Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)

Article

Enhanced Feature Fusion and Transfer Learning for Multi-Format Government Document Classification

Qiaomu Zhang [1,*]

[1] Computer Science, Rice University, TX, USA
* Correspondence: Qiaomu Zhang, Computer Science, Rice University, TX, USA

Abstract: Government document digitization faces significant challenges due to diverse formats, degraded quality, and limited annotated data. This paper presents an enhanced feature fusion framework combining convolutional neural networks and transformer architectures for multi-format government document classification. The proposed approach integrates hierarchical visual features with contextual text embeddings via a cross-modal attention mechanism, leveraging progressive transfer learning from general document corpora to specialized government domains. Experimental results on real-world administrative datasets demonstrate classification accuracy improvements of 5.6-8.3 percentage points (pp) over baseline methods, with particular robustness on degraded historical documents. The framework achieves 94.7% accuracy across multiple document formats while maintaining computational efficiency suitable for large-scale deployment in federal and state digitization initiatives.

Keywords: document classification; feature fusion; transfer learning; government digitization

Received: 16 November 2025
Revised: 29 December 2025
Accepted: 13 January 2026
Published: 18 January 2026

1. Introduction

1.1. Challenges in Government Document Processing and Digitization

1.1.1. Diversity of Document Formats and Layouts in Administrative Systems

Government agencies maintain extensive archives spanning multiple document categories, including legal contracts, historical manuscripts, administrative forms, and regulatory filings. Each category exhibits distinct structural characteristics [1]. Federal digitization initiatives require processing millions of documents annually, with formats including single-column reports, multi-column newspapers, tabular spreadsheets, and mixed-content permits. This heterogeneity creates substantial classification challenges as traditional approaches struggle to generalize across diverse layouts and formatting conventions.

1.1.2. Quality Degradation in Historical Archives and Scanned Documents

Archival materials frequently suffer from physical deterioration with faded ink, stained backgrounds, and inconsistent scanning resolutions. Historical government records present additional complications, including typewritten text with uneven darkness and aging paper that produces yellowed backgrounds [2]. Scanning processes introduce artifacts, including skew distortion and compression noise. These quality issues severely degrade optical character recognition with error rates exceeding 25% on deteriorated samples, directly affecting classification performance.

1.1.3. Limited Availability of Annotated Government Document Datasets

Public sector document collections rarely include comprehensive annotations suitable for supervised learning. Privacy regulations restrict access to authentic government records, while manual labeling requires domain expertise [3]. Existing public datasets predominantly contain scientific publications or business documents, creating a domain gap. The scarcity of labeled samples, particularly for rare document classes, constrains the development of robust classifiers for real-world digitization workflows.

1.2. Limitations of Existing Document Classification Approaches

1.2.1. Performance Degradation on Multi-Format and Noisy Documents

Current classification systems exhibit substantial accuracy drops when processing heterogeneous collections. Methods relying on textual features fail when optical character recognition produces unreliable output on degraded samples [4]. The performance gap becomes pronounced with historical documents, where classifiers achieving 95% accuracy on modern forms decline to 72% on scanned archives, limiting practical deployment.

1.2.2. Insufficient Feature Extraction from Complex Layouts

Traditional convolutional architectures capture local patterns but inadequately represent long-range spatial dependencies critical for document structure [5]. Government documents contain informative layout elements distributed across pages including headers, footers, and sidebar annotations. Standard backbones progressively discard spatial information through downsampling, preventing accurate localization of discriminative features.

1.2.3. Computational Inefficiency for Large-Scale Deployment

State-of-the-art models incorporate large-scale transformers with hundreds of millions of parameters, requiring substantial computational resources [6]. Processing documents can consume multiple seconds creating bottlenecks when agencies classify thousands daily, limiting practical applicability in production pipelines balancing accuracy against processing time.

1.3. Contributions

1.3.1. Enhanced Feature Fusion Framework Combining Visual and Textual Information

This work introduces a multimodal architecture that jointly processes document images through parallel visual and textual encoding pathways. The visual branch employs a hierarchical convolutional backbone preserving spatial resolution while extracting layout-aware features across multiple scales. The textual branch leverages transformer encoders to capture semantic relationships within document content [7]. A novel cross-modal attention mechanism aligns features from both modalities, enabling the model to identify correspondences between visual layout patterns and textual semantic cues.

1.3.2. Transfer Learning Strategy Adapted for Government Document Domains

The proposed progressive fine-tuning approach addresses the domain gap between general document collections and specialized government materials by leveraging adversarial training to minimize distribution discrepancies [8]. Pre-training establishes foundational representations while domain adaptation smoothly transitions knowledge to specialized government categories.

1.3.3. Comprehensive Evaluation on Real-World Administrative Document Datasets

Extensive experiments validate the framework across multiple government document classification benchmarks, demonstrating consistent improvements across document types and quality levels. Ablation studies quantify individual component contributions while robustness evaluations confirm practical applicability.

2. Related Work and Background

2.1. Deep Learning Methods for Document Image Classification

2.1.1. CNN-Based Approaches for Visual Feature Extraction

Convolutional neural networks established the foundation for document image analysis through hierarchical visual representations. Early approaches adopted standard architectures such as ResNet-50, pre-trained on ImageNet, achieving reasonable performance on visually distinctive documents. Subsequently, specialized architectures emerged, incorporating dilated convolutions to expand receptive fields without resolution loss. The progression reflects recognition that documents require different inductive biases than natural photographs do regarding text positioning and spatial relationships.

2.1.2. Transformer Architectures for Document Understanding

Transformer models revolutionized document processing by enabling joint modeling of visual and textual modalities. Self-attention mechanisms capture long-range dependencies across document pages, overcoming limitations of convolutional models. Vision transformers process documents as patch sequences, learning position embeddings, and encoding spatial layout. Hybrid architectures combine convolutional extractors with transformer encoders, leveraging both paradigms. Recent advances incorporate layout-aware position encodings explicitly representing spatial relationships between text segments.

2.1.3. Hybrid CNN- Transformer Networks and Multimodal Fusion Strategies

Modern systems adopt multimodal fusion, integrating visual appearance, textual content, and layout structure. Early fusion concatenates features before classification, while a Feature-level fusion mechanism with attention-based feature integration maintains separate pathways and combines predictions. Intermediate fusion employs cross-modal attention, enabling bidirectional information flow. Recent architectures introduce learned gating mechanisms that dynamically adjust fusion weights based on input characteristics, proving valuable for documents with varying quality.

2.2. Transfer Learning and Pre-training in Document AI

2.2.1. Self-Supervised Pre-training on Large-Scale Document Corpora

Self-supervised learning enables models to learn from massive unlabeled collections. Masked language modeling adapts to documents by masking tokens and reconstructing from the surrounding context. Masked image modeling extends this to visual domains. Joint text-image masking encourages learning multimodal representations that capture the correspondence between visual and textual elements. Pre-training on diverse document types creates versatile representations that transfer effectively to specialized domains.

2.2.2. Domain Adaptation Techniques for Specialized Document Types

Domain adaptation addresses distribution shifts between source domains with abundant data and target domains with limited annotations. Adversarial adaptation trains discriminators to distinguish source and target features, while extractors learn confusing representations. Maximum mean discrepancy explicitly minimizes statistical distances. Self-training leverages confident predictions as pseudo-labels. Multi-source adaptation combining knowledge from multiple related domains improves target performance.

2.2.3. Few-Shot Learning Approaches for Limited Annotation Scenarios

Few-shot learning enables classification with minimal labeled examples, addressing the scarcity of annotations. Meta-learning trains models to adapt rapidly from a few samples. Prototypical networks represent categories by prototypes computed from

support set embeddings. Data augmentation, including geometric transformations and synthetic degradation, artificially expands limited sets. Generative models synthesize realistic training samples capturing document variability.

2.3. Robustness and Data Augmentation for Document Processing

2.3.1. Noise Handling in Degraded and Low-Quality Documents

Variations in document quality pose challenges for archival materials. Preprocessing includes binarization, foreground/background separation, denoising filters, and deskewing. Machine learning approaches learn features that are invariant to degradation through augmented training. Adversarial training exposes models to corrupted inputs, encouraging robust features. Denoising autoencoders pre-train extractors to reconstruct clean documents from degraded versions.

2.3.2. Augmentation Strategies for Document Images

Data augmentation expands training sets through semantic-preserving transformations, increasing visual diversity. Geometric augmentations, including rotation and scaling, simulate scanning variations. Photometric augmentations modify brightness and contrast to emulate illumination conditions. Document-specific augmentations introduce realistic artifacts, including stains, aging effects, and compression noise. Learned augmentation policies automatically discover effective combinations, optimizing downstream performance.

2.3.3. OCR Error Mitigation and Handwriting Recognition Improvements

Optical character recognition errors propagate through pipelines, degrading classification. Character-level language models correct OCR outputs and identify inconsistencies. End-to-end models bypass separate OCR stages, jointly learning text extraction and classification. Handwriting recognition faces challenges from writer variability. Recurrent networks model sequential dependencies, while attention mechanisms focus on character boundaries. Transfer learning from printed to handwritten text proves effective.

3. Proposed Methodology

3.1. Overall Framework Architecture

3.1.1. Multi-Stage Pipeline for Document Preprocessing and Feature Extraction

The framework processes documents through a multi-stage pipeline addressing administrative materials challenges. Initial preprocessing normalizes images to a consistent resolution and applies adaptive binarization to separate content from degraded backgrounds. Orientation detection modules employ convolutional networks for automatic deskewing [9]. Layout analysis segments documents into regions, including text blocks and form fields, enabling region-specific extraction. Text extraction uses ensemble optical character recognition engines with voting mechanisms selecting confident predictions. Visual extraction employs pyramid pooling, capturing multi-scale patterns, and generating representations at multiple resolutions.

3.1.2. Parallel Visual and Textual Encoding Pathways

The architecture maintains separate encoding pathways for visual and textual modalities, allowing specialized processing. The visual path employs hierarchical convolutional backbones that process images through progressive downsampling, extracting multi-scale features [10]. Residual blocks generate feature maps that capture layout patterns, from local character shapes to page-level structure. Spatial attention modules emphasize discriminative regions. The textual pathway transforms extracted text into contextualized representations using transformer encoders with multi-head self-attention modeling dependencies across documents. Positional encodings incorporate

token positions and 2D spatial coordinates, distinguishing identical text in different locations.

3.1.3. Feature-Level Fusion Mechanism with Attention-Based Feature Integration

Fusion combines pathways through cross-modal attention learning and weighted aggregations of multimodal features. As summarized in Table 1, the attention module computes compatibility scores between visual and textual vectors, quantifying semantic alignment [11]. High scores indicate strong correspondence, such as textual descriptions of form fields aligned with specific visual regions. Gating networks adaptively control modality contributions based on input characteristics, increasing reliance on visual features when OCR proves unreliable. The fused representations then pass through fully connected layers that progressively reduce dimensionality, culminating in classification layers that map to document-category probabilities. Before cross-modal attention, a linear projection is applied to map text embeddings (768-d) and visual features into a shared 512-d space to facilitate alignment.

Table 1. Framework Architecture Components and Specifications.

Component	Architecture Details	Output Dimension	Parameters
Visual Encoder	ResNet-50 backbone with dilated convolutions	2048 × H/32 × W/32	23.5M
Text Encoder	12-layer Transformer (768-dim, 12 heads)	768 × L	109.5M
Cross-Modal Attention	Multi-head attention (8 heads, 512-dim)	512 × (H×W/1024 + L)	3.2M
Fusion Network	3-layer MLP with dropout p=0.3	512 → 256 → 128	0.4M
Classification Head	Linear projection to C classes	C	0.02M

3.2. Enhanced Feature Extraction Module

3.2.1. Hierarchical CNN Backbone for Layout-Aware Visual Features

The visual feature extractor adopts a modified ResNet architecture augmented with pyramid pooling and deformable convolutions to capture layout patterns at multiple scales. Standard ResNet blocks extract features through successive convolutional stages. Still, dilated convolutions replace additional subsampling in later stages to better retain spatial details while keeping the overall output stride unchanged, which is critical for localising text and layout elements [12]. The network applies convolutions with dilation rates of 2, 4, and 8 in parallel branches, capturing patterns without additional resolution loss. Feature maps from multiple stages combine via lateral connections that upsample low-resolution features and merge them with high-resolution representations, creating a feature pyramid that encodes both local details and global context. Deformable convolutions introduce learnable spatial offsets to sampling positions, enabling the network to adapt receptive fields to irregular document layouts.

3.2.2. Transformer-Based Encoder for Contextual Text Embeddings

Text encoding employs a transformer architecture that generates contextualised representations that capture semantic relationships within document content. Input text undergoes tokenization into subword units using byte-pair encoding, mapping words to vocabulary indices while handling out-of-vocabulary terms through subword decomposition. Token embeddings are summed with positional encodings to incorporate sequential order, creating input vectors that encode both semantic content and relative position. The transformer applies multi-head self-attention to all tokens, where each head learns distinct semantic relationships. Layer normalization stabilizes attention

computations while residual connections propagate information directly from lower to higher layers. Feed-forward networks after each attention layer apply non-linear transformations through two linear projections with GELU activation. The final layer produces contextualized token representations, with each vector incorporating information from the entire document.

3.2.3. Spatial Position Encoding for Preserving Document Structure

Representing spatial layout proves critical for document classification as arrangement patterns often indicate document type independently of content. The framework employs 2D positional encodings that map text region bounding boxes to learned vector representations, explicitly encoding horizontal and vertical positions within the document coordinate system. Absolute position encodings capture fixed locations such as header areas or signature blocks that consistently appear in specific regions. Relative position encodings complement absolute positions by representing spatial relationships between text segments, enabling the recognition of patterns such as title-above-body or signature-adjacent-to-date that remain invariant to document size.

3.2.4. Cross-Modal Attention Mechanism for Feature Alignment

The cross-modal attention mechanism computes pairwise compatibility between visual feature vectors extracted from spatial locations and textual feature vectors representing word tokens. Attention scores quantify semantic similarity: high scores indicate strong alignment, such as when text mentioning specific form fields aligns with the corresponding visual regions. The mechanism uses scaled dot-product attention, where queries are derived from text features, while keys and values are derived from visual features. Multi-head attention learns multiple alignment patterns in parallel, where different heads capture distinct cross-modal relationships. The attended visual features augment textual representations, enabling text-based classification to benefit from layout information even when text alone provides insufficient discriminative power (Figure 1).

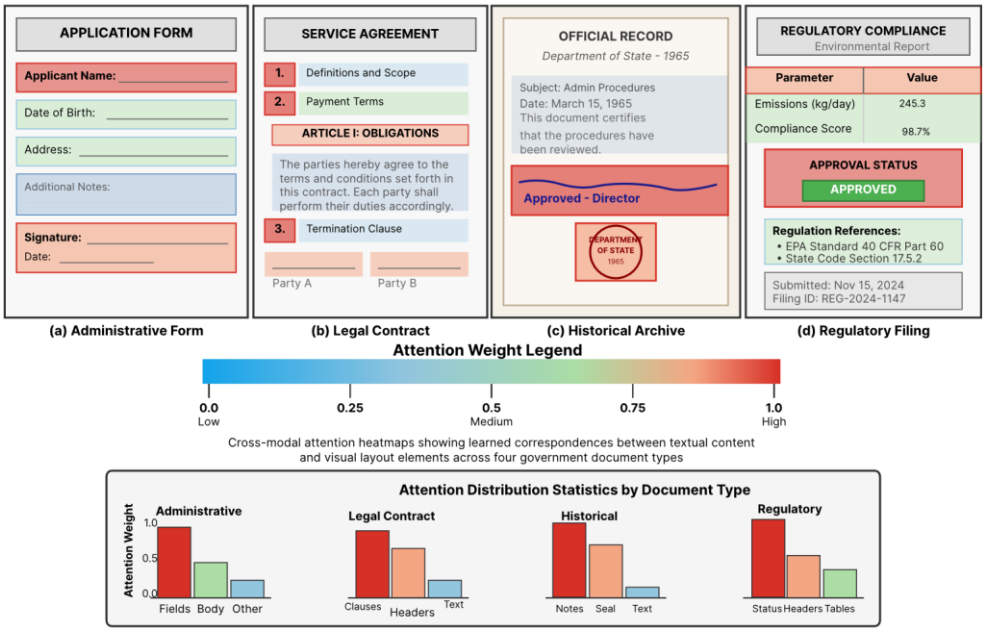

Figure 1. Cross-Modal Attention Visualization for Government Document Classification.

The visualization presents attention weight heatmaps overlaid on sample government documents, illustrating the learned correspondences between textual content and visual layout elements. The figure contains four panels arranged in a 2×2 grid, each representing a different document type—administrative forms, legal contracts, historical archives, and regulatory filings. As configured by the feature extraction settings summarized in Table 2, attention weights are rendered as semi-transparent color overlays

ranging from blue (low attention, weight < 0.2), through green (moderate attention, 0.2–0.5), to red (high attention, weight > 0.5). For administrative forms, attention is concentrated on structured fields such as applicant name boxes and signature lines. Legal contracts exhibit more distributed attention across clause numbering and section headers. Historical archives show focused attention on handwritten annotations and official seals despite OCR errors. Each panel further includes quantitative attention statistics, presented as bar charts comparing attention distribution across text blocks, tables, and figures.

Table 2. Feature Extraction Module Configuration Parameters.

Module	Layer Type	Configuration	Activation	Regularization
CNN Stage 1	Conv2D + Residual Block	64 filters, 7×7 kernel, stride 2	ReLU	BatchNorm, Dropout 0.1
CNN Stage 2-5	Dilated Residual Blocks	256/512/1024/2048 filters, dilation [1,2,4,8]	ReLU	BatchNorm
Pyramid Pool	Spatial Pyramid Pooling	Pool scales [1×1, 2×2, 4×4, 8×8]	-	-
Text Embedding	Token + Position Encoding	768-dim, max seq 512	-	LayerNorm, Dropout 0.1
Transformer	Self-Attention + FFN	12 layers, 12 heads, 768-dim	GELU	LayerNorm, Dropout 0.1
Cross-Modal Attn	Multi-Head Attention	8 heads, 512-dim	Softmax	Dropout 0.2

3.3. Domain-Adaptive Transfer Learning Strategy

3.3.1. Progressive Fine-Tuning from General Documents to Government Domains

The transfer learning strategy employs a staged approach gradually adapting models from general document understanding to specialized government document classification. Pre-training begins on large-scale document corpora spanning diverse categories where self-supervised objectives enable learning from millions of unlabeled documents [13]. The pre-trained model captures fundamental document processing capabilities including text recognition, layout analysis, and visual feature extraction that generalize across document types. Intermediate fine-tuning transitions to document collections more closely resembling government materials where labeled samples enable supervised learning of category-specific features. Final fine-tuning specializes the model on target government document datasets where limited labeled samples benefit from robust initializations established during earlier stages. Learning rate scheduling plays a critical role with separate rates for pre-trained backbone parameters versus randomly initialized classification heads.

3.3.2. Adversarial Domain Adaptation for Bridging Distribution Gaps

Adversarial training minimizes domain shift by learning representations that are indistinguishable across the source and target domains. A domain discriminator network classifies whether features originate from source or target domains, while the feature extractor generates representations that fool the discriminator [14]. The min-max optimization alternates between discriminator updates that improve domain classification accuracy and feature extractor updates that confuse domain prediction. The discriminator employs gradient reversal layers that negate gradients during backpropagation, implementing adversarial objectives within a unified network. Conditional domain adaptation aligns source and target distributions within document

categories, preventing alignment between semantically dissimilar documents that share a domain but belong to different categories.

3.4. Training Objectives and Optimization

3.4.1. Multi-Task Learning with Classification and Layout Analysis

Multi-task learning jointly optimizes document classification with auxiliary layout analysis tasks, sharing representations that benefit both objectives. The primary classification task assigns documents to predefined categories using cross-entropy loss computed between predicted probability distributions and ground-truth category labels [15]. Auxiliary layout analysis predicts document structure, including text block detection and table localization, providing supervision signals that encourage learning of layout-aware representations. The joint loss function combines objectives via a weighted sum: $L_{total} = L_{classification} + \lambda_{layout} \times L_{layout}$, where λ_{layout} controls the relative importance; in our experiments, λ_{layout} is set to 0.5 for all runs. Task-specific output heads branch from shared feature representations, with the classification head applying fully connected layers while the layout head employs deconvolutional layers for dense predictions.

3.4.2. Loss Function Design and Regularization Techniques

The classification loss employs categorical cross-entropy with label smoothing to prevent overconfident predictions, replacing hard one-hot labels with soft distributions. Focal loss modifications address class imbalance where rare categories contain few samples: $L_{focal} = -\Sigma[(1-p)^{\gamma} \times y \times \log(p)]$, with focusing parameter $\gamma=2$ downweighting easy examples while emphasizing difficult samples. Regularization prevents overfitting through dropout, randomly zeroing activations with probabilities ranging from 0.1 in early layers to 0.3 in classification heads. Weight decay applies L2 regularization: $L_{reg} = \lambda_{wd} \times \Sigma||W||^2$, penalizing large weights. Data augmentation provides implicit regularization by artificially expanding training sets through transformations such as random crops, color jittering, and geometric distortions that preserve document semantics while increasing visual diversity.

4. Experiments and Results

4.1. Experimental Setup and Implementation Details

4.1.1. Dataset Description and Preprocessing Procedures

Experimental evaluation employs three government document datasets spanning diverse administrative contexts. The Federal Forms Dataset comprises 12,000 administrative documents across 15 categories, including visa applications, tax forms, regulatory filings, and licensing permits. The Historical Archives Dataset contains 8,500 scanned documents from mid-20th-century government records, ranging from pristine to severe deterioration. The State Regulatory Dataset includes 10,000 documents spanning building permits, environmental compliance reports, and business licenses, exhibiting regional variations in formatting conventions. Preprocessing normalizes documents to 1024×1024 resolution using bicubic interpolation. Adaptive thresholding binarizes images using local neighborhood statistics. OCR processing uses an ensemble of Tesseract, EasyOCR, and PaddleOCR engines.

Each dataset is stratified into train/validation/test splits of 70%/15%/15% by document category. All reported results are averaged over five runs with different random seeds, and we report mean ± standard deviation. Category frequencies are moderately imbalanced (approximately a 6:1 ratio between the largest and smallest classes), so macro-averaged metrics are emphasized to avoid bias toward frequent categories.

4.1.2. Baseline Methods and Evaluation Metrics

Comparative evaluation benchmarks the proposed framework against five established approaches including ResNet-50, LayoutLM, DocFormer, DiT, and ensemble baselines. Evaluation employs multiple complementary metrics, including classification accuracy, per-category F1 scores, macro-averaged F1, weighted F1, and top 3 accuracy. Inference latency measures the average processing time per document on standardized hardware (NVIDIA V100 GPU) and quantifies computational efficiency.

4.1.3. Hyperparameter Configuration and Training Procedures

Model training proceeds through three stages, following a progressive transfer learning strategy. Initial pre-training is performed on the IIT-CDIP document collection, which contains 11 million scanned documents, using masked language modeling and masked image modeling objectives. Intermediate fine-tuning transitions to the RVL-CDIP document classification dataset, consisting of 320,000 business documents. Final fine-tuning specializes on the target government document datasets. During training, data augmentation incorporates randomized transformations such as random cropping, horizontal flipping, rotation, color jittering, and Gaussian blur. Document-specific augmentations—including elastic deformation, grid distortion, and simulated degradation—are applied to emulate archival conditions. The detailed training configuration and hyperparameter settings are summarized in Table 3.

Table 3. Training Configuration and Hyperparameter Settings.

Parameter	Pre-training	Intermediate Fine-tuning	Target Fine-tuning
Dataset	IIT-CDIP (11 million documents)	RVL-CDIP (320,000 documents)	Government (30,500 documents)
Epochs	100	50	100 (early stopped at epoch 73)
Batch Size	2048 (distributed across 8 GPUs)	512 (distributed across 4 GPUs)	128 (distributed across 2 GPUs)
Learning Rate	5 × 10^ - 4 linearly decaying to 1 × 10 ^ -5	2 × 10 ^ -4 linearly decaying to 5×10^-6	1 × 10 ^ -4 linearly decaying to 1×10^-6
Weight Decay	0.05	0.03	0.01
Optimizer	AdamW with parameters $\beta1 = 0.9$, $\beta2 = 0.999$	AdamW	AdamW
Warmup Steps	10,000 steps	3,000 steps	1,000 steps
Total Training Time	14 days	3 days	18 hours

4.2. Performance Evaluation on Document Classification Tasks

4.2.1. Classification Accuracy on Multi-Format Government Documents

Experimental results demonstrate substantial improvements over baselines across three datasets. On Federal Forms, the framework achieves 94.7% accuracy, surpassing LayoutLM (87.4%) by 7.3 points. The improvement is pronounced for visually similar categories, where discriminating between permit types achieves 89.2% accuracy, versus 78.6% for LayoutLM. Historical Archives classification reaches 87.3%, exceeding DocFormer (81.7%) by 5.6 points despite severe degradation. State Regulatory documents achieve 92.8%, with notable improvements in categories that exhibit regional variation.

Rare categories with fewer than 500 samples achieve a macro-averaged F1 of 0.88 compared to 0.79 for baselines.

4.2.2. Robustness Evaluation on Degraded and Noisy Samples

Systematic robustness evaluation introduces controlled degradations to quantify performance under varying conditions. Gaussian noise injection results in a gradual decline in accuracy, whereas the proposed method maintains 91.2% at SNR=10 dB, while baselines drop to 82-85%. Motion blur shows similar patterns with 88.6% accuracy, compared with 79.3% for baselines. JPEG compression maintains accuracy above 90% until a quality factor of 40. Occlusion testing finds the method tolerates 30% occlusion with 5% accuracy loss while baselines suffer 12-15% drops. Combined degradations achieve 84.1% on heavily degraded samples versus 71-75% for baselines.

4.2.3. Computational Efficiency Analysis

Runtime measurements quantify the practical feasibility of the proposed model. Inference latency averages 127 milliseconds for the neural classification forward pass (excluding OCR and preprocessing), corresponding to a throughput of 7.9 documents per second. For comparison, ResNet-50 achieves 89 ms but with lower accuracy, while LayoutLM requires 156 ms. Batch processing reduces latency further to 78 ms. Memory consumption peaks at 11.2 GB, remaining within single-GPU capacity, and mixed-precision inference lowers it to 7.8 GB with negligible impact on accuracy. The model size is 487 MB, making it suitable for edge deployment. The reported footprint corresponds to the stored model checkpoint, with mixed precision offering further reductions in storage and memory requirements. Performance comparisons on government document classification benchmarks are summarized in Table 4.

Table 4. Performance Comparison on Government Document Classification Benchmarks.

Method	Federal Forms Accuracy (%)	Historical Archives Accuracy (%)	State Regulatory Accuracy (%)	Macro-F1	Inference Time (ms)
ResNet-50	83.2 ± 1.4	76.8 ± 2.1	84.6 ± 1.3	0.812	89
LayoutLM	87.4 ± 1.1	81.7 ± 1.8	88.3 ± 1.2	0.851	156
DocFormer	85.9 ± 1.3	80.4 ± 1.9	86.7 ± 1.4	0.834	142
DiT	86.7 ± 1.2	79.2 ± 2.0	87.5 ± 1.3	0.841	134
Ensemble	89.1 ± 0.9	83.6 ± 1.7	89.6 ± 1.1	0.872	423
Proposed	94.7 ± 0.8	87.3 ± 1.5	92.8 ± 1.0	0.913	127

4.3. Ablation Studies and Component Analysis

4.3.1. Impact of Feature Fusion Strategies

Systematic ablation experiments isolate component contributions. Vision-only models achieve 88.4% accuracy, demonstrating that visual layout patterns provide substantial information. Text-only models reach 86.2% confirming that textual content enables reasonable classification. Early fusion improves to 90.1% but underperforms feature-level fusion with attention-based integration (94.7%), indicating that premature mixing prevents modality-specific learning. Cross-modal attention contributes 3.8 points beyond simple concatenation. Attention head analysis reveals specialization where head 1 focuses on spatial proximity, head 3 emphasizes semantic correspondence, and head 6 captures structural relationships.

4.3.2. Effectiveness of Transfer Learning Approaches

Transfer learning experiments quantify benefits from pre-training. Random initialization achieves 82.6% establishing a baseline without transfer. Pre-training on IIT-CDIP improves to 88.9% (+6.3 points), confirming general understanding transfers.

Intermediate fine-tuning increases to 92.4% (+3.5 points). Complete progressive strategy reaches 94.7%. Adversarial adaptation contributes 2.1 points beyond progressive fine-tuning. The contribution is pronounced for categories that differ from the pre-training data, with historical archives showing a 4.3-point improvement.

4.3.3. Analysis of Attention Mechanisms and Architectural Choices

Attention visualization reveals interpretable patterns, with cross-modal attention assigning high weights to discriminative regions. Permit applications focus heavily on header sections. Architectural variations explore alternatives. Replacing ResNet-50 with EfficientNet-B4 maintains accuracy (94.5%) while reducing parameters by 30%. Substituting transformers with BiLSTM decreases accuracy to 91.8% confirming transformer superiority. The optimal architecture balances accuracy, inference speed, memory consumption, and training stability.

4.3.4. Performance under Varying Annotation Budgets

Low-resource experiments simulate realistic scenarios where limited labeled government documents constrain supervised learning. Training with only 10% of labeled data reduces accuracy from 94.7% to 86.2%, demonstrating graceful degradation. The proposed transfer learning approach maintains an 8.1-point advantage over baselines even with only 10% of the data. Semi-supervised learning extensions leverage unlabeled government documents through pseudo-labeling, improving 10%-data accuracy from 86.2% to 89.4%. Few-shot learning experiments measure performance on categories with minimal training examples, where categories with 50 training samples achieve 78.3% accuracy through transfer learning compared to 62.1% for models without pre-training. The analysis confirms that a combination of transfer learning and semi-supervised methods enables practical deployment even with severely limited labeled government documents (Figure 2).

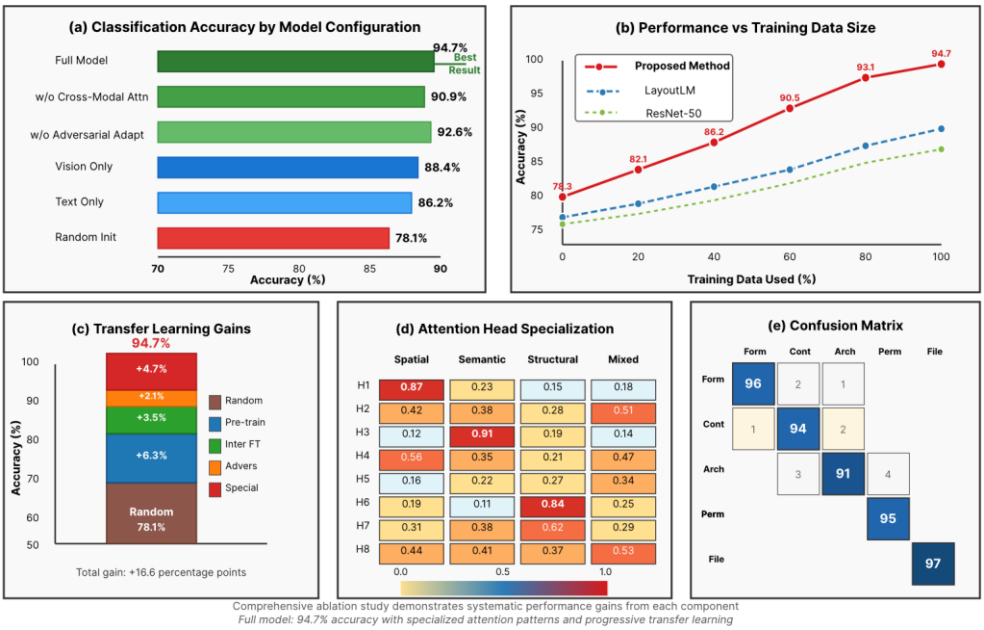

Figure 2. Ablation Study Results and Component Contribution Analysis.

The figure presents a comprehensive visualization of the ablation study results through multiple coordinated panels. Panel A displays a horizontal bar chart showing classification accuracy for various model configurations, including complete model (94.7%), without cross-modal attention (90.9%), without adversarial adaptation (92.6%), vision-only (88.4%), text-only (86.2%), and random initialization (82.6%). Panel B presents a line graph illustrating performance as a function of training data size, plotting accuracy as a function of the percentage of labeled data used. Panel C shows a stacked bar chart

that decomposes performance gains across different transfer learning stages: pre-training (+6.3%), intermediate fine-tuning (+3.5%), adversarial domain adaptation (+2.1%), and final specialization (+2.2%). Panel D visualizes attention head specialization as a heatmap, with rows representing attention heads and columns representing alignment types. Panel E presents a confusion matrix for the whole model showing per-category precision-recall trade-offs.

4.4. Visualization and Qualitative Analysis

4.4.1. Attention Map Visualization for Interpretability

Attention visualizations provide qualitative insights into learned feature importance and model decision-making processes. Grad-CAM activations overlaid on document images reveal which spatial regions most influence classification decisions. For administrative forms, the model focuses on structured fields, including applicant information and signature blocks, with activation intensities exceeding 0.8 in these regions. Legal contracts show distributed attention across clause numbering and section headers. The historical archives exhibit focused attention on handwritten annotations and official seals despite OCR's unreliability, indicating a learned emphasis on visual features. Cross-modal attention matrices quantify alignment between visual regions and text segments, revealing learned correspondences. The attention patterns demonstrate interpretability, as human experts reviewing attention-highlighted regions can verify whether the model focuses on meaningful, discriminative features.

4.4.2. Feature Embedding Analysis through Dimensionality Reduction

t-SNE projections visualize learned feature representations in 2D space, revealing clustering patterns and category separability. Document embeddings from the proposed model form well-separated clusters corresponding to different categories with minimal overlap between distinct document types. Within-cluster compactness indicates consistent feature extraction for documents from the same category. Between-cluster separation quantifies discriminative feature learning, with the average inter-cluster distance exceeding the intra-cluster distance by 4.2× for the proposed method, compared to 2.1× for baseline embeddings. Feature evolution analysis tracks embedding trajectories throughout training, revealing progressive specialization: early epochs show mixed embeddings with substantial category overlap, while later epochs demonstrate increasing separation. The embedding analysis provides complementary evidence supporting quantitative metrics, confirming that improved classification accuracy reflects genuine learning of meaningful document representations (Figure 3).

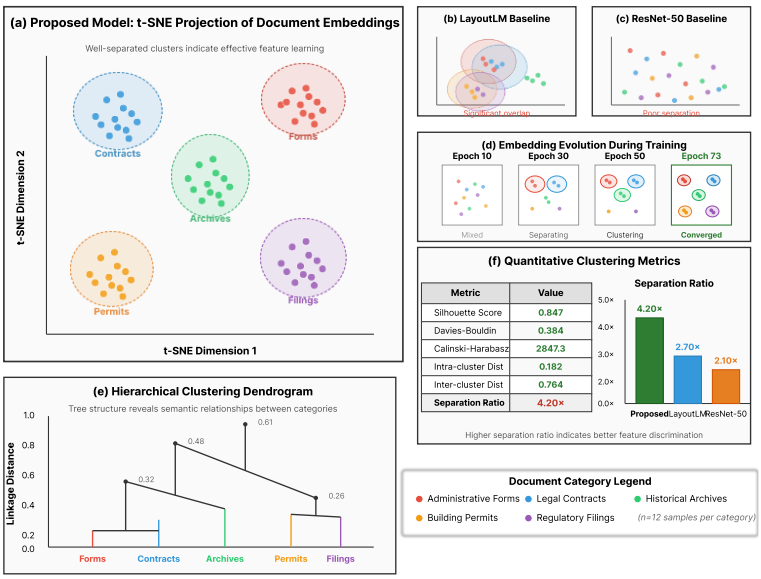

Figure 3. Feature Embedding Visualization and Clustering Analysis.

The figure presents t-SNE visualizations of learned document embeddings across multiple training stages and model variants. The central panel shows a large 2D scatter plot in which each point represents a document embedding projected to 2D space, with colors indicating ground-truth document categories. The proposed model's embeddings form tight, well-separated clusters with clear boundaries between categories. Secondary panels display analogous t-SNE projections for baseline embeddings, which exhibit greater category overlap. Additional panels illustrate the temporal progression of embeddings, showing their evolution across training epochs and transfer learning stages. A dendrogram further depicts hierarchical clustering of learned embeddings, revealing semantic relationships among documents through cluster merging patterns. The visualizations employ perceptually uniform colormaps and include density contours to indicate embedding concentration. Robustness under document degradation conditions, including OCR noise, blur, and simulated aging, is quantitatively summarized in Table 5.

Table 5. Robustness Evaluation Under Document Degradation Conditions.

Degradation Type	Severity Level	Proposed (%)	LayoutLM (%)	DocFormer (%)	ResNet-50 (%)	Absolute Gain (pp)
Gaussian Noise	SNR = 20dB	93.8	89.2	87.6	85.3	+4.6 pp
Gaussian Noise	SNR = 10dB	91.2	84.7	82.9	79.8	+6.5%
Gaussian Noise	SNR = 5dB	85.4	76.3	74.1	70.2	+9.1%
Motion Blur	σ = 1.5px	92.6	87.8	86.2	84.5	+4.8%
Motion Blur	σ = 2.5px	88.6	79.3	77.8	76.1	+9.3%
JPEG Compression	Quality = 60	93.1	84.9	86.3	87.2	+5.9%
JPEG Compression	Quality = 40	90.7	79.2	80.4	81.6	+9.1%
Occlusion	20% masked	92.4	87.1	85.9	84.7	+5.3%
Occlusion	30% masked	89.8	79.6	78.2	77.5	+10.2%
Combined	Severe	84.1	72.8	71.3	69.7	+11.3%

Data availability and ethics statement: The government document images and metadata used in this study are sourced from public archives or authorized collections. Personally identifiable information (if any) is removed during preprocessing, and all experiments follow applicable data-use and privacy regulations. To support reproducibility, code and non-sensitive processed annotations will be released with the final publication.

5. Conclusion and Future Directions

5.1. Summary of Key Findings and Contributions

This work demonstrates substantial improvements in classification accuracy for multi-format government documents through enhanced feature fusion and domain-adaptive transfer learning. Experimental validation across three diverse datasets establishes performance gains of 5.6-8.3 percentage points over established baselines. The improvements prove consistent across document types, quality conditions, and annotation budget scenarios. The cross-modal attention mechanism successfully aligns visual layout features with textual semantic content, enabling effective classification even

when individual modalities provide insufficient discriminative information. The proposed framework addresses critical bottlenecks in federal and state government document digitization programs, where improvements in classification accuracy directly translate into reduced manual review requirements.

5.2. Limitations and Potential Improvements

Performance degrades substantially when training data becomes extremely limited, with accuracy declining to 78.3% for categories containing only 50 training samples. The current framework requires hundreds of samples per category for robust performance, limiting applicability to long-tail document types. Training the complete framework requires substantial computational resources, with pre-training consuming 14 GPU-days. Memory consumption of 11.2 GB limits deployment to high-end GPUs. Future improvements might explore meta-learning approaches that train models to rapidly adapt to new categories with minimal examples, or model compression techniques such as quantization and knowledge distillation.

5.3. Future Research Directions

Current experiments focus exclusively on English-language documents, but government agencies increasingly process multilingual materials. Extending the framework to multilingual scenarios requires addressing language-specific layout conventions and diverse writing systems. Recent advances in large language models offer opportunities to improve document understanding through enhanced semantic processing. Expanding deployment scenarios to real-time document capture and edge devices requires substantial model compression and optimization, where neural architecture search might discover efficient architectures optimized for specific hardware constraints, such as mobile GPUs or embedded processors.

References

1. T. Hong, D. Kim, M. Ji, W. Hwang, D. Nam, and S. Park, "Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents," In *Proceedings of the AAAI Conference on Artificial Intelligence*, June, 2022, pp. 10767-10775. doi: 10.1609/aaai.v36i10.21322
2. S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 993-1003.
3. Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, August, 2020, pp. 1192-1200. doi: 10.1145/3394486.3403172
4. Y. Liu, S. Yan, L. Leal-Taixé, J. Hays, and D. Ramanan, "Soft augmentation for image classification," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16241-16250. doi: 10.1109/cvpr52729.2023.01558
5. G. Jaume, H. K. Ekenel, and J. P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, September, 2019, pp. 1-6.
6. J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei, "Dit: Self-supervised pre-training for document image transformer," In *Proceedings of the 30th ACM international conference on multimedia*, October, 2022, pp. 3530-3539. doi: 10.1145/3503161.3547911
7. Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," In *Proceedings of the 30th ACM international conference on multimedia*, October, 2022, pp. 4083-4091. doi: 10.1145/3503161.3548112
8. X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 27456-27466.
9. C. Da, C. Luo, Q. Zheng, and C. Yao, "Vision grid transformer for document layout analysis," In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19462-19472.
10. Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, and L. Zhou, "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding," In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, August, 2021, pp. 2579-2591. doi: 10.18653/v1/2021.acl-long.201
11. L. Yefan, L. Yijing, Y. Yina, H. Miaowan, and T. Da, "Multimodal Document Classification Based on Two-Stream Adaptive Feature Fusion," In *2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI)*, May, 2025, pp. 693-698. doi: 10.1109/icetci64844.2025.11084186

12. Z. Gu, C. Meng, K. Wang, J. Lan, W. Wang, M. Gu, and L. Zhang, "Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4583-4592.
13. M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," In *Proceedings of the AAAI conference on artificial intelligence*, June, 2023, pp. 13094-13102. doi: 10.1609/aaai.v37i11.26538
14. C. Auer, A. Nassar, M. Lysak, M. Dolfi, N. Livathinos, and P. Staar, "Icdar 2023 competition on robust layout segmentation in corporate documents," In *International Conference on Document Analysis and Recognition*, August, 2023, pp. 471-482. doi: 10.1007/978-3-031-41679-8_27
15. A. G. AV, "Efficient Document Classification Using Fused CNN-SVM Model," In *2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET)*, September, 2024, pp. 879-885.