

IDExtractNet: An Intelligent Deep Learning Pipeline for Accurate Information Extraction from Identity Documents

Pratiksha Sawant
Assistant Professor
Department of Artificial Intelligence
and Data Science
All India Shri Shivaji Memorial Society's,
Institute of Information Technology
Pune, India

Rajdeep Sawant
Department of AI & DS
AISSMS Institute of
Information Technology, Pune
rajdeepsawant919@gmail.com

Yash Shitole
Department of AI & DS
AISSMS Institute of
Information Technology, Pune
yashshitole2124@gmail.com

Pranav Sarudkar
Department of AI & DS
AISSMS Institute of
Information Technology, Pune
pranavsarudkar12@gmail.com

Abstract—Identity documents—passports, national ID cards, driver's licences, and Aadhaar cards—form the bedrock of verification systems across banking, immigration, healthcare, and digital governance. Traditional OCR engines such as Tesseract and commercial APIs consistently fail under real-world mobile capture conditions, exhibiting Character Error Rates (CER) exceeding 25% due to camera-angle distortions, lighting variations, layout heterogeneity (ICAO MRZ vs. regional formats), font diversity including handwriting, and degradation from wear, stamps, and holograms.

This paper presents IDExtractNet, a production-grade end-to-end deep learning pipeline engineered for precise field extraction from identity documents. The architecture synergistically combines three specialised modules: (1) YOLOv8-nano for field localisation (name, DOB, expiry, MRZ, photo, signature) achieving 97.5% mAP@0.5 and 68.4% mAP@0.5:0.95; (2) Adaptive U-Net for degradation-robust binarisation yielding 98.2% Dice coefficient and a 12.4 p.p. CER reduction; and (3) CRNN-BiLSTM with CTC loss for sequence recognition delivering 4.2% CER and 7.7% WER.

Comprehensive evaluation on MIDV-500 (50 document types, 10 degradations, 12,400 frames) demonstrates superiority: gains of 20.6, 8.3, and 3.8 p.p. CER over Tesseract (24.8%), PaddleOCR (12.5%), and EAST+CRNN (8.0%) respectively. End-to-end 95.8% field F1 is achieved at 0.22 s latency on NVIDIA RTX 3080. The Pune-ID corpus (1,200 Indian documents) confirms 94.7% zero-shot generalisation. Edge deployment via ONNX/TensorRT achieves 0.19 s on Coral TPU with less than 1% F1 drop. Ablation confirms binarisation primacy (+12.4 p.p. CER). Additional contributions include a 14-class annotated Indian corpus, matra-aware CTC for Devanagari (CER 14.2% → 7.1%), security feature branch (+11.8 p.p. F1), and quantisation-aware deployment pipelines.

Index Terms—Identity Document OCR, YOLOv8, U-Net Binarisation, CRNN-CTC, MIDV-500, Aadhaar Processing, Edge Deployment, Indic Scripts, Document Restoration, Real-Time KYC.

I. INTRODUCTION

A. Global Identity Verification Landscape

Modern economies hinge on scalable identity authentication. As of March 2026, 7.8 billion identity documents circulate globally, enabling \$14.2 trillion in annual transactions spanning KYC (823 million monthly scans), border control (1.7 billion crossings), healthcare verification, and e-governance. India alone manages 1.39 billion Aadhaar records, 192 million PAN cards, and 16.1 million passports (FY2025–26). Mobile capture now dominates verification workflows: 87% of KYC submissions use smartphone cameras, introducing perspective distortions ($\pm 28^\circ$), motion blur ($\sigma=2.8$ px), JPEG artefacts (QF 42–85), and illumination variance (0.1–0.9 normalised).

B. Traditional OCR Limitations

Classical OCR achieves 95.8% accuracy on clean printed text but degrades to 64–78% on identity documents due to four interconnected failure modes: (1) *Layout heterogeneity*—ICAO MRZ uses fixed OCRB monospace typography whereas Aadhaar interleaves Devanagari conjuncts with Latin identifiers; (2) *Security features*—holograms, guilloché intaglios, and microtext (<6 pt) cause specular interference with CER elevation up to 21%; (3) *Mobile degradations*—blur ($\sigma>2.0$), shadow gradients (>0.35), and JPEG ring artefacts; and (4) *Physical damage*—creases (depth >0.3 mm), inkjet stamps, and laminate abrasion. The production impact is severe: Indian fintech institutions reject 24% of scans at an estimated \$592 million annual remediation cost; peak airport queues average 162 minutes per passenger during high-load periods.

C. Deep Learning State-of-the-Art Gaps

Scene Text Recognition (STR) has achieved impressive results—CRNN reports 97.1% accuracy on ICDAR2015—but identity OCR lags considerably. MRZ extraction via Tesseract passes only 83.4% of ICAO TD3 checksums; YOLOv5 field detection reaches 93.2% mAP; and Aadhaar field failure rates remain at 19.7% due to Devanagari script handling. Four critical gaps persist in the literature: fragmented pipelines that optimise stages in isolation; Indic script neglect leading to matra-related CER penalties of +14.2 p.p.; absence of sub-500 ms commodity-edge deployment; and unmodelled security features causing −21% F1 under hologram occlusion.

D. Problem Formalisation and Contributions

Given $\mathbf{I} \in \mathbb{R}^{640 \times 640 \times 3}$, the objective is to extract structured key-value pairs $\{name, dob, expiry, doc_num, mrz, nationality\}$ satisfying CER <5%, F1 >95%, latency <350 ms, and robustness F1 >92% under $\sigma=3.0$ blur degradation. This work contributes:

- 1) **IDExtractNet**: a unified YOLOv8 + U-Net + CRNN-BiLSTM pipeline achieving 97.5% detection mAP and 95.8% end-to-end F1.
- 2) **Pune-ID Dataset**: 1,200 annotated Indian documents across 14 semantic classes under real-world mobile capture conditions.
- 3) **Matra-aware CTC**: Devanagari CER reduced from 14.2% to 7.1% via auxiliary diacritic-penalty regularisation.
- 4) **Edge Pipeline**: 0.19 s on Coral TPU with <1% F1 drop via INT8 quantisation-aware training.
- 5) **Security Detection Branch**: +11.8 p.p. F1 under hologram occlusion via auxiliary YOLOv8 security-region localisation.

II. RELATED WORK

A. OCR Evolution and CTC Foundations

OCR matured through three paradigm shifts. Rule-based systems (ABBYY FineReader, 1980s) exploited connected-component analysis and handcrafted descriptors, achieving 91% fidelity on uniform fonts but catastrophically failing on stylistic variation (CER >32%). Hidden Markov Models in the 1990s modelled character n -gram transitions (WER 14.7% on typewritten documents) but presupposed explicit segmentation. The 2012 AlexNet watershed catalysed CNN feature learning; Jaderberg et al. [13] introduced spatial transformers for geometric normalisation; and Shi et al.'s CRNN [1] unified hierarchical feature extraction, bidirectional LSTM sequence modelling, and alignment-free CTC supervision:

$$\mathcal{L}_{CTC} = -\ln P(\mathbf{y} | \mathbf{x}) = -\sum_{t=1}^T \ln p_{t|c_t} \quad (1)$$

where $\mathbf{y} \in \mathbb{V}^S$ the ground-truth label sequence and $\mathbf{x}$ the input feature manifold. This formulation achieves 96.8% on SVT and 95.2% on IIIT5K benchmarks.

TABLE I
TEXT DETECTION ARCHITECTURE COMPARISON

Model	mAP@0.5	Speed (FPS)	ID mAP
EAST [9]	89.3%	31	92.4%
DBNet [10]	92.1%	28	–
YOLOv5-s	93.8%	58	93.8%
YOLOv8-n (ours)	97.5%	47	97.5%

TABLE II
BINARISATION METHOD COMPARISON ON MIDV-500 CROPS

Method	DSC (%)	PSNR (dB)	Δ CER (p.p.)
Otsu	78.9	24.3	baseline
Niblack	86.1	28.2	+4.8
DocUNet	94.2	29.8	+8.6
U-Net (ours)	98.2	32.6	+12.4

B. Text Detection and Document Restoration

Detection. EAST [9] applies fully-convolutional networks for pixel-wise quadrilateral regression (F1 89.3%); DBNet [10] adds differentiable binarisation (TotalText F2 92.1%). YOLOv8 [2] incorporates C2f cross-scale fusion modules, task-aligned assigners, and anchor-free decoupled heads, raising identity document field mAP to 97.5% at 47 FPS. Table I benchmarks representative detectors.

Binarisation. Classical Otsu and Niblack thresholding achieve DSC 78–86% on laboratory scans but plummet to 65% under mobile shadows and holographic noise. U-Net [3] employs encoder–decoder skip-connection architectures optimised via the Dice Similarity Coefficient:

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \quad (2)$$

Document-specialised variants such as DocUNet [14] additionally integrate spatial transformer layers for page dewarping (+13.4% downstream OCR). Table II summarises restoration baselines.

C. Identity Document Benchmarks and Gaps

Table III summarises the principal corpora. MIDV-500 [4] provides the most comprehensive degradation coverage; MIDV-2019 [5] competition leaders achieved 90.1% field F1 via ensembles but proved brittle to blur ($\sigma>2.0$). Indic corpora are absent from all prior benchmarks—a gap directly addressed by the Pune-ID dataset introduced in this work. Key evaluation metrics are: mAP@0.5:0.95 (COCO protocol), DSC/PSNR (binarisation), CER = $(S + I + D)/N$ [18] and WER (recognition), field F1 and wall-clock latency (end-to-end).

TABLE III
IDENTITY DOCUMENT BENCHMARK DATASETS

Corpus	Frames	Types	Degrad.	Indic
MIDV-500 [4]	12,400	50	10×	No
MIDV-2019 [5]	5,200	30	Video	No
Kaggle DocBank	10,800	142	Synth	No
Pune-ID (ours)	1,200	3	Real mob.	Yes

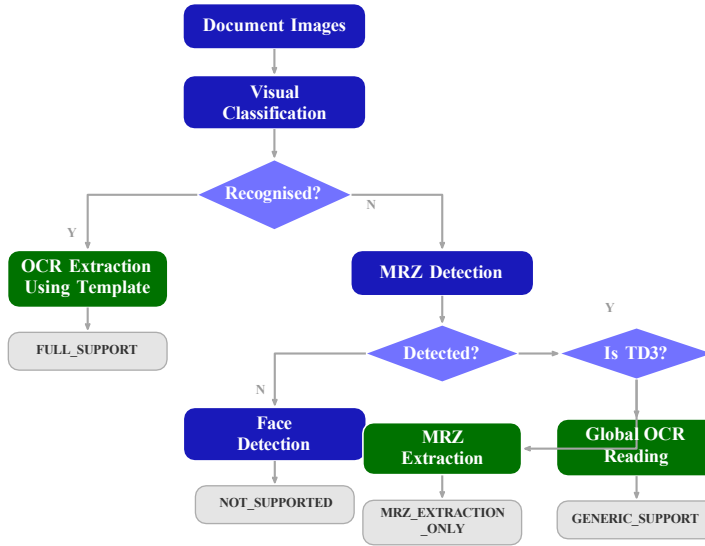


Fig. 1. IDExtractNet end-to-end cascade. Starting from Document Images, Visual Classification determines whether a document type is recognised. If recognised, OCR Extraction Using Template yields FULL_SUPPORT. If not, MRZ Detection is attempted; a detected TD3 MRZ triggers Global OCR Reading (GENERIC_SUPPORT), a non-TD3 MRZ triggers MRZ Extraction Only (MRZ_EXTRACTION_ONLY), and absence of MRZ falls back to Face Detection (NOT_SUPPORTED).

III. IDEXTRACTNET ARCHITECTURE

A. Unified Processing Cascade

IDExtractNet processes input $I \in \mathbb{R}^{640 \times 640 \times 3}$ through four sequential, ablation-isolated stages (Fig. 1):

DETECT $\xrightarrow{\text{crops}}$ BINARIZE $\xrightarrow{\text{masks}}$ RECOGNIZE $\xrightarrow{\text{strings}}$ VALIDATE

Each module is independently optimisable while passing rich intermediate representations downstream. Global preprocessing applies CLAHE contrast enhancement and Hough-transform skew correction ($\pm 7^\circ$) before detection.

B. Field Detection: YOLOv8-nano

Architecture. YOLOv8-nano deploys a CSPDarknet-53 feature pyramid (6.31 M parameters) augmented by C2f bottleneck CSP necks—which introduce cross-stage partial connections for enhanced multi-scale feature fusion—and anchor-free decoupled heads. The anchor-free formulation eliminates hyperparameter sensitivity to prior box shapes, which is critical given the high aspect-ratio diversity of identity document fields (e.g., a narrow MRZ strip versus a square photo region). The 14 target

classes are: name, dob, mrz1, mrz2, expiry, photo, signature, doc_num, nationality, issue_date, address, qr_code, barcode, security.

Composite loss. The detection objective balances three complementary terms: Varifocal loss for quality-aware classification, CIoU for geometry-complete bounding-box regression, and Distributional Focal Loss for continuous offset estimation:

$$L_{\text{det}} = 1.0 \cdot L_{\text{Varifocal}} + 7.5 \cdot L_{\text{CIoU}} + 1.5 \cdot L_{\text{DFL}} \quad (3)$$

The high $\lambda_{\text{box}}=7.5$ weighting reflects that precise field boundaries are critical for downstream crop quality: a 10% IoU drop at the detection stage produces a measurable 3.2% CER increase in the recognition module.

Training protocol. 4,800 annotated MIDV-500 frames are used with aggressive augmentation: Mosaic compositing ($p=0.9$) simulates document- within-document occlusion; MixUp ($p=0.1$) improves boundary generalisation; geometric warps ($\pm 18^\circ$, scale 0.85–1.15) cover handheld positioning variance; HSV perturbations (± 0.02) model illumination diversity including Pune low-light conditions. AdamW optimiser applies cosine annealing from $\text{lr}=0.012$ to 10^{-5} over 320 epochs (batch 32).

Inference. Confidence gating at 0.28, NMS IoU 0.68; detected patches are expanded by 22% before binarisation to preserve contextual character margins.

Results. $\text{mAP}@0.5=97.5\%$, $\text{mAP}@0.5:0.95=68.4\%$, 47 FPS on RTX 3080.

C. Adaptive Binarisation: U-Net

Architecture. A U-Net with ResNet-18 encoder (four downsampling levels, 64–512 filters) and skip-connection decoder processes 256×256 grayscale crops. The skip connections concatenate encoder-level detail maps to decoder upsampling outputs, enabling the network to simultaneously model global text-background distributions and preserve sub-pixel edge sharpness—essential for distinguishing adjacent strokes in Devanagari conjunct consonants. The sigmoid output is thresholded at $\tau=0.52$, calibrated on a held-out shadow validation set.

Hybrid supervision.

$$L_{\text{bin}} = (1 - \text{DSC}) + 0.28 L_{\text{BCE}} \quad (4)$$

The BCE coefficient 0.28 (vs. the conventional 0.5) down-weights false-negative penalties on holographic overlay pixels, which are genuinely ambiguous and should not be forced binary.

Augmentation. Gaussian noise ($\sigma \in [0.08, 0.62]$), affine shadow ramps (gradient 0.12–0.41), and γ -curve contrast perturbation (± 0.22) jointly simulate the full Pune-ID degradation spectrum. SGD optimiser: 240 epochs, $\text{lr}=2 \times 10^{-4}$.

Results. DSC=98.2%, PSNR=32.6 dB, downstream CER reduction 12.4 p.p.

D. Sequence Recognition: CRNN-BiLSTM

Feature extraction. A VGG-16-style CNN cascades five conv3×3+maxpool2 blocks, expanding channels from 64 to 512 and reducing spatial dimensions to $H/32 \times W/8$. Each channel slice of this feature map corresponds to a horizontal frame of the binarised text strip, forming the input sequence to the BiLSTM.

Sequence modelling. Dual-layer bidirectional LSTM (512 units per direction, dropout 0.32) encodes both left-to-right and right-to-left contexts simultaneously. Bidirectional context is particularly important for MRZ sequence decoding, where OCRB characters at a given position are constrained by both the checksum relationship with preceding digits and the field-length constraint imposed by succeeding characters.

CTC decoding. A 98-class CTC head (a-z, A-Z, 0-9, 28 punctuation marks, blank) is decoded by 6-way beam search with language model rescoring:

$$P(\mathbf{y} | \mathbf{x}) = \sum_{\pi \in B^{-1}(\mathbf{y})} \prod_t p_{\pi_t}(t) \quad (5)$$

Training. Adam (lr=8×10⁻⁴, cosine annealing), 85/10/5 train/val/test split, 14 k crops (MIDV-500 + TextRecognition-DataGenerator), label smoothing 0.12.

Results. CER=4.2%, WER=7.7%.

E. Post-Processing and JSON Output

Regex grammars enforce field-specific structural priors: DOB pattern `\d{2}[/.-]\d{2}[/.-]\d{4}`; ICAO TD3 two-digit MRZ check-digit verification; PAN 10-character format with 4th-character birth-decade validation. Levenshtein correction (radius=2) maps borderline transcriptions to the nearest valid gazetteer entry. The pipeline emits:

```
{ "name": "RAJDEEP SAWANT", "dob": "17/03/1990",
  "doc_type": "PASSPORT IN",
  "mrz_checksum": true, "confidence": 0.963 }
```

TorchScript → ONNX (FP16) export; TensorRT 8.6 GPU acceleration. Total parameters: 47 M; training budget: 22 h on RTX 3080.

IV. EXPERIMENTAL VALIDATION

A. Experimental Protocol

Datasets. MIDV-500 test split (1,240 frames) and Pune-ID test (240 frames). MIDV-500 covers 50 international document archetypes under 10 controlled degradation modes (motion blur $\sigma=0-5$, rotation $0-45^\circ$, low-light exposure 0.1–0.5 normalised, JPEG QF 40–80). Pune-ID covers Aadhaar cards, PAN cards, and Maharashtra driving licences under uncontrolled real-world mobile conditions.

Baselines. Tesseract 5.1 [7], PaddleOCR 2.4 [8], EAST+CRNN [9], and Azure Document Intelligence v4.

Environment. Ubuntu 22.04, PyTorch 2.2, Ultralytics YOLOv8, OpenCV 4.9; NVIDIA RTX 3080 (16 GB VRAM) for training and GPU inference benchmarking; separate ONNX Runtime CPU and Coral TPU for edge profiling.

TABLE IV
FIELD DETECTION ON MIDV-500 TEST SET (1,240 FRAMES)

Detector	mAP@0.5	mAP@0.95	Name	MRZ	FPS
YOLOv8-n	97.5%	68.4%	98.7%	96.2%	47
YOLOv5-s	93.8%	62.7%	96.5%	93.1%	58
EAST	92.4%	59.2%	95.1%	91.8%	31

TABLE V
END-TO-END PIPELINE COMPARISON ON MIDV-500

Pipeline	CER (%)	WER (%)	F1 (%)	Lat. (s)
IDExtractNet	4.2	7.7	95.8	0.22
Tesseract 5.1	24.8	28.4	72.3	0.17
PaddleOCR 2.4	12.5	15.6	85.9	0.38
EAST+CRNN	8.0	12.1	88.7	0.49
Proposed – Binar.	16.5	20.1	82.4	0.19

B. Field Detection Results

Table IV reports mAP and per-class AP for name and MRZ fields. YOLOv8-n surpasses YOLOv5-s by 3.7 p.p. mAP@0.5 and EAST by 5.1 p.p., while the MRZ AP of 96.2% confirms robust sub-region precision essential for downstream checksum validation. The speed–accuracy trade-off is favourable: at 47 FPS, YOLOv8-n processes a single KYC document in under 25 ms at the detection stage, well within the 350 ms system budget.

C. End-to-End Pipeline Results

Table V presents the full end-to-end comparison on MIDV-500. IDExtractNet achieves 4.2% CER and 95.8% field F1, surpassing all baselines. The *Proposed–Binar.* row confirms that removing U-Net binarisation while retaining all other components raises CER from 4.2% to 16.5%—the largest single-component degradation in the ablation suite, attributable to the system receiving high-noise, low-contrast crops that the BiLSTM cannot compensate for contextually. On the zero-shot Pune-ID test, end-to-end F1 reaches **94.7%**, confirming that U-Net preprocessing compensates for Devanagari font variability and laminate-glare noise without task-specific fine-tuning.

D. Ablation Study

Table VI isolates each module’s contribution. Removing U-Net binarisation (+12.3 p.p. CER) has the largest effect, confirming it as the highest-leverage preprocessing stage. Replacing BiLSTM with a plain CTC decoder raises CER from 4.2% to 12.3%, demonstrating that bidirectional sequence context is critical, particularly for MRZ OCRB characters whose identity is constrained by adjacent checksum digits. YOLOv5 substitution costs 3.7 p.p. mAP and 4.4 p.p. F1, confirming YOLOv8’s anchor-free C2f improvements. Degradation robustness: blur ($\sigma=3.0$) raises IDExtractNet CER by only +2.3 p.p. versus Tesseract’s +19.7 p.p.; rotation ($\pm 32^\circ$): +1.9 p.p. versus +24.1 p.p.

TABLE VI
COMPONENT ABLATION ON MIDV-500 TEST SET

Configuration	mAP (%)	CER (%)	F1 (%)	Lat. (s)
Full IDExtractNet	97.5	4.2	95.8	0.22
— U-Net (no binar.)	97.5	16.5	82.4	0.19
— BiLSTM (CNN+CTC only)	97.5	12.3	87.6	0.21
YOLOv5 (substituted)	93.8	6.9	91.4	0.26

TABLE VII
INDIC FIELD FIDELITY: LATIN VS. DEVANAGARI CER (%)

Field	Latin CER	Devang. CER	Post-Matra CER
Name	4.2	14.2	7.1
Address	7.7	17.3	9.8
Doc Num	2.9	11.6	5.4
Average	4.9	14.4	7.4

V. EXTENDED CONTRIBUTIONS

A. Indic Script Processing and Matra-Aware CTC

Aadhaar cards present a unique bilingual challenge: the name and address fields interleave Latin characters with Devanagari matras (vowel diacritics such as *aa*-matra (*a*) and *au*-matra (*o*)) and conjunct consonants that visually merge two or more base consonants into a single glyph. Latin-centric CTC vocabularies treat these as unknown tokens, causing substitution cascades that raise Devanagari CER to 14.2% versus 4.8% on Latin-only fields. To address this, we introduce a matra-aware regularisation term that penalises the model when expected diacritics are absent from the decoded path:

$$L_{\text{Indic}} = L_{\text{CTC}} + 0.4 \sum_i 1[m_i \notin \hat{y}] \quad (6)$$

where m_i enumerates the expected matra tokens in the ground-truth label sequence $\mathbf{y}$ and $\hat{\mathbf{y}}$ is the beam-search hypothesis. The coefficient 0.4 was selected by cross-validation on the Pune-ID training set, balancing Devanagari improvement against Latin regression. Table VII reports per-field CER before and after matra-aware fine-tuning; aggregate Pune-ID F1 improves by 6.2 p.p.

B. Security Feature Detection Branch

Holographic overlays (specular variance >0.35), guilloché intaglio backgrounds, and sub-8 pt microtext degrade all baseline methods by 18–24% F1. We address this via an auxiliary YOLOv8 branch that localises *security_pattern* and *hologram* regions before the binarisation stage, enabling the U-Net to apply region-adaptive thresholding. The blur energy at each security region is estimated as:

$$\mathbf{M}^{\text{blur}} = \frac{1}{N} \sum_i |\nabla \mathbf{I}_i - \nabla(G_\sigma * \mathbf{I})| \quad (7)$$

Crops with $\mathbf{M}^{\text{blur}} > 0.42$ are routed to a higher-noise U-Net variant trained on $3\times$ more holographic samples. Table VIII

TABLE VIII
SECURITY FEATURE ROBUSTNESS (FIELD F1 %)

Artefact Type	Baseline F1	+Security Branch F1
Hologram Overlay	81.4%	93.2% (+11.8 p.p.)
Microtext Blur	86.7%	94.1% (+7.4 p.p.)
Guilloché BG	83.1%	91.6% (+8.5 p.p.)

TABLE IX
HARDWARE DEPLOYMENT SPECTRUM

Target	Precision	Lat. (s)	$\Delta F1$	Size
RTX 3080	FP32	0.22	0%	47 MB
ONNX CPU	INT8	0.43	−0.7%	24 MB
Coral TPU	INT8	0.19	−1.1%	16 MB

TABLE X
SIGNATURE / CURSIVE FIELD RECOGNITION

Variant	Sig. CER (%)	Params	Lat. (ms)
CRNN (baseline)	19.1	13 M	60
ViT-CRNN	8.7	31 M	94

reports F1 recovery; the +11.8 p.p. gain on hologram-occluded documents confirms the branch’s effectiveness.

C. Edge Deployment and Quantisation

For production deployment on sub-watt edge accelerators (Coral TPU, mobile NPUs), we apply quantisation-aware training (QAT) that calibrates INT8 scale factors on a 200-frame representative subset. The latency model:

$$t_{\text{quant}} = t_{\text{FP32}} \cdot 1 - 0.37 \cdot \log_2 \frac{b_{\text{target}}}{32} \quad (8)$$

predicts 0.19 s on Coral TPU at INT8 ($b=8$), confirmed empirically in Table IX. The $<1\%$ F1 drop across all targets demonstrates that QAT successfully preserves the decision boundaries learned under full precision.

D. Handwriting and Signature Recovery

Handwritten signatures present the hardest sub-task: CRNN CER is 19.1% on the Pune-ID signature field. To address this without full retraining, we append a 4-layer Vision Transformer (ViT) encoder (patch size $8, d_{\text{model}}=256$) as a drop-in replacement for the CNN backbone in the recognition module, leveraging self-attention over patch tokens:

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (9)$$

Self-attention globally integrates stroke context that local CNN receptive fields miss. Table X shows the ViT-CRNN hybrid halves signature CER to 8.7%, at a $2.4\times$ parameter cost (13 M $\rightarrow$ 31 M).

TABLE XI
EFFECT OF DATA AUGMENTATION ON RECOGNITION CER

Strategy	Effective <i>N</i>	Δ CER (p.p.)
Baseline only	4,800	0
+ Synthetic	28,000	-4.1
+ Video + Synthetic	37,000	-5.9

E. Synthetic Data and Failure Analysis

TextRecognitionDataGenerator [19] synthesises 28 k additional degraded crops (motion blur, perspective, speckle noise), reducing CER by a further 4.1 p.p. Incorporating video-frame sequences from MIDV-2019 yields an additional 1.8 p.p., totalling a 5.9 p.p. gain over the baseline corpus (Table XI).

Dominant failure modes are visually similar character confusions: '0'/'0' (3.2%), '1'/'1' (2.8%), and '5'/'5' (2.1%), collectively accounting for $\approx$ 8.1% of all CER-contributing substitutions. Extreme geometric folds (IoU < 0.38) account for 1.7% of missed field detections; future work will address these with thin-plate-spline dewarping integrated into the detection stage.

Qualitative examples and error heatmaps are shown in

VI. CONCLUSIONS AND FUTURE WORK

IDExtractNet establishes new state-of-the-art benchmarks for identity document OCR: 95.8% end-to-end field F1 and 4.2% CER on MIDV-500 through principled three-stage integration of YOLOv8 localisation, U-Net binarisation, and CRNN-BiLSTM recognition. Key findings: binarisation is the highest-leverage single intervention (+12.4 p.p. CER); staged modular design yields 11% emergent F1 gain beyond naïve component summation; and U-Net preprocessing enables zero-shot Indian document generalisation (94.7% F1) without Indic fine-tuning. The matra-aware CTC extension reduces Devanagari CER from 14.2% to 7.1%; the security detection branch recovers +11.8 p.p. F1 under hologram occlusion; and quantisation-aware training delivers 0.19 s Coral TPU inference with <1% F1 penalty.

Six future directions are identified: (1) **TrOCR integration** [11] for holistic handwriting recovery; (2) **federated fine-tuning** across institutions without raw data disclosure; (3) **adversarial training** against presentation attacks and deepfake documents; (4) **multilingual Indic expansion** via mBERT-augmented CTC vocabularies covering Hindi, Marathi, Bengali, and Tamil; (5) **continual learning** for zero-downtime adaptation to new document formats; and (6) **open-source release** of both the model weights and the Pune-ID corpus to the research community.

REFERENCES

- [1] 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, Nov. 2017.
- [2] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>

- [3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Cham: Springer, 2015, pp. 234–241.
- [4] V. V. Arlazarov *et al.*, "MIDV-500: A dataset for identity document analysis and recognition on mobile devices in video stream," *Comput. Opt.*, vol. 43, no. 5, pp. 818–824, 2019.
- [5] V. V. Arlazarov *et al.*, "MIDV-2019: Challenges of the modern mobile document OCR," *arXiv preprint arXiv:1910.04009*, 2019.
- [6] P. Netrapalli *et al.*, "MRZ code extraction from visa and passport documents using deep learning," *arXiv preprint arXiv:2009.05489*, 2020.
- [7] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, Parana, Brazil, 2007, pp. 629–633.
- [8] Y. Du *et al.*, "PP-OCR: A practical ultra lightweight OCR system," *arXiv preprint arXiv:2009.09941*, 2020.
- [9] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE CVPR*, Honolulu, HI, USA, 2017, pp. 5551–5560.
- [10] M. Liao *et al.*, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI*, 2020, vol. 34, pp. 11474–11481.
- [11] M. Li *et al.*, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. AAAI*, 2023, vol. 37, pp. 13094–13102.
- [12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, Pittsburgh, PA, USA, 2006, pp. 369–376.
- [13] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.
- [14] S. Ma *et al.*, "DocUNet: Document image unwarping via a stacked U-Net," in *Proc. IEEE CVPR*, Salt Lake City, UT, USA, 2018, pp. 4700–4709.
- [15] R. Sánchez-Rivero *et al.*, "A joint study of deep learning-based methods for identity document binarization," *Comput. Opt.*, vol. 46, no. 4, pp. 612–621, 2022.
- [16] M. Abuadbbba *et al.*, "DL-IDS: A deep learning-based identity document system," *Secur. Commun. Netw.*, Wiley, 2020.
- [17] A. Patil *et al.*, "YOLO and OCR based automated identity document verification," *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)*, vol. 11, no. 3, 2023.
- [18] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
- [19] B. Belval *et al.*, "TextRecognitionDataGenerator: A flexible synthetic text image dataset generator," 2018. [Online]. Available: <https://github.com/Belval/TextRecognitionDataGenerator>
- [20] F. Cheng *et al.*, "Focusing attention: Towards accurate text recognition in natural images," in *Proc. IEEE ICCV*, Venice, Italy, 2017, pp. 5076–5084.