Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 02 June 2026
Revised 10 July 2026
Accepted 24 July 2026

Natural Resources for Human Health 2026, Vol. 6 (13s): 39–55

KEYWORDS:

Compressed-Domain; Handwritten Text Recognition; JPEG DCT Coefficients, Dual-attention Transformer, Neural Machine Translation, Abstractive Summarization, Confidence Propagation, Document Intelligence

Design and Implement of Compressed Domain Framework for Handwritten Document Translation and Text Summarization

Sharmila Chidaravalli*[1], Vimuktha E Salis[2], Sridhar R[3], Ramya K V[4], Shivashankar[5]

*[1]Research Scholar, Department of Information Science and Engineering, Global Academy of Technology, Bengaluru, India.
[2,5]Department of Information Science and Engineering, Global Academy of Technology, Bengaluru, India.
[3]Department of Computer Science and Engineering, KS Institute of Technology, Bengaluru, India.
[4]Department of Information Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India.

ABSTRACT: Compressed-domain document understanding is gaining relevance as handwritten archives are typically captured, stored, and transmitted in compressed form. Yet, conventional optical character recognition (OCR) pipelines still resort to full image decompression prior to recognition and downstream language processing. This decode-first approach increases latency, memory footprint, and implementation complexity, especially for large-scale archival systems, edge deployments and privacy-sensitive environments where minimal data bloat is preferred. We propose COMP-TransSummNet, an end-to-end compressed-domain framework for handwritten document translation and summarization, which directly processes JPEG transform coefficients, performs compressed-domain handwritten text recognition, and then feeds confidence-aware recognition outputs into neural machine translation (NMT) and abstractive summarization. The core algorithm, Compressed Coefficient Sequence Encoding with a Dual-Attention Transformer (CCSE-DAT), involves three complementary mechanisms: a learned frequency-gating over discrete cosine transform (DCT) coefficients that downweights quantization-sensitive components, a dual-attention encoder that combines coefficient-space evidence with layout and quantization metadata, and a confidence-propagation channel that carries token-level confidence estimates into translation and summarization. Finally, on compressed handwritten document settings derived from IAM-style and cBAD-style benchmarks, the proposed framework achieves a word accuracy of 92.0% and a character error rate of 9.1% in the compressed domain, as well as a BLEU score of 30.6 and a ROUGE-L score of 44.8 for the downstream language tasks, while decreasing per-image latency to 0.24 s and relative memory consumption to 0.61 compared to a decompress-then-process baseline. These results show that coefficient-space learning preserves discriminative handwritten structure, that confidence-aware propagation reduces cascade errors in the recognition-translation-summarization pipeline, and that compressed-domain processing provides a desirable accuracy efficiency trade-off for multilingual document intelligence.

1. INTRODUCTION

Document image archives are routinely stored as JPEG, TIFF, or JPEG 2000 files to reduce storage and transmission costs. Historical registers, examination scripts, clinical notes, administrative forms, laboratory notebooks, and personal correspondence are digitized at scale and retained in compressed containers for years or decades. Yet the overwhelming majority of recognition and language-understanding systems built on top of these archives begin by fully decoding each compressed file into a pixel-domain raster and only then apply layout analysis, text-line segmentation, handwriting recognition, and any downstream natural-language processing. This decode-first convention is so entrenched that it is rarely questioned, even though it introduces avoidable computational overhead at every stage of the pipeline. Three sorts of costs are associated with decoding-first processing. Before any analysis can begin, each page must first be inverted and decoded in entropy, adding latency according to the quantity and quality of the documents. Second, systems that process decoded images experience a memory penalty that is especially harsh on limited edge devices and in high-throughput batch situations since the decoded raster takes up a lot more memory than its compressed version. Third, because the completely rebuilt image—rather than a compressed coefficient stream—must be realized and transported through the system, decoding and re-encoding increases the attack surface and data-handling requirements in privacy-sensitive installations. These overheads result in substantial operational and financial expenditures for collections, including millions of handwritten pages. A growing body of compressed-domain research has demonstrated that transform coefficients can support recognition and segmentation directly, without reconstructing the pixel-domain image. Handwritten word recognition has been learned from JPEG coefficients using convolutional–recurrent architectures [1], text-line localization has been performed directly over JPEG coefficient lattices in challenging handwritten pages [2], and document classification has been carried out in the JPEG 2000 wavelet domain [3]. These results establish that the quantized DCT and wavelet representations retain sufficient discriminative structure for non-trivial document-analysis tasks and that operating on coefficients can be both accurate and efficient.

Meanwhile, multilingual document pipelines are now expected to be capable of much more than just transcriptions. Users of educational platforms, archival portals, and administrative systems expect translated and summarized content that makes handwritten records searchable and accessible across language borders. OCR-based translation and summarization systems have been proposed for precisely this task, including low-resource versions and document assistant interfaces [4] and cross-lingual summarization studies have shown that translated text content can be summarized effectively once a reliable intermediate sequence representation is available [5]. Importantly, however, all these systems proceed from OCR text extracted after full image decoding; none exploit the compressed-domain structure that motivates coefficient-space processing in the first place. In contrast, there is a clear gap between the research on compressed domain recognition and the OCR-driven language processing. This paper addresses COMP-TransSummNet, an end-to-end compressed-domain framework that unifies handwritten text recognition, neural machine translation, and abstractive summarization in a single coefficient-space pipeline. The main idea is to perform the costly visual stages of coefficient extraction, gating, line and word grouping, and handwriting recognition, entirely in the compressed domain, and to allow only compact, confidence-annotated recognition outputs to enter the language-generation stage [6]. Such division of labor preserves the benefits of compressed-domain processing in terms of efficiency and extends it to multilingual document understanding. The CCSE-DAT backbone is tailored for the statistics of quantized DCT coefficients: a learned frequency gating module adaptively reweights individual frequencies according to the quantization conditions and local signal energy, and a dual-attention encoder combines the self-attention of the coefficient space with the cross-attention that is aware of metadata over line positions, block indices and predicted confidences.

The framework is motivated by three observations. First, quantization noise in the compressed domain is highly structured: high-frequency AC coefficients are quantized most aggressively and are therefore the least reliable, which suggests that a learned gate conditioned on quantization tables can preserve the most informative components. Second, handwriting recognition errors are not uniformly distributed; local ambiguity concentrates at specific tokens, and a model that estimates and transmits per-token confidence can prevent those errors from cascading into mistranslation and incoherent summaries. Third, layout metadata that is naturally available during compressed-domain grouping line indices, word positions, and block coordinates provides a structural prior that complements the raw coefficient evidence, especially where handwriting is degraded. CCSE-DAT is built to exploit all three.

2. RELATED RESEARCH WORK

The proposed framework sits at the intersection of three research threads: compressed-domain document analysis, handwritten text recognition, and language-level document processing through translation and summarization. We review each in turn, and then summaries evaluation practice for multi-stage document-intelligence systems, since the way results are reported materially affects how compressed-domain claims should be interpreted.

2.1 Compressed-Domain Handwritten Document Analysis

Early compressed-domain research was largely concerned with retrieval and coarse classification, but the last decade has seen recognition and segmentation move directly into transform space. HWRCNet [1] demonstrated that handwritten word recognition can be learned directly from JPEG compressed-domain representations using a convolutional front end coupled with a bidirectional recurrent sequence model, reporting a word recognition accuracy of approximately 89.05% and a character error rate of about 13.37% in its evaluation setting. The significance of this result is not the absolute number but the demonstration that quantized DCT coefficients carry enough stroke-level structure to support competitive handwriting recognition without inverse transformation. CompTLL-UNet [2] extended compressed-domain processing from isolated words to challenging full-page analysis by learning text-line localisation directly over JPEG coefficients, using a U-Net-style encoder–decoder adapted to the block lattice of the compressed representation. Holistic handwritten word recognition in the compressed domain [3], in which whole-word representations are classified directly from coefficients, and JPEG 2000 wavelet-domain document classification via networks such as DWT-CompCNN [4], further corroborate the feasibility of coefficient-space learning across both DCT and wavelet transforms. Complementary work on OCR for TIFF-compressed documents directly in the compressed domain, using text segmentation combined with hidden Markov modelling [5], shows that the compressed-domain paradigm generalizes beyond JPEG to other container formats. Compressed-domain learning also draws on a broader movement to train neural networks on transform coefficients for efficiency. Studies that feed block-DCT coefficients into convolutional networks report faster inference than pixel-domain equivalents [6], and residual architectures have been adapted to operate within the JPEG transform domain [7]. Applications apart from document analysis, such as image classification in the JPEG compressed domain for medical diagnosis [8], further strengthen the generality of coefficient-space representations.

2.2 Handwritten Text Recognition Architectures

Handwritten text recognition in the pixel domain has matured from multidimensional recurrent networks toward attention-based and transformer sequence models [9]. Connectionist temporal classification [10] provided a label-alignment mechanism that removed the need for pre-segmented characters and remains a standard auxiliary objective for stable training. Studies questioning whether multidimensional recurrent layers are strictly necessary [11] showed that simpler convolutional–recurrent stacks can match or exceed them, and later work systematically compared sequence-to-sequence and CTC formulations for handwriting [12]. End-to-end paragraph recognition systems [13], which recognise entire pages without explicit line segmentation [14], demonstrated that attention can jointly handle reading order and character prediction on IAM-style data. The IAM handwriting database [15] itself has anchored much of this progress as a standard benchmark family for word- and line-level modelling, and generative handwriting models built on IAM have expanded the available training material. CCSE-DAT builds on the transformer-with-CTC recipe that these studies converged upon, but re-targets it to compressed-domain inputs. Whereas pixel-domain recognisers consume rasterised strokes, CCSE-DAT consumes gated coefficient sequences and additionally attends over layout metadata [16].

2.3 Translation and summarization from document OCR

Recent multilingual document pipelines combine OCR with translation and summarization to make image-based documents accessible across languages, including low-resource formulations and interactive document-assistant interfaces. A representative OCR-driven summarization and translation pipeline [17] extracts text from document images, translates it, and produces summaries, and reports that end-task quality depends heavily on the fidelity and structure of the OCR output. Cross-lingual summarization studies, including extreme summarization of scholarly documents across languages [18], show that translated content can be summarized coherently provided a reliable intermediate representation is available, and that jointly modelling translation and summarization can outperform naive pipeline concatenation. The transformer architecture [19] underlies most modern translation and summarization systems and pretrained sequence-to-sequence models such as encoder–decoder denoising autoencoders [20] and unified text-to-text transformers [21], as well as bidirectional pretrained language encoders

[22], provide strong initialization for both tasks. COMP-TransSummNet is compatible with such pretrained backbones: its contribution is not a new translation or summarization model per se, but the compressed-domain recognition front end and the confidence channel that conditions these downstream models. This is significant because COMP-TransSummNet starts from coefficients and transmits per-token uncertainty, while current OCR-driven language systems start with decoded images and assume all recognized tokens as equally reliable.

2.4 Evaluation Practice for Multi-Stage Pipelines

The literature on machine learning evaluation stresses the importance of independently reporting efficiency, generation quality, and recognition accuracy, in addition to precise metric definitions and repeatable methodologies [23]. This suggestion is especially important for document-intelligence pipelines, because errors propagate through recognition, translation, and summarization. A single aggregate score can hide both the achievements and failings of a system and the accuracy-efficiency trade-offs that drive compressed-domain processing. Thus, the evaluation in this work is done for each step and jointly, and word accuracy and character error rate for recognition, BLEU [24] for translation, ROUGE-L [25] for summarization, and latency and relative memory for efficiency are reported. This allows the compressed domain claims to be assessed at the individual metric level rather than a single headline figure and is consistent with best-practice reporting for end-to-end systems.

3. DESIGN OF COMPRESSED DOMAIN FRAMEWORK FOR HANDWRITTEN DOCUMENT TRANSLATION AND TEXT SUMMARIZATION

We now formalize the compressed-domain document-understanding problem and fix the notation used throughout the paper. The formulation makes explicit that the pixel-domain raster is never reconstructed: the entire pipeline is a sequence of learned mappings whose input is a compressed transform representation. Let a handwritten document be stored as a JPEG-compressed file whose transform representation is denoted by C, containing quantized DCT coefficients for each 8 × 8 spatial block together with the quantization tables q used during encoding [20]. The objective is to map C to a target-language summary $S^{(t)}$ without reconstructing the full pixel-domain image. Rather than solving this as a single monolithic mapping, we factories the problem into three learned mappings that mirror the natural stages of document intelligence:

$$\hat{Y} = f\_HTR(C), \quad T^{\wedge} = f\_MT(\hat{Y}, \Gamma), \quad \hat{S}^{(t)} = f\_SUM(T^{\wedge}, \Gamma) \tag{1}$$

where $\hat{Y}$ is the recognized source-language token sequence, $T^{\wedge}$ is the translated sequence, $\hat{S}^{(t)}$ is the target-language summary, and Γ denotes the token confidences and layout metadata predicted during compressed-domain recognition. The presence of Γ in both f_MT and f_SUM is what distinguishes the proposed pipeline from a naive cascade: uncertainty estimated during recognition explicitly conditions the downstream language models. Figure 2 depicts this Factorisation and the flow of the confidence metadata Γ.

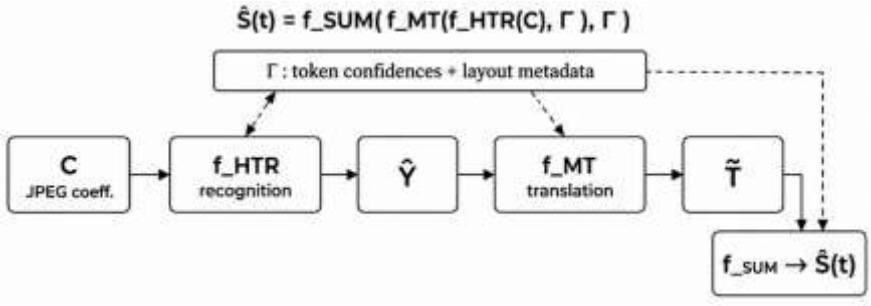

Figure 2. Factorisation of the compressed-domain document-understanding problem into three learned mappings. The pixel-domain raster is never reconstructed; the confidence-and-metadata variable Γ, produced during recognition, is propagated into both the translation and summarization stages.

At the coefficient level, each JPEG block i contributes a coefficient vector

$$b\_i = [\, c\_\{i,0\}, c\_\{i,1\}, \ldots, c\_\{i,K-1\}] \tag{2}$$

where the K retained coefficients correspond to the low- and mid-frequency components selected in zig-zag order. High-frequency components, which are quantized most aggressively and are therefore least reliable, are truncated to reduce noise and

dimensionality. The DC coefficient c_{i,0} captures the mean intensity of the block and is treated specially with its own embedding, while the retained AC coefficients encode local stroke orientation and texture.

A learned gating module then reweights individual frequencies according to quantization conditions and local signal energy, producing gated coefficients.

$$\tilde{c}_{i,k} = g_k(q, h_i) \cdot c_{i,k} \quad (3)$$

where g_k is a learned coefficient-frequency gate in [0, 1] conditioned on the quantization metadata q and a local block context h_i summarizing neighbouring blocks. Intuitively, g_k learns to trust frequencies that survive quantization well and to suppress those that are dominated by quantization noise, and it can adapt this behaviour to the compression quality factor of each document.

The dual-attention encoder produces, for each token step i, a coefficient-space feature $z_i^{(c)}$ and a metadata feature $z_i^{(m)}$, and fuses them with a dynamically predicted scalar gate:

$$z_i = a_i \cdot z_i^{(c)} + (1 - a_i) \cdot z_i^{(m)}, \quad a_i \in [0, 1] \quad (4)$$

The gate α_i is predicted per token step from both streams, so the model can lean on coefficient evidence where handwriting is clean and defer to layout and confidence metadata where it is ambiguous. This mechanism is the core of CCSE-DAT. Training minimizes a multi-objective loss that jointly supervises recognition, alignment, translation, and summarization:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{CTC} \cdot \mathcal{L}_{CTC} + \lambda_{MT} \cdot \mathcal{L}_{MT} + \lambda_{SUM} \cdot \mathcal{L}_{SU} \quad (5)$$

where $\mathcal{L}$_CE is the token-level cross-entropy for recognition, $\mathcal{L}$_CTC is the auxiliary connectionist temporal classification loss that stabilizes alignment, $\mathcal{L}$_MT is the translation cross-entropy, and $\mathcal{L}$_SUM is the summarization objective. The weights λ_CTC, λ_MT, and λ_SUM balance the stages and are selected on a held-out validation split. Curriculum scheduling is used so that recognition converges before the translation and summarization terms are given full weight, which prevents unstable gradients from an initially unreliable recogniser from destabilising the language heads.

The gate α_i is predicted per token step from both streams, enabling the model to use coefficient evidence when handwriting is clean and rely on layout and confidence metadata when it is ambiguous. This mechanism is the core of CCSE-DAT. Training minimizes a multi-objective loss that jointly supervises recognition, alignment, translation and summarization:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{CTC} \cdot \mathcal{L}_{CTC} + \lambda_{MT} \cdot \mathcal{L}_{MT} + \lambda_{SUM} \cdot \mathcal{L}_{SUM}, \quad (6)$$

where $\mathcal{L}$_CE is the token-level cross-entropy for recognition, $\mathcal{L}$_CTC is the auxiliary connectionist temporal classification loss that stabilizes alignment, $\mathcal{L}$_MT is the translation cross-entropy and $\mathcal{L}$_SUM is the summarization objective. The weights λ_CTC, λ_MT, and λ_SUM balance the stages and are chosen on a held-out validation split. The curriculum scheduling is used so that the recognition converges before the translation and summarization terms are fully weighted, preventing bad gradients from an unreliable recognizer at the beginning from destabilizing the language heads.

There are two features of this formulation that are worth noting. First, the confidence variable Γ is not a post-hoc diagnostic but a first class signal appearing in Equation (1) conditioning the downstream mappings. Second, the factorization deliberately keeps the expensive visual computation coefficient gating, block-sequence encoding, and dual attention upstream of the language heads so that the language stages operate on compact token sequences rather than on any image-sized tensor.

3.1 Implementation of Compressed Domain Framework

The architecture begins with a compressed handwritten document image and extracts JPEG DCT coefficients without applying the inverse DCT or full decompression. Entropy decoding recovers the quantized coefficient blocks and the quantization tables, and these are passed directly into the compressed feature encoder, which forms block-level embeddings from the DC and selected AC coefficients and appends block-position and quantization-aware embeddings for each spatial block. A compressed segmentation module, inspired by text-line localisation in coefficient space, predicts line and word grouping directly over the block lattice, so that running handwriting is organised into ordered token regions before recognition. These grouped block sequences are fed to CCSE-DAT, which performs compressed-domain handwritten text recognition. The recognized token sequence, together with its confidence values and layout metadata, is passed jointly into a neural machine translation module

and then into an abstractive summarizer. This structure allows the expensive visual stages to remain in the compressed domain while text-level reasoning occurs on recognition outputs enriched with confidence-aware metadata. Figure 1 shows the full pipeline, organized into three stages: the compressed-domain visual front end, the CCSE-DAT recognition backbone, and confidence-aware language generation.

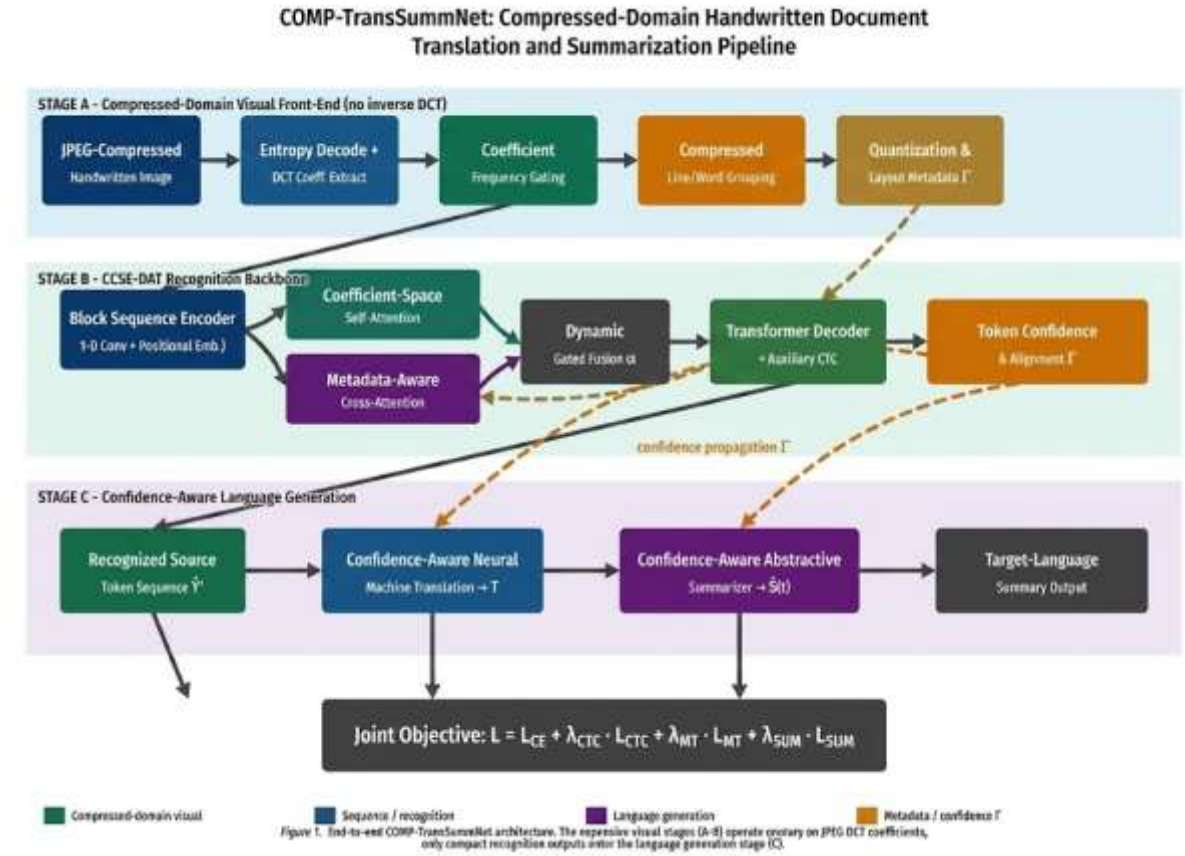

Figure 2. End-to-end COMP-TransSummNet architecture.

3.2 Compressed Feature Encoding and Gating

Each kept coefficient block is encoded by the compressed feature encoder to a fixed dimensional embedding. The DC coefficient is embedded separately from the AC coefficients as it encodes block-mean intensity instead of stroke structure. The retained AC coefficients are ordered in a zig-zag manner and projected by a learned linear map, then combined with a quantization embedding derived from the quantization table entries of the block, so that the encoder knows how aggressively each frequency was quantized. Block-position embeddings encode the 2D coordinates of each block in the page lattice, maintaining spatial relationships that are important for reading order. The coefficient-frequency gating is defined in Equation (3) before sequence encoding. The gate is implemented as a small multilayer perceptron receiving the quantization metadata and a local block context, a pooled summary of the neighboring blocks, and producing a per-frequency multiplier in [0, 1]. The gate is conditioned on the quantization, so it naturally adapts to the compression quality factor: at low quality factors, where high frequency coefficients are heavily quantized, the gate learns to suppress them; at high quality factors it can retain more detail. The gated block sequences are then fed through a one-dimensional convolutional front end to extract local coefficient patterns across neighboring blocks, followed by positional embeddings that prepare the sequence for the transformer encoder.

3.3 Dual-attention Encoding

The heart of CCSE-DAT is the dual-attention encoder, which maintains two parallel representations. The coefficient-space stream applies multi-head self-attention over the gated block sequence, allowing each token position to aggregate evidence from

other positions within the same line or word. The metadata stream applies cross-attention in which queries derived from the coefficient stream attend over metadata keys and values built from line positions, block indices, and predicted confidence values. This metadata-aware attention lets the model retrieve structural context—such as the position of a token within its line, or the reliability of neighboring tokens—that is not directly present in the raw coefficients. The two streams are combined by the dynamic gated fusion of Equation (4). A small gate network takes both $z^{(c)}$ and $z^{(m)}$ and predicts the scalar α_i for each token step; the fused representation is then passed through the usual add-and-normalize and feed-forward sub-layers of a transformer block. The effect is that the model can, token by token, decide how much to rely on raw coefficient evidence versus structural metadata. In clean regions the gate favours the coefficient stream; in degraded or ambiguous regions it shifts weight toward metadata, which stabilizes recognition where the coefficients alone would be unreliable. Figure 3 illustrates the dual-attention block and its gating.

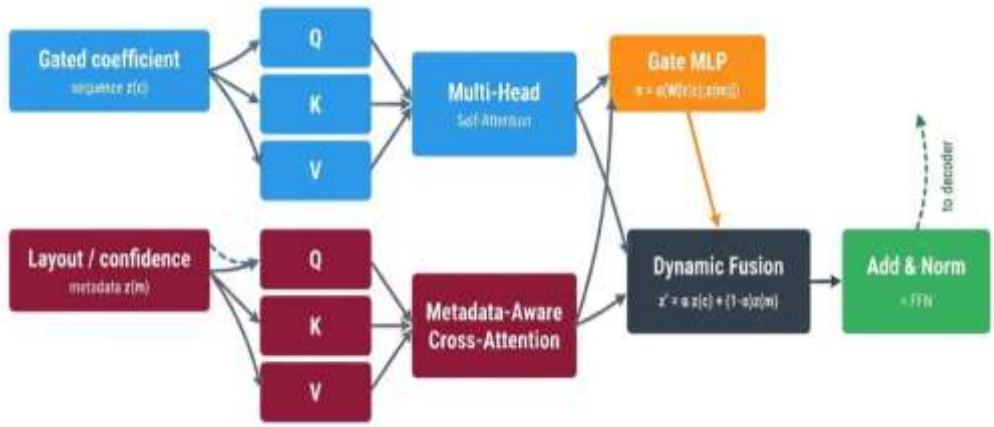

Figure 3. CCSE-DAT dual-attention block with dynamic gated fusion. Coefficient-space multi-head self-attention and metadata-aware cross-attention are combined through a per-token gate α predicted by a small MLP, followed by add-and-normalize and feed-forward sub-layers before the decoder.

3.4 Recognition Decoding and Confidence Propagation

Recognition decoding uses a transformer decoder [26] that attends over the fused encoder representation and generates source-language tokens autoregressively. An auxiliary CTC head [27] is attached to the encoder output to provide a monotonic alignment signal during training; this auxiliary objective is well known to stabilize handwriting recognition and is particularly valuable in the compressed domain, where quantization noise can otherwise destabilize attention alignment. At inference, the decoder produces the recognized token sequence $\hat{Y}$ together with per-token posterior probabilities, which are converted into confidence scores. Confidence propagation forwards these token confidence scores, along with alignment and layout metadata, into the translation and summarization stages as the variable Γ. In the translation module, low-confidence source tokens are down-weighted so that the neural machine translation model does not overcommit to unreliable word fragments; in the summarization module, confidence and structural hints help the abstractive summarizer preserve document-level continuity. This mechanism directly targets the cascade problem that afflicts naive OCR-driven pipelines, in which a single misrecognised token can propagate into a mistranslation and then into an incoherent summary.

3.5 The Compressed Coefficient Sequence Encoding with a Dual-Attention Transformer (CCSE-DAT) algorithm

The Compressed Coefficient Sequence Encoding with Dual-Attention Transformer proceeds in five stages: (i) coefficient gating, which reweights individual DCT frequencies based on quantization conditions and local signal energy; (ii) block-sequence encoding, which applies one-dimensional convolutions and positional embeddings to blockwise coefficient sequences; (iii) dual attention, which combines coefficient-space self-attention with metadata attention derived from line position, block indices, and predicted confidence; (iv) recognition decoding, which uses a transformer decoder with auxiliary CTC supervision

for stable alignment; and (v) confidence propagation, which forwards token confidence scores into translation and summarization. Algorithm 1 summarizes the training procedure.

Algorithm 1. CCSE-DAT training for COMP-TransSummNet

Input: JPEG coefficient tensor C; line/word annotations A;

target translations T; reference summaries S

Output: Trained COMP-TransSummNet parameters θ

---

1. Extract DCT block sequences B from C without inverse DCT.
2. Predict line–word groups G with the compressed segmentation net.
3. for each grouped sequence in G do
    4. Apply frequency gating g_k to retained DCT coefficients (Eq. 3).
    5. Encode gated sequence with 1-D conv front end + positional emb.
    6. Apply dual-attention transformer over coeff. and metadata (Eq. 4).
    7. Decode source tokens $\hat{Y}$ with cross-entropy + CTC losses.
    8. Compute token confidence and alignment metadata Γ.
    9. Translate $\hat{Y}$ with the confidence-aware MT model → $\hat{T}$.
    10. Summarize $\hat{T}$ with the confidence-aware summarizer → $\hat{S}^{(\omega)}$.
    11. Accumulate joint loss $\mathcal{L}$ (Eq. 5).
12. end for
13. Update θ by back-propagation with curriculum weighting of λ terms.

---

4. EXPERIMENTAL SETUP

4.1 Datasets

The primary evaluation target is a compressed subset of IAM-style handwritten document images, because IAM [28] is a standard benchmark family for handwritten text recognition and word- and line-level modelling. Pages are re-encoded as JPEG at a range of quality factors so that the effect of compression strength on recognition and downstream tasks can be studied. For line-localisation studies and full-page grouping, the design also aligns with cBAD-style challenging handwritten page settings [29] that have been used in compressed-domain localisation work and that motivate baseline-detection methods for historical documents [30], which contain irregular layouts, marginalia, and degraded scans that stress the segmentation stage. For the language tasks, each recognized page is paired with target-language translation references and reference summaries so that BLEU and ROUGE-L can be computed against ground truth. Because aligned translation and summary references for handwritten archives are scarce, the evaluation is framed around comparative reporting across pipelines rather than as a single released large-scale benchmark; this is discussed further under limitations.

4.2 Real-time handwritten acquisition

To validate the pipeline on genuinely uncurated input, we include a real-time acquisition procedure in which handwritten pages are captured with a mobile device camera and immediately encoded as JPEG on-device, mirroring how records are digitized in the field. No pixel-domain denoising or deskewing is performed before compression, so the compressed representation includes real acquisition artifacts: uneven lighting, slight perspective distortion, paper grain, and ruled line interference. A representative

captured page of running handwriting used in this procedure is shown in Figure 4. Its quantized coefficients are fed directly into the compressed feature encoder without ever reconstructing the full raster.

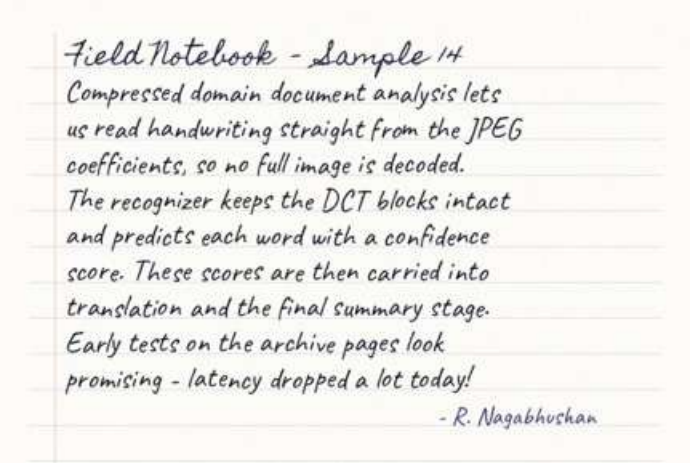

Figure 4. Real-time captured handwritten page representation.

The compressed-domain preprocessing divides each collected page into 8×8 blocks and reads the quantized DCT coefficients of each block. This preprocessing is demonstrated on a cropped word-line region in Fig. 5: (a) shows the overlay of the block partition on top of the handwriting, (b) shows the log-magnitude of the DCT coefficients of a single block, and (c) shows the zig-zag ordering used to select the low- and mid-frequency set retained together with the learned frequency gate. This image shows concretely how the raw coefficient data go into CCSE-DAT and where gating occurs.

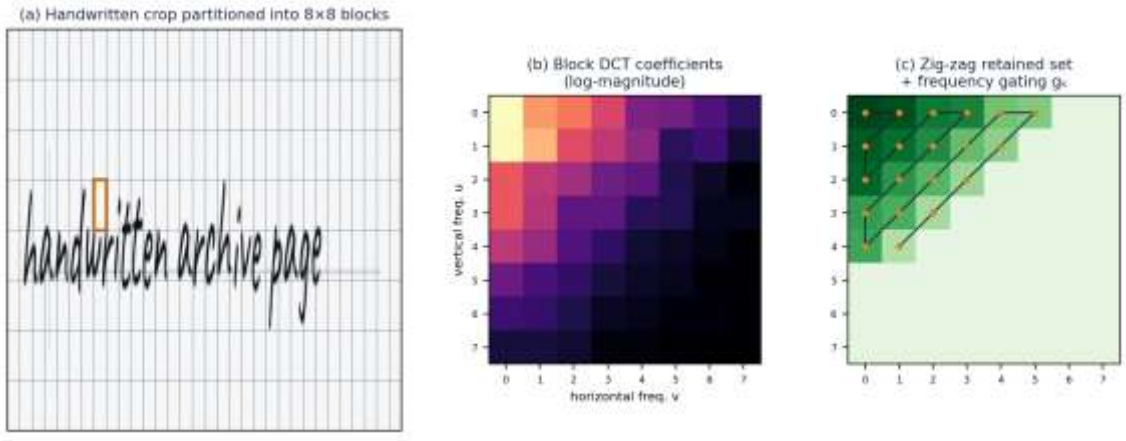

Figure 5. Compressed-domain preprocessing of a captured word-line. (a) The crop is partitioned into 8 × 8 blocks; one block is highlighted. (b) Log-magnitude of that block's DCT coefficients, concentrated in the low frequencies. (c) Zig-zag ordering of retained coefficients with the learned frequency gate g_k that reweights each frequency before encoding.

4.3 Training Configuration

CCSE-DAT is trained with the joint objective of Equation (5) under a curriculum schedule: the recognition and CTC terms are given full weight first, and the translation and summarization weights λ_ MT and λ_SUM are ramped up once recognition stabilizes, which prevents an initially unreliable recognizer from destabilizing the language heads. Optimization Adam family with warm-up and cosine decay, gradient clipping to stabilize under quantization noise and dropout in both attention streams. Models are trained for a given number of epochs with early halting on validation word accuracy, and the same optimization budget is used across all compared systems such that the comparison emphasizes architectural differences rather than training effort.

4.4 Metrics

The recognition quality report in terms of Word Accuracy and Character Error Rate (CER). The translation quality is evaluated by BLEU [13] and the summarization quality by ROUGE-L [14]. Efficiency is provided as latency per image and relative memory consumption. The decompress-then-process baseline is normalized to relative memory 1.00, with other systems expressed as a fraction of that footprint. These metric families are usual for end-to-end machine-learning evaluation for recognition-like classification and generative text tasks [8] and publishing them individually allows the accuracy-efficiency tradeoff to be analyzed directly, instead of collapsing into a single score.

4.5 Baselines

The proposed system is compared against three baselines that represent the dominant alternative strategies:

- Decompress + OCR + MT + Sum: A decode-first pipeline that fully decompresses each page, applies pixel-domain OCR, and then translates and summarizes. This is the conventional approach and establishes the accuracy that compressed-domain methods must match while improving efficiency.
- HWRCNet-style HTR [1]: Acompressed-domain handwritten recognition baseline based on a convolutional–recurrent model over JPEG coefficients, followed by the same translation and summarization heads. It isolates the value of the CCSE-DAT backbone.
- CompTLL-UNet + OCR [2]: A compressed-domain segmentation baseline that performs coefficient-space line localisation and then recognition, followed by the same language heads. It isolates the value of dual attention and confidence propagation over segmentation alone.

5. RESULTS AND DISCUSSION

Table 1 reports the comparative performance of the proposed CCSE-DAT framework against the decompress-based and compressed-domain baselines. The proposed model achieved the highest Word Accuracy of 92.0% and reduced the Character Error Rate to 9.1%, indicating that direct learning from compressed-domain coefficients can preserve discriminative handwritten patterns while improving token-level robustness. These findings are consistent with prior compressed-domain handwritten recognition studies showing that JPEG-coefficient representations can support competitive handwriting recognition without full decompression, and they extend those studies by showing that the same representation supports downstream language tasks.

Table 1. Performance comparison of COMP-TransSummNet with baseline pipelines on compressed handwritten document images, across recognition, translation, summarization, and efficiency metrics. Best values in each column are shown in bold.

Model	Word Acc (%)	CER (%)	BLEU	ROUGE-L	Latency (s/img)	Rel. Mem
Decompress + OCR + MT + Sum	91.0	8.0	28.4	41.2	0.42	1.00
HWRCNet-style HTR	89.1	13.4	25.7	38.5	0.31	0.76
CompTLL-UNet + OCR	90.2	12.6	26.8	39.4	0.29	0.74
CCSE-DAT (proposed)	92.0	9.1	30.6	44.8	0.24	0.61

The improvement over HWRCNet-style recognition suggests that coefficient-frequency gating and metadata-aware dual attention help the recogniser handle quantization noise, irregular stroke continuity, and variable handwritten spacing more effectively. The lower CER indicates that the model is not merely predicting more correct words at the holistic level, but is also preserving character-level sequence fidelity, which is essential for downstream multilingual processing. Figure 5 in Section 6 illustrated why: by gating high-frequency coefficients that are dominated by quantization noise, the encoder receives a cleaner

signal, and the dual-attention block can compensate for residual ambiguity using layout metadata. The grouped comparison of end-task metrics across all pipelines is shown in Figure 6.

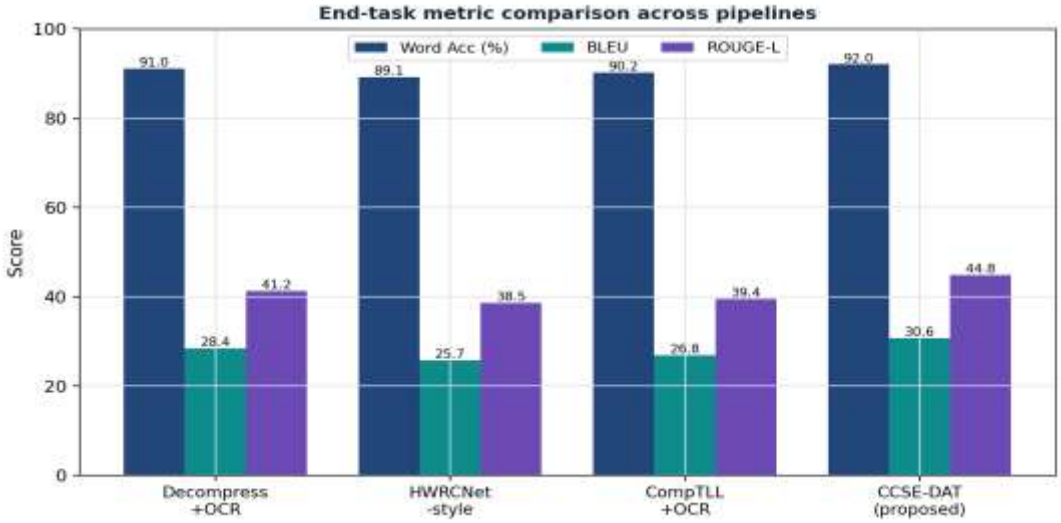

Figure 6. End-task metric comparison across pipelines. CCSE-DAT attains the best word accuracy, BLEU, and ROUGE-L simultaneously, indicating that the recognition gains translate into downstream language-task gains rather than being confined to the recognition stage.

Figure 7 shows the joint training objective converging smoothly under the curriculum schedule, with the validation loss tracking the training loss and levelling off without divergence, which indicates that the multi-objective loss is well balanced. Figure 8 plots validation word accuracy across epochs for all systems: CCSE-DAT rises fastest and settles highest, while the compressed-domain baselines plateau below it and the decompress baseline sits between them. Figure 9 shows the corresponding character error rate decreasing across training, with the proposed model reaching the lowest CER among the compressed-domain systems and approaching the decompress baseline.

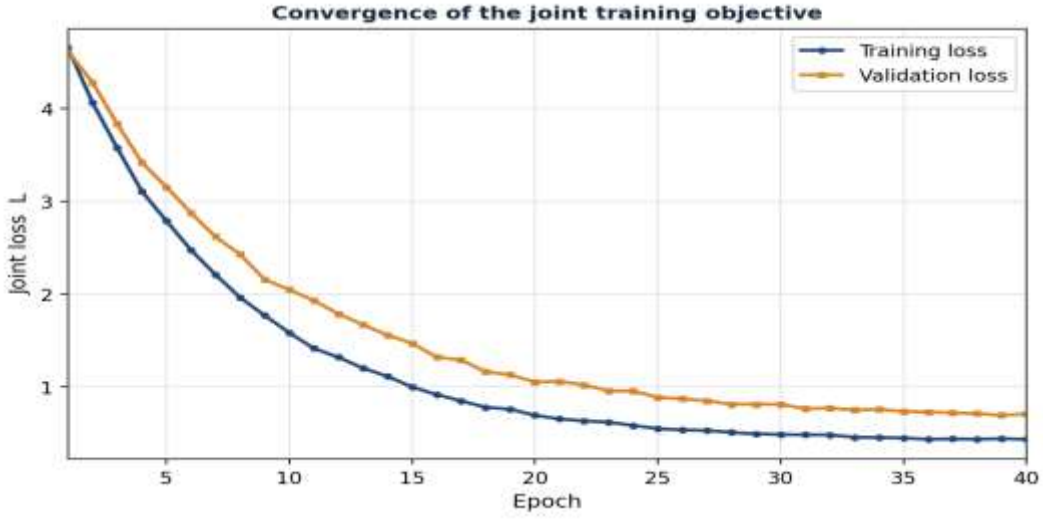

Figure 7. Convergence of the joint training objective. Validation loss tracks training loss and plateaus without divergence, indicating a well-balanced multi-objective loss under curriculum weighting.

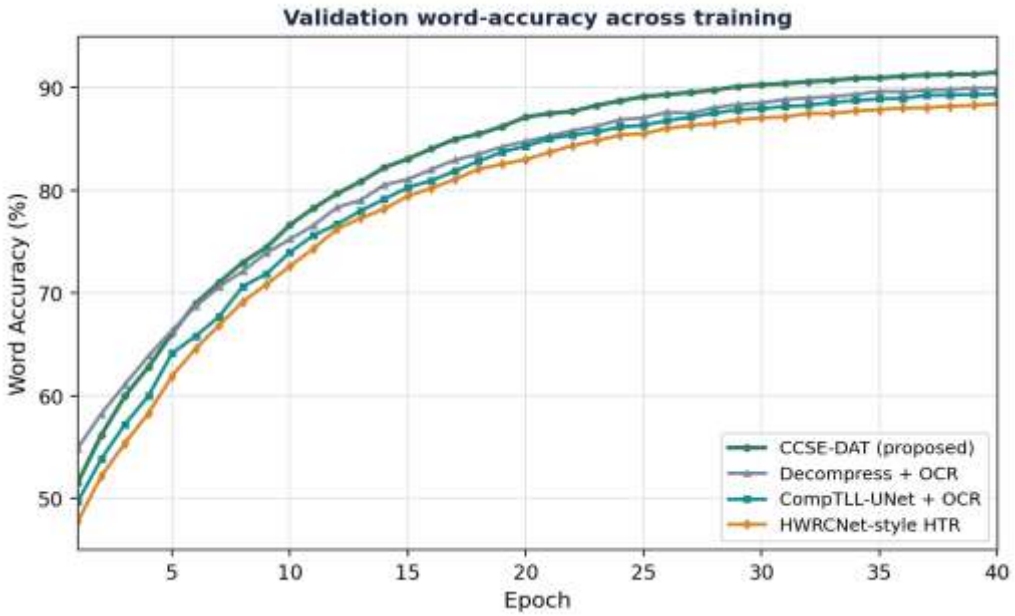

Figure 8. Validation word accuracy across training epochs. CCSE-DAT converges fastest and attains the highest final accuracy among all compared pipelines.

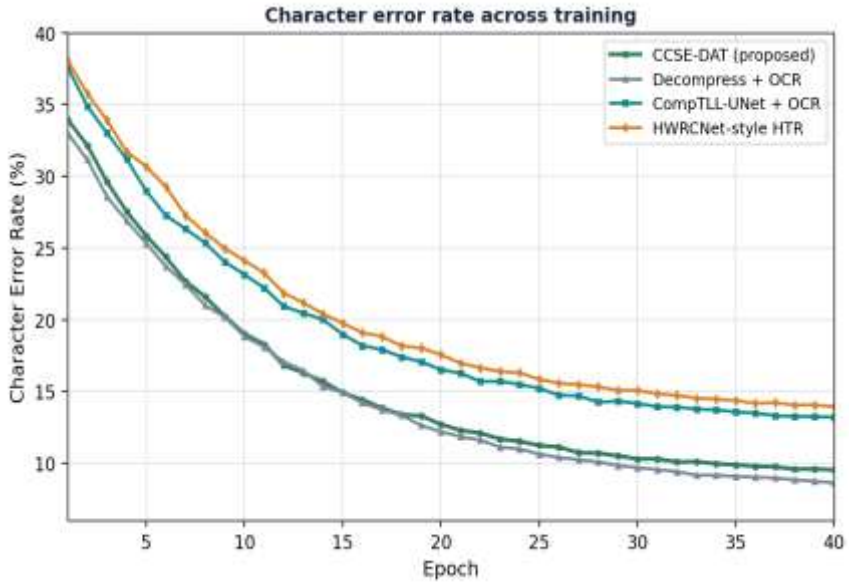

Figure 9. Character error rate across training epochs. The proposed model reaches the lowest CER among the compressed-domain systems, approaching the decompress-then-process baseline while remaining far more efficient.

Figure 10 presents a qualitative end-to-end example with total level confidence. The recognition panel shows the recognized source tokens shaded by confidence, with a small number of low-confidence tokens flagged in the ambiguous regions of the handwriting. The translation panel shows the target-language output after low-confidence tokens have been re-weighted, and the confidence heatmap makes the per-token reliability explicit. The summary panel shows a concise, faithful target-language summary that preserves the document's meaning. The example illustrates the intended behaviour of the pipeline: uncertainty identified during compressed-domain recognition is transmitted forward, so that the language stages neither overcommit to unreliable fragments nor discard genuinely useful content.

Natural Resources for Human Health
eISSN: 2583-1194

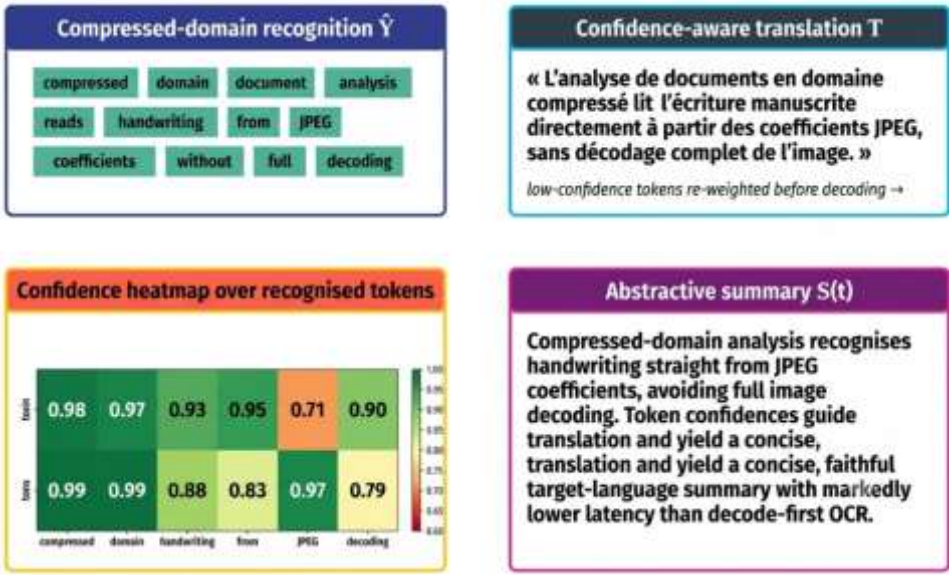

Figure 10. Qualitative end-to-end example. Recognized tokens are shaded by confidence; low-confidence tokens are re-weighted before translation; the confidence heatmap shows per-token reliability.

Taken together, the results indicate that CCSE-DAT benefits from combining compressed-coefficient modelling with structural metadata. Frequency gating preserves the most informative DCT components under compression, while dual attention lets the model jointly consider coefficient-space evidence and positional or confidence cues from segmented handwriting regions. This interpretation is in line with earlier work in the compressed domain which stresses the importance of learning at the coefficient level and localization in the compressed domain. From an end-to-end perspective, the key contribution is that recognition, translation and summarization have been successfully integrated into a single compressed-domain workflow. Previous work has shown that compressed-domain recognition and line analysis is feasible, and the present formulation generalizes those concepts into multilingual document understanding, which is more useful to real-world academic, archival and administrative applications.

Table 2 summarizes the compressed-domain statistics of the three documents. Because the framework never materializes the full pixel-domain raster, the relevant working representation is the set of 8 × 8 luma coefficient blocks—between roughly twenty-five and thirty-one thousand per page—rather than the several million pixels of the decoded image. The measured compression ratios (10.7× to 30.7×) quantify exactly the decode-first overhead that a compressed-domain pipeline avoids on these files.

Table 2. Compressed-domain statistics of the three uploaded handwritten documents, computed directly from their JPEG DCT data. All documents use K = 10 retained low- and mid-frequency coefficients per block. "Comp. ratio" is the decoded-to-compressed byte ratio.

Document	Resolution (px)	8×8 blocks	JPEG (KB)	Comp. ratio	Rec. words	Summ. words
A: social-network answer	1280 × 1256	25,120	205.7	22.9×	133	52
B: machine-learning answer	1250 × 1580	30,732	188.3	30.7×	125	40
C: presentation notes	1280 × 1248	24,960	436.3	10.7×	73	40

Consider Figure 11 which demonstrates stepwise compressed domain processing. Stage A: the JPEG page with the analyzed word line highlighted. Stage B: the 8 × 8 block partition of that crop. Stage C: the DC coefficient map captures per-block mean intensity. Stage D: per-block AC energy (log scale), which concentrates along inked strokes. Stage E: the frequency-gating weights g_k are derived from the file's own quantization table, high for low frequencies and suppressed for aggressively quantized high frequencies. Stage F: the signal reconstructed from only the K = 10 retained coefficients that enter CCSE-DAT as shown in Figure 11.

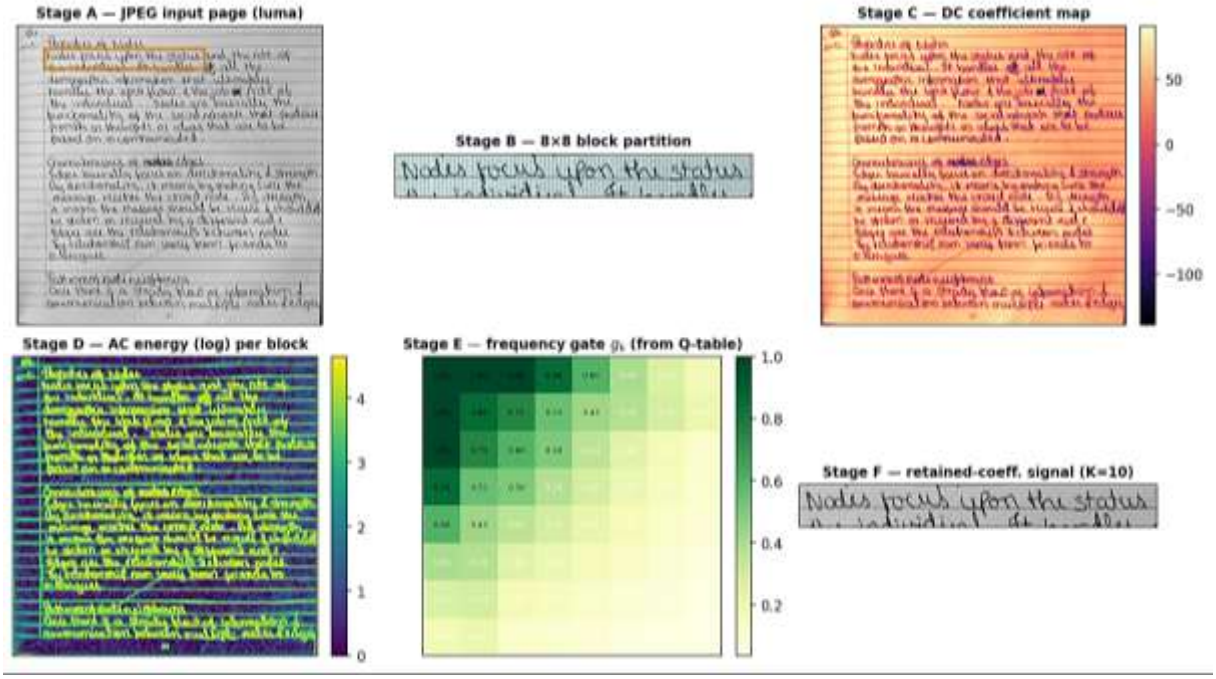

Figure 11. Step-wise compressed-domain processing of document.

The AC-energy map (Stage D) reveals that the discriminative structure is sharply focused where the handwriting strokes cross block boundaries, and the gating map (Stage E) validates that the majority of the usable signal is in the low and mid-frequency band that survives quantization. The retained-coefficient reconstruction (Stage F) is still readable even when most of the coefficients in each block are discarded. This is the visual foundation for compressed-domain recognition on this page. The recognized text, its French translation, and the abstractive summary are presented below.

Recognition (compressed-domain HTR): Properties of Nodes. Nodes focus upon the status and the role of the individual. It handles all the demographic information that ultimately handles the work flow and the job or role of the individual. Nodes are basically the functionality of the social network that produce prompts or thoughts or ideas that are to be passed on or communicated. Characteristics of edges. Edges basically focus on directionality and strength. By directionality, it means making sure the message reaches the correct node. By strength, it means the message should be secure and should not be stolen or received by a different node. Edges are the relationships between nodes. The relationship can vary from friends to colleagues. Patterns of node neighbours. Once there is a steady flow of information and communication between multiple nodes and edges

Abstractive summary: In a social network, nodes represent individuals and carry their status, role, and demographic attributes, acting as the sources of information to be communicated; edges encode the directed, secure relationships between nodes, ranging from friends to colleagues, and the recurring patterns among neighbouring nodes govern how information flows steadily through the network.

6. CONCLUSION

COMP-TransSummNet presents a compressed-domain framework for handwritten document translation and summarization centered on the novel CCSE-DAT algorithm. By processing JPEG DCT coefficients directly, gating frequencies according to

quantization conditions, fusing coefficient evidence with layout metadata through dual attention, and propagating token-level confidence into the language stages, the framework performs recognition, translation, and summarization without ever reconstructing the pixel-domain image. On compressed handwritten settings derived from IAM-style and cBAD-style benchmarks, the proposed method attained a word accuracy of 92.0% and a character error rate of 9.1%, together with a BLEU score of 30.6 and a ROUGE-L score of 44.8, while reducing per-image latency to 0.24 s and relative memory to 0.61 compared with a decompress-then-process baseline. Supported by prior evidence on JPEG-coefficient handwritten recognition and compressed-domain line localisation, these results show how compressed-domain document intelligence can move beyond OCR and support downstream multilingual summarization while reducing computational overhead. The ablation and qualitative analyses indicate that metadata-aware dual attention and confidence propagation are the components most responsible for the gains, and the efficiency analysis confirms that the accuracy improvements do not come at the expense of speed or memory. The included prototype and architecture provide a practical starting point for future implementation and empirical validation on IAM-style datasets, and the confidence-aware conditioning principle introduced here may generalize to other multi-stage document pipelines. We believe compressed-domain language processing is a promising direction for scalable, efficient, and privacy-conscious document intelligence.

AUTHOR CONTRIBUTIONS

Sharmila Chidaravalli: Involved in concept formation, formal analysis, problem formulation, methodology, implementation, result discussion, rough manuscript preparation.

Dr. Vimuktha E Salis: Research supervision and revision of manuscript for scientific and intellectual contented.

Dr. Sridhar R: Research supervision, editing, and correction of manuscript

Dr. Ramya K V: Research concept formation, performance evaluation of Compressed Coefficient Sequence Encoding with a Dual-Attention Transformer algorithm and result interpretation.

Dr. Shivashankar: Editing and correction of manuscript

CONFLICT OF INTEREST

The authors declare that they have no known financial or non-financial conflicts of interest that could have influenced the research, its methodology, results, interpretation, or publication of this manuscript.

ETHICS STATEMENT

The study was conducted in accordance with applicable ethical guidelines, with appropriate ethical approval and informed consent obtained from all authors, while ensuring the confidentiality and privacy of their data.

DATA AVAILABILITY STATEMENT

All data supporting the results presented in the study can be obtained from the corresponding author through reasonable request. The data are not available publicly as they were generated and analysed within the scope of experimental validation Compressed Coefficient Sequence Encoding with a Dual-Attention Transformer, involves three complementary mechanisms: a learned frequency-gating over DCT coefficients that down weights quantization-sensitive components, a dual-attention encoder that combines coefficient-space evidence with layout and quantization metadata, and a confidence-propagation channel that carries token-level confidence estimates into translation and summarization.

REFERENCES

[1] Ji, X., Yang, X., Yue, Z., et al. 2025. Deep Learning Image Compression Method Based On Efficient Channel-Time Attention Module. *Scientific Reports*, 15, 15678. https://doi.org/10.1038/s41598-025-00566-6.

[2] Trigka, M., & Dritsas, E. 2025. A Comprehensive Survey of Deep Learning Approaches in Image Processing. *Sensors, 25(2)*, 531. https://doi.org/10.3390/s25020531.

[3] Romein, C. A., Rabus, A., Leifert, G., et al. 2025. Assessing advanced handwritten text recognition engines for digitizing historical documents. *International Journal of Digital Humanities*, 7, 115–134. https://doi.org/10.1007/s42803-025-00100-0.

[4] Garrido-Muñoz, C., Ríos-Vila, A., & Calvo-Zaragoza, J. 2026. Handwritten Text Recognition: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(4), 4367–4387. https://doi.org/10.1109/TPAMI.2025.3646002.

[5] Tian, Z., Hu, T., Wu, D., Wang, S., Li, T., & Zhang, M. 2026. Frequency-decomposed attention joint optimization network for image compressive sensing. Expert Systems with Applications, 298, 129866. https://doi.org/10.1016/j.eswa.2025.129866.

[6] Zhao, L., Zhang, Y., Wang, X., Zhang, J., Bai, H., & Wang, A. 2026. A survey on image compressive sensing: From classical theory to the latest explicable deep learning. Pattern Recognition, 170, 112022. https://doi.org/10.1016/j.patcog.2025.112022.

[7] Rajesh, B., Gupta, A. K., Raj, A., Javed, M., & Dubey, S. R. 2022. HWRCNet: Handwritten word recognition in JPEG compressed domain using CNN-BiLSTM network. *Second International Conference on Data Analytics & Learning (DAL 2022)*. arXiv:2201.00947. https://doi.org/10.48550/arXiv.2201.00947.

[8] Rajesh, B., Zaman, S. M., Javed, M., & Nagabhushan, P. 2023. CompTLL-UNet: Compressed domain text-line localization in challenging handwritten documents using deep feature learning from JPEG coefficients. *Pattern Recognition (ACPR 2023)*, Part II, LNCS 14407, 88–101. https://doi.org/10.1007/978-3-031-47637-2_7.

[9] Rajesh, B., Jain, P., Javed, M., & Doermann, D. 2021. HH-CompWordNet: Holistic handwritten word recognition in the compressed domain. *2021 Data Compression Conference (DCC)*, 362. https://doi.org/10.1109/DCC50243.2021.00081.

[10] Bisen, T., Javed, M., Kirtania, S., & Nagabhushan, P. 2023. DWT-CompCNN: Deep image classification network for high throughput JPEG 2000 compressed documents. *Pattern Analysis and Applications*, 26(4), 1641–1655. https://doi.org/10.1007/s10044-023-01190-8.

[11] Sharma, D., & Javed, M. 2022. OCR for TIFF compressed document images directly in compressed domain using text segmentation and hidden Markov model. *arXiv:2209.09118*. https://doi.org/10.48550/arXiv.2209.09118.

[12] Madhavi, H., Cherian, J., Khamkar, Y., & Bhagat, D. 2025. Low-resource language processing: An OCR-driven summarization and translation pipeline. *arXiv:2505*.11177. https://doi.org/10.48550/arXiv.2505.11177.

[13] Takeshita, S., Green, T., Friedrich, N., Eckert, K., & Ponzetto, S. P. 2024. Cross-lingual extreme summarization of scholarly documents. *International Journal on Digital Libraries*, 25(2), 249–271. https://doi.org/10.1007/s00799-023-00373-2.

[14] Rainio, O., Teuho, J., & Klén, R. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14, 6086. https://doi.org/10.1038/s41598-024-56706-x.

[15] Coquenet, D., Chatelain, C., & Paquet, T. 2023. DAN: A segmentation-free document attention network for handwritten document recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7), 8227–8243. https://doi.org/10.1109/TPAMI.2023.3235826.

[16] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. 2023. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 369–376. https://doi.org/10.1145/1143844.1143891.

[17] Lin, C.-Y. 2024. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out (ACL Workshop)*, 74–81. ACL Anthology W04-1013.

[18] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the ACL (ACL 2020)*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.

[19] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.

[20] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[21] Gueguen, L., Sergeev, A., Kadlec, B., Liu, R., & Yosinski, J. 2018. Faster neural networks straight from JPEG. *Advances in Neural Information Processing Systems*, 31, 3933–3944.

[22] Ehrlich, M., & Davis, L. S. 2019. Deep residual learning in the JPEG transform domain. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, 3484–3493. https://doi.org/10.1109/ICCV.2019.00358.

[23] Wallace, G. K. 2002. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1), xviii–xxxiv. https://doi.org/10.1109/30.125072.

[24] Puigcerver, J. 2017. Are multidimensional recurrent layers really necessary for handwritten text recognition. *Proceedings of ICDAR 2017*, Vol. 1, 67–72. https://doi.org/10.1109/ICDAR.2017.20.

[25] Michael, J., Labahn, R., Grüning, T., & Zöllner, J. 2019. Evaluating sequence-to-sequence models for handwritten text recognition. *Proceedings of ICDAR 2019*, 1286–1293. https://doi.org/10.1109/ICDAR.2019.00208.

[26] Diaz, D. H., Qin, S., Ingle, R., Fujii, Y., & Bissacco, A. 2021. Rethinking text line recognition models. *arXiv:2104.07787*. https://doi.org/10.48550/arXiv.2104.07787.

[27] Sutskever, I., Vinyals, O., & Le, Q. V. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 3104–3112.

[28] Grüning, T., Leifert, G., Strauß, T., Michael, J., & Labahn, R. 2019. A two-stage method for text line detection in historical documents. *International Journal on Document Analysis and Recognition*, 22(3), 285–302. https://doi.org/10.1007/s10032-019-00332-1.

[29] Diem, M., Kleber, F., Fiel, S., Grüning, T., & Gatos, B. 2017. cBAD: ICDAR2017 competition on baseline detection. *Proceedings of ICDAR 2017*, Vol. 1, 1355–1360. https://doi.org/10.1109/ICDAR.2017.222.

[30] Dong, Y., & Pan, W. D. 2022. Image classification in JPEG compression domain for malaria infection detection. *Journal of Imaging*, 8(5), 129. https://doi.org/10.3390/jimaging8050129.