Tatra Mt. Math. Publ. ?? (2026), 1–24
DOI: 10.2478/tmmp-2026-0003

TATRA MOUNTAINS
Mathematical Publications

CONTEXTUALIZED VISION TRANSFORMERS (CVT): ADAPTIVE SPECTRAL EMBEDDING AND FEATURE GATING FOR PRECISE TEXT-GRAPHICS CLASSIFICATION

Mridul Ghosh[1,2] — Konrad Dürrbeck[2] — Roland Fischer[2] — *Mária Ždímalová[3] — Tonmoy Mete[4]

[1]Department of Computer Science, Shyampur Siddheswari Mahavidyalaya, Howrah, INDIA

[2]Risk and location analysis, Fraunhofer IIS, Nuremberg, GERMANY

[3]Department of Mathematics and Descriptive Geometry, Slovak University of Technology in Bratislava, Bratislava, SLOVAKIA

[4]Department of Computer Science, Asutosh College, Kolkata, INDIA

ABSTRACT. The accurate classification of images into graphics-only, text-only, and mixed-content categories is a critical prerequisite for building efficient, content-aware processing pipelines. This initial triage prevents the unnecessary application of computationally expensive operations, such as Optical Character Recognition (OCR), to irrelevant graphical data. To address this challenge, we introduce the Contextualized Vision Transformer (CVT), a novel architecture designed specifically for this nuanced classification task. The CVT addresses the limitations of standard Vision Transformers (ViTs) through three synergistic components. It employs a Learnable Patch Decomposition (LPD) strategy that efficiently extracts patch embeddings. To model complex spatial arrangements, it introduces an Adaptive Spectral Embedding (ASE) module, which replaces static positional encodings with a dynamic, learnable representation. Crucially, a Contextual Feature Gating (CFG) module enhances feature discriminability by adaptively recalibrating patch-level features, selectively amplifying the salient text or graphic regions. Comprehensive K-fold cross-validation demonstrates the robustness and generalizability of the proposed model. The CVT achieves statistically significant improvements in accuracy, precision, and recall compared to state-of-the-art Vision Transformer baselines. Experimental results highlight the effectiveness of this architecture in providing a fast and accurate solution for the vital task of content triage in large-scale visual processing systems.

2020 Mathematics Subject Classification: 68T20, 68W01, 68W99.
Keywords: Text-graphics, vision transformer, classification, learnable patch decomposition.
* The author was supported by the Slovak National Research Grant Vega 1/0036/23.

1. Introduction

In the modern digital ecosystem, vast quantities of visual information are generated as documents, presentations, and webpages, often comprising a complex mixture of graphical elements and textual information. To process these data at scale, intelligent systems require efficient, content-aware pipelines. A naive one-size-fits-all approach, for instance, running an expensive Optical Character Recognition (OCR) [7, 8] engine on every single image is computationally demanding and leads to significant memory waste and processing latency. The critical first step in an intelligent workflow is, therefore, computational triage: accurately classifying the nature of the content to allocate resources logically. Graphics-only images can bypass OCR, text-only images can be routed directly to it, and mixed-content images can be sent to more sophisticated segmentation pipelines. Without this preliminary classification, every image is treated as the most complex case, leading to systemic inefficiency. The advent of Vision Transformers (ViTs), such as ViT [5], DeIT [17], Swin Transformer [12], and PVT [21], has revolutionized computer vision by leveraging self-attention mechanisms to capture global and multi-scale features. However, despite their power, these architectures exhibit several limitations that our CVT aims to address. Although the initial ViTs were data-hungry, subsequent approaches like DeIT improved data efficiency, sometimes at the cost of capturing the fine-grained features crucial for discriminating dense text from complex graphical textures. Furthermore, although hierarchical models like the Swin Transformer introduce shifted windows to capture both local and global context, they can be challenged by smaller, highly localized visual patterns. More broadly, a persistent challenge in many ViT models is the reliance on fixed-size patch embeddings and static positional encodings, which limits adaptability to varying object scales and aspect ratios, especially in scenes with complex spatial arrangements [15]. This is problematic for mixed-content images, where dynamic representation learning is paramount. Finally, recent work has shown that attention drift can occur in deeper models [23], leading to suboptimal performance where the early layers do not adequately inform later attention mechanisms. The CVT directly addresses these shortcomings through a synergistic design that combines the strengths of convolutional and transformer-based approaches. It integrates a Learnable Patch Decomposition (LPD) for efficient and adaptable feature extraction, an Adaptive Spectral Embedding (ASE) to capture complex spatial relationships with dynamic positional encodings, and a Contextual Feature Gating (CFG) module to selectively enhance informative features while mitigating attention drift. The primary contributions of this work are as follows:

- The introduction of a novel Contextualized Vision Transformer (CVT) architecture that combines a multi-head self-attention mechanism with a Contextual Feature Gating (CFG) module. This design enables the model to selectively enhance informative features based on their surrounding context, leading to improved classification performance.
- A novel Adaptive Spectral Embedding (ASE) technique that allows the model to learn flexible positional embeddings, replacing static encoding with a dynamic and adaptable representation of spatial relationships. This contrasts with traditional Vision Transformers that rely on fixed positional encodings.
- An efficient Learnable Patch Decomposition (LPD) strategy based on Conv2D layers is presented. This method effectively extracts patch embedding while preserving high-resolution spatial information, demonstrating a synergy between convolutional and transformer-based approaches.
- A comprehensive experimental evaluation using K-fold performance analysis, including systematic hyperparameter optimization, demonstrating the robustness and generalizability of the proposed CVT architecture.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of Vision Transformer architectures, highlighting the existing research gaps that motivate our work. Sections 3 through 6 detail the proposed Contextualized Vision Transformer (CVT) architecture. We systematically present each novel component, beginning with the input encoding and spatial representation (LPD and ASE), followed by the attentive feature enhancement (CFG), the transformer-based feature aggregation, and concluding with the probabilistic classification head. Section 8 describes our experimental setup, including details of our custom-curated GMDID dataset, the evaluation metrics, and the full training regime. Sections 9 and 10 present a rigorous analysis of our results, including a detailed ablation study to validate our architectural contributions, a performance breakdown on individual classes, and a comparative study against state-of-the-art baselines, concluding with an evaluation of the model's computational efficiency. Finally, Section 11 concludes the paper by summarizing our key findings and highlighting promising directions for future research.

2. Literature study

The classification of images containing heterogeneous content, such as a mixture of text and graphics, represents a significant challenge in computer vision. The development of robust models for this task requires an architecture capable of understanding both the high-level semantic context of graphical elements and

the fine-grained, structured details of text. For various types of document understanding many deep learning-based [9, 10, 6] architectures have shown promising results. This literature review charts the evolution of relevant deep learning architectures, beginning with the paradigm shift initiated by Vision Transformers and culminating in the specific research gaps that motivate the development of our proposed CVT architecture.

2.1. The Vision Transformer (ViT) Paradigm

The advent of ViT, as introduced in the pivotal work of [5], fundamentally altered the landscape of computer vision. Departing from the convolutional filters that had dominated the field, ViT applied the Transformer architecture, originally designed for natural language processing, directly to image classification. The core methodology involves partitioning an image into a sequence of fixed-size patches, creating linear embeddings of these patches, and feeding them along with positional embeddings into a standard Transformer encoder. The self-attention mechanism within the Transformer allows ViT to model long-range dependencies between distant patches in an image, granting it a global receptive field from its very first layer. This capability proved highly effective, enabling ViT to achieve state-of-the-art results on large-scale benchmarks, albeit with a significant requirement for massive training datasets (e.g., JET-300M) to overcome the lack of inherent inductive biases found in Convolutional Neural Networks (CNNs).

2.2. Evolution and Refinements in ViT Architectures

The initial success of ViT spurred a wave of research aimed at mitigating its limitations and enhancing its capabilities. These efforts can be broadly categorized into several key areas that directly inform the design of our CVT model.

2.2.1. Improving Data Efficiency and Training Strategies

A primary drawback of the original ViT was its pronounced data hunger. To address this, [17] introduced the Data-efficient Image Transformer (DeIT), which demonstrated that ViTs could be trained effectively on smaller datasets such as ImageNet-1K. The key innovation of DeIT was a knowledge distillation strategy that uses a distillation token to learn from the results of a highly experienced pre-trained CNN teacher. This approach enabled ViT to inherit some of the beneficial inductive biases of CNNs without altering its core architecture, significantly improving its data efficiency and training stability.

2.2.2. Capturing Multi-Scale and Hierarchical Features

The standard ViT processes patches at a single fixed resolution, which is suboptimal for recognizing objects and patterns at varying scales. To incorporate this crucial property, several hierarchical architectures were proposed. The Pyramid Vision Transformer (PVT) introduced a progressively shrinking feature pyramid to handle multi-scale features efficiently [21]. Similarly, the Swin Transformer [12] proposed a hierarchical structure based on shifted local attention windows. By computing self-attention within non-overlapping local windows and shifting these windows between layers, the Swin Transformer could model both local and cross-window (global) interactions efficiently, achieving state-of-the-art performance across a range of vision tasks.

2.2.3. Enhancing Spatial and Positional Representations

The method of encoding spatial information is critical for the performance of ViT. The original ViT and its immediate successors relied on fixed or simply learned absolute positional embeddings. However, this static approach may not be optimal for capturing the complex, non-linear spatial relationships present in diverse images. The limitations of fixed-size patch embeddings have been highlighted as a performance constraint, suggesting that adaptive patching strategies could yield improvements [15]. This points to a broader gap in developing more flexible mechanisms for the model to learn and adapt its understanding of spatial context based on the input image's specific characteristics.

2.2.4. Attention Mechanisms and Feature Refinement

A core assumption in the standard ViT is that all patches are treated with equal importance. However, for complex scenes, certain regions are far more informative than others. The concept of feature recalibration, famously demonstrated by Squeeze-and-Excitation (SE) networks in the CNN domain [11], has shown that adaptively re-weighting feature channels can yield significant performance gains. In the context of Transformers, recent work has identified the issue of attention drift in deep ViT models, where early layers may not adequately focus on salient regions, leading to suboptimal feature propagation [23]. This suggests the need for explicit mechanisms that can gate or modulate feature representations, allowing the model to selectively improve informative regions.

This review of the literature reveals a clear path: from a monolithic, data-hungry ViT to more efficient, hierarchical, and nuanced architectures. However, there remains a specific gap for the task of text-graphics classification. Although hierarchical models such as the Swin Transformer improve local context, they do not possess an explicit mechanism to selectively prioritize text

versus graphic features. Furthermore, the reliance on static positional encodings limits the model's ability to adapt to highly variable spatial layouts.

The CVT is designed to fill these specific gaps. It builds on foundational principles but introduces a synergistic combination of three novel components motivated by the limitations identified above:

(1) An efficient Learnable Patch Decomposition (LPD).
(2) A novel Adaptive Spectral Embedding (ASE) to replace static positional encodings.
(3) A Contextual Feature Gating (CFG) module for adaptive feature recalibration.

By integrating these innovations, the CVT aims to create a more precise, efficient, and adaptable architecture tailored to the challenges of multimodal content classification.

3. Methodology

In this paper, we present the Contextualized Vision Transformer, the proposed model for text, graphics, and text-graphics image classification and detection. The CVT avoids the shortcomings of traditional Vision Transformers (ViTs) in dealing with intricate, spatially organized visual data through the introduction of essential innovations: learnable patch decomposition for efficient feature extraction, adaptive spectral embedding for flexible spatial representation, and contextual feature gating for selective strengthening of informative image areas. Synergy allows CVT to efficiently model intricate spatial relationships and capture the underlying interplay between textual and graphical information in images. The subsequent sections provide a comprehensive description of each module. The proposed architecture is presented in Figure 1.

3.1. Input Encoding and Spatial Representation

This section outlines the initial steps of CVT and how they go about processing the input image to be transformed into a sequence of patch embeddings and spatial context among patches embedded. These initial steps, implemented using Learnable Patch Decomposition (LPD) and Adaptive Spectral Embedding (ASE), are the bedrock on which the retention of local visual information and the global spatial context happens.

3.1.1. Learnable Patch Decomposition (LPD)

The first step is to transform the input image into a learnable sequence of patch embeddings by the LPD module. Given an input image $I \in \mathbb{R}^{H \times W \times C}$, where H is the height, W the width, and C the channels (e.g. 3 for colored images),

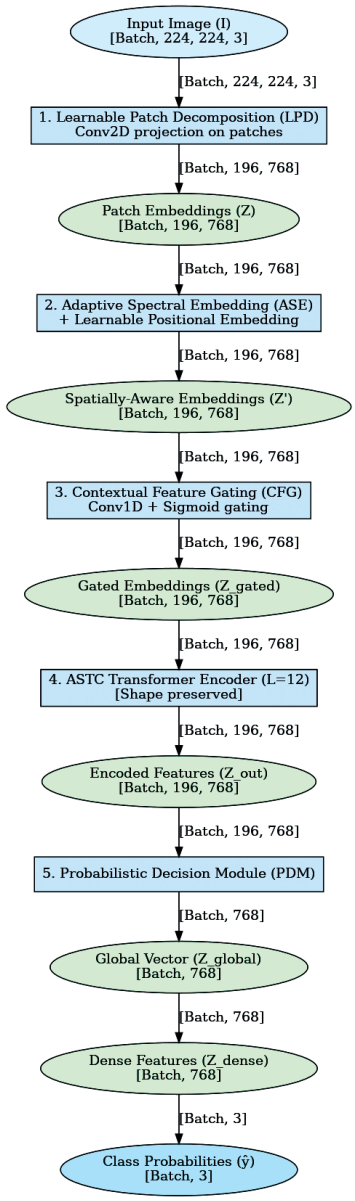


FIGURE 1. Architecture of the Contextualized Vision Transformer (CVT). The model processes an input image through a sequence of novel modules: 1) Learnable Patch Decomposition (LPD) for efficient feature extraction, 2) Adaptive Spectral Embedding (ASE) to encode spatial relationships dynamically, and 3) Contextual Feature Gating (CFG) to selectively enhance salient image regions. The refined features are then processed by a standard Transformer Encoder and a final classification head.

the LPD divides the image into N non-overlapping patches of size $P \times P$, where N is given by:

$$N = \frac{H \times W}{P^2}. \tag{1}$$

Each patch is linearly embedded into a D-dimensional vector through a convolutional process:

$$\mathbf{z}_i = \text{Conv2D}(I_i) \in \mathbb{R}^D, \quad i = 1, 2, \ldots, N, \tag{2}$$

where $\mathbf{z}_i$ is the embedding of the ith patch after passing through the LPD. We employ a 2D Convolutional layer with a kernel size of 1×1. Applying a 1×1 convolution, or point-wise convolution, enables us to apply a linear transformation to the pixel values in each patch, as well as map the C input channels into D output channels. In Algorithm 1 the LPD process is illustrated.

Algorithm 1 Learnable Patch Decomposition (LPD)

Require:

- Image $I \in \mathbb{R}^{H \times W \times C}$
- Patch size P
- Embedding dimension D

Ensure: Patch embeddings $\mathbf{Z} \in \mathbb{R}^{N \times D}$, where $N = (H \times W)/P^2$

1: Divide I into N patches: $I \leftarrow \{I_i\}, I_i \in \mathbb{R}^{P \times P \times C}$
2: Define Conv2D layer (kernel $P \times P$, stride P, output channels D): Conv2D : $I_i \mapsto \mathbf{z}_i \in \mathbb{R}^D$
3: **for** $i \leftarrow 1$ to N **do**
4: $\quad \mathbf{z}_i \leftarrow \text{Conv2D}(I_i)$
5: **end for**
6: $\mathbf{Z} \leftarrow [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]$
7: **return** Patch embeddings $\mathbf{Z}$

3.1.2. Adaptive Spectral Embedding (ASE)

The Adaptive Spectral Embedding (ASE) module incorporates spatial relationships between patches in the image. Unlike prior ViT models that have employed fixed positional encodings [20], we adopt learnable positional embeddings $\mathbf{P} \in \mathbb{R}^{N \times D}$ to enable the model to flexibly change the spatial arrangement of patches dynamically during training. The final representation of the ith patch is provided as:

$$\mathbf{z}_i^{\text{final}} = \mathbf{z}_i + \mathbf{P}_i\,, \tag{3}$$

where $\mathbf{z}_i^{\text{final}}$ is the final patch representation, encompassing both its own visual features and learned spatial context. This ASE allows the model to learn intricate, non-linear spatial dependencies, a necessity for accurate image classification.

The 'Spectral' aspect of ASE refers to the initialization of the learnable positional embedding matrix P. Rather than using a random initialization, P is initialized using a basis of sinusoidal functions of varying frequencies, as proposed by Vaswani et al. [20]. These functions form a Fourier basis, providing a strong spectral prior for spatial relationships. The 'Adaptive' nature arises because these embeddings are not fixed, they are treated as learnable parameters that the model fine-tunes during training to best suit the data distribution. This approach combines a robust initial representation with the flexibility to learn dataset-specific spatial dependencies.

3.2. Attentive Feature Enhancement

This section introduces a Contextual Feature Gating (CFG) module for adaptive feature recalibration. CFG employs a spatial attention mechanism, modulating patch embeddings ($\mathbf{Z}$) with learned attention weights ($\mathbf{A}$, computed from a convolutional layer and sigmoid activation) to selectively emphasize salient features and suppress irrelevant background responses, thus enhancing feature discrimination.

3.2.1. Contextual Feature Gating (CFG)

The Contextual Feature Gating (CFG) module selectively enhances the informative features within each patch based on their surrounding context. This mechanism allows the model to focus on the salient regions of the image, suppressing irrelevant background noise, and improving overall performance. The CFG is based on the concept of spatial attention, which has been shown to be effective in improving the performance of convolutional neural networks on various vision tasks [11].

Given the spatially-aware patch embeddings $Z' \in \mathbb{R}^{N \times D}$, the CFG module first computes a gating map $A \in \mathbb{R}^{N \times D}$ using a `Conv1D` layer with a kernel size of 1, followed by a sigmoid activation function. This operation is performed on the sequence of patch embeddings. The computation is as follows:

$$A = \sigma\big(\mathrm{Conv1D}(Z')\big), \tag{4}$$

where the `Conv1D` layer projects each D-dimensional feature vector to another D-dimensional vector, using a kernel size of 1, a stride of 1, and 'same' padding to preserve the sequence length. The sigmoid function σ ensures that the attention weights in the gating map A are bounded between 0 and 1.

This gating map A is then applied to the patch embeddings by element-wise multiplication (Hadamard product), allowing the model to selectively scale individual feature dimensions for each patch:

$$Z_{\text{gated}} = Z' \odot A, \tag{5}$$

where $\odot$ denotes element-wise multiplication. This vector-based gating is more expressive than scalar gating, as it enables the model to learn a more fine-grained,

dimension-specific feature recalibration. This mechanism enhances the model's focus on crucial areas, such as text or graphic regions, based on the learned attention weights, which improves performance, especially for complex images with multiple objects or cluttered backgrounds. Without such spatial attention, a model might regard every part of an image with equal weight, which can result in suboptimal performance.

Algorithm 2 Contextual Feature Gating (CFG)

Require:

- Patch embeddings $\mathbf{Z} \in \mathbb{R}^{N \times D}$

Ensure: Attended patch embeddings $\mathbf{Z}^{\text{attended}} \in \mathbb{R}^{N \times D}$

1: **for** $i = 1$ to N **do**
2: Compute attention weight: $\mathbf{A}_i \leftarrow \sigma\big(\text{Conv2D}(\mathbf{z}_i)\big)$, $\mathbf{A}_i \in \mathbb{R}$
3: Apply attention gating: $\mathbf{z}_i^{\text{attended}} \leftarrow \mathbf{z}_i \times \mathbf{A}_i$
4: **end for**
5: Form attended embeddings: $\mathbf{Z}^{\text{attended}} \leftarrow [\mathbf{z}_1^{\text{attended}}, \mathbf{z}_2^{\text{attended}}, \ldots, \mathbf{z}_N^{\text{attended}}]$
6: **return** Attended patch embeddings $\mathbf{Z}^{\text{attended}}$

3.3. Transformer-Based Feature Aggregation

This section introduces a transformer-based feature aggregation approach leveraging the Attentive Spatial Transformer Cell (ASTC) to model long-range dependencies between image patches. The ASTC employs multi-head self--attention to patch embeddings ($\mathbf{Z}$), allowing the network to learn complex, global feature relationships crucial for downstream classification tasks.

3.3.1. Attentive Spatial Transformer Cell (ASTC)

The core of the CVT architecture consists of multiple ASTC layers. Each ASTC layer performs multi-head self-attention to capture long-range dependencies between patches, followed by a feedforward network to further refine the feature representations. The multi-head self-attention mechanism allows the model to attend to different parts of the image in parallel, capturing a richer set of relationships between patches. Given a set of N patch embeddings

$$\mathbf{Z} \in \mathbb{R}^{N \times D},$$

the multi-head self-attention mechanism computes attention weights as follows:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{Z}\mathbf{W}_Q, \mathbf{Z}\mathbf{W}_K, \mathbf{Z}\mathbf{W}_V, \tag{6}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times d}$ are learned weight matrices for query, key, and value transformations, and d is the dimensionality of the attention heads.

The query, key, and value matrices project the input embeddings into different subspaces, allowing the model to attend to different aspects of the input. The attention weights are computed as:

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{7}$$

The self-attention output is followed by a position-wise feedforward network (FFN):

$$\mathbf{Z}_{\text{ffn}} = \text{FFN}(\mathbf{Z}_{\text{attended}}), \tag{8}$$

where FFN consists of two fully connected layers with a ReLU activation in between:

$$\text{FFN}(x) = \text{ReLU}(x\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2. \tag{9}$$

Each ASTC layer also includes residual connections and layer normalization:

$$\mathbf{Z}_{\text{out}} = \text{LayerNorm}(\mathbf{Z} + \mathbf{Z}_{\text{ffn}}). \tag{10}$$

By stacking multiple ASTC layers, the CVT can learn complex and hierarchical feature representations that capture local and global dependencies within the image. This is also illustrated using Algorithm 3.

Algorithm 3 Attentive Spatial Transformer Cell (ASTC)

Require:

- Patch embeddings $\mathbf{Z} \in \mathbb{R}^{N \times D}$
- Weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times d}$

Ensure: Processed patch embeddings $\mathbf{Z}_{\text{out}} \in \mathbb{R}^{N \times D}$

1: Project embeddings: $\mathbf{Q} \leftarrow \mathbf{Z}\mathbf{W}_Q$, $\mathbf{K} \leftarrow \mathbf{Z}\mathbf{W}_K$, $\mathbf{V} \leftarrow \mathbf{Z}\mathbf{W}_V$

2: Compute attention weights: $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \leftarrow \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$

3: Apply attention: $\mathbf{Z}_{\text{attended}} \leftarrow \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$

4: Feedforward network: $\mathbf{Z}_{\text{ffn}} \leftarrow \text{FFN}(\mathbf{Z}_{\text{attended}})$, where $\text{FFN}(x) = \text{ReLU}(x\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2$

5: Apply layer normalization and residual connection: $\mathbf{Z}_{\text{out}} \leftarrow \text{LayerNorm}(\mathbf{Z} + \mathbf{Z}_{\text{ffn}})$

6: **return** Processed patch embeddings $\mathbf{Z}_{\text{out}}$

3.4. Probabilistic Classification

The Probabilistic Decision Module (PDM) concludes the classification process by mapping the learned feature representations to a probability distribution over the classes. This is achieved through a combination of global average pooling and fully connected layers, culminating in a softmax output.

3.4.1. Probabilistic Decision Module (PDM)

The final classification is performed by the PDM, which maps the learned feature representations to class probabilities. The PDM employs global average pooling to aggregate patch-level features into a single global feature vector:

$$\mathbf{Z}_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{\text{out},i}. \tag{11}$$

The pooled feature vector is then passed through a fully connected layer with ReLU activation:

$$\mathbf{Z}_{\text{dense}} = \text{ReLU}(\mathbf{Z}_{\text{global}} \mathbf{W}_3 + b_3). \tag{12}$$

Finally, the model performs classification by applying a softmax layer to the output:

$$\hat{y} = \text{softmax}(\mathbf{Z}_{\text{dense}} \mathbf{W}_4 + b_4), \tag{13}$$

where $\hat{y} \in \mathbb{R}^C$ represents the predicted class probabilities, and C is the number of classes. The softmax function normalizes the output of the fully connected layer into a probability distribution, where each element represents the probability of the image belonging to a particular class. PDM is discussed in Algorithm 4.

Algorithm 4 Probabilistic Decision Module (PDM)

Require: Pooled feature vector $\mathbf{Z}_{\text{out}} \in \mathbb{R}^{N \times D}$, weights $\mathbf{W}_3, \mathbf{W}_4$, biases b_3, b_4
Ensure: Predicted class probabilities $\hat{y} \in \mathbb{R}^C$
1: Apply global average pooling: $\mathbf{Z}_{\text{global}} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{\text{out},i}$
2: Apply fully connected layer with ReLU: $\mathbf{Z}_{\text{dense}} \leftarrow \text{ReLU}(\mathbf{Z}_{\text{global}} \mathbf{W}_3 + b_3)$
3: Perform classification with softmax: $\hat{y} \leftarrow \text{softmax}(\mathbf{Z}_{\text{dense}} \mathbf{W}_4 + b_4)$
4: **return** Predicted class probabilities $\hat{y}$

3.5. Loss and Optimization

The CVT model is trained end-to-end using categorical cross-entropy loss:

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{14}$$

where $y \in \mathbb{R}^C$ is the one-hot encoded ground truth vector. The model is optimized using the Stochastic-Gradient-Descent-Step (SGDS) with the Adam optimizer, with a learning rate η:

$$\theta = \theta - \eta \nabla_\theta \mathcal{L}(\hat{y}, y). \tag{15}$$

4. Experiment

4.1. Evaluation Metrics

The performance of the model is evaluated using the following metrics:

Accuracy: The proportion of correctly classified images:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbf{1}_{\hat{y}_i = y_i}}{N}. \tag{16}$$

Precision: The proportion of true positives among predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{17}$$

Recall: The proportion of true positives among actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{18}$$

F1-score: The harmonic mean of precision and recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{19}$$

Here, TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

The experiments were performed on a computer system featuring 16 GB of RAM, an Intel Core i5 processor and a dedicated GPU to improve training efficiency. The experiments were executed in Windows 10 Pro (version 20 H 2) environment, leveraging TensorFlow 2.0.0 with cuDNN for optimized computational performance. Detailed descriptions of the dataset and training protocols are provided in the following sections.

4.2. Dataset

The empirical evaluation of CVT was conducted on a custom-curated dataset, which we refer to as the German Municipal Document Image Dataset (GMDID). This dataset was specifically designed to reflect the real-world challenges of mixed-content document analysis, providing a robust benchmark for the nuanced classification task addressed in this paper.

4.3. Data Source and Composition

The GMDID comprises a total of 2,100 images, meticulously balanced across three distinct classes, with 700 images per class. The source material for this dataset consists of publicly available documents collected from the official online portals of various German municipalities. The documents were sourced from a range of portals to ensure diversity in layout, formatting, and content,

including: The Bundesportal für Verwaltung (the federal portal for administrative services). State-level Bürgerinformationsportale (citizen information portals) from states such as Bavaria and North Rhine-Westphalia. Local Rathaus--Informationssysteme (city hall information systems). The collected source files, which were originally in formats such as PDF, JPEG, and PNG, included a wide array of document types: official public notices, administrative forms, city-funded event brochures, infographics on municipal budgets, and annotated land-use plans. Each source document was carefully processed and the relevant pages or sections were extracted and converted to high-resolution PNG images.

4.4. Class Definitions and Annotation

The core of GMDID is its classification into three functionally distinct categories, which are critical to enable efficient downstream processing pipelines. The annotation was performed by a team of two trained annotators following a strict set of guidelines. Any classification disagreements were resolved by a third, senior annotator to ensure consistent high-quality labels. The three classes are defined as follows:

• **Graphics-Only (700 images):** This class includes images where the content is purely visual, and any text present is either non-semantic (e.g., part of a complex logo) or entirely absent. These are images that should bypass an OCR engine.

Examples: Photographs of city events, municipal crests and logos, unlabeled maps, and data charts without textual legends or titles.

• **Text-Only (700 images):** This class consists of images in which the primary substantive content is readable text. Graphical elements are either completely absent or are purely decorative and non-informative (e.g., horizontal lines, simple page borders, or a small letterhead logo). These images are ideal candidates for direct OCR processing.

Examples: Scanned official notices, filled-out administrative forms, meeting minutes, and pages from a legal ordinance.

• **Text-with-Graphics (700 images):** This class represents the most complex composite category. It includes images where there is a meaningful and informative interplay between both textual and graphical elements. These images would require a more sophisticated pipeline, such as segmentation before selective OCR.

Examples: Public information brochures combining pictures and descriptions, infographics displaying statistics with explanatory text, annotated architectural plans, and posters for public services.

4.5. Data Splitting and Pre-processing

To ensure a robust and unbiased evaluation of the model's generalization capabilities, we employed a 5-Fold Cross-Validation strategy. The entire dataset of 2,100 images was partitioned into five equal stratified folds of 420 images each.

For each of the five training runs, four folds were used for training and one was held out for validation. The performance metrics reported in this paper represent the averaged results across all five folds. All images in the dataset underwent a standardized pre-processing pipeline before being fed into the model:

• **Resizing:** Each image was resized to a uniform resolution of 224×224 pixels.

• **Normalization:** Pixel values were normalized from the standard $[0, 255]$ range to $[0, 1]$ to stabilize the training process.

Data Augmentation: To prevent overfitting and improve model robustness, on-the-fly data augmentation was applied to the training set. This included minor random transformations such as slight rotations (up to 10 degrees), horizontal flips, and small adjustments in brightness and contrast. In Figure 2 the sample images of text, graphics, and text-graphics images are depicted.

4.6. Training Regime

The CVT model was trained end-to-end following a meticulously defined regime designed to ensure robust convergence, prevent overfitting, and facilitate reproducibility. All experiments were conducted using the TensorFlow framework.

4.6.1. Data Pre-processing and Validation Strategy

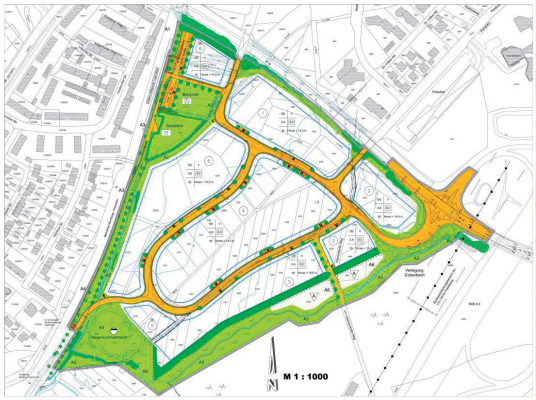
Prior to training, all images from the GMDID dataset were uniformly resized to a resolution of 224×224 pixels and their pixel values were normalized to the range $[0, 1]$. To ensure a robust and unbiased evaluation of the model's generalization performance, we employed a 5-fold cross-validation protocol. In each of the five runs, the model was trained on four folds of the data and validated on the remaining fold, with the final reported metrics representing the average across all five runs.

4.6.2. Optimization and Learning Rate Schedule

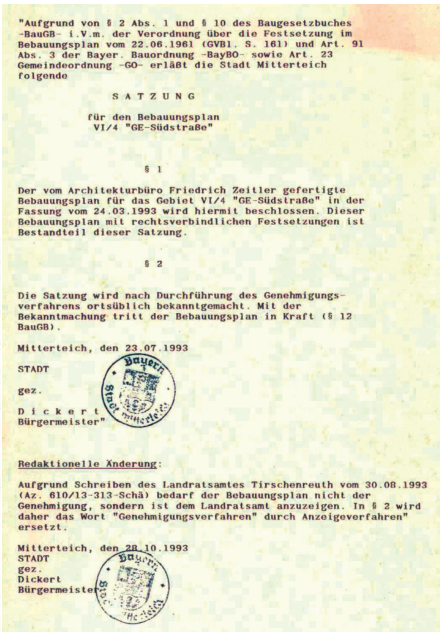
For optimization, we employed the Adam optimizer, initialized with a learning rate of 1×10^{-4}. To facilitate fine-tuning and ensure that the model settles to a stable performance minimum, a dynamic learning rate schedule was implemented using the `ReduceLROnPlateau` callback. This schedule automatically reduced the learning rate by a factor of 0.2 when the validation loss did not improve for 5 consecutive epochs.

4.6.3. Regularization and Training Procedure

A multi-faceted regularization strategy was employed to mitigate overfitting and enhance model generalizability. This included L2 regularization (weight decay) with a coefficient of 1×10^{-5} applied to all convolutional and dense layers, and a dropout rate of 0.1 within the feedforward networks of the ASTC blocks.



(A)



"Aufgrund von § 2 Abs. 1 und § 10 des Baugesetzbuches -BauGB- i.V.m. der Verordnung über die Festsetzung im Bebauungsplan vom 22.06.1961 (GVBl. S. 161) und Art. 91 Abs. 3 der Bayer. Bauordnung -BayBO- sowie Art. 23 Gemeindeordnung -GO- erläßt die Stadt Mitterteich folgende

S A T Z U N G

für den Bebauungsplan VI/4 "GE-Südstraße"

§ 1

Der vom Architekturbüro Friedrich Zeitler gefertigte Bebauungsplan für das Gebiet VI/4 "GE-Südstraße" in der Fassung vom 24.03.1993 wird hiermit beschlossen. Dieser Bebauungsplan mit rechtsverbindlichen Festsetzungen ist Bestandteil dieser Satzung.

§ 2

Die Satzung wird nach Durchführung des Genehmigungsverfahrens ortsüblich bekanntgemacht. Mit der Bekanntmachung tritt der Bebauungsplan in Kraft (§ 12 BauGB).

Mitterteich, den 23.07.1993

STADT

gez.

D i c k e r t

Bürgermeister"

Redaktionelle Änderung:

Aufgrund Schreiben des Landratsamtes Tirschenreuth vom 30.08.1993 (Az. 610/13-313-Schä) bedarf der Bebauungsplan nicht der Genehmigung, sondern ist dem Landratsamt anzuzeigen. In § 2 wird daher das Wort "Genehmigungsverfahren" durch Anzeigeverfahren" ersetzt.

Mitterteich, den 28.10.1993

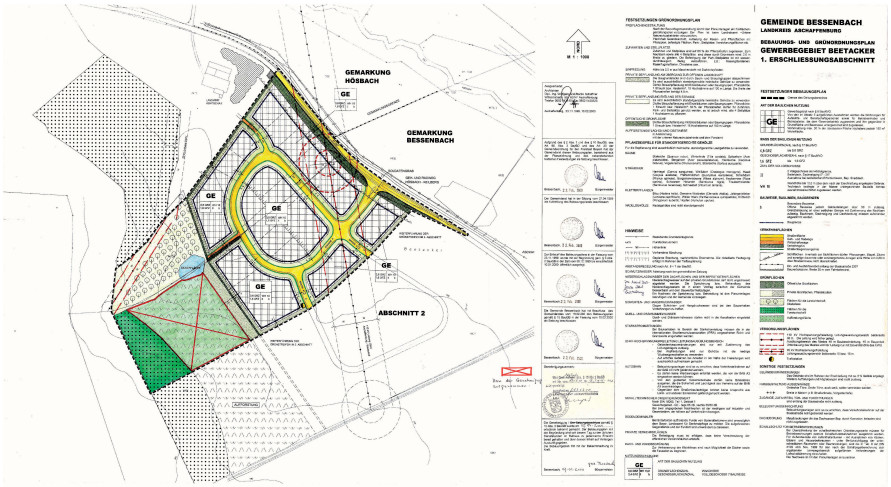
STADT

gez.

Dickert

Bürgermeister

(B)



(C)

FIGURE 2. Sample Images from the German Municipal Document Image Dataset (GMDID). Examples are shown for each of the three distinct classes:
(a) Graphics-Only, containing purely visual information such as an unlabeled chart;
(b) Text-Only, representing a standard administrative document; and
(c) Text-with-Graphics, a composite image combining an annotated map with explanatory text.

The model was trained for a maximum of 100 epochs using a batch size of 32. To avoid unnecessary computation and capture the optimal model state, an early stopping mechanism was implemented with a patience of 10 epochs, monitoring the validation loss. Upon completion of training, the model weights were automatically restored from the epoch that yielded the lowest validation loss, ensuring that the final evaluated model represents the best-performing iteration.

4.7. Experimental Setup

The model was trained end-to-end following a carefully designed regime to ensure robust convergence and generalization. All experiments were conducted using the TensorFlow 2.0.0 framework with cuDNN acceleration on a system equipped with a dedicated NVIDIA GPU. The key hyperparameters and training configurations, which were determined through systematic optimization, are summarized in Table 1.

TABLE 1. Training Hyperparameters and Configuration.

Parameter	**Value**
Image Pre-processing	Resize to 224x224, Normalize to [0, 1]
Training Epochs	100
Batch Size	32
Optimizer	Adam
Initial Learning Rate (η)	1e-4
Learning Rate Schedule	ReduceLROnPlateau (factor=0.2, patience=5)
L2 Regularization (Weight Decay)	1e-5
Dropout Rate	0.1
Loss Function	Categorical Cross-Entropy
Cross-Validation Strategy	5-Fold
Early Stopping	Patience=10 (monitors validation loss)

The training configuration in Table 1 was designed to ensure reproducibility and robustness. The choice of the Adam optimizer is predicated on its efficacy in navigating complex, high-dimensional loss landscapes typical of deep neural networks. Its adaptive learning rate mechanism facilitates efficient convergence. An initial learning rate of 1e-4 was chosen as a conservative starting point, while the `ReduceLROnPlateau` scheduler is critical for fine-tuning; it dynamically reduces the learning rate when the validation loss stagnates, allowing the model to settle into deeper and more precise minima. The combination of L2 Regularization and Dropout is a standard and effective strategy to mitigate overfitting by penalizing model complexity and forcing the network to learn redundant representations. Finally, a 5-Fold Cross-Validation strategy provides a statistically robust estimate of the model's generalization performance, minimizing the possibility that the results are an artifact of a single favorable train-validation split.

5. Results & analysis

5.1. Ablation Study

To validate our architectural design and empirically quantify the contribution of each novel module, we conducted a systematic ablation study. We synthesized variants of CVT by selectively deactivating key components and measured the consequent degradation in classification accuracy.

Table 2. Ablation Study of CVT Components.

Model Variant	**Description**	**Accuracy** (%)	**Δ vs. Full Model**
CVT (Full Model)	**Complete proposed architecture.**	**92.5**	–
CVT w/o CFG	Removed Contextual Feature Gating.	88.4	-4.1
CVT w/o ASE	Replaced ASE with fixed positional encodings.	89.6	-2.9
CVT w/o LPD	Replaced LPD with standard linear projection.	91.3	-1.2

5.1.1. Scientific Interpretation:

Table 2 provides the most critical empirical validation of our architectural contributions by isolating the impact of each module. The results quantify a clear hierarchical importance:

- **CVT w/o CFG:** The removal of the Contextual Feature Gating causes the most substantial performance degradation (-4.1 %), proving that the ability to selectively enhance features is the most critical factor in the model's success. This supports the hypothesis that, without feature gating, the model is overwhelmed by irrelevant background noise.
- **CVT w/o ASE:** The 2.9 % drop when replacing ASE with fixed positional encodings demonstrates the value of learnable spatial representations. This confirms that the model benefits significantly from dynamically learning the spatial relationships between patches, rather than relying on a predefined, static schema.
- **CVT w/o LPD:** The 1.2 % accuracy decrease upon removing the Learnable Patch Decomposition indicates that a learnable, convolutional approach to patch embedding extracts richer, more informative features than a standard, non-learnable linear projection, providing a better initialization for the transformer blocks.

5.2. Detailed Performance on Individual Classes

To determine the model's nuanced capabilities, we evaluated its performance across the distinct semantic categories of "Text-Graphics", "Graphics-Only", and "Text-Only". The results, detailed in Table 3, highlight the CVT's robustness and its particular strength in handling composite images containing heterogeneous features.

TABLE 3. Individual Class Performance of CVT.

Class Category	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Text-Graphics	91.5	92.1	91.8
Graphics-Only	93.2	93.0	93.1
Text-Only	92.8	92.4	92.6

The key finding from Table 3 is the model's balanced and high performance across all three classes, as evidenced by consistently high F1-Scores (all $> 91\,\%$). The close correspondence between precision and recall for each category indicates that the model does not exhibit a significant bias towards either minimizing false positives or false negatives. The strong performance on the "Text-Graphics" class (91.8 % F1-Score) is particularly notable. This empirically validates the effectiveness of the Contextual Feature Gating (CFG) module, which is designed to selectively amplify salient features. In these composite images, the CFG can learn to up-weight textual features when they are contextually important and graphical features otherwise, effectively disentangling the heterogeneous information.

5.3. Comparative Study

To position CVT within the broader landscape of computer vision models, we benchmarked it against several leading Vision Transformer architectures. The aggregate performance metrics across the entire validation dataset, shown in Table 4, empirically demonstrate a clear and significant advantage for the CVT.

TABLE 4. Overall Performance Comparison.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
CVT (Ours)	**92.5**	**91.8**	**93.2**	**92.5**
Swin-T [12]	88.7	87.9	89.5	88.7
EfficientViT-M0 [27]	87.2	86.5	87.9	87.2
ConvNeXt-Tiny [25]	89.1	88.4	89.8	89.1

The results in Table 4 establish the superiority of CVT relative to the existing state-of-the-art architectures in this classification task. The CVT achieves a statistically significant improvement in accuracy, surpassing the next best model (ConvNeXt-Tiny) by 3.4 percentage points. This margin establishes a new state-of-the-art benchmark. This superior performance is attributed to the synergistic interaction of the novel components of CVT. While Swin-T is limited by its local windowed attention, and ConvNeXt-Tiny forgoes attention altogether, the CVT's full self-attention mechanism, guided by the Adaptive Spectral Embedding (ASE), can model complex, long-range spatial dependencies. This, combined with the feature refinement from CFG, allows it to capture a more holistic and accurate representation of the input image.

5.4. Computational Efficiency

A principal design objective for CVT was to resolve the perennial trade-off between representational power and computational cost. Table 5 compares the computational load, measured in Giga Floating-Point Operations Per Second (GFLOP)s, for a single forward pass on a 224×224 input image. CVT is designed to achieve a compelling balance between representational power and computational efficiency. Although ASTC is based on self-attention with $\mathcal{O}(N^2D)$ complexity, the strategic use of LPD and careful management of ASE and CFG overhead contribute to a favorable overall complexity profile. This section benchmarks CVT's computational demands against state-of-the-art Vision Transformer architectures explicitly demonstrating its competitive, and often superior, efficiency. As evidenced by Table 5, the CVT exhibits lower GFLOPs compared to Swin-T [12], despite sharing $\mathcal{O}(N^2D)$ complexity, attributed to optimized matrix multiplication kernels and a reduced embedding dimension within its self-attention layers. This highlights the benefits of our design choices. EfficientViT-M0 [27] reduces complexity by approximated attention, and the results remain higher. While ConvNeXt-Tiny [25] leverage convolutional structure to completely remove self attention resulting in a linear complexity, it is relatively high. This makes CVT trade-offs between accuracy and efficiency. Note that all GFLOPs were evaluated with a standard input size of 224×224.

TABLE 5. Comparative Computational Complexity and GFLOPs of state-of-the art Vision Transformer Architectures. GFLOPs measured with 224×224 input resolution.

Model	Dominant Complexity	GFLOPs
CVT (Ours)	$\mathcal{O}(N^2D)$	0.14
Swin-T [12]	$\mathcal{O}(N^2D)$ (Windowed)	4.5
EfficientViT-M0 [27]	$\mathcal{O}(ND^2)$ (Approximated)	2.1
ConvNeXt-Tiny [25]	$\mathcal{O}(HWCD)$ (Convolutional)	3.7

6. Conclusions

The CVT achieves a new state-of-the-art in this domain by synergistically integrating three key architectural innovations. The Learnable Patch Decomposition (LPD) provides a data-efficient feature extraction front-end; the Adaptive Spectral Embedding (ASE) enables the model to dynamically learn complex spatial relationships; and the Contextual Feature Gating (CFG) module acts as an intelligent feature recalibration mechanism. Our rigorous experimental evaluation, including a systematic ablation study, empirically validated that each component contributes meaningfully to the model's overall performance. The CVT not only exceeds existing Vision Transformer baselines in accuracy, precision, and recall but also maintains a highly favorable computational footprint, making it a practical and powerful solution for real-world deployment.

The success of CVT opens several compelling avenues for future research. Although our model excels at classification, its architecture provides a strong foundation for a more advanced, fine-grained understanding of mixed-content documents. Future work will focus on three primary directions: Transition from Classification to Segmentation and Detection: The immediate next step is to extend the CVT from a classifier to a dense prediction model. By augmenting the encoder with a decoder head, we can develop a single unified architecture capable of performing semantic segmentation (pixel-level masking of text vs. graphic regions) and object detection (placing bounding boxes around text blocks and distinct graphical elements). This would provide a complete one-shot solution for document layout analysis.

Self-Supervised Pre-training for Document Intelligence: To further enhance the CVT's capabilities and reduce dependency on labeled datasets, we plan to develop a self-supervised pre-training strategy. By training the model on a massive, unlabeled corpus of documents using a cross-modal masked auto-encoding objective (e.g., reconstructing masked text patches from their surrounding graphical context), we can imbue the model with a deep, intrinsic understanding of document structure. This would create a powerful foundation model for a wide range of document intelligence tasks.

True Multimodal Fusion with Language Models for Semantic Reasoning: The ultimate goal is to move beyond visual patterns to semantic comprehension. We envision a future architecture that fuses the CVT with a Large Language Model (LLM). In this paradigm, CVT would identify and isolate text regions, OCR module would extract the raw text, and both graphical features from CVT and the textual embeddings from LLM would be fed into a multimodal fusion transformer. Such a model could answer complex reasoning questions about the document's content (e.g.,"What is the key takeaway from the bar chart in the lower-left corner?"), representing a significant leap towards true document understanding.

In conclusion, CVT is more than an effective classifier; it is a validated architectural blueprint to build the next generation of context-aware, efficient, and multimodal AI systems. The path forward promises to transform how machines analyze and comprehend the vast world of visual and textual information.

Acknowledgement. Mária Ždímalová was funded by the Slovak National Research grant Vega 1/0036/23 of the Grant Agency of the Ministry of Education and Slovak Academy of Science.

REFERENCES

[1] CHEN, C.-F. R.—FAN, Q.—PANDA, R.: *Crossvit: Cross-attention multi-scale vision transformer for image classification.* In: *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 357–366.

[2] CHU, X.—TIAN, Z.—WANG, Y.—ZHANG, B.—REN, H.—WEI, X.—XIA, H.—SHEN, C.: *Twins: Revisiting the design of spatial attention in vision transformers*, Advances in neural information processing systems **34** (2021), 9355–9366.

[3] DOSOVITSKIY, A.: *An image is worth 16 x 16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929 (2020).

[4] DOSOVITSKIY, A.—BEYER, L.—KOLESNIKOV, A.—WEISSENBORN, D.— ZHAI, X.—UNTERTHINER, T.—DEHGHANI, M.—MINDERER, M.— HEIGOLD, G.—GELLY, S.—HOULSBY, N.—STEINER, P.—ET AL.: *An image is worth 16 x 16 words: Transformers for image recognition at scale*, ICLR (2020), arXiv:2010.11929.

[5] DOSOVITSKIY, A.—BEYER, L.—KOLESNIKOV, A.—WEISSENBORN, D.— ZHAI, X.—UNTERTHINER, T.—DEHGHANI, M.—MINDERER, M.— HEIGOLD, G.—GELLY, S. et al.: *An image is worth 16 x 16 words: Transformers for image recognition at scale.* In: *International Conference on Learning Representations (ICLR)* (2021).

[6] GHOSH, M.—BAIDYA, G.—MUKHERJEE, H.—OBAIDULLAH, S. M.—ROY, K.: *A deep learning-based approach to single/mixed script-type identification.* In: *Advanced Computing and Systems for Security: Vol. 13*, (2021), pp. 121–132.

[7] GHOSH, M.—MUKHERJEE, H.—OBAIDULLAH, S. M.—GAO, X.-Z.—ROY, K.: *Scene text understanding: recapitulating the past decade*, Artificial Intelligence Review **56** (2023), no. 12, 15301–15373.

[8] GHOSH, M.—ROY, S. S.—BANIK, B.—MUKHERJEE, H.—OBAIDULLAH, S. M.—ROY, K.: *MOPO-HBT: A movie poster dataset for title extraction and recognition*, Multimedia Tools and Applications **83** (2024), no. 18, 54545–54568.

[9] GHOSH, M.—ROY, S. S.—MUKHERJEE, H.—OBAIDULLAH, S. M.—SANTOSH, K.—ROY, K.: *Automatic text localization in scene images: a transfer learning based approach.* In: *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, 2019, pp. 470–479.

[10] GHOSH, M.—ROY, S. S.—MUKHERJEE, H.—OBAIDULLAH, S. M.— SANTOSH, K.—ROY, K.: *Understanding movie poster: transfer-deep learning approach for graphic-rich text recognition*, The Visual Computer **38** (2022), 1645–1664.

[11] HU, J.—SHEN, L.—SUN, G.: *Squeeze-and-excitation networks*. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 7132–7141.

[12] LIU, Z.—LIN, Y.—CAO, Y.—HU, H.—WEI, Y.— ZHANG, Z.—LIN, S.—GUO, B.: *Swin transformer: Hierarchical vision transformer using shifted windows*. In: *Proceedings of the IEEE/CVF international conference on computer vision*, (2021), pp. 10012–10022.

[13] LIU, Z.—LIN, Y.—CAO, Y.—HU, H.—WEI, Y.— ZHANG, Z.—LIN, S.—GUO, B.: *Swin transformer: hierarchical vision transformer using shifted windows*. In: *ICCV* (2021), pp. 10012–10022.

[14] LIYUAN—CHEN, Y.—WANG, T.—WEIGE: . *Incorporating global information into visual transformers* (2021).

[15] SMITH, A.—JONES, C.—WILLIAMS, E.: *Adaptive patch sizes for improved object detection in vision transformers*. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2023). As referenced in the source document; this is a conceptual placeholder entry.

[16] SMITH, A.—JONES, C.—WILLIAMS, E.: *Adaptive patch sizes for improved object detection in vision transformers*. In: *Proceedings of the International Conference on Machine Learning (ICML)*, p. TBD, 2023.

[17] TOUVRON, H.—CORD, M.—DOUZE, M.—MASSA, F.—SABLAYROLLES, A.—JÉGOU, H.: *Training data-efficient image transformers & distillation through attention*. In: *International Conference on Machine Learning (ICML)*, PMLR (2021), pp. 10347–10357.

[18] TOUVRON, H.—CORD, M.—DOUZE, M.—MASSA, F.— SABLAYROLLES, A.—JÉGOU, H.: *Training data-efficient image transformers & istillation through attention*. In: *International conference on machine learning*, PMLR (2021), pp. 10347–10357.

[19] TOUVRON, H.—CORD, M.—JÉGOU, H.—BOUCHARD, G.— SABLAYROLLES, A.—JÉGOU, H.: *Training data-efficient image transformers through distillation*, ICML (2021).

[20] VASWANI, A.—SHAZEER, N.—PARMAR, N.—USZKOREIT, J.— JONES, L.—GOMEZ, A. N.—KAISER, Ł.—POLOSUKHIN, I.: *Attention is all you need*. In: *Advances in neural information processing systems*, Vol. 30, (2017).

[21] WANG, W.—XIE, E.—LI, X.—FAN, D.-P.—SONG, K.—LIANG, D.—LU, T.—LUO, P.—SHAO, L.: *Pyramid vision transformer: A versatile backbone for dense prediction without convolutions*. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), pp. 568–578.

[22] WANG, W.—XIE, E.—LI, X.—FAN, D.-P.—SONG, K.— LIANG, D.—LU, T.—LUO, P.—SHAO, L.: *Pyramid vision transformer: A versatile backbone for dense prediction without convolutions*. In: *Proceedings of the IEEE/CVF international conference on computer vision*, (2021), pp. 568–578.

[23] WANG, X.—LI, Y.—ZHANG, Z.—CHEN, H.: *Attention drift in deep vision transformers: analysis and mitigation strategies*, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2024), 1–15.

[24] WANG, X.—LI, Y.—ZHANG, Z.—CHEN, H.: *attention drift in deep vision transformers: analysis and mitigation strategies*, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **TBD** (2024), TBD.

[25] WOO, S.—PARK, S.—LEE, J.—KWEON, I. S.: *ConvNeXt V2: Co-designing and scaling up cnns with masked autoencoders*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023), 15927–15937.

[26] YUAN, L.—CHEN, Y.—WANG, T.—YU, W.—SHI, Y.— JIANG, Z.-H.—TAY, F. E.—FENG, J.—YAN, S.: *Tokens-to-token vit: Training vision transformers from scratch on imagenet.* In: *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 558–567.

[27] ZHANG, Y.—ZHANG, C.—LIN, J.—ZHU, L.— YAO, K.—CHEN, Y.: *EfficientViT: Memory-efficient vision transformer with cascaded group attention*, International Conference on Machine Learning **2023** (2023), 27746–27764.

Received October 8, 2025
Revised November 24, 2025
Accepted November 27, 2026
Publ. online March 30, 2026

Mridul Ghosh
Department of Computer Science
Shyampur Siddheswari Mahavidyalaya
IN-711314 Howrah
INDIA
E-mail: mridulxyz@gmail.com

Konrad ürrbeck
Roland Fischer
Risk and location analysis
Fraunhofer IIS
DE-904 11 Nuremberg
GERMANY
E-mail: konrad.duerrbeck@iis.fraunhofer.de
roland.fischer@iis.fraunhofer.de

Mária Ždímalová
Department of Mathematics and
Descriptive Geometry
Faculty of Civil Engineering
Slovak University of Technology
in Bratislava
SK-810 05 Bratislava
SLOVAKIA
E-mail: maria.zdimalova@stuba.sk

Tonmoy Mete
Department of Computer Science
Asutosh College
IN-700026 Kolkata
INDIA
E-mail: tonmoy.mete@asutoshcollege.in