

Deep Learning and Transformer Based Models For Power Transformer Nameplate Information Extraction: A Systematic Comparative Review

Amit kumar Subhashchandra Vishwakarma¹, Ganesh Mhatre², Samadhan Patil³, Dhananjay Labde⁴

¹ Students of Masters of Power System, Yadavrao Tasgaonkar Institute of Engineering & Technology, Karjat.

² Assistant Professor at Department of Mechatronics Engineering, New Horizon Institute of Technology & Management.

³ Head of Department at Department of Electrical Engineering, Yadavrao Tasgaonkar Institute of Engineering & Technology, Karjat.

⁴ Assistant Professor at Department of Electrical Engineering, Yadavrao Tasgaonkar Institute of Engineering & Technology, Karjat.

Corresponding Author(s) Email Id : akvish.4198@gmail.com

Abstract:

Power-transformer nameplates record the ratings and identification data required for asset registers, loading studies, protection settings and electrical audits, yet field photographs of these plates are frequently degraded by specular glare, corrosion, low contrast, perspective distortion and small or engraved characters. This paper presents a systematic comparative review of deep-learning methods for extracting structured information from power-transformer and related electrical-equipment nameplates. Following a documented search of six bibliographic databases with explicit inclusion and exclusion criteria, 53 sources were retained and classified against a five-stage pipeline: image enhancement, text detection, text recognition, structured field extraction and engineering validation. Detection models (CTPN, EAST, CRAFT, DBNet), recognition models spanning CNN–RNN–CTC, attention-based, Transformer-based (e.g., PARSeq, TrOCR) and multimodal document-understanding architectures (LayoutLM family, Donut) are compared using source-reported results, each stated together with its dataset, test-set size, protocol and metric definition. Three principal findings emerge. First, domain-specific nameplate studies use small private datasets and non-standard metric definitions, and some headline figures were not obtained on nameplates at all (e.g., the 98.2% recognition rate of a CTPN–CRNN nameplate pipeline was measured on ICDAR 2017), so these figures cannot be ranked against one another or against public benchmarks. Second, public-benchmark results indicate architectural trends but do not transfer directly to nameplates. Third, the reviewed nameplate studies concentrate on detection and recognition; field association, unit normalization and engineering plausibility are rarely, if ever, evaluated. Because a single-character error such as “11 kV” read as “1.1 kV”, or “Dyn11” read as “Dyn1”, can alter an engineering decision, an engineering-aware evaluation framework is synthesized, comprising field-level exact match, an all-critical-fields-correct rate, normalized numeric error, cross-field consistency checks and confidence-based abstention. Research gaps and a literature-derived evaluation protocol for future studies are identified.

Keywords: Power transformer nameplate, optical character recognition, scene text detection, scene text recognition, Transformer, key information extraction, systematic review, asset management.

I. INTRODUCTION

Power transformers are long-lived, capital-intensive assets, and their nameplates are the primary on-site record of the quantities required to operate and maintain them. IEC 60076-1 specifies the information to be carried on the rating plate, including rated power, rated voltages, rated currents, rated frequency, connection symbol (vector group), short-circuit impedance, type of cooling and insulation levels [1]. Utilities and industrial users transcribe these data into asset registers during commissioning, audits and condition assessments. Manual transcription is slow for large, geographically dispersed fleets and is prone to error, particularly for legacy units whose plates are faded, painted over or corroded. Nameplate data also serve directly as inputs to engineering calculations; for example, the equivalent-circuit parameters of induction machines have been estimated from nameplate data alone [2]. Automated extraction of nameplate data from photographs is therefore of practical interest for asset management, loading studies and protection coordination.

Nameplate reading differs from both conventional scene-text recognition and document understanding. The text is compact, tabular and domain-specific, and it contains many non-semantic alphanumeric tokens, such as serial numbers, vector groups (e.g., Dyn11), cooling codes (e.g., ONAN/ONAF) and compound ratings (e.g., 33/11 kV), for which natural-language priors are of limited use. Image acquisition conditions are also adverse: metallic plates produce specular highlights; outdoor units exhibit dirt, paint fade and corrosion; characters may be stamped or engraved rather than printed; and handheld photographs are frequently non-frontal. Several domain studies have applied deep learning to this problem, including a CTPN–CRNN pipeline whose detection stage was evaluated on 800 real electrical-nameplate images [3], an improved DBNet detector for power nameplates [4], exposure correction for nameplate images [5], a recognition approach reporting 87.71% accuracy on 1,200

power-equipment nameplate images [6], and a recurrent recognizer reporting 99.30% alphanumeric character accuracy on a 30-image test set [7]. A recent systematic review of computer-vision information extraction from electrical assets covers a broader range of asset types and tasks and notes that foundation vision–language models are not yet established in electrical-equipment inspection [8]. Across these studies, datasets are largely private and metrics are defined differently, which prevents a direct ranking of methods.

Terminology. To avoid ambiguity, “power transformer” denotes the electrical apparatus and “Transformer” (capitalized) denotes the self-attention neural architecture [9]. The title therefore refers to both: the application domain is power-transformer nameplates, and the methodological focus is the progression from convolutional–recurrent models to Transformer-based and multimodal models.

This review differs from [8] in scope and emphasis: it concentrates on nameplate information extraction, compares models at the level of individual pipeline stages, and treats engineering-field correctness as the primary evaluation target. Its contributions are as follows:

- 1) a transparent literature search and selection methodology, reported following PRISMA 2020 [10] (Section II);
 - 2) a five-stage taxonomy of nameplate information extraction, with a mapping of the reviewed domain studies to the stages they address (Section III);
 - 3) a stage-wise review of detection and recognition models that separates convolutional–recurrent, attention-based, Transformer-based and multimodal approaches (Sections IV–VI);
 - 4) a contextualized comparison of source-reported results, stating dataset, domain, test-set size, protocol, metric definition and input resolution, and an explicit rubric for the qualitative suitability ratings (Section VII);
 - 5) an engineering-aware evaluation framework for nameplate extraction (Section VIII); and
 - 6) a synthesis of research gaps and a literature-derived protocol for future comparative studies (Section IX).
- Section X concludes the paper.

II. REVIEW METHODOLOGY

A. Research Questions

The review was structured around four research questions (RQs):

RQ1: Which deep-learning architectures have been applied, or are directly applicable, to each stage of nameplate information extraction?

RQ2: What performance has been reported, and on which datasets, under which protocols and with which metrics?

RQ3: How do the architectural families differ in robustness to nameplate-specific degradations and in computational cost?

RQ4: How is, and how should, extraction quality be evaluated at the level of engineering fields rather than characters?

B. Information Sources and Search Strategy

Six bibliographic sources were searched: IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Scopus and Google Scholar. Because many foundational computer-vision papers are published in CVF and AAAI proceedings, the CVF Open Access repository and arXiv were consulted to obtain full texts of peer-reviewed papers; arXiv-only preprints were not included unless no peer-reviewed version existed and the work is widely used as a benchmark resource. Backward and forward citation searching (snowballing) was performed on the domain studies and on the benchmark-analysis paper of Baek et al. [11]. Search strings combined three concept blocks with the Boolean AND operator (Table I). The publication window for application and model literature was January 2012 to September 2026; the search was last updated in September 2026. Foundational methods published before 2012 (LSTM, CTC) were admitted only through snowballing, as methodological background.

TABLE I SEARCH CONCEPT BLOCKS (COMBINED WITH AND; TERMS WITHIN A BLOCK COMBINED WITH OR)

Block	Search terms
B1 Domain	“nameplate” OR “name plate” OR “rating plate” OR “power equipment label” OR “transformer nameplate” OR “electrical equipment nameplate”

B2 Task	"text detection" OR "text recognition" OR "OCR" OR "optical character recognition" OR "scene text" OR "information extraction" OR "key information extraction"
B3 Method	"deep learning" OR "convolutional" OR "recurrent" OR "attention" OR "Transformer" OR "CTPN" OR "CRNN" OR "DBNet" OR "EAST" OR "CRAFT" OR "layout" OR "vision-language"
Model search	B2 AND B3 without B1, restricted to peer-reviewed venues, to identify the original sources of models and benchmarks used by domain studies

The domain query (B1 AND B2 AND B3) targets nameplate studies; the model query (B2 AND B3) targets the original model, dataset and benchmark papers.

C. Eligibility Criteria

Inclusion criteria. IC1: peer-reviewed journal article or conference paper (or a standard or dataset paper that defines a benchmark used by included studies). IC2: the work proposes or evaluates a deep-learning method for at least one pipeline stage (S1–S5, Section III) on nameplates or electrical-equipment labels, or is the original source of a model, dataset or evaluation protocol used by such work. IC3: the work reports quantitative results or a method description sufficient for reproduction. IC4: written in English.

Exclusion criteria. EC1: purely classical OCR (template matching, hand-crafted features) without a deep-learning component, unless used as a stated baseline. EC2: related but structurally different tasks (licence-plate recognition, meter-dial reading, product-label recognition), unless the method is later applied to nameplates. EC3: abstracts, patents, theses, white papers, and web pages. EC4: duplicate publications; when a conference paper was superseded by an extended journal version, only the most complete version was retained. EC5: studies whose reported results could not be traced to a defined dataset.

D. Selection Procedure and Duplicate Removal

Records were exported to a reference manager. Duplicates were removed first by exact DOI match and then by matching normalized title, first-author surname and publication year. Screening proceeded in two stages: title and abstract screening against IC1–IC4 and EC1–EC3, followed by full-text assessment against all criteria. Screening was performed by the author. The selection protocol is summarized in Table II. Applying the criteria excluded, for example, general text-classification and embedding studies without an OCR component, retrieval systems for natural images, CAPTCHA breaking, LLM sentiment-analysis comparisons, and non-peer-reviewed preprints (EC2 and EC3).

TABLE II SUMMARY OF THE SEARCH AND SELECTION PROTOCOL (REPORTED FOLLOWING PRISMA 2020 [10])

Protocol element	Description
Databases searched	IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Scopus, Google Scholar
Supplementary sources	CVF Open Access and arXiv (full texts of peer-reviewed papers); backward and forward citation searching
Search window	January 2012 to September 2026 (earlier foundational works via citation searching only)
Screening	Two stages: title/abstract, then full text; performed by the author
Duplicate handling	DOI match, then normalized title + first author + year; most complete version retained (EC4)
Numerical data	Extracted from results tables or text of each original publication; unverifiable values excluded
Sources included	53 (composition in Table III)

E. Data Extraction and Classification

Each included source was coded on four dimensions: (a) the pipeline stage(s) it addresses (S1–S5); (b) its architecture family (proposal-, regression- or segmentation-based detection; CTC-, attention-, Transformer- or multimodal-based recognition and extraction); (c) its evidence type: domain evidence on nameplate data (D), evidence on public general-purpose benchmarks (G), foundational method (F), or dataset/standard (S); and (d) for every reported number, the dataset, test-set size, training data, input resolution, hardware and metric

definition. Each numerical value quoted in this review was checked against the results table or abstract of the original publication; where the original did not state a parameter, it is marked “n.s.” (not stated). The composition of the resulting corpus is given in Table III.

TABLE III COMPOSITION OF THE REVIEWED CORPUS (n = 53)

Category	Stage	Evidence	n	Sources
Nameplate / electrical-equipment OCR studies	S1–S3	D	5	[3]–[7]
Related reviews and surveys	S1–S4	n/a	4	[8], [12]–[14]
Text detection models	S2	G	7	[15]–[21]
Text recognition models	S1, S3	G	15	[22]–[36]
Benchmark analyses	S3	G	2	[11], [37]
Synthetic training data	S2–S3	G	2	[38], [39]
Foundational sequence and attention methods	S3	F	4	[9], [40]–[42]
Document understanding / key information extraction	S4	G	4	[43]–[46]
Post-OCR correction and hypothesis re-ranking	S4	G	2	[47], [48]
Benchmark datasets	S2–S4	S	6	[49]–[54]
Standards and engineering use of nameplate data	S5	S/D	2	[1], [2]

Evidence: D = domain nameplate data; G = general public benchmark; F = foundational method; S = dataset or standard. The PRISMA reporting guideline [10] is cited for method only and is not counted.

Two features of the corpus are notable. Only five of the 53 sources report experiments on nameplate OCR, and all five use datasets that are not publicly released; in one of them the recognition stage is evaluated on a public benchmark rather than on nameplate images [3]. Conversely, the model literature is evaluated almost entirely on public scene-text and document benchmarks. Much of the analysis that follows therefore concerns how evidence from general benchmarks can, and cannot, be transferred to nameplates.

III. TAXONOMY AND COMPARISON FRAMEWORK

A. Five-Stage Pipeline

The reviewed literature can be organized as a five-stage pipeline (Fig. 1): S1, image enhancement and rectification; S2, text detection; S3, text recognition; S4, structured field extraction (field association and semantic normalization); and S5, engineering validation. Each stage has a distinct input, output and appropriate metric, and errors propagate forward: a nameplate system may fail to produce a correct engineering record even when character recognition is accurate, for example because a value is associated with the wrong label or its unit is lost. The established scene-text literature likewise cautions against comparing systems whose training data, lexicons and test protocols differ [11]. Accordingly, this review keeps detection F-measure, word/character recognition accuracy and structured-field accuracy separate rather than merging them into a single score. Table IV formalizes the taxonomy.

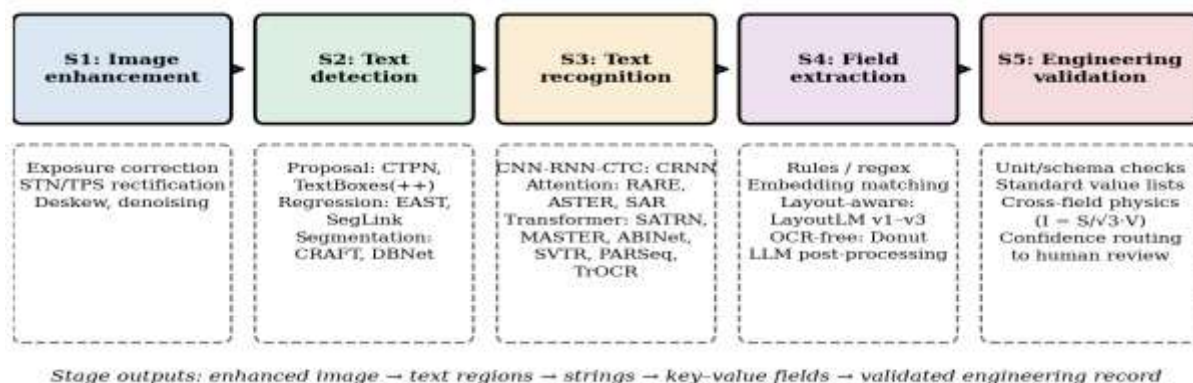


Fig. 1. Five-stage taxonomy of nameplate information extraction synthesized from the reviewed literature, with representative model families at each stage.

TABLE IV TAXONOMY OF APPROACHES FOR NAMEPLATE INFORMATION EXTRACTION

Stage	Representative methods	Input → Output	Advantages	Limitations	Stage metric
S1 Enhancement	Exposure correction [5]; STN/TPS rectification [25], [26]	Image → enhanced image	Improves visibility; reduces perspective distortion	May introduce artifacts; benefit to OCR seldom quantified	PSNR, SSIM, PI; ΔCER downstream
S2 Detection	CTPN, TextBoxes++, SegLink, EAST, CRAFT, DBNet [15], [17]–[21]	Image → text regions	Localizes text lines/words	Sensitive to glare, small and dense text	Precision, recall, F-measure; FPS
S3 Recognition	CRNN; RARE/ASTER; SAR, SRN, ABINet; SATRN, MASTER, SVTR, PARSeq, TrOCR	Cropped region → string	Segmentation-free, variable-length reading	Domain shift; language-prior over-correction	Word accuracy, CER, WER
S4 Field extraction	Rules/regex; embeddings [13], [14]; LayoutLM v1–v3, Donut [43]–[46]; LLM post-processing [12], [48]	Strings + boxes (or image) → key-value fields	Links values to labels; tolerates label variation	Needs key-value annotations; higher compute	Field P/R/F1, exact match
S5 Validation	Unit/schema checks; standard values [1]; cross-field consistency; confidence calibration	Fields → validated record or review flag	Catches implausible values before reporting	Requires domain rules; rarely evaluated	All-critical-fields-correct; accuracy-coverage

B. Stage Coverage of Domain Studies

Table V maps the nameplate-specific studies to the stages they address. The mapping is based on the methods and results reported in each publication. Every domain study addresses at least one of S1–S3, but none reports a quantitative evaluation of S4 or S5. The step from transcribed text to a validated engineering record, which determines whether the output is usable in an audit, is therefore the least studied part of the pipeline.

TABLE V PIPELINE STAGES ADDRESSED BY NAMEPLATE-SPECIFIC STUDIES

Study	S1	S2	S3	S4	S5	Remarks
Wang et al. [3]	–	●	●	–	–	CTPN detection on 800 nameplate images; CRNN recognition evaluated on ICDAR 2017
Zhao et al. [4]	–	●	–	–	–	Improved DBNet detector
Wu et al. [5]	●	○	○	–	–	Image-quality gains measured; OCR benefit shown qualitatively only
Zhong et al. [6]	–	●	●	–	–	Recognition accuracy on 1,200 test images
Zhang et al. [7]	–	–	●	–	–	LSTM-based character recognition; 30 test images
Alvarez-Murillo et al. [8]	○	○	○	○	–	Systematic review of electrical-asset vision (66 studies)

● = addressed and evaluated quantitatively on nameplate data; ● = evaluated quantitatively, but on non-nameplate data; ○ = discussed or shown qualitatively; – = not addressed.

C. Stage-Specific Evaluation Criteria

Detection (S2) is conventionally reported as precision, recall and F-measure at an intersection-over-union threshold, typically 0.5. Recognition (S3) is reported as word accuracy or as character error rate (CER) and word error rate (WER):

$$\text{CER} = (S + D + I) / N \quad (1)$$

where S, D and I are the numbers of character substitutions, deletions and insertions required to transform the prediction into the reference, and N is the number of reference characters; WER is defined analogously over words. Structured extraction (S4) is reported as field-level precision, recall and F1, and exact-match accuracy. Engineering validation (S5) requires metrics that reflect the consequences of errors; these are developed in Section VIII. Latency (mean and 95th percentile), parameter count and memory footprint are reported alongside all stages, since they constrain deployment on laptops, mobile devices or edge hardware.

IV. IMAGE ENHANCEMENT AND TEXT DETECTION

A. Image Enhancement and Rectification (S1)

Image quality is a primary cause of failure in nameplate reading. Wu et al. constructed a nameplate exposure dataset of 5,000 images generated from 1,000 original photographs at five simulated exposure levels (4,000/200/800 training/validation/test split) and reported improvements of 5.6%, 5.1% and 7.96% in PSNR, SSIM and perceptual index, respectively, relative to the LMPEC baseline [5]. The authors argue that corrected images facilitate subsequent text detection and recognition, but the downstream effect is illustrated qualitatively rather than measured with a detector or recognizer. The quantitative effect of enhancement on nameplate OCR accuracy is therefore still undetermined. Geometric rectification is addressed in the recognition literature by spatial transformer networks with thin-plate-spline (TPS) transformations, learned jointly with the recognizer in RARE [25] and ASTER [26]; these reduce the effect of oblique viewing angles, which are common in handheld nameplate photographs.

B. Text Detection (S2)

Proposal-based detectors. CTPN detects text lines by predicting fixed-width vertical proposals and connecting them with a recurrent layer [15]. Its line-oriented design suits approximately horizontal nameplate rows, and it is the detector used in the domain pipeline of Wang et al. [3]. On 800 real nameplate images (8:1:1 split), that study reports precision 0.85, recall 0.82 and F-measure 0.67 for CTPN, compared with 0.80/0.80/0.64 for EAST and 0.76/0.75/0.65 for RRPN. The reported F-measures are lower than both precision and recall, which is inconsistent with the usual harmonic mean and indicates a different weighting; they are therefore quoted as reported. The original paper reports an F-measure of 0.88 on ICDAR 2013 [49] and 0.61 on the more difficult ICDAR 2015 incidental-text benchmark [50], at approximately 0.14 s per image; the drop reflects its horizontal-

line assumption. TextBoxes [16] adapts a single-shot object detector to text with long default boxes, and TextBoxes++ [17] extends it to oriented quadrilaterals. SegLink [18] decomposes text into oriented segments and links, allowing detection of long, multi-oriented lines.

Regression-based detectors. EAST directly regresses rotated rectangles or quadrilaterals from a fully convolutional network and reports an F-score of 0.782 on ICDAR 2015 at 13.2 frames per second (FPS) at 720p resolution [19]. The single-stage design simplifies deployment; however, direct regression is constrained by the receptive field for long text lines and does not explicitly model individual characters.

Segmentation-based detectors. CRAFT predicts character-region and inter-character affinity maps, trained with weak character-level supervision, and groups characters into words [20]. It reports H-means of 95.2, 86.9 and 82.9 on ICDAR 2013, ICDAR 2015 and MSRA-TD500 [51], respectively, and 8.6 FPS on ICDAR 2013. Its character-level representation is, by design, tolerant of uneven spacing, embossing and partial occlusion. DBNet inserts a differentiable binarization module into a segmentation network so that the binarization threshold is learned [21]. With a ResNet-18 backbone and a 736-pixel short side, it reports an F-measure of 82.8 at 62 FPS on MSRA-TD500 and 82.3 at 48 FPS on ICDAR 2015 (single GTX 1080 Ti). The segmentation formulation accommodates arbitrarily shaped text boundaries, and the lightweight backbone reduces computation. Zhao et al. adapted DBNet specifically to power nameplates [4].

Comparative reading. ICDAR 2015 is the only benchmark on which all four detector families report results, and its incidental images (blurred, low-resolution and captured without deliberate framing) resemble handheld nameplate photographs more closely than the focused images of ICDAR 2013. On this benchmark the reported F-measures increase from 0.61 (CTPN) to 0.782 (EAST), 0.823 (DBNet-R18) and 0.869 (CRAFT). The ordering indicates the capability gained by oriented and segmentation-based formulations. Notably, the ordering of CTPN and EAST is reversed on the predominantly horizontal nameplate images of [3], which illustrates that benchmark rankings do not transfer directly to nameplates; the associated speed figures, however, were obtained on different hardware and input sizes and are indicative only.

V. TEXT RECOGNITION: FROM CNN–RNN TO TRANSFORMER MODELS

Recognition architectures have progressed through four broad generations (Fig. 2): convolutional–recurrent models with CTC decoding; rectification- and attention-based encoder–decoders; models with explicit semantic or language reasoning; and Transformer-based recognizers. Multimodal document-understanding models, which fuse text, layout and image, extend this progression towards structured extraction and are discussed in Section VI. Table VI summarizes the characteristics of each generation that matter for nameplates.

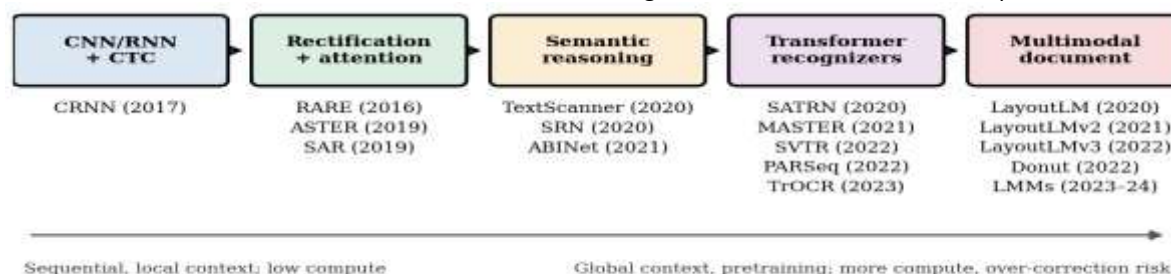


Fig. 2. Progression of text-recognition architectures reviewed in this paper, from CNN–RNN–CTC models to Transformer-based and multimodal document models.

A. CNN–RNN–CTC Recognizers

CRNN combines convolutional feature extraction, a bidirectional LSTM [40] for sequence modeling and connectionist temporal classification (CTC) [41] for alignment-free decoding [23]. It does not require character segmentation, handles variable-length sequences and is compact (approximately 8.3 M parameters as reported in [23]), which makes it suitable for tokens such as “33/11 kV”, “ONAN/ONAF” and “Dyn11”. Its limitations are an implicit assumption of approximately horizontal text, recurrent computation that limits parallelism, and short-range context modeling. CRNN is used in the domain pipeline of [3], where replacing the VGG16 backbone with ResNet and DenseNet raised the recognition rate from 91.3% to 93.2% and 98.2%, at 0.054, 0.058 and 0.060 s per sample; these figures were obtained on the ICDAR 2017 dataset, not on nameplate images. Fully convolutional CTC recognizers without recurrent layers, such as that of Yousef et al., remove the sequential bottleneck of the BiLSTM while retaining alignment-free decoding [24]. Zhang et al. applied an LSTM-based recurrent recognizer to power-equipment nameplates, training on 9,300 images (300 field photographs and 9,000 web images, augmented) and reporting 99.30% alphanumeric and 99.90% Chinese-character accuracy on

30 test images, at 2.15 ms per image [7]. Baek et al. showed that the transformation, feature-extraction, sequence-modeling and prediction stages each affect accuracy, speed and memory differently, so that recognizers should be reported by full configuration rather than by model name alone [11].

B. Rectification and Attention-Based Recognizers

Attention-based encoder–decoders replace CTC with a decoder that attends to relevant image regions at each output step. RARE [25] and ASTER [26] precede the recognizer with a learned TPS rectification network, improving tolerance to perspective distortion and curved text. SAR applies two-dimensional attention over ResNet features, avoiding the need to flatten irregular text into a one-dimensional sequence [27]. TextScanner predicts character segmentation and order maps and reads characters in parallel [28]. These models improve robustness to geometric distortion but retain autoregressive decoding and can suffer attention drift on long strings.

C. Semantic-Reasoning Recognizers

A further group adds explicit linguistic context. SRN introduces a global semantic reasoning module [29], and ABINet decouples a vision model from a bidirectional language model that iteratively refines the prediction [31]. Such models recover occluded or degraded characters when the surrounding context is predictive, as in natural-language words. For nameplates the benefit is conditional: many critical tokens are non-semantic, and a language prior trained on ordinary words can replace a correctly perceived but unusual token with a more frequent one. Zhang et al. showed that attention-based recognizers over-rely on learned context: their accuracy drops sharply on out-of-vocabulary words even when the images are clear and undistorted, a problem they mitigate with context-based contrastive learning [36]. Almost every nameplate token (ratings, serial numbers, codes) is out-of-vocabulary in this sense. OCRBench, a 29-dataset evaluation of large multimodal models, similarly reports that these models perform poorly on non-semantic character strings [37].

D. Transformer-Based Recognizers

The Transformer replaces recurrence with self-attention [9], and the Vision Transformer applies the same mechanism to image patches [42]. SATRN uses a two-dimensional self-attention encoder for text of arbitrary shape [30]. MASTER replaces recurrent components with multi-aspect non-local self-attention [32]. SVTR removes the separate sequence model and recognizes text with a single visual model that mixes local and global features [33]. PARSeq trains an ensemble of autoregressive models with permutation language modeling, which enables both context-free (non-autoregressive) and context-aware decoding with iterative refinement in a single model [34]. It reports 91.9% word accuracy when trained on synthetic data and 96.0% when trained on real data, on the combined set of six standard benchmarks (7,672 cropped words, including IIIT5K [52] and ICDAR 2013/2015), with approximately 23.6 M parameters for the base model. TrOCR couples a pretrained image-Transformer encoder with a pretrained text-Transformer decoder [35]; the Large variant (558 M parameters) reports word accuracies of 94.1%, 96.1% and 88.1% on IIIT5K, SVT and ICDAR 2015 (1,811-word protocol), and a word-level F1 of 96.58 on the printed-receipt SROIE task [54].

For nameplates, two properties are relevant. First, PARSeq’s context-free decoding mode allows the language prior to be switched off for non-semantic fields such as serial numbers and vector groups, providing a direct way to control over-correction. Second, TrOCR’s printed-text results on SROIE are closer to the visual regime of printed nameplates than scene-text results are, although neither benchmark contains metallic, reflective or engraved plates. The principal limitations are computational cost (especially for TrOCR), sensitivity to domain shift without fine-tuning data, and token substitution caused by pretrained language priors.

TABLE VI PROGRESSION OF RECOGNITION ARCHITECTURES AND NAMEPLATE-RELEVANT CHARACTERISTICS

AR = autoregressive; NAR = non-autoregressive.

Generation	Representative models	Sequence modeling / decoding	Pretraining	Nameplate-relevant strength	Principal risk
CNN–RNN–CTC	CRNN [23]	BiLSTM; CTC	None (synthetic data)	Compact; alignment-free; easy to profile	Horizontal-text assumption; short context
Rectification + attention	RARE, ASTER, SAR, TextScanner	Attention decoder, 1-D or 2-D	None	Tolerates perspective and curvature	Attention drift; sequential decoding

Semantic reasoning	SRN, ABINet	Explicit language model; iterative refinement	Language model on text corpora	Recovers degraded characters in words	Over-corrects non-semantic tokens
Transformer	SATRN, MASTER, SVTR, PARSeq, TrOCR	Self-attention; AR or NAR decoding	Synthetic/real text; image and text pretraining (TrOCR)	Parallel, global context; transfer learning	Compute; domain shift; token substitution
Multimodal document	LayoutLM v1–v3, Donut	Joint text–layout–image Transformer	Large document corpora	Direct key–value output	Needs labeled fields; generative errors

VI. STRUCTURED FIELD EXTRACTION AND SEMANTIC NORMALIZATION

Nameplate automation is complete only when transcribed text has been converted into typed key value fields with units. This stage is where the reviewed domain literature is thinnest (Table V).

A. Rule-Based and Embedding-Based Mapping

The simplest approach applies regular expressions for numeric values with units (voltage, power, current, frequency, impedance), controlled vocabularies for categorical fields (cooling type, vector group) and spatial proximity between label and value boxes. Dense text embeddings extend matching beyond exact strings, allowing label variants such as “Rated Power”, “Capacity” and “kVA Rating” to be mapped to a single field [13], [14]. Embedding similarity is designed to capture semantic proximity; it is not designed to discriminate numeric magnitude or unit prefix, so “11 kV” and “1.1 kV” may be close in embedding space. Embeddings can also support recognition directly: Bansal et al. fuse word-image and text embeddings to select among multiple recognizer hypotheses, improving word accuracy [47]. Embeddings are therefore suited to label matching and hypothesis re-ranking, while values require deterministic parsing and validation.

B. Layout-Aware and OCR-Free Document Models

LayoutLM jointly pretrains on text and two-dimensional token positions and was evaluated on form understanding and receipt information extraction [43]. LayoutLMv2 and LayoutLMv3 add visual features and unified text–image masking [44], [45]. Donut removes the external OCR stage and maps a document image directly to a structured output sequence with a Transformer encoder–decoder [46]. The key information extraction protocols of FUNSD (199 annotated scanned forms) [53] and SROIE (scanned receipts) [54] illustrate how field-level evaluation is conducted. A nameplate is, in effect, a small semi-structured document of label–value pairs, often arranged as a table, so these models are applicable in principle. No nameplate key-information dataset exists, however, and receipts and forms differ from nameplates in material, font, degradation and vocabulary; the transfer of these models to nameplates has not yet been demonstrated.

C. Post-OCR Correction and Large Language Models

This subsection separates published evidence from proposed system designs. Published evidence: the survey of Nguyen et al. shows that post-OCR correction can reduce recognition errors but can also introduce substitutions when context is misleading [12]. Thomas et al. reported that an instruction-tuned Llama 2 model reduced character error rate by 54.51% on 19th-century British newspaper text, compared with 23.30% for a BART baseline [48]. OCRBench reports weaknesses of large multimodal models on non-semantic character strings, handwriting and multilingual text [37], and the review of [8] finds that foundation vision–language models are not yet established for electrical-equipment inspection. Implication: the positive evidence for LLM-based correction concerns natural-language text, whereas nameplate fields are predominantly numeric or code-like. No study in the reviewed corpus reports the effect of LLM post-processing on nameplate field accuracy.

Proposed designs. Applied systems increasingly route OCR output to a general-purpose LLM service for field mapping and validation. On current evidence, such a component should be treated as a hypothesis to be tested rather than as an established improvement. Its contribution should be isolated by an ablation on a common test set (OCR only; OCR with rule-based normalization; OCR with embedding retrieval; and OCR with LLM-assisted mapping), using the field-level metrics of Section VIII. Supplying the language model with text coordinates and recognition confidences, and constraining its output to a typed schema, limits the scope for unsupported substitutions.

VII. COMPARATIVE ANALYSIS OF REPORTED PERFORMANCE

A. Source-Reported Results in Context

Table VII lists representative source-reported results together with the information required to interpret them. The table is not a ranking. Rows that share a dataset and metric (for example, the ICDAR 2015 detection rows) are comparable to a first approximation, subject to differences in backbone, input size and hardware. Rows with different datasets or tasks are not comparable, and the domain-specific rows are not comparable with each other because their datasets and metric definitions differ.

TABLE VII SOURCE-REPORTED PERFORMANCE WITH DATASET, PROTOCOL AND METRIC CONTEXT

Model (ref.)	Stage	Dataset (domain; test size)	Training / protocol	Metric	Input / hardware	Reported result
CTPN [3]	S2	Electrical nameplates (private); 800 images, 8:1:1 split (≈ 80 test)	Compared with EAST and RRPN on same split	P / R / F-measure (weighted, $\alpha = 0.5$)	n.s.	0.85 / 0.82 / 0.67
CRNN-DenseNet [3]	S3	ICDAR 2017 (public; not nameplates)	Backbone comparison (VGG16, ResNet, DenseNet)	Recognition rate (correct / total)	0.060 s/sample	98.2% (VGG16: 91.3%)
Recognition model [6]	S2+S3	Power-equipment nameplates (private); 1,200 images	n.s.	Recognition accuracy	n.s.	87.71%
LSTM-based RNN [7]	S3	Power-equipment nameplates (private); 30 images	9,300 training images (300 field + 9,000 web), augmented	Character accuracy	2.15 ms/image	99.30% (alphanumeric); 99.90% (Chinese)
LMPEC-based correction [5]	S1	Nameplate exposure set; 800 test images (1,000 originals $\times$ 5 exposures)	4,000/200/800 split	PSNR, SSIM, PI (relative)	n.s.	+5.6%, +5.1%, 7.96% better PI vs. LMPEC
CTPN [15]	S2	ICDAR 2013 (focused scene; 233) / ICDAR 2015 (incidental; 500)	Standard splits	F-measure	GPU; ≈ 0.14 s/image	0.88 / 0.61
EAST [19]	S2	ICDAR 2015 (incidental; 500)	Standard split	F-score	720p	0.782; 13.2 FPS
CRAFT [20]	S2	ICDAR 2013 / ICDAR 2015 /	Weak character supervision	H-mean	n.s.; 8.6 FPS on IC13	95.2 / 86.9 / 82.9

		MSRA-TD500 (multi-oriented; 200)				
DBNet-R18 [21]	S2	MSRA-TD500 (200) / ICDAR 2015 (500)	Standard splits	F-measure	736 px short side; GTX 1080 Ti	82.8 @ 62 FPS / 82.3 @ 48 FPS
PARSeq-S [34]	S3	Six scene-text benchmarks combined (7,672 cropped words)	Synthetic vs. real training data	Word accuracy	≈23.6 M parameters	91.9% (synthetic); 96.0% (real)
TrOCR-Large [35]	S3	IIIT5K (3,000) / SVT (647) / ICDAR 2015 (1,811 words)	Synthetic + benchmark data	Word accuracy	558 M parameters	94.1 / 96.1 / 88.1%
TrOCR-Large [35]	S3	SROIE printed receipts (word-level task)	Fine-tuned	Word-level F1	558 M parameters	96.58

n.s. = not stated in the source. Detection F-measures are at IoU ≥ 0.5 under each benchmark's official protocol. Speeds were measured on different hardware and are indicative only.

Three observations follow. First, the domain figures require careful reading. The 98.2% recognition rate associated with the CTPN–CRNN nameplate pipeline was measured on ICDAR 2017, not on nameplates, while its nameplate evaluation covers detection only, on approximately 80 test images [3]. The 99.30% of [7] rests on a 30-image test set. With 30 independent test items, even a perfect score has a 95% Clopper–Pearson lower confidence bound of approximately 88%, so small test sets cannot establish generalization across manufacturers, plate materials and capture conditions. Second, a model that scores lower on a generic benchmark may perform better on nameplates after domain fine-tuning, because the vocabulary, layout and character set are constrained; conversely, high accuracy on clean cropped words does not guarantee reliable localization of small text on a reflective plate. Third, no reviewed study reports results at the field level. The unit of comparison appropriate to this application, namely the full image-to-fields pipeline, has not yet been used in the literature.

B. Qualitative Suitability: Rating Rubric

Because no common nameplate benchmark exists, Tables VIII and IX provide qualitative suitability ratings. These ratings are the authors' synthesis. They are derived from the architectural properties and source-reported results described above, using the explicit rules below, and were not measured on nameplate data. The rules are intended to make the ratings reproducible and open to challenge.

Detectors. Orientation/skew: High if the model outputs quadrilaterals or polygons and reports results on a multi-oriented benchmark (ICDAR 2015 or MSRA-TD500); Medium if orientation is supported only partially; Low if output is restricted to horizontal boxes. Degraded/incidental text: based on the reported ICDAR 2015 F-measure, used as a proxy for blurred, low-resolution, unframed images: High ≥ 0.80 , Medium 0.70–0.80, Low < 0.70 . Small/dense text: High if the output is a character- or pixel-level segmentation map; Medium if boxes are regressed or proposed at feature-map resolution. Compute: from the reported throughput: Low ≥ 30 FPS, Medium 8–30 FPS, High < 8 FPS.

Recognizers. Irregular text: High if the model includes learned rectification or two-dimensional attention/self-attention; Low if it reads a one-dimensional sequence without rectification. Contextual recovery: High with an explicit language model or iterative refinement; Medium with implicit context (BiLSTM, attention decoder or

global mixing). Over-correction risk on non-semantic tokens: High if a pretrained or explicit language prior cannot be disabled at inference; Medium if it can be disabled or is only implicit; Low if there is no language prior. Compute: from the reported parameter count where available (R), with Low < 30 M, Medium 30–100 M and High > 100 M; otherwise estimated from the architectural composition (A).

TABLE VIII RUBRIC-BASED SUITABILITY OF DETECTORS FOR NAMEPLATE IMAGES (AUTHOR-DERIVED)

Detector	Orientation / skew	Degraded text (IC15 F)	Small / dense text	Compute	Basis
CTPN [15]	Low	Low (0.61)	Medium	High (≈7 FPS)	Horizontal line proposals
EAST [19]	High	Medium (0.782)	Medium	Medium (13.2 FPS)	Rotated-box regression
CRAFT [20]	High	High (0.869)	High	Medium (8.6 FPS)	Character region + affinity maps
DBNet-R18 [21]	High	High (0.823)	High	Low (48–62 FPS)	Segmentation, learned binarization

TABLE IX RUBRIC-BASED SUITABILITY OF RECOGNITION AND EXTRACTION MODELS (AUTHOR-DERIVED)

Model	Irregular text	Contextual recovery	Over-correction risk	Compute	Suggested role
CRNN [23]	Low	Medium	Low	Low (R)	Baseline for regular rows
ASTER [26]	High	Medium	Medium	Medium (A)	Oblique or curved text
ABINet [31]	High	High	High	Medium (A)	Degraded natural-language labels
SVTR [33]	High	Medium	Low	Low–Medium (A)	Efficient recognizer without language prior
PARSeq-S [34]	High	High	Medium	Low (R)	Accuracy-oriented; prior can be disabled
TrOCR-Large [35]	High	High	High	High (R)	Transfer learning on printed text
Donut [46]	High	High	High	High (A)	OCR-free key-value extraction

R = parameter count reported in the source; A = estimated from architectural composition. Ratings indicate expected relative behavior and are hypotheses to be tested on nameplate data.

C. Model-Selection Trade-Offs

A valid comparison must distinguish three kinds of performance: detection quality, transcription quality and engineering-field correctness. A detector may localize every text row yet miss a short segment such as the vector group; a recognizer may achieve a low CER yet corrupt a safety-relevant digit; a semantic mapper may convert a correct transcription into an incorrect field. Architecture selection is therefore a multi-objective problem. For audit applications, field exact match and the ability to abstain are the primary criteria, CER/WER and detector F-measure are diagnostics, and latency, memory and retrainability are constraints.

Within these constraints the literature suggests the following. CTPN remains a relevant reference because it anchors the domain literature [3], and on predominantly horizontal nameplate images it outperformed EAST and RRPN in the only published comparison [3]; its reported ICDAR 2015 result, however, indicates limited tolerance of oblique capture. EAST offers a simple single-stage design with moderate speed. CRAFT reports the highest ICDAR 2015 H-mean among the reviewed detectors and represents text at character level, which is advantageous for fragmented or unevenly spaced characters, at lower throughput. DBNet combines arbitrary-shape segmentation with the highest reported throughput among the reviewed detectors, and it is the only one

of the general-purpose detectors that has been adapted specifically to power nameplates [4]. Among recognizers, CRNN offers the simplest, most easily profiled baseline, whereas PARSeq and TrOCR offer stronger contextual modeling at the cost of requiring careful control of language priors on engineering tokens. These statements describe reported characteristics and are hypotheses for nameplate evaluation, not claims that any single model is universally superior.

VIII. ENGINEERING-AWARE EVALUATION OF NAMEPLATE EXTRACTION

The feature that most clearly distinguishes nameplate extraction from generic scene-text recognition is that the cost of an error depends on its engineering meaning, not on its edit distance. This section develops that point into an evaluation framework.

A. Consequences of Character-Level Errors

Reading “11 kV” as “1.1 kV” is a single-character insertion, a small CER contribution, but it assigns the transformer to the wrong voltage class by a factor of ten. Reading “Dyn11” as “Dyn1” is a single-character deletion, but the clock number defines the phase displacement between windings in multiples of 30° [1]: Dyn11 corresponds to 330° (the low-voltage side leading by 30°) and Dyn1 to 30° (lagging by 30°). Transformers with different phase displacements cannot be operated in parallel, so this error affects switching decisions directly. Table X lists the critical fields, typical errors, their consequences, and validation rules that can detect them.

TABLE X CRITICAL NAMEPLATE FIELDS, REPRESENTATIVE ERRORS AND VALIDATION RULES

Field	Typical form	Representative critical error	Engineering consequence	Validation rule
Rated power	500 kVA; 31.5 MVA	31.5 → 3.15; kVA ↔ MVA	Order-of-magnitude error in capacity and loading studies	Unit in {kVA, MVA}; consistency with rated current
Rated voltages (HV/LV)	33/11 kV; 11000/433 V	11 → 1.1; 433 → 43.3	Wrong voltage class; incorrect protection settings	Standard voltage list; HV > LV; ratio consistency
Rated currents	26.2/666.7 A	Decimal-point loss	Incorrect loading and protection limits	$I \approx S/(\sqrt{3} \cdot V)$ within tolerance
Vector group	Dyn11; YNd1; Yyn0	Dyn11 → Dyn1	Phase displacement 330° vs. 30°; parallel-operation error	Grammar of winding letters and clock number 0–11
Short-circuit impedance	4.5%; 10.0%	4.5 → 45	Incorrect fault-level and load-sharing calculations	Plausible range for the rating class
Frequency	50 Hz; 60 Hz	50 → 50 / 60	Mismatch with system frequency	Value in {50, 60}
Cooling type	ONAN; ONAF	ONAN ↔ ONAF	Wrong rating stage used for loading	Controlled vocabulary of cooling codes
Serial number	Alphanumeric	O ↔ 0, I ↔ 1, S ↔ 5	Record linked to wrong asset	Exact match only; no language-model correction

B. Cross-Field Consistency

Nameplate fields are physically related, which provides a validation signal unavailable in generic OCR. For a three-phase transformer the rated current on each side satisfies

$$I_r = S_r / (\sqrt{3} \cdot V_r) \quad (2)$$

where S_r is the rated power and V_r the rated line voltage of that winding. For a 500 kVA, 11/0.433 kV unit, (2) gives approximately 26.2 A and 666.7 A for the HV and LV windings. If the HV voltage is misread as 1.1 kV, (2) yields about 262 A, which is inconsistent with the printed 26.2 A, and the error can be flagged even though each field is individually well formed. Similar checks apply to the voltage ratio and to the agreement between the vector-group letters and the winding connections stated elsewhere on the plate.

C. Field-Level Metrics

Let N be the number of test nameplates, C the set of critical fields, $y_{i,f}$ the normalized ground-truth value of field f on plate i , and $\hat{y}_{i,f}$ the prediction. Let $\delta[\cdot]$ equal 1 when its argument is true and 0 otherwise. The following metrics are proposed:

- **Field exact match** after normalization of units and formatting:

$$EM_f = (1/N) \sum_i \delta[\hat{y}_{i,f} = y_{i,f}] \quad (3)$$

- **All-critical-fields-correct rate**, the fraction of plates for which every critical field is correct, i.e., the proportion of records that could enter an asset register without correction:

$$ACFC = (1/N) \sum_i \prod_{f \in C} \delta[\hat{y}_{i,f} = y_{i,f}] \quad (4)$$

- **Normalized numeric error** for continuous ratings, $|\hat{y} - y|/|y|$, reported together with the rate of order-of-magnitude errors, $\delta[|\log_{10}(\hat{y}/y)| \geq 1]$, which isolates decimal-point and unit-prefix failures.
- **Unit correctness, schema completeness** and the **consistency-violation rate** under rules such as (2).
- **Selective-prediction performance**: field accuracy plotted against coverage as low-confidence fields are routed to human verification, and the expected calibration error of the confidence scores.

In practice, silent errors are more costly than explicit uncertainty. A system with slightly lower coverage that reliably identifies its uncertain fields may be more useful than a nominally more accurate system that produces confident but incorrect values. A field can be accepted automatically when visual recognition confidence, schema validity and engineering-range checks agree; otherwise it should be flagged, and the original image crop displayed beside the extracted value for rapid verification. None of the reviewed nameplate studies reports metrics of this kind (Table V).

IX. RESEARCH GAPS AND FUTURE DIRECTIONS

A. Identified Gaps

G1 : No public benchmark. All five domain studies use private data. General OCR datasets contain street scenes, receipts and forms, not metallic nameplates. A useful public benchmark would contain full images, text polygons, transcriptions, field labels, label-value links, normalized ground-truth records with units, and degradation metadata (illumination, blur, viewing angle, corrosion, character height). Training, validation and test splits should be disjoint by manufacturer template to prevent near-duplicate layouts from leaking across splits. Synthetic data have enabled large-scale scene-text training [22], [38], [39]; synthetic nameplates would additionally need to preserve nameplate syntax, units, vector-group codes, manufacturer layouts and realistic degradation, and should supplement rather than replace real-world validation.

G2 : Degradation-stratified evaluation. Results are reported as single aggregate figures. Robustness should be reported as the performance drop between clean and degraded subsets, stratified by degradation type. The downstream effect of enhancement (S1) on OCR accuracy, asserted but not measured in [5], should be quantified.

G3 : Structured extraction and validation. Stages S4 and S5 are essentially unevaluated for nameplates. Transfer of layout-aware and OCR-free models [45], [46] to nameplates, and the engineering-aware metrics of Section VIII, are open research directions.

G4 : Evidence on language-model post-processing. Published gains from LLM-based correction concern natural-language text [48], and multimodal models are weak on non-semantic strings [37]. Controlled ablations on nameplate fields are needed before LLM components can be recommended.

G5 : Calibration and human-in-the-loop operation. No reviewed study reports confidence calibration or accuracy-coverage behavior, although these determine whether a system can be deployed safely.

G6 : Reporting and reproducibility. Domain studies seldom report input resolution, split composition, hardware or metric definitions, and do not always state which dataset a headline figure refers to, which prevents comparison (Table VII).

B. Literature-Derived Evaluation Protocol for Future Studies

The gaps above suggest a protocol that future comparative studies could adopt. It is derived from the reviewed literature and is offered as a recommendation; no experimental results are reported in this review.

- 1) **Candidate pipelines**, evaluated under identical data splits, field schema and post-processing: P1, CTPN + CRNN, the domain reference [3]; P2, EAST + CRNN, a fast oriented baseline; P3, DBNet + CRNN, an efficient segmentation-based baseline; and P4, DBNet + PARSeq or DBNet + TrOCR, representing Transformer-based recognition. CRAFT + PARSeq can be added where perspective distortion or irregular spacing is prominent, and Donut as an OCR-free comparison.
- 2) **Data**: multiple manufacturers, plate materials and layouts; controlled augmentation (rotation, perspective, motion blur, noise, under- and overexposure, local glare); polygon or box annotations, line or word transcriptions, and a structured ground-truth record with units for each image.
- 3) **Metrics**: detection precision, recall and F-measure; CER and WER; field-level precision, recall, F1 and exact match; ACFC; normalized numeric and order-of-magnitude error; mean and 95th-percentile latency; model size and memory; performance drop under degradation; bootstrap confidence intervals.
- 4) **Ablation of S4**: OCR only; with rule-based normalization; with embedding retrieval; with LLM-assisted mapping, all on the same test set.
- 5) **Error taxonomy**: missed text, character substitution, incorrect grouping, incorrect field association, unit-normalization error and validation failure.
- 6) **Reproducibility record**: image resolution, preprocessing parameters, split composition, augmentation probabilities, optimizer and learning rate, backbone initialization, decoding parameters, random seeds, batch size and hardware; archived checkpoints and annotation schema where permitted.

X. CONCLUSION

This review examined deep-learning methods for power-transformer nameplate information extraction through a documented search and selection process and a five-stage taxonomy. The reviewed literature supports four conclusions. First, performance cannot be reduced to a single ranking: domain-specific figures are obtained on small private datasets with unstandardized protocols, and in at least one case the headline recognition rate was measured on a public benchmark rather than on nameplates, while public-benchmark results measure related but different tasks. Second, public benchmarks show consistent architectural trends: segmentation-based detectors such as CRAFT and DBNet report stronger results on incidental, multi-oriented text than proposal-based CTPN, and Transformer-based recognizers such as PARSeq and TrOCR report higher accuracy than CNN–RNN–CTC models. However, these trends remain to be confirmed on nameplates. Third, language priors that help on natural-language text can harm non-semantic engineering tokens, so their use should be controllable and evaluated. Fourth, the reviewed nameplate studies stop at detection and recognition; field association, unit normalization and engineering validation, which determine whether an extracted record is usable, are largely unevaluated. Progress in the field therefore depends less on selecting the highest-scoring model than on a public nameplate benchmark and on engineering-aware, field-level evaluation of the kind outlined in Sections VIII and IX.

REFERENCES

- [1] Power Transformers – Part 1: General, IEC Standard 60076-1, Ed. 3.0, 2011.
- [2] M. I. Mossad, M. Azab, and A. Abu-Siada, “A novel evolutionary technique to estimate induction machine parameters from name plate data,” in Proc. 22nd Int. Conf. Electrical Machines (ICEM), Lausanne, Switzerland, 2016.
- [3] Y. Wang, Y. Du, Z. Pang, J. Qi, and H. Ge, “The text detection and recognition method for electrical nameplates based on deep learning,” in Electronic Engineering and Informatics: Proc. 5th Int. Conf. Electronic Engineering and Informatics (EEI 2023), 2024, pp. 589–598, doi: 10.3233/ATDE240123.
- [4] Y. Zhao, Y. Xiong, and W. Li, “Text detection of power nameplate based on improved DBNet,” in Proc. China Automation Congress (CAC), 2023, pp. 4788–4793, doi: 10.1109/CAC59555.2023.10451112.
- [5] H. Wu, Y. Liu, Z. Jin, and Y. Zhou, “Exposure image correction of electrical equipment nameplate based on the LMPEC algorithm,” PLoS ONE, vol. 19, no. 6, Art. no. e0300792, 2024, doi: 10.1371/journal.pone.0300792.
- [6] Y. Zhong, G. Ruan, E. Abozinadah, and J. Jiang, “Least-squares method and deep learning in the identification and analysis of name-plates of power equipment,” Appl. Math. Nonlinear Sci., vol. 7, no. 1, pp. 103–112, 2022, doi: 10.2478/amns.2021.1.00055.

- [7] X. Zhang, W. Bai, and H. Cui, "Optical character recognition of power equipment nameplate for energy systems based on recurrent neural network," *Front. Energy Res.*, vol. 9, Art. no. 834283, 2022, doi: 10.3389/fenrg.2021.834283.
- [8] C. Alvarez-Murillo, A. Ospina-Bedoya, J. W. Branch-Bedoya, G. D. Zapata Madrigal, and R. García Sierra, "Computer vision-based information extraction from electrical assets: A systematic literature review," *Sensors*, vol. 26, no. 16, Art. no. 5009, 2026, doi: 10.3390/s26165009.
- [9] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [10] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.
- [11] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF ICCV*, 2019, pp. 4715–4723, doi: 10.1109/ICCV.2019.00481.
- [12] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of post-OCR processing approaches," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–37, 2021.
- [13] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A survey of text representation and embedding techniques in NLP," *IEEE Access*, vol. 11, pp. 36120–36146, 2023.
- [14] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: A review," *Artif. Intell. Rev.*, vol. 56, no. 9, pp. 10345–10425, 2023.
- [15] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. ECCV*, 2016, pp. 56–72.
- [16] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI*, vol. 31, no. 1, pp. 4161–4167, 2017.
- [17] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, 2018, doi: 10.1109/TIP.2018.2825107.
- [18] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE CVPR*, 2017, pp. 2550–2558.
- [19] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE CVPR*, 2017, pp. 5551–5560, doi: 10.1109/CVPR.2017.283.
- [20] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE/CVF CVPR*, 2019, pp. 9365–9374, doi: 10.1109/CVPR.2019.00959.
- [21] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI*, vol. 34, no. 7, pp. 11474–11481, 2020, doi: 10.1609/aaai.v34i07.6812.
- [22] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, pp. 1–20, 2016.
- [23] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017.
- [24] M. Yousef, K. F. Hussain, and U. S. Mohammed, "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks," *Pattern Recognit.*, vol. 108, Art. no. 107482, 2020.
- [25] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE CVPR*, 2016, pp. 4168–4176, doi: 10.1109/CVPR.2016.452.
- [26] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, 2019, doi: 10.1109/TPAMI.2018.2848939.
- [27] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proc. AAAI*, vol. 33, no. 1, pp. 8610–8617, 2019.
- [28] Z. Wan, M. He, H. Chen, X. Bai, and C. Yao, "TextScanner: Reading characters in order for robust scene text recognition," in *Proc. AAAI*, vol. 34, no. 7, pp. 12120–12127, 2020.
- [29] D. Yu, X. Li, C. Zhang, T. Liu, J. Han, J. Liu, and E. Ding, "Towards accurate scene text recognition with semantic reasoning networks," in *Proc. IEEE/CVF CVPR*, 2020, pp. 12113–12122.
- [30] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in *Proc. IEEE/CVF CVPR Workshops*, 2020, pp. 546–547.
- [31] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proc. IEEE/CVF CVPR*, 2021, pp. 7098–7107.

- [32] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao, and X. Bai, "MASTER: Multi-aspect non-local network for scene text recognition," *Pattern Recognit.*, vol. 117, Art. no. 107980, 2021, doi: 10.1016/j.patcog.2021.107980.
- [33] Y. Du, Z. Chen, C. Jia, X. Yin, T. Zheng, C. Li, Y. Du, and Y.-G. Jiang, "SVTR: Scene text recognition with a single visual model," in *Proc. IJCAI*, 2022, pp. 884–890, doi: 10.24963/ijcai.2022/124.
- [34] D. Bautista and R. Atienza, "Scene text recognition with permuted autoregressive sequence models," in *Proc. ECCV*, 2022, pp. 178–196, doi: 10.1007/978-3-031-19815-1_11.
- [35] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. AAAI*, vol. 37, no. 11, pp. 13094–13102, 2023, doi: 10.1609/aaai.v37i11.26538.
- [36] X. Zhang, B. Zhu, X. Yao, Q. Sun, R. Li, and B. Yu, "Context-based contrastive learning for scene text recognition," in *Proc. AAAI*, vol. 36, no. 3, pp. 3353–3361, 2022.
- [37] Y. Liu et al., "OCR Bench: On the hidden mystery of OCR in large multimodal models," *Sci. China Inf. Sci.*, vol. 67, no. 12, Art. no. 220102, 2024, doi: 10.1007/s11432-024-4235-6.
- [38] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv:1406.2227*, 2014.
- [39] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE CVPR*, 2016, pp. 2315–2324.
- [40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [41] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.
- [42] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [43] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in *Proc. ACM SIGKDD*, 2020, pp. 1192–1200.
- [44] Y. Xu et al., "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding," in *Proc. ACL-IJCNLP*, 2021, pp. 2579–2591.
- [45] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for document AI with unified text and image masking," in *Proc. ACM Multimedia*, 2022, pp. 4083–4091.
- [46] G. Kim et al., "OCR-free document understanding transformer," in *Proc. ECCV*, 2022, doi: 10.1007/978-3-031-19815-1_29.
- [47] S. Bansal, P. Krishnan, and C. V. Jawahar, "Improving word recognition using multiple hypotheses and deep embeddings," in *Proc. 25th Int. Conf. Pattern Recognition (ICPR)*, 2021, doi: 10.1109/ICPR48806.2021.9412417.
- [48] A. Thomas, R. Gaizauskas, and H. Lu, "Leveraging LLMs for post-OCR correction of historical newspapers," in *Proc. 3rd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING 2024*, Torino, Italy, 2024, pp. 116–121.
- [49] D. Karatzas et al., "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, 2013, pp. 1484–1493.
- [50] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in *Proc. ICDAR*, 2015, pp. 1156–1160.
- [51] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE CVPR*, 2012, pp. 1083–1090.
- [52] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *Proc. BMVC*, 2012.
- [53] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "FUNSD: A dataset for form understanding in noisy scanned documents," in *Proc. ICDAR Workshops*, 2019, doi: 10.1109/ICDARW.2019.10029.
- [54] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar, "ICDAR2019 competition on scanned receipt OCR and information extraction," in *Proc. ICDAR*, 2019, pp. 1516–1520, doi: 10.1109/ICDAR.2019.00244.