Confidence-Aware Ensemble and Long-Word Refinement for Artistic Text Recognition

Lucas A. Dias*, Henrique A. Schulz*, Rafaela de Miranda*,
Guilherme L. Peres*, Pedro L. Bittencourt*, and Rayson Laroca*

*Pontifical Catholic University of Paraná, Curitiba, Brazil

*{dias.azevedo,henrique.schulz,r.miranda2,guilherme.peres,pedro.bittencourt}@pucpr.edu.br
*rayson@ppgia.pucpr.br

***Abstract*—Artistic Text Recognition (ATR) remains challenging because word images often combine decorative fonts, curved layouts, object-like characters, clutter, and severe distortions. This paper studies WordArt-V1.5 as a standardized benchmark for this setting and evaluates recent scene and artistic text recognizers under a common protocol. We propose a confidence-aware ensemble that combines SVTRv2, PARSeq, and MAERec after fine-tuning on the official training split. The ensemble selects predictions using the minimum confidence over disagreement positions, emphasizing characters that separate competing hypotheses. For long words, where a single character error can invalidate the whole prediction, we add a targeted refinement stage based on Needleman-Wunsch alignment and lexicon-guided correction. On the WordArt-V1.5 Test B split, the proposed system reaches 89.90% Word Recognition Accuracy, improving the best individual fine-tuned model by 1.77 percentage points. The long-word refinement produces a modest global gain, but improves the targeted long-word subset by 2.72 percentage points. Finally, an error analysis of all remaining mistakes shows that 48.8% are associated with labeling issues, visual ambiguity, or illegible samples, highlighting the value of diagnostic reporting for future ATR benchmarks and models. Our source code is available at *https://github.com/lucas-azdias/Artistic-Text-Recognition/*.**

Fig. 1. Examples from WordArt-V1.5, illustrating decorated characters, curved text, stylized fonts, and non-standard layouts that make ATR substantially different from conventional scene text reading [2].

I. Introduction

Artistic Text Recognition (ATR) aims to read words whose appearance is deliberately shaped by visual design rather than by readability alone [1]–[3]. Unlike regular scene text, artistic text may include non-standard fonts, strong color variation, objects replacing characters, and deformations that place samples close to the human readability limit, as shown in Fig. 1. These properties make ATR a stress test for modern recognizers because errors often arise from both visual ambiguity and language-level uncertainty.

The task is related to Scene Text Recognition (STR), but the domain gap is non-trivial. Many STR benchmarks focus on natural images affected by perspective distortion, blur, occlusion, or low resolution [4]–[8], challenges also observed in specialized text-recognition applications such as license plate recognition [9], [10]. WordArt-style images add difficulty because letters are embedded in artistic compositions, not merely degraded by acquisition conditions [1], [11]. Thus, recognizers that perform well on regular benchmarks may still fail when characters are stretched, merged with drawings, stylized as objects, or intentionally distorted to create a visual identity. This is relevant to posters, advertisements, logos, book covers, social media images, and other media where text and design are inseparable.

The WordArt-V1.5 dataset, introduced in the *ICDAR 2024 Competition on Artistic Text Recognition* [2], provides a standardized setting for this problem. It contains 6,000 training images and a 6,000-image test set split evenly into Test A and Test B. The fixed competition partitions enable controlled comparisons between general-purpose STR recognizers, ATR-oriented architectures, and ensembles, addressing the issue that inconsistent protocols can make model comparisons difficult to interpret [2], [8]. They also help separate model failures from annotation noise, visual ambiguity, and limited visual evidence, which is essential when some samples are difficult even for human readers.

This paper studies WordArt-V1.5 as a reproducible benchmark for current recognizers and proposes a confidence-aware ensemble tailored to this setting. Rather than introducing a new backbone, it addresses a practical question: *how far can a compact fusion of strong recent recognizers go under fixed training, validation, and test partitions?* We evaluate three

Accepted for presentation at SIBGRAPI 2026. The final published version will be available on IEEE Xplore.

widely adopted STR models, MAERec [11], SVTRv2 [12], and PARSeq [13], fine-tune them under the same protocol, combine them through a disagreement-aware confidence rule, and add a refinement stage for long words. The long-word stage is motivated by the word-level metric, since longer words are more likely to contain at least one character error that invalidates the full prediction.

The design separates three aspects of the problem. First, the protocol evaluates representative recognizers under the same split, following the recommendation that STR methods should be compared under consistent datasets, metrics, and implementation assumptions [8]. Second, the ensemble rule focuses on disagreement positions rather than averaging confidence across the whole word, prioritizing the characters that resolve competing predictions. Third, the long-word module is activated only when word-level evaluation is more sensitive to insertions, deletions, and isolated low-confidence characters, reducing unnecessary dictionary intervention on short or confident words.

The contributions are threefold. First, we report a standardized benchmark of representative recognizers on WordArt-V1.5 using the official competition splits and Word Recognition Accuracy (WRA) metric. Our code is available at *https://github.com/lucas-azdias/Artistic-Text-Recognition/* to support reproducibility beyond the reported scores. Second, we propose a confidence-aware ensemble with targeted long-word refinement based on sequence alignment and lexicon-guided correction. Third, we provide an error analysis of all 303 remaining Test B mistakes, separating model errors from labeling issues, ambiguity, and illegible samples. Together, these results provide both a competitive recognition system and diagnostic evidence about the current limits of ATR on WordArt-V1.5.

The rest of this paper is organized as follows. Section II reviews related work and the WordArt-V1.5 protocol. Section III describes model selection, fine-tuning, the ensemble rule, and long-word refinement. Section IV reports quantitative results, ablations, and error analysis, while Section V summarizes the findings, discusses limitations, and outlines future directions.

II. Related Work and WordArt-V1.5 Protocol

A. Scene and Artistic Text Recognizers

Scene text recognition has progressed from convolutional and recurrent sequence models toward transformer, masked-reconstruction, and refined Connectionist Temporal Classification (CTC)-based approaches [8], [14], [15]. ASTER [16] combined rectification and attention to handle irregular layouts, while ViTSTR [17] showed that a compact Vision Transformer (ViT) can recognize text without recurrent decoding or explicit rectification.

Recent methods are particularly relevant to ATR, where stylization can substantially alter character shapes, spacing, and boundaries. CornerTransformer [1] employs contour-guided representations to recognize arbitrary and artistic text; PARSeq [13] explores multiple decoding orders through permutation language modeling; MAERec [11] learns robust visual representations through masked autoencoding; and SVTRv2 [12] improves CTC-based recognition through multi-size resizing, feature reorganization, and semantic guidance. These complementary architectures provide a diverse set of candidates for both benchmarking and ensemble construction. Their published results on widely used STR benchmarks are summarized in Table I.

TABLE I
Representative recognizers considered in the benchmark.

Model	Word Recognition Accuracy (WRA) (%)				
	IIIT5k	SVT	IC15	SVTP	U14M-B[a]
ViTSTR [17]	88.4	87.7	72.6	81.8	–
CornerTransformer [1]	95.9	94.6	86.3	91.5	–
PARSeq [13]	99.1	97.9	89.6	95.7	84.3[b]
MAERec [11]	98.5	97.8	89.5	94.4	85.2
SVTRv2 [12]	99.2	98.0	91.1	99.0	86.1

[a] Union14M-Benchmark.
[b] Reported by Du *et al.* [12], since Union14M-Benchmark was not available when PARSeq was published.

B. WordArt-V1.5 Protocol

WordArt-V1.5 extends the WordArt dataset introduced with CornerTransformer [1] and was released for the *ICDAR 2024 Competition on Artistic Text Recognition* [2]. It contains word images from posters, greeting cards, covers, billboards, advertisements, and other graphical designs, where artistic intent may alter character shape, spacing, orientation, and texture. The adopted partitions are summarized in Table II.

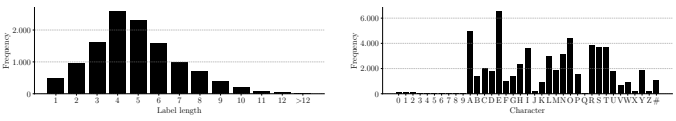
TABLE II
WordArt-V1.5 protocol used in this paper.

Split	Images	Use in this paper
Train	6,000	Fine-tuning
Test A	3,000	Validation and model selection
Test B	3,000	Final evaluation

The official metric is Word Recognition Accuracy (WRA), which ignores case and symbols and counts a prediction as correct only when the normalized full word matches the ground truth [2]. Let W denote the number of evaluated words and W_r the number of correctly recognized words:

$$WRA = \frac{W_r}{W}\,. \tag{1}$$

Any incorrect, missing, or inserted character invalidates the full prediction. Hence, longer labels provide more opportunities for word-level errors, motivating the analysis in Fig. 2.



(a) Label length distribution. (b) Character frequency distribution.

Fig. 2. WordArt-V1.5 label statistics, with symbols grouped into the # class.

We preserve the official partitions: public pretrained checkpoints are fine-tuned on Train, Test A is used for validation and model selection, and Test B is reserved for final reporting. The public leaderboard is used only for context, not as evidence of an official competition entry.

C. Public Baselines on WordArt-V1.5

Public WordArt-V1.5 systems frequently combine complementary models and objectives [2]. Examples include supervised, semi-supervised, and self-supervised components in *OCR for WordArt*; ensembling and character-to-character distillation in *ViettelAI-OCR*; a character-level contrastive adaptation of MAERec in *Let Me See*; and a PARSeq, MAERec, and CornerTransformer ensemble in *iPad_OCR*. Rather than reproducing a large leaderboard-oriented pipeline, we construct a compact and reproducible baseline using public checkpoints, a fixed fine-tuning protocol, and a transparent fusion rule.

III. Confidence-Aware Ensemble

A. Candidate Models and Fine-Tuning

The proposed system, shown in Fig. 3, combines three recognizers selected from five candidates. All candidates were first evaluated on Test A using pretrained checkpoints, as reported in Table III. The strongest and most complementary models were SVTRv2 [12], PARSeq [13], and MAERec [11], with Test A WRA values of 85.80%, 83.97%, and 83.40%, respectively. ViTSTR [17] and CornerTransformer [1] reached 78.80% and 73.80%, and were not included in the final ensemble. This keeps the ensemble compact while preserving diversity across CTC, autoregressive, and masked-reconstruction-based recognizers.

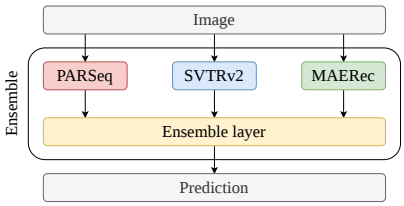


Fig. 3. Overview of the proposed ensemble, which combines PARSeq, SVTRv2, and MAERec through a confidence-aware aggregation layer.

TABLE III
Candidate model performance on Test A, used as the validation set in this paper, before and after fine-tuning.

Model	Pretrained WRA (%)	Fine-tuned WRA (%)
SVTRv2 [12]	85.80	86.23
PARSeq [13]	83.97	84.53
MAERec [11]	83.40	84.53
ViTSTR [17]	78.80	–
CornerTransformer [1]	73.80	–

The decision to use an ensemble is supported by the Test A error overlap in Fig. 4. Among 677 samples missed by at least one selected model, only 287 errors were shared by all three models, corresponding to 42.39% shared errors. If an oracle could always choose a correct prediction whenever at least one model was right, the theoretical Test A upper bound would be 90.43%, substantially higher than any individual model. This gap indicates that many mistakes are model-specific rather than inherent to all recognizers. It motivated a fusion rule that focuses on disagreement positions rather than average confidence over the full sequence.

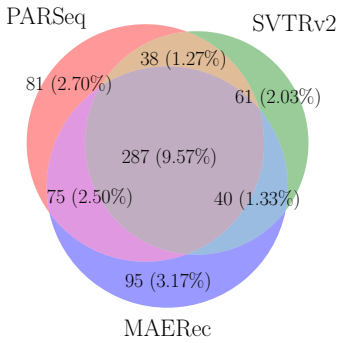


Fig. 4. Error overlap among the selected pretrained models on Test A.

Each selected recognizer was fine-tuned on the official training split, while Test A was used exclusively for validation, hyperparameter selection, and checkpoint selection. The main configurations selected through this validation procedure are summarized in Table IV. All models were optimized with AdamW and trained for at most 50 epochs, with the best checkpoint selected according to the WRA obtained on Test A.

TABLE IV
Main fine-tuning configurations selected based on validation performance on Test A. All models used the AdamW optimizer.

Configuration	MAERec	PARSeq	SVTRv2
Initial learning rate	1×10^{-5}	7×10^{-4}	5×10^{-5}
Batch size	64	384	128
Scheduler	CosineAnnealingLR	OneCycleLR	CosineAnnealingLR

Fine-tuning employed online geometric and photometric augmentations to reduce sensitivity to distortions and appearance variations [8], [16]. These augmentations included rotations, translations, scaling, shearing, affine and perspective transformations, pyramid resizing, photometric and color perturbations, Gaussian noise, motion blur, generic blur, and random erasing. After fine-tuning, SVTRv2 improved from 85.80% to 86.23% WRA on Test A, while PARSeq and MAERec both reached 84.53%. These validation results guided the selection of the final models and configurations, without using Test B for parameter tuning.

Experiments were implemented in Python with PyTorch and executed on Google Colab with an NVIDIA L4 GPU, 22.5 GB of memory, and 53 GB of system memory. This environment was sufficient to fine-tune the selected recognizers and evaluate the ensemble without multi-GPU infrastructure.

B. Disagreement-Aware Confidence Rule

Let $M = \{m_1, m_2, \ldots, m_K\}$ be the set of recognizers. Each recognizer m_k produces a predicted sequence $\hat{\mathbf{y}}_k =$

$[\hat{y}_{k,1}, \ldots, \hat{y}_{k,n}]$ and confidence scores $\mathbf{l}_k = [l_{k,1}, \ldots, l_{k,n}]$. We first identify positions at which at least two models disagree:

$$D = \{j \mid \exists p, q,\ \hat{y}_{p,j} \neq \hat{y}_{q,j}\}\ . \tag{2}$$

The score of model m_k is the minimum confidence over disagreement positions. When all predictions agree, the minimum over the whole sequence is used:

$$S_k = \begin{cases} \min_{j \in D} l_{k,j}, & \text{if } |D| > 0, \\ \min_j l_{k,j}, & \text{otherwise.} \end{cases} \tag{3}$$

The ensemble output is the sequence predicted by the model with the largest score:

$$\hat{\mathbf{y}}_{ens} = \hat{\mathbf{y}}_{k^*}, \quad k^* = \arg\max_k S_k\ . \tag{4}$$

This rule penalizes low-confidence characters exactly where recognizers disagree, which is more targeted than selecting a prediction by global average confidence. A model can therefore keep a high score when it is uncertain only on characters shared by all predictions, but it is penalized when uncertainty occurs at a position that changes the final word. Scores are normalized by each model's mean character confidence to reduce calibration differences across recognizers, since modern neural networks can produce confidence values that are not directly comparable across architectures [18], [19]. The rule remains conservative: by default, it selects one complete model hypothesis rather than synthesizing a new word through character-level voting. This avoids creating implausible sequences from locally confident but globally inconsistent characters. To assess the contribution of this scoring rule, Table V compares the proposed minimum-weighted strategy with majority voting and maximum-weighted selection on Test A. The minimum-weighted strategy achieves the highest WRA, supporting the use of the least-confident disagreement position to select among the complete-model hypotheses.

TABLE V
ABLATION OF ENSEMBLE SCORING STRATEGIES ON TEST A.

Strategy	WRA (%)
Majority voting	85.67
Maximum weighted	87.20
Minimum weighted	**87.77**

C. Long-Word Refinement

The full-word metric makes long labels sensitive to isolated character errors. If each character has an independent error probability p, the probability of a correct word of length n is $(1-p)^n$. Thus, small character-level uncertainty accumulates quickly as n increases. In WordArt-V1.5, labels with at least nine characters represent 749 samples in the full dataset and 184 samples in Test B.

For words with at least nine characters, we apply a two-stage refinement using a threshold chosen empirically on the validation set. First, the predictions of SVTRv2, PARSeq, and MAERec are aligned with the Needleman-Wunsch algorithm [20], [21]. The aligned sequences enable character-level voting even when insertions or deletions shift positions across models. Gap positions inherit the confidence of the nearest available neighboring character. Second, when the aligned word contains at least one character with confidence below 60%, a lexicon correction is considered. The lexicon stage searches the `wlist_match3` English word list [22] using bigram-based Levenshtein similarity [23]. A correction is accepted only when the best candidate reaches at least 70% similarity, reducing the risk of replacing a confident artistic word with an unrelated dictionary entry, a known concern in unconstrained scene text [8].

The refinement is not post-processing applied to every output. It targets the subset where the metric, the label-length distribution, and the observed disagreements indicate a specific weakness. This design is important because dictionary correction can be harmful when the image contains a brand name, a stylized invented word, or a proper noun that is absent from the lexicon. Restricting the module to long words with low-confidence evidence helps recover likely character-level mistakes while preserving the original model prediction in most cases.

IV. RESULTS AND ANALYSIS

A. Overall Recognition Accuracy

Table VI reports the final Test B results. The proposed ensemble reaches 89.90% WRA, outperforming the best individual fine-tuned model by 1.77 percentage points. SVTRv2 remains the strongest individual model, but PARSeq and MAERec provide alternative hypotheses that the ensemble can exploit when SVTRv2 is uncertain. The contextual comparison with public WordArt-V1.5 systems shows that the compact ensemble is within the range of strong specialized pipelines.

TABLE VI
FINAL TEST B RESULTS AND CONTEXTUAL WORDART-V1.5 BASELINES.

Method	WRA (%)	Method	WRA (%)
MAERec [11] fine-tuned	87.50	iPad_OCR [2]	89.27
PARSeq [13] fine-tuned	87.60	Let Me See [2]	89.77
SVTRv2 [12] fine-tuned	88.13	ViettelAI-OCR [2]	90.77
Proposed ensemble	**89.90**	Ocr For WordArt [2]	91.07

This comparison is informative because all entries use the same Test B split and WRA metric, the type of controlled protocol needed for meaningful text-recognition comparisons [2], [8]. The proposed system is not intended to replace larger leaderboard-oriented pipelines. Instead, it provides a transparent baseline for future ATR studies and shows that a compact three-model system can recover a relevant portion of the errors left by a single strong recognizer. In this sense, the main result is not only the final score, but also the gap between individual fine-tuned models and the confidence-aware ensemble under the same validation and test protocol.

Fig. 5 illustrates both successful and failed cases. In some samples, the ensemble selects a correct hypothesis even when another recognizer produces a visually similar but wrong word.

In others, all recognizers converge to the same wrong word or assign high confidence to an incorrect interpretation, showing that many remaining errors arise from stylized shapes with multiple plausible readings.

(a) Correct predictions. (b) Incorrect predictions.

Fig. 5. Qualitative examples from the final ensemble, showing (a) cases where confidence-aware fusion selects the correct prediction and (b) cases where all models remain confused.

B. Effect of Long-Word Refinement

The refinement stage is intentionally targeted, so its global gain is modest. As shown in Table VII, the complete refinement improves Test B WRA from 89.67% to 89.90%, a gain of 0.23 percentage points over all 3,000 test samples. Within the 184-sample long-word subset, however, the gain is 2.72 percentage points. This targeted improvement is meaningful because long artistic words concentrate a difficult failure mode: a visually plausible prediction can be rejected by WRA due to a single shifted, inserted, or low-confidence character.

TABLE VII
ABLATION OF THE LONG-WORD REFINEMENT ON TEST B.

Refinement	WRA (%)	Long-word WRA[a] (%)
None	89.67	75.00 (–)
Needleman-Wunsch only	89.73	75.00 (+0.00 pp)
Dictionary only	89.80	76.63 (+1.63 pp)
Needleman-Wunsch + dictionary	**89.90**	**77.72 (+2.72 pp)**

[a] Gain measured only on Test B words with at least nine characters.

The minimum word-length threshold was selected through an ablation on Test A. Increasing the threshold from three to eight characters progressively reduced the refinement coverage from 91.13% to 11.50%, while increasing the WRA from 87.00% to 87.93%. We selected a threshold of nine characters, corresponding to 5.70% coverage and 87.73% WRA, to restrict the module to the intended long-word cases. Higher thresholds of 10 and 11 characters reduced coverage to 2.53% and 1.13%, respectively, without providing consistent accuracy gains.

The ablation should therefore be interpreted on its target subset. As the module is activated for only 6.13% of Test B, even a substantial correction rate inside that subset appears small when averaged across all samples. This is expected and should not be interpreted as evidence that the module is ineffective. Needleman-Wunsch alignment is most helpful when models disagree on insertions or deletions, while the dictionary stage is most useful when the aligned result contains a localized low-confidence character. The combined alignment and dictionary strategy improves the intended cases without degrading the remaining predictions.

C. Qualitative and Error Analysis

Fig. 6 shows representative error cases, and Table VIII summarizes all 303 Test B mistakes. Model errors account for 51.2% and mostly involve cursive characters, heavy stylization, missing strokes, or misleading decorative components. Interestingly, the remaining 48.8% are not purely model failures: *labeling errors* account for 33.0%, ambiguity for 10.9%, and illegibility for 5.0%. This result shows why diagnostic reporting matters, since the same WRA score may combine solvable recognition failures with samples whose ground truth or visual evidence is debatable.

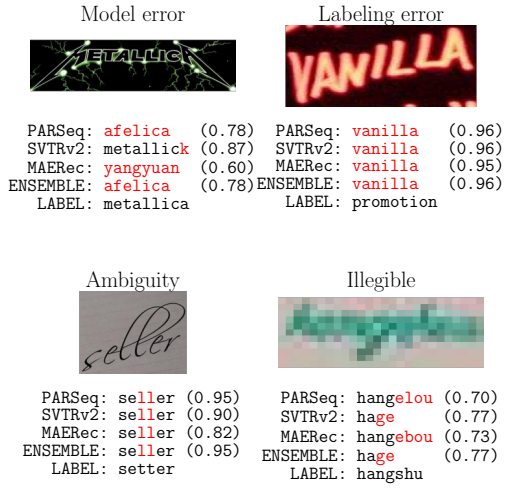


Fig. 6. Examples from the manual error analysis, covering model errors, labeling errors, ambiguous words, and visually illegible samples.

TABLE VIII
MANUAL CLASSIFICATION OF THE 303 TEST B ERRORS OF THE PROPOSED ENSEMBLE.

Error type	Count	Share (%)
Model error	155	51.2
Labeling error	100	33.0
Ambiguity	33	10.9
Illegible sample	15	5.0

The labeling-error group is important for benchmark interpretation. When the visual content clearly differs from the label, a correct model prediction may still be counted as an error by the official metric. Ambiguous samples create a different problem, where the annotation may be plausible but another reading is also visually defensible. Illegible samples provide insufficient evidence for a reliable word-level decision. Separating these categories does not invalidate WordArt-V1.5; it clarifies its limits and helps prevent attributing every remaining error to model weaknesses.

The analysis suggests that WordArt-V1.5 is valuable not only for ranking methods but also as a diagnostic benchmark. It exposes complementary model errors, reveals the fragility of word-level scoring on long labels, and contains ambiguous cases that motivate uncertainty-aware evaluation. Future reporting should therefore include official WRA, targeted subset

ablations, and qualitative error categories whenever possible. Such reporting makes progress easier to interpret, especially when new methods obtain small global gains on a dataset where a non-negligible fraction of errors may be annotation-related or visually underdetermined.

V. Conclusions

This paper presented a confidence-aware ensemble for Artistic Text Recognition (ATR) on WordArt-V1.5. The system combines SVTRv2, PARSeq, and MAERec after fine-tuning, selects predictions through a disagreement-aware confidence rule, and applies targeted refinement to long words using sequence alignment and lexicon-guided correction. On Test B, the proposed ensemble achieved 89.90% WRA, improving the best individual fine-tuned model by 1.77 percentage points. Although the long-word refinement yielded a modest global improvement, it increased accuracy by 2.72 percentage points on its targeted subset, showing the benefit of addressing isolated character errors under word-level evaluation.

The error analysis also highlights limitations that are not captured by recognition accuracy alone. Among the 303 remaining Test B errors, 48.8% were associated with labeling issues, visual ambiguity, or illegible samples. These findings suggest that progress in ATR depends not only on stronger recognizers, but also on better confidence calibration, uncertainty estimation, language-aware decoding, and diagnostic evaluation, particularly for samples near the limits of human readability [8], [18].

Future work will explore learned fusion strategies and the use of generative models to expand ATR training data with controlled styles, long words, rare characters, and challenging distortions [24]–[26]. Vision-Language Models (VLMs) may further support this process by filtering generated samples according to readability and label consistency [27], [28]. Future benchmarks could also report complementary measures such as long-word accuracy and uncertainty-aware rejection, helping distinguish genuine recognition failures from ambiguous or low-readability cases.

Acknowledgments

The authors thank the *Pontifícia Universidade Católica do Paraná* (PUCPR) for the financial support that made their participation in the conference possible.

References

[1] X. Xie, L. Fu, Z. Zhang, Z. Wang, and X. Bai, "Toward understanding wordart: Corner-guided transformer for scene text recognition," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 303–321.

[2] X. Xie, L. Deng, Z. Zhang, Z. Wang, and Y. Liu, "ICDAR 2024 competition on artistic text recognition," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2024, pp. 301–314.

[3] T. Do, T. Tran, K. Le, D.-D. Le, and T. D. Ngo, "Skeleton-guided artistic text recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, Dec 2025.

[4] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 1457–1464.

[5] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *Proc. BMVC*, 2012.

[6] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 569–576.

[7] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. ICDAR*, 2015, pp. 1156–1160.

[8] 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. ICCV*, 2019, pp. 4715–4723.

[9] R. Laroca, V. Estevam, G. J. P. Moreira, R. Minetto, and D. Menotti, "Advancing multinational license plate recognition through synthetic and real data fusion: A comprehensive evaluation," *IET Intelligent Transport Systems*, vol. 19, no. 1, p. e70086, 2025.

[10] R. Laroca, V. Nascimento, D. Kim, S. Chung, S. Bae, U. Seo, S. Oh, C. M. Phung, M. G. Vo, X. Ye, Y. Du, Y. Su, Z. Chen, S. Heo, H. Lee, K. Na, K. V. Vu Nguyen, S. T. Pham, D. N. N. Phung, T. P. Le, V. N. Vo Tran, and D. Menotti, "ICPR 2026 Competition on low-resolution license plate recognition," in *International Conference on Pattern Recognition (ICPR)*, Aug 2026, pp. 256–275.

[11] Q. Jiang, J. Wang, D. Peng, C. Liu, and L. Jin, "Revisiting scene text recognition: A data perspective," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 20 486–20 497.

[12] Y. Du, Z. Chen, H. Xie, C. Jia, and Y.-G. Jiang, "SVTRv2: CTC beats encoder-decoder models in scene text recognition," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2025, pp. 20 147–20 156.

[13] D. Bautista and R. Atienza, "Scene text recognition with permuted autoregressive sequence models," in *Proc. ECCV*, 2022, pp. 178–196.

[14] 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.

[15] 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," in *Proc. ICPR*, 2016, pp. 1368–1373.

[16] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.

[17] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2021, pp. 319–334.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.

[19] R. Laroca, L. A. Zanlorensi, V. Estevam, R. Minetto, and D. Menotti, "Leveraging model fusion for improved license plate recognition," in *Iberoamerican Congress on Pattern Recognition*, Nov 2023, pp. 60–75.

[20] A. Willis, D. Morse, A. Dil, D. King, D. Roberts, and C. Lyal, "Improving search in scanned documents: Looking for ocr mismatches," 2009, Technical Report.

[21] S. Eger, "Sequence alignment with arbitrary steps and further generalizations," *Information Sciences*, vol. 237, pp. 287–304, 2013.

[22] K. Vertanen. (2025) Big english word lists. [Online]. Available: https://www.keithv.com/software/wlist/

[23] G. Kondrak, "N-gram similarity and distance," in *String Processing and Information Retrieval*, 2005, pp. 115–126.

[24] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. CVPR*, 2016, pp. 2315–2324.

[25] R. Rombach, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 674–10 685.

[26] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei, "TextDiffuser-2: Unleashing the power of language models for text rendering," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 386–402.

[27] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[28] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. NeurIPS*, 2023.