpose OCR systems and VLMs. The evaluation pipeline is structured into four stages: acquiring the input image, extracting the ground-truth (GT) bounding box of the meter region, forwarding this cropped region to the selected models, and finally, decoding the predicted reading. Figure 1 illustrates this workflow.

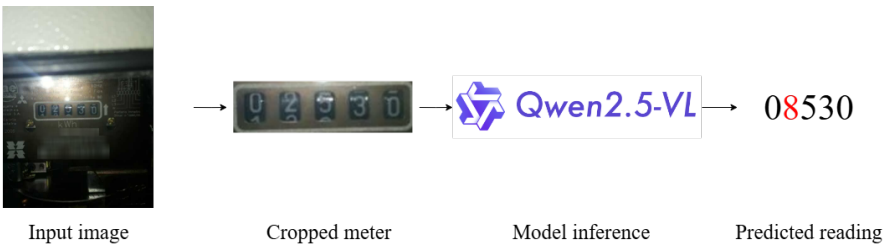


Figure 1. Overview of our experimental pipeline

3.1. Datasets

The models were evaluated on two public datasets, UFPR-AMR [Laroca et al. 2019] and Copel-AMR [Laroca et al. 2021]. UFPR-AMR consists of unconstrained images of meters acquired under varying capture conditions and lighting, whereas Copel-AMR serves as a significantly larger and more challenging benchmark, with meters heavily degraded. Examples of meters from both datasets are shown in Figure 2.

Table 1. Distribution of images across the training, validation, and test splits for the evaluated datasets.

Dataset	Total Images	Train	Validation	Test
UFPR-AMR	2,000	800	400	800
Copel-AMR	12,500	5,000	2,500	5,000

To ensure a direct and valid comparison with the supervised methods, we adopted the same test split proposed by the original authors. The distribution of images across these splits for both datasets is detailed in Table 1.



Figure 2. Examples of meters from the UFPR-AMR and Copel-AMR datasets.

3.2. Model Evaluation

To strictly evaluate the recognition capabilities of the models and isolate them from potential meter detection errors, the counter region is first cropped from the full image using the ground-truth bounding box coordinates provided in each dataset. These cropped meter images are then directly fed into the evaluated models. Laroca et al. [Laroca et al. 2021] demonstrated that their YOLO-based detector achieves near-perfect localization, meaning that missed detections have little effect on final recognition performance. Furthermore, as noted in their experiments, rectifying the counter regions using the ground-truth coordinates improved the baseline recognition rate from 95.87% to 96.53% [Laroca et al. 2021]. By utilizing the aforementioned data splits, the results for the AMR-specific models can be extracted directly from the established benchmarks in [Laroca et al. 2021]. This allows us to evaluate the zero-shot capabilities of the architectures themselves, rather than the performance variations caused by different detection pipelines or domain-specific fine-tuning.

The general-purpose OCR systems (EasyOCR [JaidedAI 2020], Keras-OCR [Morales 2020], PyTesseract [Smith 2007], PARSeq [Bautista and Atienza 2022]) and

VLMs (GOT-OCR 2.0 [Wei et al. 2024], DeepSeek-VL2 [Wu et al. 2024], Qwen2.5-VL [Bai et al. 2025]) are evaluated locally. All local experiments are performed using default settings and original weights, without domain-specific fine-tuning. For the VLMs, the following text prompt is provided alongside the cropped image to guide the zero-shot recognition: *"Transcribe only the meter reading. Return only digits. If the meter is unreadable, faulty, or illegible, return an empty or 1 character output."* The local experiments are executed on a machine equipped with an AMD Ryzen 5 7600 CPU, 32 GB of RAM, and an NVIDIA RTX 4070 SUPER (12 GB VRAM).

3.3. Metrics

Recognition performance is quantified using Word Accuracy (WA) and Character Error Rate (CER). Word Accuracy considers a prediction correct if and only if the entire predicted string perfectly matches the ground-truth sequence. Conversely, CER provides a character-level metric, calculating the Levenshtein distance [Levenshtein et al. 1966] between the predicted string and the ground truth. It is formally defined as

$$CER = \frac{S + D + I}{N}, \tag{1}$$

where S, D, and I denote the number of character substitutions, deletions, and insertions, respectively, and N represents the total number of characters in the ground-truth string. Lower CER values indicate better performance, with 0 representing a perfect prediction.

To analyze the computational overhead, we evaluate inference latency (measured in milliseconds per image), Frames Per Second (FPS), alongside the operational cost represented by giga floating-point operations (GFLOPs) and giga multiply-accumulate operations (GMACs) per inference. Deep learning models were profiled using a standardized input resolution of 224×224 pixels, with the exception of GOT-OCR 2.0, which enforces a 1024×1024 input resolution. Legacy OCR engines, whose architectures do not rely entirely on standard deep learning computational graphs, cannot be accurately profiled via standard tensor operations, their GFLOPs and GMACs are reported as Not Applicable (N/A).

4. Results and Discussion

As shown in Table 2, the results reveal a performance gap between domain-specific supervised architectures and general-purpose zero-shot models for the AMR task. The best supervised baseline outperformed the zero-shot models, achieving an average WA of 95.87%. In contrast, the best-performing zero-shot model, DeepSeek-VL2, achieved only a 64.86% average WA. This discrepancy is particularly pronounced on the Copel-AMR dataset, where severe real-world degradation causes a sharp decline in accuracy across most zero-shot models. PyTesseract and Keras-OCR performed poorly on these constrained layouts, registering average accuracies below 15%. These findings indicate that the broad pre-training of VLMs and generic OCR systems is currently insufficient to overcome the domain shift presented by the highly degraded visual distribution of utility meters.

Beyond recognition accuracy, practical AMR deployment also depends on computational efficiency. Table 3 details the inference frame rate and computational complexity

for each approach. AMR systems typically operate on low-power edge devices and require high-throughput processing. Qwen2.5-VL and DeepSeek-VL2 require 5,388.2 and 1,719.9 GFLOPs, respectively, and achieve fewer than 5 FPS, making them impractical for real-time deployment. While this extreme latency might be acceptable in specific, non-time-critical scenarios, such as asynchronous cloud processing or human-in-the-loop manual inspections where an operator can wait, their adoption would only be justified if accompanied by near-perfect accuracy. The high computational cost of VLMs is not justified by their current accuracy, as they remain significantly less reliable than lightweight, domain-specific models.

Table 2. Recognition performance of the evaluated models.

Model	UFPR-AMR (WA %)	Copel-AMR (WA %)	Average WA %	Average CER
Laroca et al. (Fast-OCR) [Laroca et al. 2021] †	**94.75**%	**96.98**%	**95.87**%	-
Laroca et al. (CR-NET) [Laroca et al. 2019] †	94.25%	95.20%	94.73%	-
YOLOv4 (608×608) [Bochkovskiy et al. 2020] †	94.00%	95.33%	94.67%	-
Fast-YOLOv4-SmallObj + Calefati et al. [Calefati et al. 2019] †	93.87%	95.15%	94.51%	-
Liao et al. (larger input size) [Liao et al. 2019] †	92.25%	94.27%	93.26%	-
YOLOv4 (416×416) [Bochkovskiy et al. 2020] †	91.63%	93.75%	92.69%	-
Fast-YOLOv4-SmallObj + TRBA [Baek et al. 2019] †	93.87%	90.80%	92.34%	-
Liao et al. [Liao et al. 2019] †	91.37%	92.25%	91.81%	-
Laroca et al. (Multi-task) [Laroca et al. 2019] †	89.12%	87.02%	88.07%	-
Fast-YOLOv4-SmallObj + CRNN [Shi et al. 2016] †	89.62%	84.07%	86.85%	-
Fast-YOLOv4-SmallObj + Gómez et al. [Gómez et al. 2018] †	88.25%	83.93%	86.09%	-
DeepSeek-VL2	68.38%	61.34%	64.86%	0.11
PARSeq	83.94%	38.02%	60.98%	0.26
Qwen2.5-VL	74.03%	42.74%	58.39%	0.17
GOT-OCR 2.0	79.67%	34.26%	56.97%	0.27
EasyOCR	39.40%	23.80%	31.60%	0.53
PyTesseract	10.16%	18.24%	14.20%	0.79
Keras-OCR	4.14%	18.54%	11.34%	0.98

† Results extracted from [Laroca et al. 2021].

Table 3. Computational cost and FPS of the evaluated models.

Model	FPS	GFLOPs	GMACs
Laroca et al. (Fast-OCR)	55.00	-	-
Laroca et al. (CR-NET)	41.00	-	-
Qwen2.5-VL	0.15	5388.2	2694.1
DeepSeek-VL2	4.34	1719.9	859.9
GOT-OCR 2.0	3.75	754.5	377.2
Keras-OCR	2.94	13.4	6.7
PARSeq	84.03	4.5	2.3
EasyOCR	39.78	0.4	0.2
PyTesseract	12.56	N/A	N/A

Table 4 details the primary visual artifacts that disrupt zero-shot recognition. Real-world meters frequently present display reflections, dirt, and mechanical digits in mid-transition, which generalist models fail to interpret correctly. For instance, GOT-OCR 2.0 predicted "0.707" for the ground truth "07077". This suggests a linguistic bias where the model, pre-trained on natural text, overfits to decimal patterns when encountering specular reflections that mimic punctuation. Similarly, mechanical counters presenting partially rotated digits introduce structural ambiguity. In the case where the ground truth is "08449", the visual overlap during the mechanical transitions (e.g. between digits 4 and 5 and between digits 9 and 0) causes recognition failures across the evaluated models, resulting in truncated or entirely arbitrary sequences. When presented with faulty

or illegible meters, some large models produce spurious outputs. Qwen2.5-VL returns the sequential string "12345678", while GOT-OCR 2.0 returns contextually irrelevant text such as "OPEN". Unlike domain-specific pipelines that implicitly learn the structural priors of meter counters, zero-shot models frequently misinterpret shadows or lighting anomalies as valid alphanumeric characters.

Table 4. Comparison of predictions generated by the evaluated models. Correct predictions are highlighted in bold.

Image	GT	DeepSeek-VL2	EasyOCR	GOT-OCR 2.0	Keras-OCR	PARSeq	PyTesseract	Qwen2.5-VL
	**19896**	**19896**	9896	9896	5GO5	**19896**	CS	**19896**
	**14540**	2547		1555		2195907	T	3473
	**07077**	**07077**		0.707	LINE	**07077**		**07077**
	**08449**	84	84R	08428	ENAOA	08423		0845
	**faulty/illegible**	-	-	OPEN	-	MON	-	12345678
	**92566**	92567	9256	92564	215169	92567		92567
	**02734**	**02734**	8273	FREE	DTT	02134		2734
	**07996**	7996	7996	1996	1916	**07996**		7996
	**21412**	**21412**	212	21712	ZZA1	**21412**		**21412**
	**07075**	**07075**	01015	**07075**	OLO5	**07075**	O7075	**07075**
	**00823**	**00823**	**00823**	**00823**	ODOTS	**00823**	N	**00823**
	**09484**	9484	0948	**09484**	9O84	**09484**		9484
	**23555**	**23555**	223555	**23555**	V355	**23555**		**23555**
	**16572**	**16572**	**16572**	**16572**	16ST7	**16572**	T	**16572**
	**05366**	**05366**	**05366**	**05366**	DETEE	**05366**	045366	**05366**
	**06985**	**06985**	**06985**	**06985**	89O65	**06985**	**06985**	**06985**

Considering accuracy and computational efficiency, Figure 3 illustrates the speed-to-accuracy trade-off for the evaluated models. Positioned in the optimal upper-left region of the plot, PARSeq provides the most viable balance among the zero-shot baselines. It achieved an average WA of 60.98% while sustaining real-time inference with a computational cost of 4.5 GFLOPs. Nevertheless, while lightweight scene text recognizers like PARSeq successfully meet edge hardware constraints, their recognition rates remain fundamentally inadequate for utility billing systems, reinforcing the necessity of optimized, domain-specific architectures for AMR. Ultimately, these findings indicate that while general-purpose models may reduce the need for costly data annotation, they currently fail to meet the operational standards required for AMR. The visual domain shift from natural-scene text to degraded industrial displays is too severe for off-the-shelf deployment. Until zero-shot models reliably handle structural constraints, noise, and hallucinations, supervised fine-tuning remains essential for real-world AMR.

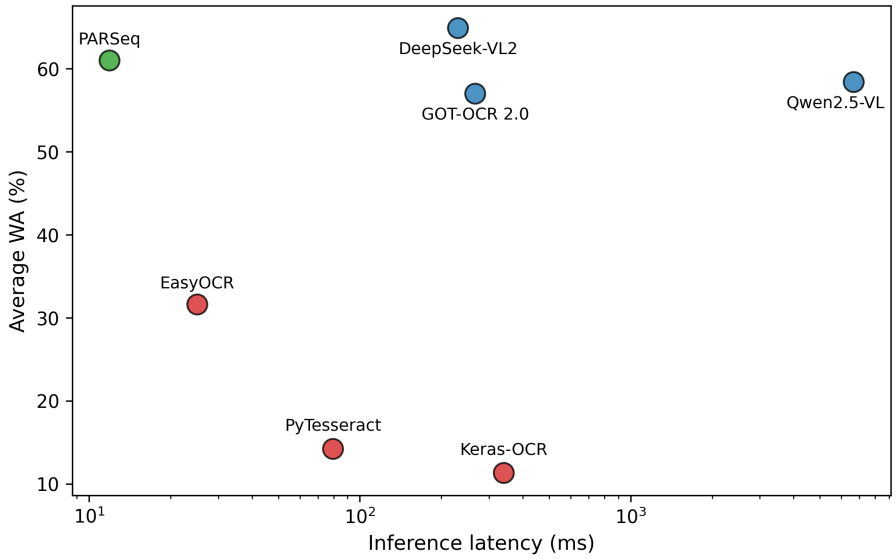


Figure 3. Trade-off between average WA and inference latency (log scale).

5. Conclusion

This work presented an evaluation of the zero-shot capabilities of VLMs and general-purpose OCR systems for AMR. The results demonstrate that while general-purpose architectures have advanced significantly in natural scene text recognition, they currently fail to achieve the reliability required for utility reading. The visual domain shift from unconstrained natural images to unconstrained utility-meter imagery, characterized by severe reflections, mechanical digit transitions, and low-contrast displays, remains a problem for models without domain-specific training.

Overall, supervised AMR pipelines still significantly outperform the most advanced zero-shot models. The top-performing supervised model reached an average WA of 95.87%, while the best zero-shot framework, DeepSeek-VL2, achieved only 64.86%. This performance gap confirms that general-purpose pre-training is insufficient for real-world utility meter reading, particularly in images containing mechanical "half-digits" and severe perspective distortions. Moreover, VLMs introduce a high computational overhead without providing superior accuracy for the AMR task. Qwen2.5-VL requires over 1,190 times more GFLOPs than PARSeq while achieving an average WA that is 2.59 percentage points lower. The high computational overhead and low throughput observed across the evaluated VLMs limit their suitability for edge-based utility meter reading. While lightweight models like PARSeq offer a more viable balance between efficiency and zero-shot performance, their accuracy remains fundamentally inadequate for production environments.

As future work, we intend to improve the reliability of zero-shot and few-shot regimes without fully reverting to costly, domain-specific training. We plan to investigate whether prompt engineering can reduce hallucinations by incorporating task context, such as fixed digit lengths, leading-zero rules, and expected character sets, together with few-shot examples of reflections and mechanical "half-digits". As a subsequent step, we intend to investigate the implementation of constrained decoding or structured outputs to provide safer transcriptions by enforcing digit-only constraints and explicit abstention

when readings are uncertain. We also aim to explore hybrid pipelines that combine geometric priors, such as counter rectification and illumination normalization. Lastly, we plan to include additional datasets and evaluate other VLMs.

References

Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., and Khan, F. S. (2025). Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S. J., and Lee, H. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4715–4723.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. (2025). Qwen2.5-VL technical report.

Bautista, D. and Atienza, R. (2022). Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Calefati, A., Gallo, I., and Nawaz, S. (2019). Reading meter numbers in the wild. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE.

Carvalho, R., Melo, J., Graça, R., Santos, G., and Vasconcelos, M. J. M. (2023). Deep learning-powered system for real-time digital meter reading on edge devices. *Applied Sciences*, 13(4):2315.

Gómez, L., Rusinol, M., and Karatzas, D. (2018). Cutting sayre's knot: reading scene text without segmentation. application to utility meters. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 97–102. IEEE.

Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., and Wang, J. (2024). AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1932–1940.

JaidedAI (2020). EasyOCR: Ready-to-use OCR. `https://github.com/JaidedAI/EasyOCR`.

Laroca, R., Araujo, A. B., Zanlorensi, L. A., De Almeida, E. C., and Menotti, D. (2021). Towards image-based automatic meter reading in unconstrained scenarios: A robust and efficient approach. *IEEE Access*, 9:67569–67584.

Laroca, R., Barroso, V., Diniz, M. A., Gonçalves, G. R., Schwartz, W. R., and Menotti, D. (2019). Convolutional neural networks for automatic meter reading. *Journal of Electronic Imaging*, 28(1):013023–013023.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710. Soviet Union.

Liao, S., Zhou, P., Wang, L., and Su, S. (2019). Reading digital numbers of water meter with deep learning based object detector. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 38–49. Springer.

Morales, F. (2020). Keras-OCR: A packaged and flexible version of the CRAFT text detector and keras CRNN recognition model. `https://github.com/faustomorales/keras-ocr`.

Mueez, A., Baranwal, A., Chaj-Mejia, J., Bhatia, G., Voelker, J. T., and Vyas, S. (2026). Vision-language models for analog gauge reading: An empirical study of specialization, transfer and reliability. *arXiv preprint arXiv:2608.17723*.

Rodríguez, M., Berdugo, G., Jabba, D., Calle, M., and Jimeno, M. (2014). Hd-mr: A new algorithm for number recognition in electrical meters. *Turkish Journal of Electrical Engineering and Computer Sciences*, 22(1):87–96.

Shi, B., Bai, X., and Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Shu, D., Ma, S., and Jing, C. (2007). Study of the automatic reading of watt meter based on image processing technology. In *2007 2nd IEEE Conference on Industrial Electronics and Applications*, pages 2214–2217.

Smith, R. (2007). An overview of the tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Spichkova, M., Van Zyl, J., Sachdev, S., Bhardwaj, A., and Desai, N. (2019). Easy mobile meter reading for non-smart meters: Comparison of aws rekognition and google cloud vision approaches. *arXiv preprint arXiv:1910.12617*.

Tong, R., Liu, J., Wang, T., Hu, X., Liu, S., Wang, L., and Xu, J. (2026). Does bigger mean better? comparative analysis of CNNs and biomedical vision-language models in medical diagnosis. In *2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, pages 1–6.

Vanetti, M., Gallo, I., and Nodari, A. (2013). Gas meter reading from real world images using a multi-net system. *Pattern Recognition Letters*, 34(5):519–526.

Wang, F., Weng, C., Wang, X., Chen, Z., Zhao, Z., and Tang, J. (2025). Dialbench: Towards accurate reading recognition of pointer meter using large foundation models. *arXiv preprint arXiv:2511.21982*.

Wang, Y., Chen, Q., Hong, T., and Kang, C. (2019). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 10(3):3125–3148.

Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al. (2024). General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. (2024). DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Zapata-Sierra, A. J., Salmerón-Manzano, E., Alcayde, A., Zapata-Castillo, M. L., and Manzano-Agugliaro, F. (2023). The scientific landscape of smart water meters: A comprehensive review. *Water*, 16(1):113.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124.