MBNA Publishing House Constanta 2026



Proceedings of the International Scientific Conference SEA-CONF

SEA-CONF PAPER • OPEN ACCESS

Dataset-Driven Evolution and Complexity Distribution in Scene Text Detection Methods

To cite this article: Milutinovici Sorin, Racuciu Ciprian, Proceedings of the International Scientific Conference SEA-CONF 2026, pg. 93-107.

Available online at www.anmb.ro

ISSN: 2457-144X; ISSN-L: 2457-144X

doi: 10.21279/2457-144X-26-012
SEA-CONF© 2025. This work is licensed under the CC BY-NC-SA 4.0 License

Dataset-Driven Evolution and Complexity Distribution in Scene Text Detection Methods

Milutinovici Sorin[1,*], Racuciu Ciprian[2,]
[1]"Titu Maiorescu" University, Faculty of informatics, sorin.milutinovici@prof.utm.ro
[2] Military Technical Academy, Doctoral School
*** sorin.milutinovici@prof.utm.ro**

Abstract. This paper investigates the relationship between architectural design and dataset evolution in the field of scene text detection. While progress is often viewed as a linear sequence of architectural improvements, we argue that methods are primarily shaped by the geometric complexity of the benchmarks they target. We propose a two-stage conceptual framework that understands detection as a composite of a prediction stage (the model) and a reconstruction stage (post-processing). To quantify this, we introduce the Structural Complexity Quotient (SCQ), a heuristic metric that evaluates methods based on prediction density, feature aggregation, representational degrees of freedom, and post-processing burden. Our analysis reveals that increasing dataset complexity—from horizontal to multi-oriented and curved text—does not merely necessitate more powerful models but instead drives a redistribution of structural complexity between the predictive model and the reconstruction pipeline..
Keywords: scene text detection; geometric representation, dataset complexity

1. Introduction

Scene text detection has evolved into a sophisticated field characterized by architectures optimized for precision, resilience, and computational speed. The objective of scene text detection is the identification of text regions in an image, without assigning semantic content, which is handled in a subsequent recognition stage.

The development of scene text detection methods follows two parallel influences. On one hand, it inherits ideas from general object detection, where successive architectures such as Faster R-CNN (Ren et al., 2015), YOLO (Redmon et al., 2016), SSD (Liu et al., 2016), DETR (Carion et al., 2020) and many more define dominant paradigms. On the other hand, it is driven by the evolution of datasets, which progressively introduce more complex geometric configurations of text and the shift from English-centric to multilingual and script-heavy datasets (e.g., Chinese or Arabic). In practice, new methods often emerge by adapting the latest object detection framework to text detection, followed by modifications to the output representation and post-processing to accommodate the specific structure of text.

This leads to a recurring development pattern: newly introduced datasets expose limitations of existing models, triggering the adaptation of more recent object detection frameworks and the introduction of more complex outputs and reconstruction procedures to handle the new geometric constraints.

Early adaptations of object detection methods included Deep Features for Text Spotting (Jaderberg et al., 2014), TextBoxes (Liao et al., 2016) and TextBoxes++ (Liao et al., 2018). These methods retain the core object detection architecture while introducing adjustments such as elongated anchors, oriented boxes, or mask outputs. However, they remain constrained by the assumption that objects are compact

and well approximated by rectangular regions, which is often not true for text instances that may be elongated, fragmented, or curved.

As datasets evolved from predominantly horizontal text (e.g., ICDAR 2013) to multi-oriented (ICDAR 2015), and later to curved and irregular text (Total-Text, CTW1500), new methods introduced representations better aligned with these geometric properties. Approaches such as EAST (Zhou et al., 2017), CRAFT (Baek et al., 2019), PSENet (Wang et al., 2019a), PAN (Wang et al., 2019b), DBNet (Liao et al., 2020), and TextSnake (Long et al., 2018) moved away from direct region prediction and instead produce dense or relational outputs that require reconstruction through grouping, expansion, or geometric decoding procedures.

More recently, transformer-based methods have been introduced, again following developments in general object detection. While typically retaining convolutional backbones for feature extraction, models such as DETR (Carion et al., 2020) and its adaptation TextDETR (Zhu et al., 2021) reformulate detection as a set prediction problem, eliminating components such as anchor design and non-maximum suppression. This continues the same pattern: a new detection paradigm is adopted, while adjustments are made to the representation and decoding process to handle text-specific geometry. Transformer-based methods were amongst the first that were developed less as a need to tackle a yet more complicated dataset and more as a reaction to a paradigm shift hat has redefined state-of-the-art performance across the entire deep learning landscape.

While developments in scene text detection are often framed as a linear sequence of architectural breakthroughs (Zhu et al., 2020), this perspective overlooks the fundamental role of dataset geometry. Methods are rarely superseded; instead, they align with specific regimes of geometric and linguistic complexity—ranging from horizontal English text to dense, multilingual curves. These challenges do not simply demand 'more power,' but rather necessitate a specialized distribution of complexity between a model's prediction stage and its subsequent reconstruction stage.

In many frameworks, post-processing is presented as a secondary detail. In practice, however, the model and its reconstruction logic are interdependent; the model must be specifically trained to generate signals that the decoder can interpret. Increasing the expressiveness of a model's output—such as moving from bounding boxes to vector fields—does not eliminate complexity but merely redistributes it. While region-based methods concentrate complexity within the model for easy decoding, dense and field-based approaches shift the burden to the reconstruction stage.

To better contextualize this relationship, we propose the **Structural Complexity Quotient (SCQ)** as a taxonomic guide. While our objective is to move toward a more rigorous, quantitative index, the current lack of standardized reporting regarding computational constraints—such as inference latency and memory overhead—across the full spectrum of reviewed methodologies necessitates a qualitative framework at this stage. In this capacity, the SCQ serves to categorize how different architectures distribute their structural effort between instance prediction and geometric reconstruction. By distinguishing the model's ability to localize a text object from its capacity to synthesize its specific shape, this approach highlights that progress in the field is not a simple linear advancement, but rather a series of structural adaptations to the increasing geometric variety of modern benchmarks. This framework provides a practical lens for evaluating current state-of-the-art methods and offers a useful heuristic for the design of future detection pipelines.

2. Scene Text Detection as a Two-Stage Process

Scene text detection systems can be understood as a composition of two distinct but interdependent stages: a **prediction stage**, in which a model produces intermediate outputs from an input image, and a **reconstruction stage**, in which these outputs are transformed into final text instances.

Given an input image I, a detection system can be expressed as:

$$Y = R(M(I)) \tag{1}$$

where *M* denotes the predictive model and *R* denotes the reconstruction procedure. The output Y represents the set of detected text instances, typically encoded as geometric objects.

The output of the model is not, in general, the result but an intermediate representation whose interpretation depends on the reconstruction stage. The complexity of a scene text detection system is therefore not confined to the model *M*, but is distributed across both components.

2.1. The Prediction Stage

The prediction stage extracts a structured set of outputs from the input image, defining the granularity of the model's initial spatial reasoning. The type and shape of these outputs vary significantly depending on the way the detection problem is formulated. The density of information generated by the backbone prior to instance formation (henceforth named **Prediction Density**) varies from relatively low, in region-based methods (ex: Faster R-CNN (Ren et al., 2015), TextBoxes (Liao et al., 2016), and TextBoxes++ (Liao et al., 2018) to very high in dense prediction methods, where the backbone produces spatially distributed outputs, such as probability maps, affinity maps, or geometric fields. Examples include EAST (Zhou et al., 2017), CRAFT (Baek et al., 2019), PSENet (Wang et al., 2019a), and DBNet (Liao et al., 2020).

Field-based approaches extend this idea further by encoding geometric relationships at the pixel level, such as direction vectors or distances to boundaries. This includes direction vectors in TextSnake (Long et al., 2018), frequency-domain coefficients in FCENet (Zhu et al., 2021), and morphological expansion maps in MorphText (Xu et al., 2024). In these formulations, the model output is a mathematical abstraction of the text's shape. These representations are highly expressive—allowing for the capture of extreme curvature and irregular boundaries—they leave a significant effort for the reconstruction stage - text instances must be 'synthesized' from these descriptors with mathematically intensive processes that are highly sensitive to noise, weak signals, or ambiguous local patterns.

Transformer-based methods like TextDETR (Zhu et al., 2021) replace the 'predict-and-filter' workflow with a set-prediction strategy. Instead of generating thousands of overlapping candidates that require Non-Maximum Suppression (NMS), the model uses object queries to directly identify a specific set of text instances. The 'reasoning' required to group pixels into complex shapes is handled internally by the model's attention layers rather than by external post-processing scripts. Consequently, while the final decoding is simple, the architectural complexity is moved into the transformer's internal circuitry.

2.2. The Reconstruction Stage

The reconstruction stage is responsible for interpreting raw predictive outputs and mapping them onto explicit geometric regions, such as bounding boxes, polygons, or parametric curves. Rather than a monolithic post-processing step, this stage can be viewed more as a structural continuum composed of three distinct components:

Feature Aggregation (N): Often referred to as the "Neck", this component handles the fusion of multi-scale semantic information. The complexity of the neck (N) varies from simple pass-through layers to sophisticated Feature Pyramid Networks (FPN) or Transformer Encoders. These structures are critical for maintaining spatial fidelity and ensuring that text instances are accurately captured across varying scales.

Representational Head (H): The "Head" dictates the geometric language of the detection. Its complexity is determined by the degrees of freedom needed to define a text instance. While a standard regression head for horizontal boxes is structurally minimal, heads designed for arbitrary geometry - such as those utilizing high-order Fourier descriptors or iterative query-based decoders - represent a significant increase in the model's internal reasoning burden.

Reconstruction and Consolidation (D): The final component involves the algorithmic logic required to generate localized text instances. For explicit representations, this logic is often reduced to basic filtering (e.g., non-maximum suppression). For implicit or relational representations, however, this stage involves computationally intensive procedures such as pixel aggregation, geometric tracing, or iterative scale expansion.

2.3. Distribution of Complexity

The formulation above leads to a key observation: the complexity of a scene text detection system is distributed between the prediction stage and the reconstruction stage, and different methods allocate this complexity differently. In methods that produce explicit, instance-level predictions, the model M is responsible for capturing most of the structure of the problem. The reconstruction stage R is correspondingly simple. In contrast, methods that rely on dense or implicit representations shift part of the structural burden to the reconstruction stage. The model produces richer but less explicit outputs, and the reconstruction procedure must perform additional work to extract coherent instances.

This distribution reflects a fundamental design choice. Increasing the expressiveness of the model's output does not eliminate complexity; it redistributes it across the system. Importantly, this redistribution does not necessarily reduce the complexity of the model itself: the underlying convolutional architecture typically remains comparable in size and capacity, while the nature of its outputs and the associated reconstruction procedures change.

3. Structural Complexity Scoring

While the inherent complexity of an architecture is difficult to quantify precisely, we propose a heuristic scoring strategy designed to approximate the 'structural weight' of the pipeline. We attribute a score between 1 and 5 to the four key stages - prediction, feature aggregation, representation, and reconstruction with the hope to facilitate a more nuanced comparison between disparate methods.

3.1. Scoring strategy

Prediction Density and Type (P) This component reflects both the spatial density of predictions and the richness of supervision:

1. Global regression (e.g., a single bounding box per region)
2. Coarse anchor-based regression (e.g., early anchor-based detectors)
3. Sparse or point-based prediction (e.g., key point-style detection)
4. Dense pixel-wise prediction (e.g., segmentation or probability maps)
5. Multi-task dense prediction (e.g., joint prediction of complementary fields such as regions and affinities)

*Neck Complexity (N)*This component captures the complexity of feature aggregation and multi-scale fusion:

1. Direct path (no neck) (direct use of backbone features)
2. Simple skip connections or single-scale processing (linear fusion)
3. Standard feature pyramid (top-down hierarchical fusion)
4. Bi-directional or heavy multi-scale fusion or re-weighting mechanisms
5. Iterative or recursive refinement mechanisms

Head and Representation Degrees of Freedom (H) This component measures the expressiveness of the geometric representation:

1. Low-dimensional (e.g., horizontal bounding boxes)
2. Moderate-dimensional (e.g., rotated rectangles or quadrilaterals)
3. Structured representations (e.g., centerlines with local attributes)
4. High-order parametric representations (e.g., Fourier descriptors)
5. Generative or query-based representations (e.g., transformer decoders)

Post-processing Complexity (D) This component reflects the complexity of reconstructing instances from predictions:

1. Minimal processing (e.g., thresholding and non-maximum suppression)
2. Moderate grouping (e.g., box merging or polygon NMS)
3. Efficient region extraction (e.g., differentiable binarization or fast connected components)
4. Iterative expansion or region growing (e.g., kernel expansion, watershed)

5. Complex reconstruction (e.g., graph-based grouping or inverse geometric transforms)

3.2. Scoring calculation

The overall structural complexity of a method is quantified using the **Structural Complexity Quotient (SCQ)**

$$SCQ = \frac{(P \times H)+(N+D)}{10} \quad (2)$$

The formulation reflects two distinct contributions:

- The product P × H captures the **information density** of the model. Dense predictions combined with high-dimensional representations increase the intrinsic difficulty of the prediction task. We use a product to reflect that prediction density and representational degrees-of-freedom act as mutual constraints**.** In a dense prediction paradigm, any increase in the complexity of the output representation is amplified across the entire spatial grid, leading to a combinatorial expansion of the state space the model must learn
- The sum N+D captures the **infrastructure cost**, including feature aggregation and post-processing required to obtain final detections. These two components are treated as additive, as they represent sequential stages of the pipeline where complexity scales linearly rather than exponentially

Finally, we divide the result by a constant factor of 10 to normalize the SCQ into a manageable value. This scaling does not alter the relative ranking of the methods but serves to produce a human-readable coefficient, where values typically range from 0.4 (minimalist architectures) to 3.0 (mathematically intensive representations), facilitating a clearer visualization of the complexity-to-performance trade-off.

This formulation emphasizes that increasing representational capacity, especially when coupled with dense prediction, leads to a disproportionate increase in structural complexity, while architectural and reconstruction components contribute additively.

3.3. Scoring results

Scoring results are presented in the Table 2. In attributing the score components, the following design choices have been made:

DeepDirectRegression (P=1, N=1, D=1): This method regresses coordinates directly from a global feature map, it has no dense predictions or complex post-processing

TextBoxes (D=1) vs. TextBoxes++ (D=2): Both methods are anchor-based but we assign a higher D to TextBoxes++ because it shifts the representation to multi-oriented quadrilaterals, which requires more complex filtering and refinement.

WordSup, WeText, MOTD-FCN, and STD-H (P=4): These methods are based on dense pixel-level or segmentation-based mechanisms. Even if the final output collapses to bounding boxes, the model must still perform intense spatial reasoning during the prediction stage.

CTPN (N=3): Although its basic parameters may suggest a lower score, CTPN adds the additional weight of the Bi-LSTM (sequence modeling) used to connect text snippets, which is a more complex architecture than standard pure-CNN regression.

EAST (P=4, H=2, D=2): EAST maintains a relatively low-dimensional representation by regressing simple geometric offsets, even when using a dense FCN-style prediction, Also the reconstruction is relatively straightforward.

DBNet (H=2, D=3): DBNet D score is elevated to reflect the algorithmic complexity of the Differentiable Binarization process compared to simple thresholding.

CRAFT (P=5): This architecture is assigned the maximum prediction density because it explicitly performs multi-task dense prediction (region maps and affinity maps).

FCENet (P=4, H=4, D=5): This method receives one of the highest overall scores. The complexity is driven by the use of high-order Fourier descriptors and the mathematical intensity of the Inverse Fourier Transform required to reconstruct text contours.

PSENet (N=5, D=4): PSENet gains complexity through its infrastructure. The high scores in the Neck and Post-processing stages are motivated by its iterative "Progressive Scale Expansion" logic.

TextDETR (H=5): As a Transformer-based model, it utilizes a generative, query-based decoder. While this simplifies post-processing (D=2) by removing the need for NMS, it shifts the computation into the model's internal attention mechanisms.

MorphText (D=4): This method achieves a high score due to its multi-stage morphological reconstruction pipeline. While its representational complexity is lower than fully structured curve models like TextSnake, the cost of assembling instances through morphological operations is significant

Table 2. SCQ values and their base coeficients for some of the scene text detection methods.

Method	P	N	H	D	SCQ	Method	P	N	H	D	SCQ
DeepDirectRegression (He et al, 2017)	1	1	2	1	0.4	LRANet (Su et al, 2024)	3	4	4	4	2
TextBoxes (Liao et al, 2016)	3	2	2	1	0.9	FFPN (Wu et al, 2024)	4	4	2	3	2.1
TextBoxes++ (Liao et al, 2016)	3	2	2	2	1	TCAM (Xiao et al, 2019)	4	4	2	3	2.1
WordSup (Hu et al, 2017)	4	2	2	2	1.2	PAN (Wang et al., 2019b)	4	4	2	3	2.1
WeText (Tian et al, 2017)	4	2	2	2	1.2	Mask R-CNN (Huang et al, 2019)	4	4	3	2	2.1
Motd-FCN (Zhang et al, 2016)	4	2	2	2	1.2	CRAFT (Baek et al., 2019)	5	3	3	3	2.1
STD-H (Long et al, 2014)	4	2	2	2	1.2	FCENet (Zhu et al. 2021)	4	3	4	5	2.1
CTPN (Tian et al, 2016)	3	3	2	2	1.3	PSENet (Wang et al., 2019a)	4	4	2	4	2.2
EAST (Zhou et al, 2017)	4	3	2	2	1.6	TextDETR (Zhu et al., 2021)	3	5	5	2	2.2
DBNet (Liao et al., 2020)	4	3	2	3	1.7	MorphText (Xu et al., 2024)	4	4	3	4	2.3
PixelLink (Deng et al, 2018)	3	4	2	4	1.8	TextSnake (Long et al., 2018)	5	3	3	5	3.2
TextField (Xu et al, 2018)	4	3	3	4	1.9						

Our SCQ analysis identifies three distinct regimes of method design. **Light** methods (SCQ ≤ 1.0) dominate horizontal detection with minimal overhead. The **Medium** regime (SCQ 1.1–1.7) represents the shift toward dense, multi-oriented prediction. Finally, the **Heavy** regime (SCQ ≥ 1.8) marks the current frontier of scene text detection, where models must internalize complex structural reasoning or employ high-intensity mathematical reconstructions to resolve arbitrary geometric curvature.

4. Datasets and Annotation Geometry

The development of scene text detection methods is closely tied to the datasets on which they are trained and evaluated. Unlike many other areas of computer vision, where benchmark datasets tend to converge toward a common annotation format, scene text detection datasets have evolved in a fragmented manner, reflecting changing assumptions about what constitutes a text instance and how it should be represented.

Early datasets, such as ICDAR 2003 and ICDAR 2005, were relatively small and contained well-captured, predominantly horizontal text (Lucas et al., 2003; Lucas et al., 2005). Their annotations consisted of axis-aligned rectangular bounding boxes at the word level. As a result, early detection

methods were implicitly designed around these constraints, treating text as a compact object that could be localized with a small number of geometric parameters.

As datasets grew in size and diversity, new challenges emerged. ICDAR 2013 expanded the scale of evaluation but largely preserved the same annotation model (Karatzas et al., 2013). A more significant shift occurred with ICDAR 2015, where annotations moved to oriented quadrilaterals to accommodate multi-oriented text in natural scenes (Karatzas et al., 2015). This change reflects a fundamental limitation of axis-aligned representations: they cannot accurately capture rotated or perspective-distorted text without including a lot of background. Subsequent datasets introduced further complexity, not only in terms of appearance but also in geometric structure. Datasets such as MSRA-TD500 (Yao et al., 2012) and ICDAR 2017 MLT (Nayef et al., 2017) incorporate multi-oriented and multilingual text, often at the line level rather than the word level, requiring representations that can work with elongated and densely packed instances. Later datasets, including Total-Text (Ch'ng & Chan, 2017), CTW1500 (Liu et al., 2019), and ICDAR 2019 ArT (Chng et al., 2019), move beyond rectangular approximations entirely and provide polygonal annotations that explicitly model curved and irregular text shapes. At the same time, other datasets emphasize different aspects of the problem. COCO-Text (Veit et al., 2016) focuses on cluttered natural scenes with significant background variation, while LSVT (Sun et al., 2019), ReCTS (Zhang et al., 2019), and RCTW-17 (Shi et al., 2017) introduce large-scale datasets with dense annotations and a strong presence of non-Latin scripts. More recent datasets such as TextOCR (Singh et al., 2021) and HierText (Long et al., 2022) further increase annotation density and introduce hierarchical relationships between characters, words, and lines.



Figure 1. Various annotation types for scene text detection datasets: a) axis aligned rectangles (ICDAR 2013 - Karatzas et al., 2013), b) quadrilaterals (ICDAR 2015 - Karatzas et al., 2015), c) polygons (CTW-1500 - Liu et al., 2019), d) rotated rectangles (MSRA-TD500 - Yao et al., 2012), e) character level annotations (KAIST 2011 - Jung et al., 2011), f) segmentation mask (KAIST 2011 - Jung et al., 2011)

Each dataset reflects a specific set of design choices regarding annotation granularity, geometric representation, language coverage, and scene complexity. Consequently, there is no single dataset that captures all aspects of the problem simultaneously. A dataset that provides precise polygonal annotations for curved text may lack large-scale multilingual coverage, while a dataset with dense hierarchical annotations may rely on simpler geometric representations. The cost of annotation, the

intended application domain, and the historical context in which the dataset was created all influence these trade-offs.

Table 1. Annotations for various datasets used in scene text detection.

Ground truth geometry	Dataset
Axis-aligned rectangles	SynthText (Gupta et al., 2016) ICDAR 2003 (Lucas et al., 2003) ICDAR 2013 (Karatzas et al., 2013) COCO-Text (Veit et al., 2016) SVT (Epshtein et al., 2010) TextOCR (Singh et al., 2021) HierText (Long et al., 2022) UnrealText (Long & Yao, 2020)
Oriented rectangles / quadrilaterals	ICDAR 2015 (Karatzas et al., 2015) ICDAR 2017 MLT (Nayef et al., 2017) ICDAR 2019 MLT (Nayef et al., 2019) MSRA-TD500 (Yao et al., 2012) LSVT (Sun et al., 2019) ReCTS (Zhang et al., 2019) RCTW-17 (Shi et al., 2017)
Arbitrary polygons	ICDAR 2019 ART (Chng et al., 2019) Total-Text (Ch'ng & Chan, 2017) CTW1500 (Liu et al., 2019)

The datasets summarized in Table 1 illustrate this progression, grouping them according to the geometric form of their annotations. This table highlights the increasing complexity of the detection task and provides a basis for analysing how different methods distribute complexity between prediction and reconstruction.

5. Complexity vs performance

To evaluate the empirical performance of the surveyed methods, we aggregate their reported F1-scores across five benchmark datasets that represent the primary evolutionary milestones in scene text detection. These benchmarks encompass a broad range of challenges, from the predominantly horizontal text of ICDAR 2013 to the complex, multi-oriented instances found in ICDAR 2015 and MSRA-TD500. Additionally, we include results from the multi-lingual ICDAR 2017 (MLT) and the curved-text focused CTW1500 to capture the state of the art in handling arbitrarily shaped text. The following table provides a comprehensive summary of these performance metrics, establishing a baseline for how different architectural approaches contend with increasing geometric and linguistic complexity across various standard benchmarks. The F1 scores for the datasets are available in Table 3.

Table 3 Performance of Scene Text Detection methods on various datasets (results from respective publications)

Method	ICDAR2013	ICDAR2015	MSRA-TD500	ICDAR2017	CTW1500
DeepDirectRegression (He et al, 2017)	86	81	74	-	-
TextBoxes (Liao et al, 2016)	85	-	-	-	-
TextBoxes++ (Liao et al, 2016)	88	81.7	-	-	-
WordSup (Hu et al, 2017)	90.34	78.16	-	-	-
WeText (Tian et al, 2017)	86.9	-	-	-	-
MOTD-FCN (Zhang et al, 2016)	83	54	74	-	-
STD-H (Long et al, 2014)	84.3	64.7	75.9	-	-
CTPN (Tian et al, 2016)	88	60.9	-	-	56.9
EAST (Zhou et al, 2017)	82.7	78.2	76.1	-	60.4
DBNet (Liao et al., 2020)	87	87.3	84.9	74.7	83.4
PixelLink (Deng et al, 2018)	-	83.7	-	-	-
TextField (Xu et al, 2018)	-	84.1	81.3	-	81.4
LRANet (Su et al, 2024)	-	87.4	-	89.2	89
FFPN (Wu et al, 2024)	-	85.3	79.7	67.7	-
TCAM (Xiao et al, 2019)	-	84.6	-	70.2	-
PAN (Wang et al., 2019b)	-	-	84.1	-	83.7
Mask R-CNN (Huang et al, 2019)	-	85	-	74.3	85
CRAFT (Baek et al., 2019)	-	86.9	82.9	-	83.5
FCENet (Zhu et al. 2021)	-	86.2	-	-	85.5
PSENet (Wang et al., 2019a)	-	-	83.3	72.2	82.2
TextDETR (Zhu et al., 2021)	90.3	88.5	86.7	76.5	85.1
MorphText (Xu et al., 2024)	-	-	87	78.3	86.5
TextSnake (Long et al., 2018)	-	-	78.3	-	75.6

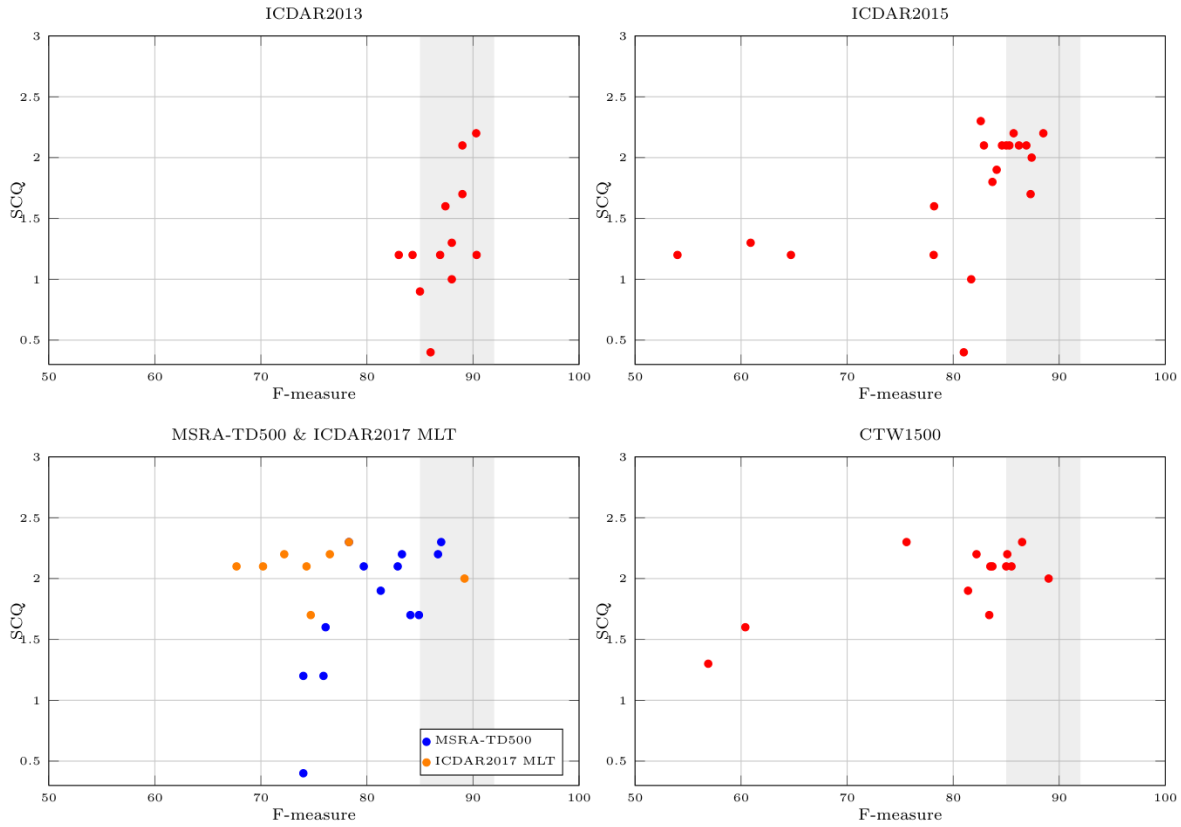


Figure 2. Relationship between detection performance (F1-score) and structural complexity SCQ factor across datasets. Gray band represents the F1-score interval at which the performance would be acceptable (85 to 92).

The plots in Figure 2 illustrate the relationship between detection performance (F1-score) and structural complexity (SCQ) across datasets of increasing difficulty. A consistent pattern emerges for simpler datasets, such as ICDAR2013, high performance is achieved across a wide range of model complexities. Methods with relatively low SCQ already reach F-measures close to the upper bound, and increasing structural complexity yields only marginal improvements, if any. This indicates that, for well-constrained scenarios with predominantly horizontal text, the problem does not require highly expressive representations or complex reconstruction pipelines.

In contrast, ICDAR2015 exhibits a broader spread along both axes. While low-complexity methods still achieve reasonable performance, the highest F-measures are increasingly associated with models of moderate to high SCQ. This reflects the transition to multi-oriented text, where simple geometric assumptions are no longer sufficient. The performance gains in this regime are not uniform, however; methods with similar SCQ values often exhibit comparable results, suggesting that beyond a certain level of structural sophistication, improvements are incremental rather than transformative.

The combined plot for MSRA-TD500 and ICDAR2017 MLT further reinforces this trend. Both datasets introduce additional challenges—long text lines in arbitrary orientations (MSRA-TD500) and multi-lingual, highly variable text (MLT). Here, the lower-complexity methods are clearly separated from the higher-performing ones, and the latter cluster at higher SCQ values. At the same time, the dispersion within the high-complexity region remains limited: once a sufficient level of representational and reconstruction capability is reached, different architectural choices lead to relatively small

performance differences. The distinction between the two datasets is also visible, with MLT showing slightly wider variability, consistent with its higher intrinsic difficulty.

Finally, the CTW1500 plot, which focuses on curved text, shows the strongest dependency on structural complexity. Methods relying on simple geometric representations perform significantly worse, while those employing structured or high-order representations and more elaborate reconstruction mechanisms achieve substantially higher F-measures. The concentration of high-performing methods at elevated SCQ values supports the view that curved text detection inherently requires richer representations and more complex processing pipelines. However, even in this setting, performance tends to plateau within the high-complexity regime, indicating diminishing returns once the necessary representational capacity is in place.

Taken together, these observations support a central claim: improvements in scene text detection are closely tied to the alignment between dataset complexity and representational capacity, rather than to a uniform progression in model architecture. Increasing structural complexity is beneficial primarily when demanded by the geometric and semantic variability of the data; otherwise, simpler models remain competitive.

Finally, it is essential to note that many recent Transformer-based architectures integrate recognition capabilities directly into the pipeline, becoming end-to-end text spotters. By 'recognizing' the text as it is localized, these methods render several traditional post-processing steps—such as explicit Orientation Correction Stages (OCS), complex spatial rectification and OCR modules redundant. Within the SCQ framework, this represents the ultimate shift: the burden of correcting for extreme or irregular geometry is no longer an external algorithmic task (D) but is instead moved into the model's internals (H), streamlining the transition from pixels to structured text.

6. Conclusions

This work examined scene text detection methods as a function of dataset complexity and representation design, rather than as a sequence of architectural innovations. By modelling detection as a two-stage process and introducing the Structural Complexity Quotient (SCQ) as an evaluative heuristic, we established a structural approach for analysing how complexity is distributed across the pipeline.

The analysis reveals that different methods are best understood in terms of how they allocate structural burden. Approaches based on explicit geometric predictions concentrate complexity within the model, while dense and field-based methods shift part of this burden to reconstruction procedures. Increasing representational expressiveness does not eliminate complexity but redistributes it across the pipeline.

Empirical comparisons across datasets show that performance is strongly dependent on the relation between representation and dataset geometry. On simpler datasets, methods with relatively low structural complexity achieve performance comparable to more complex approaches. As dataset complexity increases—through orientation, curvature, or multilingual variability—higher-complexity representations become necessary. However, within each regime, performance tends to cluster among methods with similar structural characteristics, indicating diminishing returns beyond a certain level of complexity.

These observations illustrate that scene text detection evolves through strategic adaptations to increasingly complex datasets, often reusing existing architectural paradigms while modifying representations and reconstruction strategies. Transformer-based approaches extend this trend by shifting the burden of structural reasoning into the model's head. However, this does not eliminate the underlying trade-offs. Even when utilizing convolutional backbones for feature extraction, the transformer layers lack the geometric priors (such as anchor boxes) that guide traditional detectors. Consequently, they require substantially larger datasets to learn to associate features into coherent text instances and to master coordinate regression. Furthermore, they impose significant computational overhead during inference; the quadratic complexity of self-attention mechanisms and high VRAM requirements often necessitate specialized hardware to maintain competitive latency..

From a practical perspective, method selection should be guided by dataset characteristics and application constraints, rather than by architectural novelty alone. Future work should therefore focus not only on improving model performance, but on designing representations and reconstruction procedures that align with the geometric structure of the target data.

References

1. Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9365–9374. https://doi.org/10.1109/CVPR.2019.00959
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. Proceedings of the European Conference on Computer Vision (ECCV), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
3. Chng, C. K., Chan, C. S., & Liu, C. L. (2019). ICDAR2019 robust reading challenge on arbitrary-shaped text (RRC-ArT). 2019 International Conference on Document Analysis and Recognition (ICDAR), 1571–1576.
4. Ch'ng, C. K., & Chan, C. S. (2017). Total-Text: A comprehensive dataset for scene text detection and recognition. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 935–942. https://doi.org/10.1109/ICDAR.2017.157
5. Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. CoRR, abs/1801.01315, 2018
6. Gupta, A., Vedaldi, A., & Zisserman, A. (2016). Synthetic data for text localisation in natural images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2315–2324. https://doi.org/10.1109/CVPR.2016.254
7. Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding. Wordsup: Exploiting word annotations for character based text detection. CoRR, abs/1708.06720, 2017.
8. He, W., Zhang, X., Yin, F., & Liu, C. (2017). Deep Direct Regression for Multi-oriented Scene Text Detection. 2017 IEEE International Conference on Computer Vision (ICCV), 745-753.
9. Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Reading text in the wild with convolutional neural networks. arXiv preprint arXiv:1412.1842
10. Jung, K., In Kim, K., & K. Jain, A. (2004). Text information extraction in images and video: a survey. Pattern Recognition, 37(5), 977–997. doi:10.1016/j.patcog.2003.10.012
11. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., et al. (2013). ICDAR 2013 robust reading competition. 2013 12th International Conference on Document Analysis and Recognition (ICDAR), 1484–1493. https://doi.org/10.1109/ICDAR.2013.221
12. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., et al. (2015). ICDAR 2015 competition on robust reading. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 1156–1160. https://doi.org/10.1109/ICDAR.2015.7333942
13. Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2018). TextBoxes++: A single-shot oriented scene text detector. IEEE Transactions on Image Processing, 27(8), 3676–3690. https://doi.org/10.1109/TIP.2018.2825107
14. Liao, M., Zou, Z., Wan, Z., Yao, C., & Bai, X. (2023). Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1), 919–931. doi:10.1109/TPAMI.2022.3155612
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9905 LNCS, 21–37.
16. Liu, Y., Jin, L., Zhang, S., Luo, C., & Zhang, S. (2019). Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recognition, 90, 337–345. https://doi.org/10.1016/j.patcog.2019.02.002

17. Long, S., & Yao, C. (2020). UnrealText: Synthesizing realistic scene text images from the unreal world. Proceedings of the AAAI Conference on Artificial Intelligence, 11456–11463.
18. Long, S., He, X., & Yao, C. (2014). Scene text detection via holistic, multi-channel prediction. Proceedings of the International Conference on Pattern Recognition (ICPR), 3475–3480.
19. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., & Raptis, M. (2022). Towards End-to-End Unified Scene Text Detection and Layout Analysis. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1039–1049. doi:10.1109/CVPR52688.2022.00112
20. Long, S., Ruan, J., Zhang, W., He, X., Wu, W., & Yao, C. (2018). TextSnake: A flexible representation for detecting text of arbitrary shapes. Proceedings of the European Conference on Computer Vision (ECCV), 20–36.
21. Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., & Young, R. (2003). ICDAR 2003 robust reading competitions. In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) (pp. 682–687). IEEE. https://doi.org/10.1109/ICDAR.2003.1227801
22. Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., & Young, R. (2005). ICDAR 2005 text locating competition results. In Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR 2005) (pp. 80–84). IEEE. https://doi.org/10.1109/ICDAR.2005.104
23. Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., & Matas, J. (2017). ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017) (pp. 1454–1459). IEEE. https://doi.org/10.1109/ICDAR.2017.237
24. Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khlif, W., … Ogier, J.-M. (2019). ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition — RRC-MLT-2019. 2019 International Conference on Document Analysis and Recognition (ICDAR), 1582–1587. doi:10.1109/ICDAR.2019.00254
25. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779–788. doi:10.1109/CVPR.2016.91
26. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28, 91–99.
27. Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. CoRR, abs/1807.01544, 2018.
28. Shi, B., Bai, X., & Yao, C. (2017). RCTW-17: Robust reading challenge on reading Chinese text in the wild. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017) (pp. 1429–1434). IEEE. https://doi.org/10.1109/ICDAR.2017.233
29. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T., & Natarajan, P. (2021). TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8801–8811.
30. Su, Y., Chen, Z., Shao, Z., Du, Y., Ji, Z., Bai, J., … Jiang, Y.-G. (2024). LRANet: towards accurate and efficient scene text detection with low-rank approximation network. Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. doi:10.1609/aaai.v38i5.28302
31. Sun, Y., Ni, Z., Wang, C., Liu, X., Luo, C., Wang, W., & Jin, L. (2019). Chinese street view text: Large-scale Chinese text reading with partially supervised learning. In Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV 2019) (pp. 9086–9095). IEEE. https://doi.org/10.1109/ICCV.2019.00918
32. Tian, S., Lu, S., & Li, C. (2017). WeText: Scene Text Detection under Weak Supervision. 2017 IEEE International Conference on Computer Vision (ICCV), 1501-1509.
33. Tian, Z., Huang, W., He, T., He, P., & Qiao, Y. (2016). Detecting text in natural image with connectionist text proposal network. In Proceedings of the European Conference on Computer Vision (ECCV 2016) (pp. 56–72). Springer.
34. Veit, A., Matera, T., Neumann, L., Matas, J., & Belongie, S. (2016). COCO-Text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140.
35. Wang, W., Xie, E., Li, X., et al. (2019a). Shape robust text detection with progressive scale expansion network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9336–9345. https://doi.org/10.1109/CVPR.2019.00956
36. Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., … Shen, C. (2019b). Efficient and Accurate Arbitrary-Shaped Text Detection With Pixel Aggregation Network. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 8439–8448. doi:10.1109/ICCV.2019.00853
37. Xiang Li, Wenhai Wang, Wenbo Hou, Ruo-Ze Liu, Tong Lu, and Jian Yang. Shape robust text detection with progressive scale expansion network. CoRR, abs/1806.02559, 2018.
38. Xiao, Y., Xue, M., Lu, T., Wu, Y., & Palaiahnakote, S. (2019). A Text-Context-Aware CNN Network for Multi-oriented and Multi-language Scene Text Detection. 2019 International Conference on Document Analysis and Recognition (ICDAR), 695–700. doi:10.1109/ICDAR.2019.00116
39. Xu, C., Jia, W., Wang, R., Luo, X., & He, X. (2023). MorphText: Deep Morphology Regularized Accurate Arbitrary-Shape Scene Text Detection. Trans. Multi., 25, 4199–4212. doi:10.1109/TMM.2022.3172547
40. Xu, C., Jia, W., Wang, R., Luo, X., & He, X. (2024). MorphText: Deep morphology regularized arbitrary-shape scene text detection. Proceedings of the AAAI Conference on Artificial Intelligence.
41. Yao, C., Bai, X., Liu, W., Ma, Y., & Tu, Z. (2012). Detecting texts of arbitrary orientations in natural images. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 1083–1090. doi:10.1109/CVPR.2012.6247787
42. Yirui Wu, Lilai Zhang, Hao Li, Yunfei Zhang, and Shaohua Wan. Feature fusion pyramid network for end-to-end scene text detection. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 23(11), November 2024.
43. Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, and Xiang Bai. Textfield: Learning A deep direction field for irregular scene text detection. CoRR, abs/1812.01393, 2018.
44. Zhang, R., Zhou, Y., Jiang, Q., Song, J., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., Bai, X., Shi, B., Karatzas, D., & Lu, S. (2019). ICDAR 2019 robust reading challenge on reading Chinese text on signboard — ReCTS. 2019 International Conference on Document Analysis and Recognition (ICDAR), 1516–1521. https://doi.org/10.1109/ICDAR.2019.00244
45. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., & Bai, X. (2016). Multi-Oriented Text Detection with Fully Convolutional Networks. CoRR, abs/1604.04018. Retrieved from http://arxiv.org/abs/1604.04018
46. Zhida Huang, Zhuoyao Zhong, Lei Sun, and Qiang Huo. Mask r-cnn with pyramid attention network for scene text detection. pages 764–772, 01 2019.
47. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). EAST: An efficient and accurate scene text detector. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

48. Zhu, X., & Zhang, Z. (2021). Transformer-based end-to-end scene text recognition. 2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA), 1691–1695. doi:10.1109/ICIEA51954.2021.9516154
49. Zhu, Y., Yao, C., & Bai, X. (2020). Scene text detection and recognition: The deep learning era. International Journal of Computer Vision, 129(1), 161–184. https://doi.org/10.1007/s11263-020-01369-0
50. Zhu, Y., Yao, C., & Bai, X. (2021). Fourier contour embedding for arbitrary-shaped text detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3123–3131.