Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France - August, 2025
Paper No. MVML 128
DOI: 10.11159/mvml25.128

Contrastive Self-Supervised Learning for Packaging Artwork Layer Classification with Text and Image Features

Anshul Verma[1], Pooja Bandal[2], Zohreh Hirbodvash[3], Wei-Yin Chien[4], Dhanush Dharmaretnam[5]
[1,2,3,4,5] SGS & Co
2 Dorchester Ave, Toronto, Canada
{anshul.verma; pooja.bandal; zohreh.hirbodvash; weiyin.chien; dhanush.dharmarentnam}@sgsco.com

Abstract – Creating product packaging relies heavily on digital artworks, typically in multi-layered PDFs. Each layer in these artworks represents critical elements, such as text, brand logos, nutrient panels, die-lines, dimensions, varnish areas, and graphics, essential for accurate and consistent printing. However, the absence of standardized layer organization poses significant challenges to maintaining the quality and consistency of these digital artworks. To address this, we introduce a self-supervised learning framework for automated packaging artwork layer classification. We pre-train our model using contrastive learning methods, specifically simple contrastive learning (SimCLR) and Momentum Contrast (MoCo), leveraging unlabelled data. By strategically managing augmentations during pre-training, our framework extracts discriminative features, including shape, colour, and text, mapping semantically similar features into a shared representation space. Notably, MoCo enhanced feature learning and generalization through increased negative sample diversity. Additionally, this paper proposes a multi-modal classification approach that integrates image encoder with text embeddings. Experimental results show that our multi-modal model achieves a 91% accuracy rate, outperforming traditional machine learning models by 2% average accuracy. These findings highlight the model's ability to enhance classification performance, particularly in classifying visually similar artwork layers. We demonstrate the effectiveness of multi-modal architectures incorporating text, alongside visual features, for improved classification accuracy. This pre-training approach, particularly when combined with text embeddings, significantly boosts classification accuracy to 97% for complex artworks, representing a 5-10% improvement over baseline models. Our proposed solution streamlines production workflows through faster and more accurate layer identification. Ultimately, our research offers a scalable pathway towards standardizing packaging artwork management, improving consistency, reducing errors, and enhancing overall printing process efficiency.

***Keywords*:** packaging artwork, multi-modal model, self-supervised learning

1. Introduction

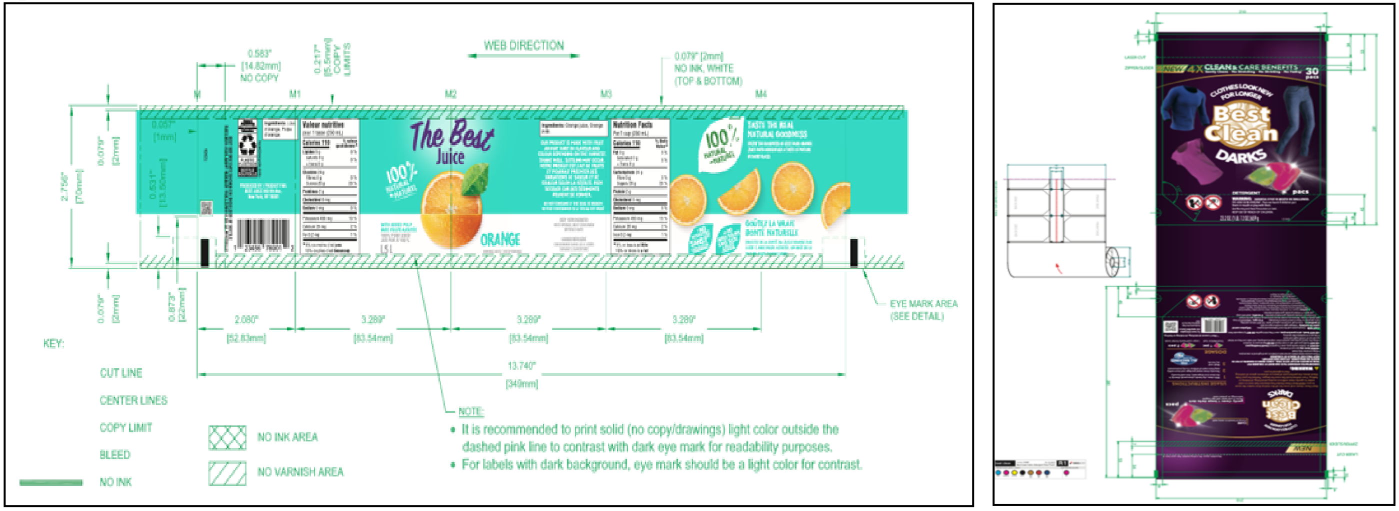
In the packaging industry, manual digital artwork quality control is a critical bottleneck, as it is both error-prone and costly. These manual processes often lead to significant material wastage, production delays, and inconsistencies in printed packaging, making them unsustainable in an increasingly fast-paced, quality-driven market [1]. Human validators, responsible for multiple stages of artwork verification, can introduce errors such as misaligned text, incorrect alignment, or inconsistent colour usage due to reliance on subjective judgment [2, 3]. As digital packaging files grow in complexity and volume, manual checks require large teams of specialized personnel, exacerbating inefficiencies and costs [2].

The need for more efficient, reliable, and cost-effective quality control methods has driven the shift towards automation in the packaging industry. Automated systems, capable of verifying and classifying artwork layers offer a solution to reduce human involvement while maintaining accuracy. These systems can detect issues such as dimensional inaccuracies, misaligned graphics, improper text placement, and colour mismatches, improving both the speed and precision of the verification process. However, despite advances in computer vision and deep learning, current models for automated artwork verification face challenges when classifying complex, overlapping, or intricate layers. These challenges stem from the limitations of vision models in accurately distinguishing fine details, such as intricate shapes and graphics, which are prone to misclassification in traditional models [4].

In recent years, data classification has become a fundamental task in machine learning, with applications in fields such as medical diagnosis, fraud detection, and image recognition. However, traditional classification models face challenges, including overfitting and class imbalance, particularly when applied to artwork layer classification. While convolutional neural networks (CNNs) and vision transformers (ViTs) have been widely explored in image classification, this study applies them in a novel context—artwork layer classification—offering insights into their effectiveness for this specialized task. Our approach aims to improve address these challenges, we propose an automated layer classification solution that enhances digital artwork verification by leveraging deep learning and self-supervised learning techniques, specifically contrastive learning methods such as SimCLR and MoCo. Furthermore, we incorporate a multi-modal architecture that includes text processing. Our approach aims to improve the accuracy and reliability of layer classification by enabling models to understand the complex interactions and dependencies between different design elements. This solution not only complements human validators but also optimizes production workflows, reduces errors, and ensures higher quality packaging prints, offering a scalable and efficient pathway for the packaging industry.

2. Background

The packaging industry has increasingly adopted automation to address the inefficiencies in manual digital artwork quality control. Computer vision techniques, particularly CNNs, have shown promise in classifying visual elements in digital artwork. CNNs, for instance, have been used in various applications to detect and classify specific artwork features such as logos, text, and images [5, 6]. However, when applied to illustrated in Figure 1, current CNN-based systems often struggle with the complexity of overlapping layers, especially in accurately classifying fine details like small text or intricate graphics [4, 7].



WEB DIRECTION
0.583" [14.82mm] NO COPY
0.217" [5.5mm] COPY LIMITS
0.079" [2mm] NO INK, WHITE (TOP & BOTTOM)
M1
M2
M3
M4
2.756" [70mm]
0.079" [2mm]
0.531" [13.50mm]
0.079" [2mm]
0.873" [22mm]
2.080" [52.83mm]
3.289" [83.54mm]
3.289" [83.54mm]
3.289" [83.54mm]
13.740" [349mm]
EYE MARK AREA (SEE DETAIL)
The Best Juice
ORANGE
KEY:
CUT LINE
CENTER LINES
COPY LIMIT
BLEED
NO INK
NO INK AREA
NO VARNISH AREA
NOTE:
• It is recommended to print solid (no copy/drawings) light color outside the dashed pink line to contrast with dark eye mark for readability purposes.
• For labels with dark background, eye mark should be a light color for contrast.
Best Clean
DARKS

(a)

(b)

Fig. 1: Dummy digital artwork images used in packaging printing industry, demonstrating various layers in packaging artwork. Flattened artwork (a) A dummy orange juice product with different overlapping artwork layers (layers include die-line, die-info, dimensions, codes, varnish), (b) A dummy detergent product's artwork, displaying all the layers in the artwork (layers include die-line, die-info, codes, varnish, artwork, shirt-tail, dimensions).

Recent advancements in ViTs offer an alternative by capturing global image dependencies, providing advantages over traditional CNNs [8, 9, 10]. However, these models also face challenges in classifying complex ambiguous layers due to intricate design overlaps [11, 12]. To address these issues, self-supervised learning techniques have been explored, reducing reliance on large labelled datasets by learning meaningful representations from unlabelled data [13]. Approaches such as SimCLR and MoCo have shown success in improving model performance in vision tasks [14, 15].

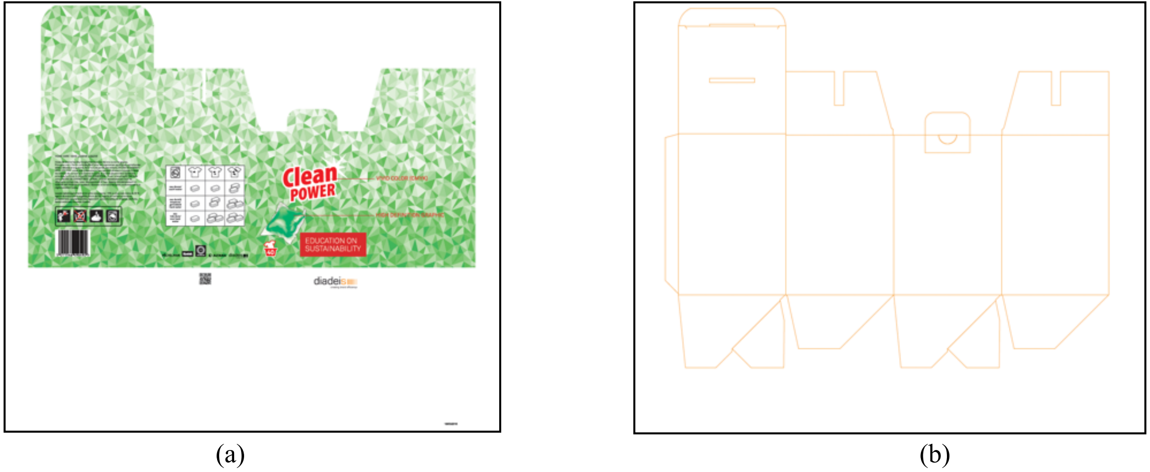
A key challenge in digital packaging is its multi-modal nature, which requires integrating both visual and texture data. Packaging artworks often contain crucial textual information, such as product descriptions, legal disclaimers, ingredient lists, and branding slogans, which are essential for accurate layer classification. Incorporating texture data alongside visual features enhances the model's ability to interpret artwork content and context. Research suggests that multi-modal learning, combining text and visual data, can significantly improve classification accuracy by capturing complex relationships between design elements [16]. This is particularly valuable in packaging artwork where textual information (e.g., product descriptions, legal text, dimensions, shirt-tail) is crucial for accurate layer classification. Furthermore, multi-modal learning can help resolve ambiguities in visual data by leveraging the semantic information provided by the text. By effectively combining textual and visual features, we can develop more robust and accurate classification models, crucial for automating quality control in the packaging industry.

3. Methodology

3.1. Dataset

Due to the lack of publicly available datasets for artwork layer classification, we created a custom dataset containing 1,361 images of multi-layered packaging artworks. This dataset consists of multi-layered packaging dummy artwork images simulating realistic packaging designs. This approach avoided the use of proprietary client data for training. The dummy artworks represent the realistic designs and standards present in the production artwork images, incorporating various layers commonly found in packaging designs, such as text layers, image layers, and varnish layers. Each layer corresponds to several distinct classes. To ensure accurate labelling, we collaborated with industry experts who provided valuable insights into the essential elements requiring categorization. The images were annotated with multiple classes, reflecting the common overlapping layers present in packaging designs.

The dataset includes the following layer classes: 1. *artwork* (background, graphics, barcode, brand logo, text, nutrition panel), 2. *shirt-tail / legend* (legend containing colour, job, and printer information), 3. *die-line* (lines indicating the artwork die printing area), 4. *die-info* (information layer detailing the die and artwork dimensions), 5. varnish (protective film layer), 6. dimensions (layers illustrating side dimensions), 7. *codes* (barcodes, QR codes) and 8. *others* (miscellaneous layers). Due to the industry's lack of standardized artwork layering practices, we approached this as a multi-label classification problem, where each image can contain multiple overlapping layers. For instance, a single image might include both artwork and die-line layers. Figure 2 illustrates sample images showcasing some layers within packaging artwork.



(a)

(b)

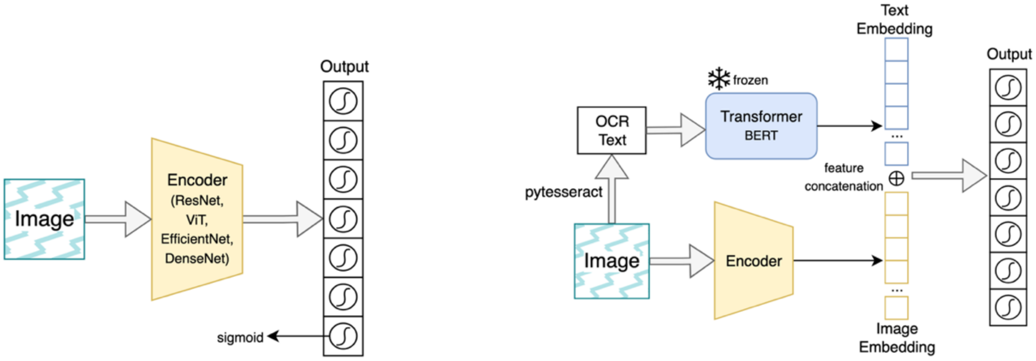
Fig. 2: A few examples of different layers withing artwork images which we aim to classify to automate digital artwork quality check. These layers contain different type of information including, (a) *artwork, codes, others*, and (b) *die-line*.

In addition to the labelled dataset, we also collected 2,000 unlabelled images. These images were used to pre-train the backbone networks (CNN and ViT) using self-supervised learning techniques, which are essential in improving the performance of models trained on limited labelled data. By leveraging this unlabelled data, we could enhance the feature learning process and help the model generalize better to unseen data.

3.2. Baseline Experiments

To establish a performance baseline, we trained our models using only the 1,361 labelled images, initializing the weights with ImageNet pre-trained weights [17, 18]. Given the multi-label nature of the task, we used binary cross-entropy as the loss function, which calculates errors independently for each class, allowing the models to classify multiple layers within an image. To enhance model robustness, we applied data augmentation techniques, including random rotations (up to 20 degrees), and a random weighted combination of the original image with a random noise matrix (noise weight: 0.0-0.1). Random noise augmentation was specifically introduced to reduce reliance on texture-based features, encouraging the model to focus on structural variation within the layers. All baseline models were trained using 3-fold cross-validation for 50 epochs, with the Adam optimizer and a fixed learning rate of 0.0001. We employed ResNet-18, ResNet-50, EfficientNet-B0, EfficientNet-B2, and DenseNet-121 as CNN backbones, and ViT-Base as the Vision Transformer backbone [8, 9, 10, 19].

Performance was evaluated using standard classification metrics: *precision*, *recall*, *F1-score*, and *accuracy*. These metrics were calculated for each layer class and then averaged to assess overall model performance. To ensure robustness and prevent dataset bias, we employed 3-fold cross-validation. Figure 3(a) illustrates the baseline training architecture.



(a) (b)

Fig. 3: Training architecture for baseline and multi-modal experiments for artwork layer classification. (a) *baseline experiments*, with different CNN and ViT encoders, and (b) *multi-modal experiments*, with all the CNN and ViT encoders along with fixed/frozen (weights not trained) text-encoder BERT encoding OCR text in an artwork image extracted using *pytesseract*.

3.3. Multi-modal Approach with Text Embeddings

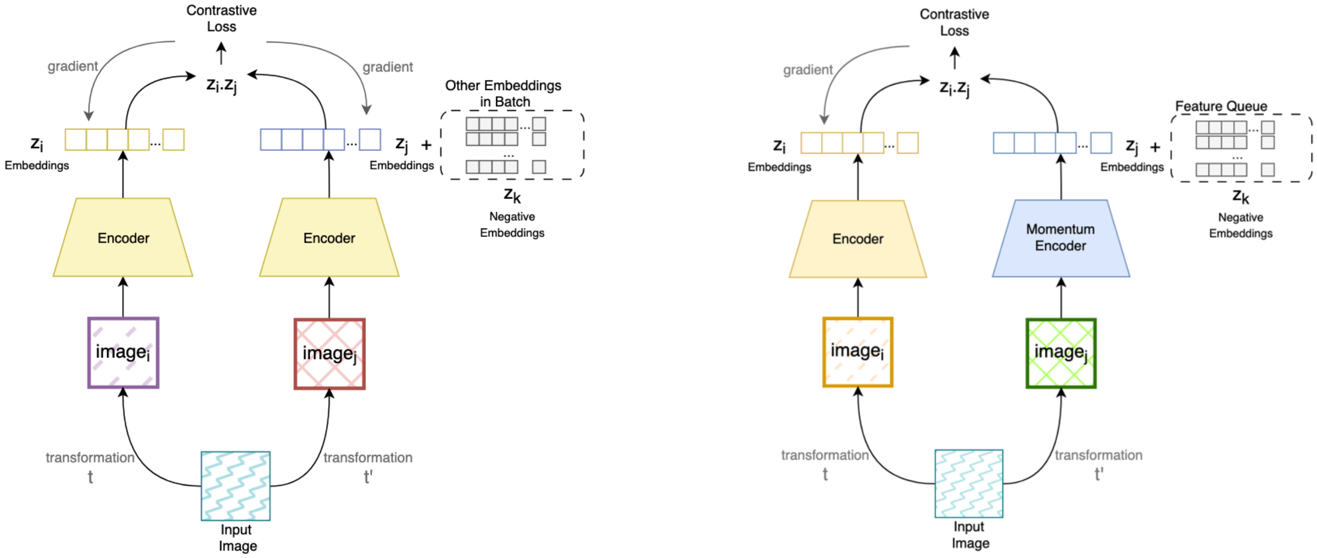
To improve baseline performance, we incorporated text information, recognizing its significance in artwork layer classification. We implemented a multi-modal model that integrates a pre-trained BERT model as a text encoder [20]. Text extracted from layer images using pytesseract [21] was passed into the BERT encoder. The BERT model's weights (pre-trained on Book Corpus and Wikipedia [22, 23]) were frozen and therefore remained unchanged during training. CNN and ViT embedding were concatenated with BERT text embeddings and then passed through a fully connected layer for artwork layer classification. The CNN and ViT weights were trained similarly to the baseline models, using 3-fold cross validation, the same augmentation, and initial ImageNet weights. We trained the model for 50 epochs using binary cross-entropy loss, with the Adam optimizer and a fixed learning rate of 0.0001. Figure 3(b) presents multi-modal architecture.

3.4. Self-supervised pre-training

To utilize the 2,000 unlabelled images, we pre-trained models using self-supervised learning techniques SimCLR and MoCo. The same CNN and ViT architectures from baseline experiments were also used for pre-training. SimCLR and MoCo use the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss, shown in Equation (1), to minimize the distance between representations of positive pairs (zi and zj). A positive pair is defined as two transformations of the same image, while a negative pair consists of an image transformation against every other image's transformation. In SimCLR, the number of negative pairs is 2N-1, where N is the batch size. In contrast, MoCo leverages a queue with its length set as a hyperparameter, to select negative pairs (queue-length: 2000), allowing larger negatives even with smaller batch sizes. A key distinction of MoCo is the use of a momentum encoder, which updates its parameters using a moving average of the query encoder's parameters at each training step. The momentum coefficient for the momentum encoder update was set to 0.9999. In Equation (1), *sim* represents the cosine similarity score $\left(\text{i.e. } sim\left(z_i, z_j\right) = \frac{(z_i . z_j)}{\| z_i \| * \| z_k \|}\right)$, where z_i, z_j are embedding vectors generated by image$_i$ and image$_j$ in Figure 4. The term $\mathbb{I}_{[k \neq i]}$ represents the indicator function, which is 1 if $k \neq i$ and 0 otherwise.

$$l_{i,j} = -\log \frac{\exp\left(\mathrm{sim}\left(z_i, z_j\right)\right)/\tau}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp\left(\mathrm{sim}\left(z_i, z_k\right)/\tau\right)} \tag{1}$$

The temperature parameter (τ) for the NT-Xent loss used in SimCLR and MoCo was set to 0.07. Figure 4 illustrates the key differences between the two pre-training approaches, highlighting momentum encoder and queue mechanism used in MoCo.



(a) (b)

Fig. 4: Self-supervised learning pre-training used to train the encoders for extracting key features from artwork images. (a) SimCLR, uses the same encoder to generate embeddings for two augmented views of the same image, comparing encodings by selecting pairs from the mini-batch, and (b) MoCo, uses a momentum encoder to encode the key image, which is drawn from the mini-batch and queue. The momentum encoder is updated using a moving average of the query encoder's parameters.

For SimCLR and MoCo, positive pairs were generated using augmentations including: rotation (up to 30 degrees), horizontal and vertical flips, colour jitter (brightness, contrast and saturation), Gaussian blurring (5x5 kernel), grayscale conversion and a random weighted combination of the original image with a random noise matrix. The noise weight was randomly selected to be between 0.0 and 0.1. Each augmentation was applied with a probability of 0.2. The Adam optimizer, with a learning rate 0.0001 was used for pre-training and the models were trained for 100 epochs with batch size of 64. This pre-training phase significantly improved the model's ability to detect and classify complex, overlapping layers within packaging artworks. Following pre-training, the pre-trained vision encoders were fine-tuned on the labelled layer classification dataset, incorporating text embeddings as shown in Figure 3(b). Only vision encoder weights were fine-tuned for multi-label artwork layer classification, using binary cross-entropy loss with the Adam optimizer and a fixed learning rate of 0.00001 for 10 epochs. As in previous experiments, models were trained using 3-fold cross validation to ensure robustness.

4. Results and Discussions

In this section, we present the results of our experiments on the multi-layer packaging artwork classification task. We evaluated the performance of baseline models, multi-modal models incorporating text, and multi-modal models leveraging pre-trained image encoders alongside text.

4.1. Baseline Performance

Table 1 summarizes the baseline performance of selected models on multi-label artwork layer classification. All models were fine-tuned from ImageNet pre-trained weights, and the reported metrics are averaged across all classes and folds using 3-fold cross-validation.

Table 1: Baseline model performance for multi-label layer classification, presenting average metrics across all classes and folds using 3-fold cross-validation.

Model	Precision	Recall	F1-score	Accuracy
ResNet-18	0.8684	0.8774	0.8729	0.8775
ResNet-50	0.8937	0.8663	0.8798	0.8957
EfficientNet-B0	0.8501	0.7939	0.8201	0.8710
EfficientNet-B2	0.8568	0.7856	0.8197	0.8626
DenseNet-121	0.8694	0.8605	0.8648	0.8777
ViT-Base 16-bit	0.8843	0.8625	0.8733	0.8893

ViT-Base 16-bit achieved the highest precision and accuracy, with an accuracy of 88.93% for artwork layer classification.

4.2. Text-embedding concatenation Performance

Table 2 presents the performance of multi-modal models on the layer classification dataset, incorporating text embeddings. All metrics are averaged across all classes and folds using 3-fold cross-validation.

Table 2: Performance of models for multi-label layer classification using text embeddings, presenting average metrics across all classes and folds using 3-fold cross-validation. [↑] and [↓] indicate improved and decreased performance, respectively, compared to the baseline experiment in Table 1.

Model	Precision	Recall	F1-score	Accuracy
ResNet-18	0.8725 [↑]	0.8974 [↑]	0.8828 [↑]	0.9021 [↑]
ResNet-50	0.8971 [↑]	0.8776 [↑]	0.8892 [↑]	0.9144 [↑]
EfficientNet-B0	0.8691 [↑]	0.8320 [↑]	0.8481 [↑]	0.8876 [↑]
EfficientNet-B2	0.8765 [↑]	0.8107 [↑]	0.8343 [↑]	0.8791 [↑]
DenseNet-121	0.8751 [↑]	0.8609 [↑]	0.8679 [↑]	0.8819 [↑]
ViT-Base 16-bit	0.8691 [↓]	0.8894 [↑]	0.8779 [↑]	0.8942 [↑]

The results indicate that combining text embeddings with image embeddings generally improved model performance for artwork layer classification, notably for ResNet-50, which achieved the highest precision, F1-score, and accuracy. While the multi-modal approach enhanced the differentiation of layers such as shirt-tail, artwork, die-line, die-info, and dimensions, as expected. However, ViT-Base 16-bit showed minimal improvements with text embedding, possibly due to increased model complexity and limited training data.

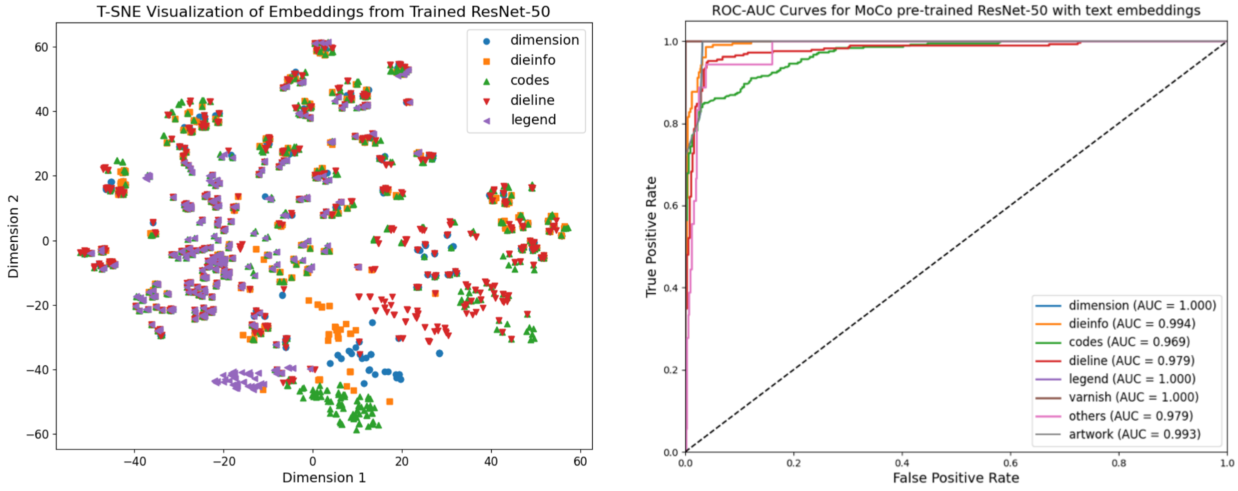
4.3. Pre-training Performance

Table 3 presents the performance of multi-modal models for artwork layer classification, utilizing text embeddings and pre-trained image encoders (SimCLR and MoCo). All metrics are averaged across classes and folds using 3-fold cross-validation. The results demonstrate the significant impact of image encoder pre-training. MoCo consistently outperformed

SimCLR, likely due to its smaller batch size during self-supervised learning. Notably, ResNet-50 pre-trained with MoCo achieved the highest recall, F1-score, and accuracy.

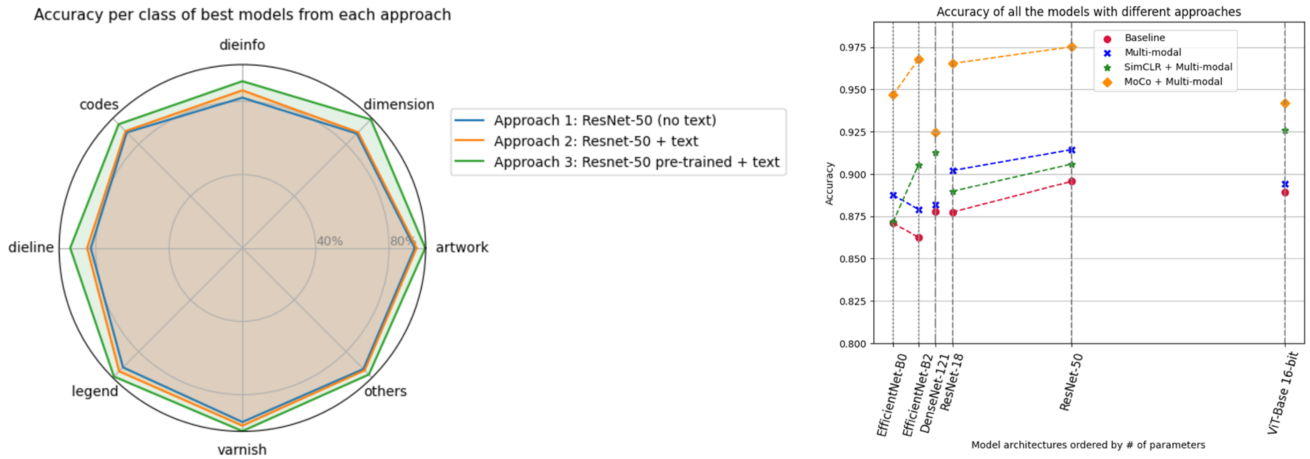
Table 3: Performance of multi-modal models utilizing text embeddings and pre-trained image encoders for multi-label layer classification. Image encoders were pre-trained with SimCLR and MoCo. Metrics are averaged across all classes and folds using 3-fold cross-validation. [↑] and [↓] indicate improved and decreased performance, respectively, compared to the baseline experiment in Table 1 and **Bold** shows the model with best performance across all the experiments.

Model	Pre-training	Precision	Recall	F1-score	Accuracy
ResNet-18	SimCLR	0.8869 [↑]	0.8722 [↓]	0.8758 [↑]	0.8898 [↑]
	MoCo	0.9429 [↑]	0.9565 [↑]	0.9369 [↑]	0.9653 [↑]
ResNet-50	SimCLR	0.8819 [↓]	0.8905 [↑]	0.8766 [↑]	0.9059 [↑]
	MoCo	0.9629 [↑]	**0.9673** [↑]	**0.9647** [↑]	**0.9752** [↑]
EfficientNet-B0	SimCLR	0.8627 [↑]	0.8379 [↑]	0.8531 [↑]	0.8715 [↑]
	MoCo	0.9677 [↑]	0.9278 [↑]	0.9373 [↑]	0.9467 [↑]
EfficientNet-B2	SimCLR	0.8784 [↑]	0.8794 [↑]	0.8778 [↑]	0.9055 [↑]
	MoCo	0.9683 [↑]	0.9545 [↑]	0.9514 [↑]	0.9678 [↑]
DenseNet-121	SimCLR	0.8858 [↑]	0.8936 [↑]	0.8876 [↑]	0.9127 [↑]
	MoCo	0.9385 [↑]	0.9038 [↑]	0.9182 [↑]	0.9245 [↑]
ViT-Base 16-bit	SimCLR	0.8926 [↑]	0.8895 [↑]	0.8923 [↑]	0.9259 [↑]
	MoCo	**0.9764** [↑]	0.9287 [↑]	0.9576 [↑]	0.9417 [↑]



(a) (b)

Fig 5: (a) T-SNE projection of ResNet-50 embeddings for top-5 occurring classes in the labelled dataset, pretrained via MoCo self-supervised learning, showing the model's ability to distinguish between artwork layer classes without fine-tuning. (b) ROC-AUC curve of each class with MoCo pre-trained ResNet-50 model with text embedding for layer classification.



(a) (b)

Fig 6: (a) Average accuracy of different class of the best model using different approach from Table 1, Table 2 and Table 3, and (b) accuracy comparison of various models trained using different methods ordered by number of parameters.

Figure 5(a) illustrates the T-SNE projection of ResNet-50 embeddings (MoCo pre-trained, without fine-tuning) for the five most frequent classes, demonstrating the model's ability to distinguish classes based solely on pre-trained features. Figure 5(b) displays the ROC-AUC curves for each class using the best model (MoCo pre-trained ResNet-50 with text embeddings). Figure 6(a) shows the average accuracy per class for the best-performing models across different approaches, and Figure 6(b) compares the overall accuracy of all models, confirming that ResNet-50 achieves the best balance between accuracy and parameter count.

5. Conclusion

In this study, we addressed the challenging task of multi-layer packaging artwork classification, a crucial step in automated packaging design analysis. We explored a range of deep learning architectures, demonstrating that baseline models, fine-tuned on ImageNet pre-trained weights, achieved promising results. We demonstrated the effectiveness of self-supervised pre-training using SimCLR and MoCo, where incorporating random noise in data augmentation and during positive pair generation enabled the model to robustly learn features for artwork layer classification. Notably, ResNet-50, pre-trained with MoCo and fine-tuned with concatenated text embeddings on text from layer image, achieved the highest overall performance, showcasing the power of combining self-supervised learning with multi-modal data. The combination of multi-modal data and self-supervised learning, allowed the model to leverage both visual and textual information, leading to better feature extraction and classification performance.

Our study demonstrates the significant value of multi-modal learning, self-supervised pre-training, and random noise incorporation for complex visual classification tasks, particularly within niche datasets like multi-layer packaging artwork. By integrating image and textual data, we achieved improved layer classification accuracy, paving the way for more efficient and automated packaging design analysis. This approach offers a design automation system solution that reduces dependency on large labelled datasets, benefiting real-world scenarios by improving efficiency and lowering human error in industries requiring complex visual pattern identification. Furthermore, the successful integration of self-supervised pre-trained visual embeddings with textual embeddings opens new avenues for tackling complex visual problems. Future research should focus on validating the model's robustness with larger and more diverse datasets, exploring additional data modalities like layer names and cross-layer context, and investigating more sophisticated text embedding models. These advancements not only enhance packaging artwork analysis but also have the potential to improve automation in other design-oriented machine learning tasks.

Acknowledgements

This research was made possible through the collaborative efforts of the authors and SGS&Co. We extend our sincere appreciation to SGS&Co for their provision of essential resources and data, as well as their ongoing support throughout the project. All contributors to this study were employees of SGS&Co, and the research was conducted using company-provided computational infrastructure.

References

[1] L. D. Vemuri and N. V. Gupta, “A review on electronic art work management,” *International Journal of Pharmacy and Pharmaceutical Sciences,* vol. 5, no. 3, 2013.

[2] G. P. Ball, R. Shah and K. D. Wowak, “Product competition, managerial discretion, and manufacturing recalls in the US pharmaceutical industry,” *Journal of Operations Management,* vol. 58, pp. 59--72, 2018.

[3] S. Kumar, “A knowledge based reliability engineering approach to manage product safety and recalls,” *Expert Systems with Applications,* vol. 41, no. 11, pp. 5323--5339, 2014.

[4] D. Heinke, P. Wachman, W. van Zoest and E. C. Leek, “A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision,” *Vision Research,* vol. 189, pp. 81--92, 2021.

[5] D. M. Montserrat, Q. Lin, J. Allebach and E. Delp, “Scalable logo detection and recognition with minimal labeling,” *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR),* pp. 152--157, 2018.

[6] S. C. Hoi, X. Wu, H. Liu, Y. Wu, H. Wang, H. Xue and Q. Wu, “Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks,” *arXiv preprint arXiv:1511.02462,* 2015.

[7] K. Hermann, T. Chen and S. Kornblith, “The origins and prevalence of texture bias in convolutional neural networks,” *Advances in Neural Information Processing Systems,* vol. 33, pp. 19000--19015, 2020.

[8] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, “Densely connected convolutional networks,” *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 4700--4708, 2017.

[9] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 770--778, 2019=8.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and others, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929,* 2020.

[11] D. Heinke, P. Wachman, W. van Zoest and E. C. Leek, “A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision,” *Vision Research,* vol. 189, pp. 81--92, 2021.

[12] Q. Hou, R. Xia, J. Zhang, Y. Feng, Z. Zhan and X. Wang, “Learning visual overlapping image pairs for SfM via CNN fine-tuning with photogrammetric geometry information,” *International Journal of Applied Earth Observation and Geoinformation,* vol. 116, p. 103162, 2023.

[13] M. Benčević, M. Habijan, I. Galić and A. Pizurica, “Self-supervised learning as a means to reduce the need for labeled data in medical image analysis,” *2022 30th European Signal Processing Conference (EUSIPCO),* pp. 1328--1332, 2022.

[14] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, “A simple framework for contrastive learning of visual representations,” *International conference on machine learning,* pp. 1597--1607, 2020.

[15] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 9729--9738, 2020.

[16] A. Abdelhamed, M. Afifi and A. Go, “What do you see? enhancing zero-shot image classification with multimodal large language models,” *arXiv preprint arXiv:2405.15668,* 2024.

[17] A. Paszke, “Pytorch: An imperative style, high-performance deep learning library,” *arXiv preprint arXiv:1912.01703,* 2019.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” *2009 IEEE conference on computer vision and pattern recognition,* pp. 248--255, 2009.

[19] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *International conference on machine learning,* no. PMLR, pp. 6105--6114, 2019.

[20] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),* pp. 4171--4186, 2019.

[21] R. Smith, “An overview of the Tesseract OCR engine,” *Ninth international conference on document analysis and recognition (ICDAR 2007),* vol. 2, pp. 629--633, 2007.

[22] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),* pp. 4171--4186, 2019.

[23] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” *Proceedings of the IEEE international conference on computer vision,* pp. 19--27, 2015.

[24] M. Hasan, “Vision Eagle Attention: a new lens for advancing image classification,” *arXiv preprint arXiv:2411.10564,* 2024.

[25] N. Kapila, J. Glattki and T. Rathi, “CNNtention: Can CNNs do better with Attention?,” *arXiv preprint arXiv:2412.11657,* 2024.