IJWR

Unlocking Book Genre from Covers: A Multimodal Approach to Book Genre Prediction

Reza Toosi[a]*, Alireza Hosseini[b], Ramin Toosi[b] , Mohammad Ali Akhaee[b]*

[a] Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran; rtoosi81@gmail.com

[b] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran; {arhosseini77, r.toosi, akhaee}@ut.ac.ir

ABSTRACT

In today's visually driven market, book cover design plays a crucial role in conveying a work's narrative and thematic essence. A book cover is a multimodal entity, consisting of various visual and textual elements. While conventional recommendation systems have often overlooked the semantic richness of cover imagery, prior work attempting to incorporate textual information relied on OCR to extract text from covers. However, these raw tokens capture only a fraction of the cover's meaning and often miss deeper thematic and narrative cues. Recognizing these limitations, we leverage the advanced knowledge accumulated in VLMs to derive a more comprehensive representation, using this knowledge to add it as an additional feature to the system. In this paper, we use VLM-generated descriptions and integrate these rich descriptions as a new textual feature. Our enhanced corpus comprises 57,000 book covers across 30 genres (1,900 per genre), each annotated with both raw imagery and VLM-generated narrative summaries. We fuse two state-of-the-art vision encoders (ViT and VisionMamba) with a text encoder that processes these VLM descriptions. Experimental results demonstrate a Top 1 accuracy of 63.31% and a Top 3 accuracy of 83.03%, marking a substantial improvement over the state-of-the-art variant and underscoring the value of VLM-derived context in multimodal genre classification.

***Keywords*— *Book Cover Analysis, Book Genre Prediction, Multimodal Learning, Vision Transforme, Mamba*.**

1. Introduction

In the modern, image driven world, cover designs are widely understood to be far more than decorative elements [1]; they function as potent conveyors of meaning and emotion. Across literature, music, and the visual arts, the combination of imagery, typography, and composition delivers vital signals that influence audience perceptions and form expectations about a work's tone and substance. The visual impression a cover creates can shape mood, imply genre, and even guide consumer decisions, underscoring its significance across multiple creative disciplines.

The importance of cover art has historically been recognized not only in the publishing industry but also in other media domains. Studies on album covers [2] [3] [4], film posters [5] [6] [7], and paintings have repeatedly demonstrated the ability of visual design to communicate artistic vision [8] [9] [10]. In book cover analysis, early research often utilized traditional neural network models such as AlexNet and LeNet to extract visual features [11]. More recently, multimodal approaches have gained prominence, combining visual and textual data through methods such as deep multimodal architectures and vision transformers to enhance genre classification performance [12]. Within this field, existing studies on book genre classification by cover can generally be grouped into three primary categories.

We posit that a book cover comprises two core elements: a visual component, encompassing graphical and aesthetic attributes, and a textual component, containing typographic features and embedded textual information. For the visual stream, our method leverages two cutting-edge architectures Vision Transformer (ViT) and VisionMamba [13].

http://dx.doi.org/10.22133/ijwr.2025.525259.1285

Citation R. Toosi, A. Hosseini, R. Toosi, M.A. Akhaee, "Unlocking Book Genre from Covers: A Multimodal Approach to Book Genre Prediction ", *International Journal of Web Research*, vol.8, no.3,pp.73-81, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.525259.1285.

*Coressponding Author

Article History: Received: 21 April 2025; Revised: 12 June 2025; Accepted: 18 June 2025 .

ViT is adept at modeling the global context of an image, enabling effective representation of its overall structure and layout.

On the other hand, the Mamba architecture [14] which has emerged as a powerful force in computer vision in recent years, demonstrated exceptional performance across a variety of tasks [15] [16]. Building upon this success, we use VisionMamba-a hybrid model that combines CNNs with Mamba state space modules and self-attention mechanisms. This fusion allows it to extract nuanced, multidirectional patch features that are often overlooked by conventional methods, ensuring a comprehensive representation of the visual content. In parallel, to further enrich the textual modality and capture nuances beyond raw optical character recognition (OCR) tokens, we leverage the vision language models **(**VLM) to produce narrative descriptions of each book cover. This substitution yields richer semantic cues such as thematic motifs, color palettes, and implied objects that enhance our multimodal fusion strategy. The textual features are then addressed using the sentence transformer model [17], specifically miniLM-L6, which parses and robustly encodes semantic information from these VLM generated descriptions.

The features extracted from these two modalities are subsequently fused into a unified feature vector, which is then provided as input to a final classifier that predicts the book's genre. This integrated approach not only exploits the complementary strengths of visual and textual representations but also mitigates the limitations associated with single modality systems. By combining the capabilities of ViT, VisionMamba, the VLM for richer text generation, and miniLM-L6 for encoding, our framework overcomes previous shortcomings and demonstrates marked improvements in genre recognition accuracy, paving the way for a more comprehensive and precise book genre classification system.

This work is an extended version of our earlier conference paper, which focused solely on OCR based textual feature extraction for book genre classification [18]. In the present work, we significantly enhance the textual modality by incorporating vision language model (VLM) generated descriptions, leading to richer semantic representations and improved classification accuracy.

2. Related Work

2.1. Image-Based Approaches

Methods that depend solely on visual information derived from cover imagery have been the subject of multiple investigations [19] [11].

In one such study, Iwana et al. [11] concentrated exclusively on book covers, utilizing a pre-trained AlexNet [20] model on the ImageNet dataset [21]. Their approach achieved a Top-1 prediction accuracy of 24.7% and a Top-3 accuracy of 40.3%. Buczkowski et al. [19] two convolutional neural network (CNN) architectures were developed: a simpler model (N1) with three convolutional layers, max pooling, and fully connected layers, and a deeper VGG-inspired model (N2) incorporating dropout regularization. Both models were trained on a dataset of 14 relabeled genres derived from GoodReads cover images. N1 achieved superior performance with and 61% accuracy, outperforming N2 was 58% accuracy.

2.2. Text-Based Approaches

Approaches centered on textual information make use of the rich semantic content present in cover text, and have been examined in numerous studies [6] [22] [23] [24].

Gupta et al. [22] concentrated on text-based methods by transforming book texts into feature matrices using term frequency-inverse document frequency (TF-IDF), reducing dimensionality with principal component analysis (PCA), and applying an AdaBoost classifier, which achieved an accuracy of 92.88% when incorporating unlabeled data. Kundalia et al. [6] proposes a transfer learning approach using InceptionV3 to predict movie genres from posters. A balanced dataset of 30,000 posters (12 genres, 2,500 each) was created to address class imbalance. The model, trained on single label data but predicting Top-3 genres, achieved 84.82% accuracy on a multi-label test set, leveraging pre-trained features and fully connected layers for classification.

2.3. Multimodal Approaches

Recent advancements have increasingly emphasized the integration of both visual and textual cues [25]. Biradar et al. [26] introduced a model that combines cover imagery with title text, employing the Xception network for extracting image features and GloVe embeddings for textual representation. Working with a dataset of 6,800 book covers across five genres, their method achieved an accuracy of 87.2%. Building on this line of research, Rasheed et al. [12] proposed a multimodal deep learning framework that merges an EXplicit interActive Network (EXAN) for text processing with a CNN enhanced by Gram and SE layers for image classification, attaining classification accuracies of 69.09% on a Latin book dataset and 38.12% on an Arabic book dataset.

Haraguchi et al. [27] explored how text design elements influence book genre classification. Their study revealed that semantic features alone yielded a baseline accuracy of 45.46%, while incorporating design related features increased accuracy to 48.45%. Similarly, Patel and Aggarwal's BGCNet [28] illustrates the potential of advanced multimodal

architectures. Their model integrates dual visual processing streams one focused on local detail capture and the other on global context alongside a robust text encoder, resulting in a Top-1 accuracy of 52.5% and a Top-3 accuracy of 74.6% on the BookCover30 dataset. In another line of investigation, Bielawski et al. [29] assessed the performance of contrastive language image pre-training (CLIP), a vision language model trained via multimodal contrastive learning, for genre classification in both movies and books, reporting an F1-score of 68.7%. Existing multimodal techniques for book genre classification often employ a straightforward combination of text embeddings with visual features obtained from CNNs or conventional transformer-based models. However, such approaches frequently demonstrate constrained accuracy. A key limitation lies in their inability to fully capture the breadth of visual signals embedded in book covers particularly directional cues that are essential for identifying design patterns and stylistic subtleties and in their underutilization of the synergistic relationship between visual and textual information.

Our previous work [18] proposed an OCR based multimodal model, integrating visual features with text extracted directly from book covers. While effective, this approach was limited by the shallow semantic depth of OCR tokens. The current study extends that work by replacing OCR text with VLM generated descriptions, thereby capturing richer thematic and stylistic cues.

3. Proposed Method

In the proposed framework, we present a multimodal classification model that draws on both textual and visual information from book covers to determine their respective genres. Text appearing on the cover is first transformed into a rich semantic embedding. For visual processing, the model incorporates two complementary encoders: the first employs the MambaVision framework, introduced by Hatamizadeh et al. [13], to extract fine-grained, multidirectional patch level features, while the second applies a Vision Transformer (ViT) [30] to capture broader, global contextual patterns within the image. The outputs from these two visual encoders are then combined with the textual embedding via a fusion mechanism, which concatenates the features into a single joint representation. This unified vector is subsequently fed into a final classification layer to predict the book's genre. An overview of the architecture is illustrated in Figure 1.

3.1. Text Encoder

To capture the semantic content from the book cover, we first generate a descriptive paragraph using a VLM. This paragraph encapsulates the visual and thematic elements of the cover, such as objects, color palettes, and implied motifs.

The generated description $\boldsymbol{T}$ is then tokenized using the tokenizer associated with the MiniLM-L6-v2 pre-trained model [17]. This lightweight model produces a compact semantic representation of the description. Formally, after tokenization, the input $\boldsymbol{T}$ is transformed into a sequence of tokens $\{t_1, t_2, \ldots, t_n\}$. Equation (1) defines the processing of these tokens by the text encoder to produce a sequence of hidden states.

$$E = \textit{all-MiniLM-L6-v2}(\{t_1, t_2, \ldots, t_n\}) \quad (1)$$

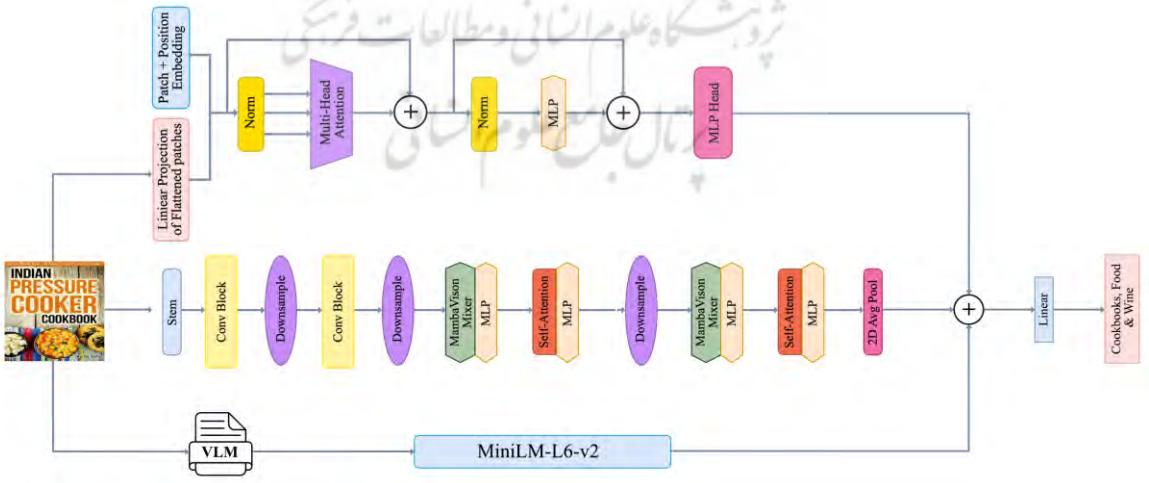


Figure. 1. Overview of the proposed multimodal architecture integrating a text encoder with dual visual encoders MambaVision and Vision Transformer (ViT).

We select the hidden state corresponding to the first token (i.e., the [CLS] token) as the feature vector. Equation (2) specifies this selection.

$$f_{text} = E_{[CLS]} \quad (2)$$

This feature vector, with dimensionality d_{text}, encapsulates the semantic information of the cover's text.

3.2. Vision Encoder

Our architecture captures complementary visual representations through two distinct components: the MambaVision encoder and the Vision Transformer (ViT) encoder.

MambaVision Encoder: Let $\boldsymbol{I \in R^{H \times W \times 3}}$ represent the input image. The MambaVision encoder begins with a stem layer that reduces the spatial resolution of $\boldsymbol{I}$ by applying two successive 3 × 3 convolutional layers with a stride of 2. This transformation produces a feature map $\boldsymbol{F \in R^{H' \times W' \times C}}$, where $\boldsymbol{H'}$ and $\boldsymbol{W'}$ are the reduced height and width, and $\boldsymbol{C}$ denotes the number of channels in the embedding space.

The resulting feature map $\boldsymbol{F}$ is further processed by the MambaVision Mixer, which operates through two parallel pathways. In the first pathway, $\boldsymbol{F}$ undergoes a linear projection followed by convolution and a non-linear activation function (e.g., SiLU [31]). A selective scan operation [14] is then applied to model sequential dependencies. The output from this pathway is denoted as $\boldsymbol{F_1}$, as expressed in Equation (3).

$$F_1 = Scan\ (\sigma\ (Conv\ (Linear(F)))) \quad (3)$$

In the second pathway, $\boldsymbol{F}$ undergoes a comparable transformation but omits the selective scan stage. The resulting output, $\boldsymbol{F_2}$, is expressed in Equation (4):

$$F_2 = \sigma\ (Conv\ (Linear(F))) \quad (4)$$

The outputs from the two pathways $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ are then concatenated and passed through an additional linear projection to generate the final feature representation, as shown in Equation (5):

$$F_{out} = Linear\ (Concat\ (F_1, F_2)) \quad (5)$$

The final output $\boldsymbol{F_{out}}$ serves as the MambaVision feature vector, denoted by $\boldsymbol{f_{mamba}}$.

ViT Encoder: Running in parallel with the MambaVision encoder, the Vision Transformer (ViT) encoder is designed to extract global visual context. Built upon the vision backbone of the CLIP model, the ViT encoder first partitions the input image $\boldsymbol{I}$ into fixed size patches. Each patch is then linearly projected into an embedding space and enriched with positional encodings, producing a sequence of patch embeddings.

This sequence is passed through multiple transformer layers, each comprising multi-head self-attention mechanisms and feed forward networks, enabling the model to capture long range dependencies and holistic structural information. The final global feature representation, $\boldsymbol{F_{vit}}$, is obtained from the transformer's pooler output, as expressed in Equation (6).

$$F_{vit} = ViT\ (I)_{pooler} \quad (6)$$

This representation encapsulates the overall context of the image.

Fusion and Classification

The outputs from the text encoder, MambaVision encoder, and ViT encoder are concatenated to create a single unified feature vector. This operation is expressed in Equation (7).

$$f_{combined} = [\ f_{text}\ ;\ f_{mamba}\ ;\ f_{vit}\ ]. \quad (7)$$

To more effectively merge the multimodal information, the concatenated feature vector can be passed through a series of fully connected layers. Let $\boldsymbol{L}$ represent the number of hidden layers, each containing mmm neurons. As shown in Equation (8), the computation begins by setting $h_0 = f_{combined}$ and for each layer $i=1,\ldots,$ L_i the intermediate representations are calculated accordingly.

$$h_0 = f_{combined}\ ,\ h_i = ReLU(W_i h_{i-1} + b_i)\ ,\quad i = 1,\ldots,L. \quad (8)$$

Finally, a linear classifier maps the final hidden representation to the genre logits. Equation (9) defines this mapping.

$$\hat{y} = W_{cls}\ h_L + b_{cls} \quad (9)$$

Here, $\hat{\boldsymbol{y}}$ denotes the predicted logits corresponding to the book genre classes. This fusion classification pipeline exploits the complementary strengths of the textual and visual streams, resulting in a more robust and accurate genre prediction process.

4. Experiments and Results

This section presents a detailed overview of our experimental setup, including implementation details, dataset description, comparative analysis with existing models, and an ablation study to assess the contributions of different components in our model.

4.1. Dataset

Our evaluation is conducted on the balanced BookCover30 dataset [11], which contains 57,000 book covers evenly distributed across 30 distinct genres (1,900 titles per genre). Each book is assigned to exactly one genre, creating a perfectly balanced class distribution and making the dataset a strong benchmark for assessing classification performance. Figure 2 illustrates examples from the dataset, along with the classification outcomes across genres, including predicted genre labels and their respective confidence scores.

For the textual modality, we explore two distinct approaches to feature extraction from the book covers. The first approach uses traditional OCR to extract the visible text directly from the cover this represents the text that is literally seen. To ensure high quality textual inputs from this method, we first processed all OCR outputs by ranking every extracted word and sentence according to its confidence score, retaining only the top 25 % most reliable tokens. We further constrained each sample to at most 12 words corresponding to the mean valid length to avoid overloading our model with extraneous text. Any terms identified as non-English (comprising under 2% of the total corpus) were excluded to reduce noise and maintain language consistency. This rigorous filtering pipeline guarantees that subsequent analyses based on OCR are driven by the clearest, most pertinent visible cover text.

Visual concepts implicitly conveyed by the cover design, As shown in Figure 3, leveraging the knowledge accumulated in VLMs this represents the concepts embedded within the visual elements. To achieve this, we employed a VLM to automatically generate concise, human readable description for every book cover and replace the limited OCR output with the returned description for this feature set. These VLM derived summaries capture thematic cues, plot highlights, and stylistic nuances that go beyond bare metadata or simple visible text, providing an alternative, richer textual feature set derived from the visual elements. By integrating these descriptions into our feature fusion pipeline, we can assess how this deeper descriptive context impacts performance.

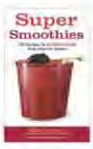
4.2. Implementation Details

Our implementation was developed in PyTorch, employing a multimodal framework that integrates a text encoder (MiniLM-L6-v2) with two visual encoders a ViT model (adapted from CLIP, openai/clip-vit-base-patch32) and the MambaVision architecture. All modules were initialized with pre-trained parameters. The final classification head comprises two fully connected layers of 256 units each, mapping the fused multimodal representation to 30 genre categories.

Training was carried out for a maximum of 20 epochs using the AdamW optimizer with both weight

Figure. 2. Sample distribution of data and classification outcomes across various book genres. The accompanying table displays predicted genre labels alongside their confidence scores.

The book cover features a young girl, presumably Red Riding Hood, in a red hooded cloak, holding a basket. She appears to be running, and the background depicts a dark and mysterious forest scene. The title "RED" is prominently displayed in a stylized font above her. Above the title is the author's name, "Liesl Shurtliff" and a reference to another book she wrote. At the top right, partially hidden in the trees, is the head of a wolf, hinting at the story's antagonist. Below the illustration is the tagline: "The true story of Red Riding Hood." The artwork is framed within a circular border, and the edges of the cover have a red, textured background with hints of golden light.



The book cover is for "Super Smoothies: 100 Recipes to Supercharge Your Immune System" by Ellen Brown. The title is prominently displayed in a bold, reddish-brown font. Below the title, in smaller letters, is the subtitle. The focal point of the cover is a glass of a red-colored smoothie, presumably made with berries, garnished with a skewer holding a few blackberries. The bottom of the cover features the author's name in the same reddish-brown color. A band of solid reddish-brown color forms the lower border of the cover, providing a backdrop for the author's name.

The book cover is titled "My Indian Kitchen" in bold, stylized white lettering with "my" in a yellow script. The subtitle reads "Preparing Delicious Indian Meals Without Fear or Fuss" in smaller yellow text. To the left, a vertical collage displays various spices and ingredients in individual compartments, including seeds, star anise, cardamom pods, and dried chilies. On the right, a small portrait of a man named Hari Nayak, presumably the author, is placed above a picture of a plate of vibrant yellow rice. The plate is set atop a blue textured background with patterns, along with a stone bowl and naan-like bread. The book's publisher, Tuttle, is indicated at the bottom left.

Figure. 3. sample of VLM description

decay and gradient clipping, alongside an adaptive learning rate scheduler initialized at a few × 10^{-5}. Early stopping was applied if validation performance failed to improve for five consecutive epochs. During fine-tuning, only the classification layers and fusion component were updated, while the pre-trained visual backbones were kept frozen.

To further enrich the textual modality with descriptions of visual content, we utilized the Gemini 2.0 Flash API with the query: "This is a book cover. Describe it in a paragraph." The generated descriptions were incorporated as auxiliary textual features. All experiments were executed on an NVIDIA RTX 3090 GPU with a batch size of 64.

4.3. Comparative Analysis

To evaluate the effectiveness of our approach relative to existing methods, we performed a comparative study. Table I reports the Top-1 and Top-3 accuracy scores obtained by different models.

As shown in Table II, our method achieves the highest performance, with a Top-1 accuracy of 63.31% and a Top-3 accuracy of 83.03%. These results surpass those of conventional baselines as well as more advanced ensemble techniques and recent state-of-the-art solutions [27].

Different genres exhibit distinct cover conventions-ranging from minimalistic art in literary fiction to bold iconography in science fiction. By dissecting our model's accuracy gains on a per genre basis, we observe that certain categories (e.g., Teen & Young Adult improve 28.7%, Religion & Spirituality improve 17%, Humor & Entertainment improve 26%) benefit disproportionately from VLM generated descriptions. In Figure 4 presents these genre specific improvements in Top 1 accuracy.

Furthermore, the class-wise performance detailed in Table III illustrates that our approach consistently delivers robust results across different genres. The analysis reveals notable improvements in challenging categories-demonstrating the model's effectiveness in handling the diverse and balanced distribution of genres present in the dataset.

4.4. Ablation Study

To gain deeper insight into the role of each modality and the effectiveness of our fusion mechanism, we performed an ablation study comparing multiple configurations of our model, as detailed in Table II.

The results show that incorporating textual features alongside visual representations leads to a substantial performance increase. For example, pairing the ViT with the text encoder already yields competitive results; However, introducing the VisionMamba branch tailored to capture fine-grained, localized patterns further enhances redictive accuracy.

This multimodal fusion harnesses both global structuralcues and salient local details, producing a richer feature representation and superior classification performance.

In our feature fusion pipeline, we replaced every OCR derived text field with its corresponding VLM description. This straightforward substitution allows us to isolate the impact of semantic richness on model accuracy. The Table IV compares classification results when using (1) visual features + Cover.

OCR versus visual features + VLM descriptions, highlighting the performance delta attributable to our new annotations.

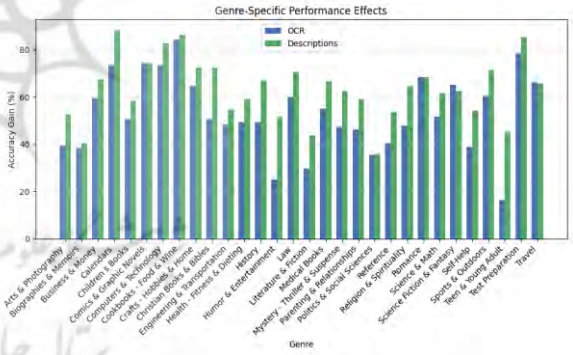


Figure. 4. comparing the impact of OCR and VLM across all genres on model accuracy

Table 1. Results of the ablation study comparing different model variants and their classification performance.

Model Variant	*Parameters (M)*	*Top-1(%)*	*Top-3(%)*
VisionMamba	32	27.23	47.82
ViT	87	50.54	74.29
MiniLM-L6-v2 (OCR) [18]	22	38.30	55.43
MiniLM-L6-v2 (VLM)	22	54.69	74.45
VisionMamba + MiniLM-L6-v2 (OCR) [18]	54	45.31	64.92
VisionMamba + MiniLM-L6-v2 (VLM)	54	62.27	82.01
ViT + MiniLM-L6-v2 (OCR) [18]	109	52.22	73.70
ViT + MiniLM-L6-v2 (VLM)	109	62.52	82.23
VisionMamba + ViT + MiniLM-L6-v2 (OCR)[18]	141	52.98	75.27
VisionMamba + ViT + MiniLM-L6-v2 (VLM)	141	63.31	83.03

Table 2. Comparative Analysis of Different Models

Model Name	*Top-1(%)*	*Top-3(%)*
LeNet[11]	13.5	27.8
AlexNet[11]	24.7	40.3
ResNet – LSTM[28]	47.3	64.7
ResNet – GloVe[28]	52.5	74.6
ResNet+YOLO – GloVe[28]	52.7	74.4
ResNet50 Ensemble[32]	40.0	66.0
MobileNet-V1	27.2	46.4
MobileNet-V2	23.6	42.0
Inception-V2	26.2	45.6
DenseNet-161	39.99	-
GoogleNet	40.21	-
VGGNet-16	25.6	45.6
VGG-19	41.27	-
RNN-LSTM	41.5	61.6
ResNet-50	42.21	-
ResNet-101	42.01	-
ResNet-152	43.11	-
Inception-ResNet-v2	43.96	-
ResNeXt-50	44.13	-
ResNeXt-101	45.19	-
ResNeXt-101 with fusion	47.44	-
Full model (all text design)[27]	48.45	68.90
Baseline (only semantic)[27]	45.46	67.00
Our's Model + OCR [18]	**52.98**	**75.27**
Our's Model + VLM	**63.31**	**83.03**

Table 3. Class-wise Performance (Description)

Class Name	*Top-1(%)*	*Top-3(%)*
Arts & Photography	52.63	82.11
Biographies & Memoirs	40.53	71.58
Business & Money	67.37	84.74
Calendars	88.42	94.74
Children's Books	58.42	81.05
Comics & Graphic Novels	74.21	87.89
Computers & Technology	82.63	92.11
Cookbooks - Food & Wine	86.32	97.89
Crafts - Hobbies & Home	72.63	85.79
Christian Books & Bibles	72.63	84.74
Engineering & Transportation	54.74	76.32
Health - Fitness & Dieting	58.95	84.21
History	66.84	84.21
Humor & Entertainment	51.58	75.79
Law	70.53	84.21
Literature & Fiction	43.68	81.05
Medical Books	66.67	86.77
Mystery - Thriller & Suspense	62.63	79.47
Parenting & Relationships	58.95	83.68
Politics & Social Sciences	35.79	66.84
Reference	53.68	72.63
Religion & Spirituality	64.74	86.84
Romance	68.42	81.58
Science & Math	61.58	78.95
Science Fiction & Fantasy	62.63	77.37
Self-Help	54.21	80.53
Sports & Outdoors	71.58	89.47
Teen & Young Adult	45.26	78.95
Test Preparation	85.26	94.74
Travel	65.79	84.74

5. Conclusion

This study presents a multimodal framework that integrates both visual and textual information for book genre classification. The framework combines MambaVision and Vision Transformer encoders for visual features with a MiniLM-L6-v2 text encoder.

Unlike our earlier work [18], which relied solely on OCR derived text, the extended version presented here employs VLM generated descriptions to capture richer semantic cues, leading to substantial accuracy gains across multiple genres.

Our analysis shows that the VLM descriptions particularly benefit genres where visual storytelling elements are prominent (e.g., Teen & Young Adult, Humor & Entertainment, Religion & Spirituality), with accuracy improvements exceeding 25% in some categories. This highlights the importance of bridging visual and textual modalities using high-level semantic representations rather than raw text extraction alone. For future research, we propose three directions:

Multi-label classification: Extending the model to predict multiple genres per book, reflecting real-world publishing trends.

Table 4. Class-wise Results: OCR vs. VLM

Class Name	OCR[18]	VLM
Arts & Photography	39.32	52.63
Biographies & Memoirs	38.42	40.53
Business & Money	59.47	67.37
Calendars	73.68	88.42
Children's Books	50.53	58.42
Comics & Graphic Novels	74.21	74.21
Computers & Technology	73.68	82.63
Cookbooks - Food & Wine	84.21	86.32
Crafts - Hobbies & Home	64.74	72.63
Christian Books & Bibles	50.53	72.63
Engineering & Transportation	48.42	54.74
Health - Fitness & Dieting	49.47	58.95
History	49.47	66.84
Humor & Entertainment	25.26	51.58
Law	60.00	70.53
Literature & Fiction	29.63	43.68
Medical Books	55.03	66.67
Mystery - Thriller & Suspense	47.37	62.63
Parenting & Relationships	46.32	58.95
Politics & Social Sciences	35.53	35.79
Reference	40.53	53.68
Religion & Spirituality	47.89	64.74
Romance	68.42	68.42
Science & Math	51.58	61.58
Science Fiction & Fantasy	65.26	62.63
Self-Help	38.95	54.21
Sports & Outdoors	60.53	71.58
Teen & Young Adult	16.53	45.26
Test Preparation	78.42	85.26
Travel	66.16	65.79

Domain adaptation: Evaluating performance on different datasets, such as non-English covers or independent publisher collections, to test robustness.

Generative augmentation: Leveraging large language models not only for description but also for generating synthetic cover–description pairs to enhance training diversity.

Overall, this work demonstrates that integrating advanced vision language models into multimodal pipelines can substantially improve genre classification performance, paving the way for richer content-based recommendation systems.

Declarations

Conflict of interest

The author declares that no conflicts of interest exist.

Acknowledgements

We extend our gratitude to the Adak Vira Iranian Rahjoo (Avir) company for their invaluable assistance with this study.

References

[1] C. A. Kratz, "On Telling/Selling a Book by Its Cover," Cultural Anthropology, vol. 9, no. 2, pp. 179–200, May 1994, https://doi.org/10.1525/can.1994.9.2.02a00020.

[2] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images using deep features", *arXiv:1707.04916*, 16-Jul-2017. https://doi.org/10.48550/arXiv.1707.04916

[3] Y. Yu, S. Luo, S. Liu, H. Qiao, Y. Liu, and L. Feng, "Deep attention based music genre classification," Neurocomputing, vol. 372, pp. 84–91, Jan. 2020, https://doi.org/10.1016/j.neucom.2019.09.054.

[4] A. Dorochowicz and B. Kostek, "Relationship between album cover design and music genres," *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Poland, Sep. 2019, pp. 93–98, https://doi.org/10.23919/spa.2019.8936738.

[5] G. Barney and K. Kaya, "Predicting genre from movie posters", *Machine Learning*, vol. 229, 2019. https://cs229.stanford.edu/proj2019spr/report/9.pdf

[6] K. Kundalia, Y. Patel, and M. Shah, "Multi-label Movie Genre Detection from a Movie Poster Using Knowledge Transfer Learning," *Augmented Human Research*, vol. 5, no. 1, Dec. 2019, https://doi.org/10.1007/s41133-019-0029-y.

[7] T. Behrouzi, R. Toosi, and M. A. Akhaee, "Multimodal movie genre classification using recurrent neural network," Multimedia Tools and Applications, vol. 82, no. 4, pp. 5763–5784, Jul. 2022, https://doi.org/10.1007/s11042-022-13418-6.

[8] Y. Zeng, Y. Gong, and X. Zeng, "Controllable digital restoration of ancient paintings using convolutional neural network and nearest neighbor," Pattern Recognition Letters, vol. 133, pp. 158–164, May 2020, https://doi.org/10.1016/j.patrec.2020.02.033.

[9] E. Cetinic and S. Grgic, "Genre classification of paintings," *2016 International Symposium ELMAR*, Zadar, Croatia, 2016, pp. 201–204, https://doi.org/10.1109/elmar.2016.7731786.

[10] S. Agarwal, H. Karnick, N. Pant, and U. Patel, "Genre and Style Based Painting Classification," *2015 IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, 2015, pp. 588–594, https://doi.org/10.1109/wacv.2015.84.

[11] B. K. Iwana, S. T. R. Rizvi, S. Ahmed, A. Dengel, and S. Uchida, "Judging a Book By its Cover", arXiv [cs.CV], 28-Oct-2016. https://doi.org/10.48550/arXiv.1610.09204

[12] A. Rasheed, A. I. Umar, S. H. Shirazi, Z. Khan, and M. Shahzad, "Cover-based multiple book genre recognition using an improved multimodal network," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, no. 1, pp. 65–88, Sep. 2022, https://doi.org/10.1007/s10032-022-00413-8.

[13] A. Hatamizadeh and J. Kautz, "MambaVision: A hybrid Mamba-Transformer vision backbone", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 25261-25270.

[14] A. Gu and T. Dao, 'Mamba: Linear-time sequence modeling with selective state spaces', arXiv [cs.LG], 01-Dec-2023. https://doi.org/10.48550/arXiv.2312.00752

[15] A. Hosseini, A. Kazerouni, S. Akhavan, M. Brudno, and B. Taati, "SUM: Saliency Unification Through Mamba for Visual Attention Modeling," *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1597–1607, Feb. 2025, https://doi.org/10.1109/wacv61041.2025.00163.

[16] M. M. Rahman, A. A. Tutul, A. Nath, L. Laishram, S. K. Jung, and T. Hammond, "Mamba in vision: A comprehensive survey of techniques and applications", arXiv [cs.CV], 03-Oct-2024. https://doi.org/10.48550/arXiv.2410.03105

[17] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence embeddings using Siamese BERT-networks', arXiv [cs.CL], 27-Aug-2019. https://doi.org/10.48550/arXiv.1908.10084

[18] R. Toosi, A. Hosseini, R. Toosi, and M. A. Akhaee, "Judge a Book by its Cover: A Multimodal Approach to Book Genre Prediction," *2025 11th International Conference on Web Research (ICWR)*, 2025, pp. 200–204, https://doi.org/10.1109/icwr65219.2025.11006189.

[19] P. Buczkowski, A. Sobkowicz, and M. Kozlowski, "Deep Learning Approaches towards Book Covers Classification," *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*, 2018, pp. 309–316, https://doi.org/10.5220/0006556103090316.

[20] D. Pye, "Content-based methods for the management of digital music," *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, Istanbul, Turkey, 2000, pp. 2437-2440, vol.4, https://doi.org/10.1109/icassp.2000.859334.

[21] M. Z. Afzal et al., "Deepdocclassifier: Document classification with deep Convolutional Neural Network," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015, pp. 1111–1115, https://doi.org/10.1109/icdar.2015.7333933.

[22] S. Gupta, M. Agarwal, and S. Jain, "Automated Genre Classification of Books Using Machine Learning and Natural Language Processing," *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2019, pp. 269–272, https://doi.org/10.1109/confluence.2019.8776935.

[23] J. R. Doyle and P. A. Bottomley, "Dressed for the Occasion: Font-Product Congruity in the Perception of Logotype," *Journal of Consumer Psychology*, vol. 16, no. 2, pp. 112–123, Jan. 2006, https://doi.org/10.1207/s15327663jcp1602_2.

[24] P. W. Henderson, J. L. Giese, and J. A. Cote, "Impression Management using Typeface Design," *Journal of*

Marketing, vol. 68, no. 4, pp. 60–72, Oct. 2004, https://doi.org/10.1509/jmkg.68.4.60.42736.

[25] H. Chiang, Y. Ge, and C. Wu, "Classification of book genres by cover and title", *Computer science: class report*, 2015. https://cs229.stanford.edu/proj2015/127_report.pdf

[26] G. R. Biradar, R. JM, A. Varier, and M. Sudhir, "Classification of Book Genres using Book Cover and Title," *2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT)*, Visakhapatnam, India, 2019, pp. 72–723, https://doi.org/10.1109/icisgt44072.2019.00031.

[27] D. Haraguchi, B. K. Iwana, and S. Uchida, "What Text Design Characterizes Book Genres?," *Document Analysis Systems*, pp. 165–181, 2024, https://doi.org/10.1007/978-3-031-70442-0_10.

[28] S. H. Patel and D. Aggarwal, "BGCNet: A novel deep visual textual model for book genre classification", 2020. https://doi.org/10.13140/RG.2.2.35108.09604/1

[29] R. Bielawski, "Assessing and efficiently leveraging the generalisation abilities of multimodal models", Doctoral dissertation, Université Paul Sabatier-Toulouse III.

[30] A. Dosovitskiy *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv [cs.CV], 22-Oct-2020. https://doi.org/10.48550/arXiv.2010.11929

[31] A. Paul, R. Bandyopadhyay, J. H. Yoon, Z. W. Geem, and R. Sarkar, "SinLU: Sinu-Sigmoidal Linear Unit," *Mathematics*, vol. 10, no. 3, p. 337, Jan. 2022, https://doi.org/10.3390/math10030337.

[32] S. Kjartansson and A. Ashavsky, "Can you judge a book by its cover". Stanford CS231N, 2017,https://cs231n.stanford.edu/reports/2017/pdfs/814.pdf

Reza Toosi received his B.Sc. degree in Computer Engineering from Golestan University, Iran, in 2025. During his undergraduate studies, he developed a strong foundation in computational systems and software engineering principles. His research interests encompass machine learning and deep learning methodologies, with a focus on exploring innovative approaches to artificial intelligence and data driven solutions. He is particularly interested in advancing the theoretical and practical applications of neural networks and algorithmic learning systems.

Alireza Hosseini received his B.Sc. degree in Electrical Engineering from the Iran University of Science and Technology and completed his M.Sc. in Telecommunication Systems at the University of Tehran in 2024. He is currently a Research Assistant at the Computation and Communication Lab, University of Tehran, and an AI Developer at the AVIR AI Center. His research interests include machine learning, deep learning, and signal processing, with a particular focus on saliency map prediction, large vision models, and implicit neural representations.

Ramin Toosi received his B.Sc. and M.Sc. degrees in Electrical Engineering from Shahid Beheshti University and the University of Tehran, Iran, in 2014 and 2016, respectively. He completed his Ph.D. in Telecommunication Engineering at the University of Tehran in 2024. His research interests encompass machine learning, deep learning, computational neuroscience, multimedia security, and image and video analysis. He is currently a Senior Researcher at the Brain Computing Lab, University of Tehran.

Mohammad Ali Akhaee (S07M07) received the B.Sc. degree in Electronics and Communications Eng. from the Amirkabir University of Technology, Tehran, Iran, and the M.Sc. and Ph.D. degrees from the Sharif University of Technology, Tehran, Iran, in 2005 and 2009, respectively. He is currently an Assistant Professor with the College of Engineering and the Director of the Secure Communication Laboratory, University of Tehran, Tehran, Iran. He has authored or coauthored more than 60 papers, and holds one Iranian patent. His research interests include the area of signal processing, in particular multimedia security, data hiding, and machine learning. Prof. Akhaee was the Technical Program Chair of EUSIPCO 11 and the Executive Chair of ISCISC 14 and Financial Chair of RTEST18. He received the Governmental Endeavour Research Fellowship from Australia in 2010 and the Governmental award from Ministry of Information and Communication Technology (ITC) from Iran in 2017.