DiCLS: Exploring Cross-Modal Alignment through Leaf Disease Classification

JinZhou Tang
School of Intelligent Systems Engineering, Sun Yat-sen University
ShenZhen, China
tangjzh23@mail2.sysu.edu.cn

Abstract

Methods for bridging the gap between image and language through cross-modal training have emerged as a significant area of research in recent years. In this paper, we present DiCLS (Disease Classification), a novel approach that aims to dive deeper into the dynamic mechanisms underlying this semantic-rich cross-modal alignment. We argue that multilabel classification can be considered a subset of the phase grounding task and draw inspiration from GLIP to inform the architecture of our model. Additionally, we introduce experimental loss functions and model structures to evaluate their efficacy on a disease leaf classification dataset. Though the research findings indicate that the effects of these attempts are quite limited, we are still able to gain deeper insights in this field. Code is released at https://github.com/ThreebodyDarkforest/DiCLS.

1. Introduction

In the realm of image classification, convolutional and transformer models have emerged as the prevailing approaches since the inception of AlexNet. These models have demonstrated remarkable prowess in various image-related tasks.

More recently, some methods, such as CLIP [15], have introduced novel loss functions that enable the contrasive learning of image and text representations. This breakthrough has not only facilitated cross-modal alignment but has also unlocked new avenues for leveraging the combined power of visual and textual information.

1.1. Motivation

As more complex tasks have emerged, the integration of control conditions into models has become necessary. In this context, the use of natural language as a control condition has gained significant attention. Early models [19] often adopted more formal and abstract methodologies for constructing control conditions. However, with the advancements in language models and transformer-based architectures [22], researchers have started focusing on utilizing natural language as an input control condition.

It is important to note that while simple tasks like image classification typically do not necessitate the incorporation of textual infomation, as discriminative models are predominantly employed for such purposes, the potential performance enhancement that may arise from the implementation of textual information remains an intriguing area that needs further research.

1.2. Insights

Training method and data quality impact cross-modal alignment significantly. Different training methodologies facilitate alignment at various semantic levels. For instance, CLIP/GLIP maximizes similarity between image-text pairs, aligning them in a shared latent space. Flamingo/MiniGPT [1, 26] projects image features into the language backbone for image semantics comprehension. SimVLM/VirTex [4, 24] use joint visual-textual input in transformers to learn fused representations. However, these approaches require extensive pre-training on high-quality datasets for effective results.

Cross-modal alignment enables learning of semantic correspondence between image and text. In the context of supervised learning using a substantial dataset of image-text pairs, the model inherently captures the underlying correspondence between images and text. This is evidenced by the model's ability to predict high degrees of similarity or assign greater attention weights to image-text pairs that share similar semantic content.

Relative to the training methodology, the model architecture is less critical. The achievement of cross-modal alignment often requires a vast amount of training data for pre-training [15]. In this regard, the adoption of advanced and resource-efficient feature extractors as backbones tends to yield more notable outcomes compared to relying on conventional tricks employed in other domains. Moreover, the design of efficient fine-tuning methods specific to different

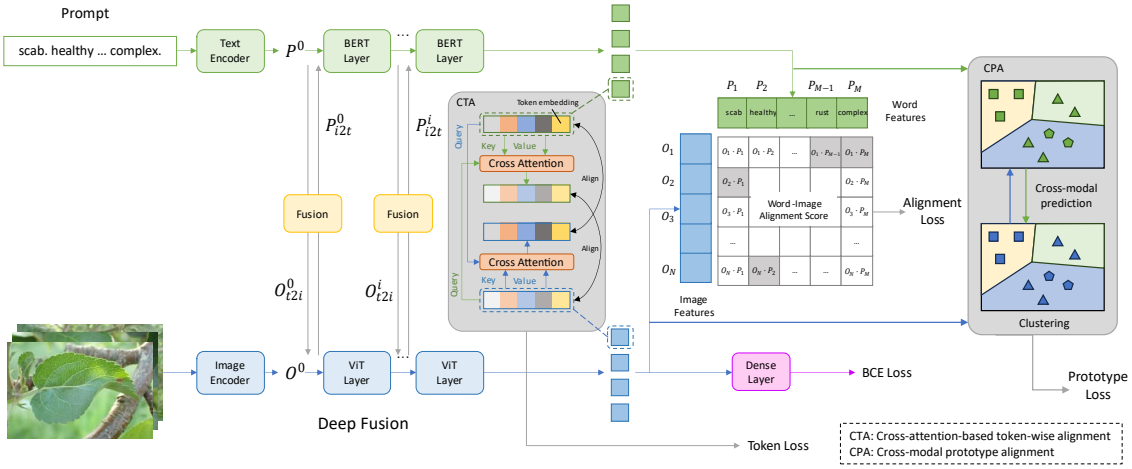


Figure 1. **DiCLS architecture overview.** Drawing inspiration from GLIP [11] and MGCA [23], we jointly train the visual and language backbones in DiCLS to predict accurate image-label pairs. Additionally, we incorporate cross-modal fusion at different stages to learn language-aware image features. Specifically, the cross-modal attention deep fusion integrates textual semantic embeddings into image features before alignment. The Cross-attention-based Token Alignment (CTA) implicitly models the alignment between different image patches and textual word tokens. Moreover, the Cross-modal Prototype Alignment (CPA) aligns image and text features at a more abstract semantic level.

downstream tasks becomes particularly crucial in optimizing the overall performance of the model.

1.3. Contributions

In this paper, we propose DiCLS, as shown in Fig. 1, to strike a further exploration and understanding of the dynamic mechanisms behind this cross-modal alignment. By leveraging a model architecture inspired by GLIP and a set of experimental loss functions, we conducted thorough testing and analysis of our model across multiple evaluation metrics. When we pre-train Swin Transformer on ImageNet [2] and BERT [5] on BookCorpus and internet data, the model achieves 97.2% Accuracy and 96.7% Precision on Leaf Disease dataset.

- We propose that multilabel classification is a subset of the phase grounding.
- We evaluated the impact of different backbones on cross-modal alignment.
- We employed multiple experimental loss functions in our study.

2. Related Works

Traditional image classification systems, such as ResNet, VGG, and EfficientNet [8, 18, 20], are trained to map image features to class labels. However, these systems often overlook the semantic information in textual labels. In contrast, DiCLS transforms multilabel classification into the phase grounding task without localization, enabling a language context-aware contrastive learning framework that effectively utilizes the semantic information present in the labels.

Contrastive learning In recent years, the fusion of visual and language modalities has become the mainstream approach in image recognition. Models like ConVIRT [25] and CLIP employ cross-modal contrastive learning on a large number of image-text pairs to achieve open-vocabulary image classification. The objective of contrastive learning is to learn an embedding space wherein positive instances are close to each other, while negative pairs are far apart. Building upon this training paradigm, our DiCLS model learns rich semantic and language-aware features from image-text pairs, yielding excellent performance on a leaf disease classification dataset.

Cross-modal attention Another common approach to integrate images and text is through cross-modal attention, a

variant of cross-attention [22]. Recent studies [9, 10, 13, 14] have demonstrated the effectiveness of attention mechanisms in improving system performance. Cross-modal attention is typically represented as:

$$\text{softmax}(QK^T/\sqrt{d}) \cdot V \tag{1}$$

where $Q = W_Q X, K = W_K Y$, and $V = W_V Y$. Here, W_Q, W_K, and $W_V \in R^{(d,n)}$ are learnable parameters, while $X \in \mathbb{R}^{n,c}$ and $Y \in \mathbb{R}^{n,c}$ represent the input features. This structure allows the system to capture the strengths and relationships between different channel features in X and Y. Leveraging this characteristic, our DiCLS model utilizes cross-modal attention for effective fusion of multi-modal information.

Vision & Language transformers Transformer models have gained significant attention in recent years due to their remarkable scalability, modality compatibility, and parallelizability, making them a focal point in various research domains [5, 6, 13]. Transformers designed for different modalities, such as language and vision, often exhibit consistent model structures. The key differentiating factor lies in the preprocessing of different modalities to conform to the input format of Transformers. For instance, visual Transformers like ViT and VAN [6, 7] partition images into patches and encode them as tokens before incorporating them into the Transformer structure. Similarly, language Transformers like BERT and GPT [5, 16] encode words as tokens before integrating them into the Transformer architecture. DiCLS leverages both language and vision transformers as backbones to harness the advantages of both transformer models.

3. Approach

Both multi-label classification and phase grounding aim to learn the correspondence between image features and labels. However, they differ in two aspects: multi-label classification neither require localization prediction boxes, nor involve identifying phases in text. In Sec. 3.1, we propose a method to convert multi-label classification into phase grounding. Sec. 3.2 and Sec. 3.3 detail the feature extraction from text and images, and the alignment of image and text information, respectively.

3.1. Transform task

Traditional image classification systems typically employ an image encoder, denoted as Enc^v, to extract image features, represented as $O \in \mathbb{R}^{N,d}$, from input image batch I which is similar to the process of Fig. 1(Bottom-left). These features are then transformed using a linear layer with weight matrix $W \in \mathbb{R}^{c,d}$, yielding a set of probability values $p^v \in \mathbb{R}^{N,c}$, through the application of the softmax(or sigmoid) function. The workflow can be succinctly described as follows:

$$O = Enc^v(I),\ p^v = \text{softmax}(OW^T) \tag{2}$$

$$L(p^v; Y) = -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{c} y_{jk}\log(p^v_{jk}) \tag{3}$$

Here, Y denoted as one-hot ground truth vector $(y_{j1}, y_{j2}, \cdots, y_{jc}),\ \forall k,\ y_{jk} \in \{0, 1\}$. While p^i refers to the predicted probability vector $(p^v_{j1}, p^v_{j2}, \cdots, p^v_{jc})$, where $\forall k,\ p^i_{jk} \in [0, 1]$.

It is worth noting that Enc^i can be any image encoder, such as ViT, ResNet, or VGG .etc. Furthermore, when considering multi-label classification, the calculation method remains identical to the one described above.

However, this approach does not utilize the textual semantics of the labels, but rather treats them as distinctive markers for different categories. To address this limitation, we concatenate the labels in a textual form, creating a simple prompt similar to the following

Prompt = "Detect: scab, healthy, ..., rust."

to convert multi-label classification to phase grounding task, which involves grounding the phases mentioned in the prompt within the images through an implicit localization process. Further, improved prompts can be created by providing more detailed class descriptions and leveraging pre-trained language models' preferences.

To fully leverage the semantic information in the text, we introduce a language encoder, Enc^v, to extract context-aware word-wise textual features $P \in \mathbb{R}^{n,d}$. These features are then used to compute alignment scores $S_{align} \in \mathbb{R}^{N,n}$ between words and images as follows

$$O = Enc^v(I),\ P = Enc^l(T) \tag{4}$$

$$S_{align} = \text{sigmoid}(OP^T) \tag{5}$$

Here, N denotes for batch-size, while n represents the number of tokens in the input sequence T. Hence, the similarity scores, denoted as S_{align}, are computed between each word (or token) in the prompt and a batch of images' global feature.

Next, the ground truth for this batch can be straightforwardly obtained by converting the labels into token masks by expanding it from $\{0, 1\}^{N,c}$ to $\{0, 1\}^{N,n}$. Specifically, for each image-label pair, it suffices to mark the corresponding tokens that matches the label in the sequence T obtained through prompt tokenizer, as the mask $M \in \{0, 1\}^{N,n}$ for it. And then we can apply cross entropy to compute loss the same as Eq. (3)

$$L_{ALN} = \text{CrossEntropy}(S_{align}; M) \tag{6}$$

During inference, we average token probabilities as the phrase probability. Then, we keep the formulation the same as the multi-label classification.

Notably, within the same model architecture, we have implemented and compared the performance of the two aforementioned training paradigms, by separately evaluate p^v and S_{align} as output logits, which will be discretely describe in Sec. 4. In Sec. 4.2, we present interesting findings derived from this comparison. Specifically, we conducted tests using the BCE Loss and Alignment Loss as the primary loss functions, as illustrated in Fig. 1.

3.2. Feature extraction

Feature extraction plays a critical role in the overall model architecture as it determines the upper limit of the model's performance. This is because the richness and expressiveness of the features extracted from images or text by the image/language encoder will determine the ease with which the classifier can distinguish different image characteristics.

In the specific case of DiCLS, the adoption of DeiT, ResNet, and Swin Transformer as vision backbones, while BERT and GPT2 as language backbones, aims to explore the potential of cross-modal alignment across diverse architectural paradigms.

3.3. Cross-modal alignment

As the focal point of this paper, the modules employed for cross-modal alignment are designed as interchangeable structures to facilitate a comprehensive analysis of their functionality. These modules primarily include the Word-Image Alignment Loss, which aligns images with label text; the Deep Fusion, which merges textual and semantic language features; the Cross-attention-based Token Alignment, which aligns token features; and the Cross-modal Prototype Alignment, which aligns image and text features in a higher-level semantic space.

Language-aware Deep Fusion Inspired by Flamingo, we designed a Bidirectional Cross-Modal Attention module XMA-B as shown in Fig. 2. It takes the features of both images and text as inputs and produces a fused feature as output. Formally, when we use BERT as language encoder and DeiT as image encoder, the fusion module can be represented as follows

$$O^i_{t2i}, P^i_{i2t} = \text{XMA-B}(O^i, P^i),\ i \in \{0, 1, \cdots, L-1\} \quad (7)$$

$$O^{i+1} = \text{ViTLayer}(O^i + O^i_{t2i}) \quad (8)$$

$$P^{i+1} = \text{BERTLayer}(P^i + P^i_{i2t}) \quad (9)$$

where L represents the number of ViTLayers present in the deep fusion module. BERTLayer refers to the additional BERT Layers incorporated on top of the pre-trained BERT. O^0 denotes the visual features obtained from the vision backbone, while P^0 represents the token features derived from the language backbone (BERT). The cross-modality communication is established through the cross-modality attention module (XMA-B), as described in Eq. (7). Subsequently, the process involves the fusion of single modality and is updated according to Eq. (8) and Eq. (9).

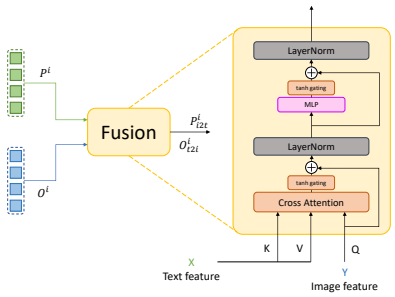


Figure 2. **Architecture of the fusion module XMA-B.** Employing a Transformer-like architecture, the design integrates image and text features through cross-modal attention, facilitating the fusion of information across different modalities.

Specifically, for the cross-attention layer, we employ Eq. (1) to compute the attention weights. Furthermore, we incorporate a set of learnable tanh gate matrices to introduce nonlinearity to the module, enabling the ability to capture more semantic-rich features.

Cross-attention-based Token Alignment Transformer-based models are typically designed to process tokens, representing image patches or text words. This facilitates token-wise alignment, as specific text segments often show stronger associations with specific visual patterns, and vice versa. However, these associations are not explicitly captured in the ground truth data, necessitating the use of unsupervised training methods.

Inspired by MGCA, Instead of directly computing the cosine similarity of different tokens, we propose to calculate the soft matching between generated visual and text tokens with the cross-attention mechanism. Formally, $O^L = \{o^L_1, o^L_2, \cdots, o^L_n\}$, $P^L = \{p^L_1, p^L_2, \cdots, p^L_3\}$, where $\forall i, o^L_i \in \mathbb{R}^d, p^L_i \in \mathbb{R}^d$. By attend o^L_i to each $p^L_j, j \in [1, n]$, we assume its corresponded cross-modal text embedding z^L_i

$$z^L_{i,j} = \sum_{j=1}^{n} O(\text{X-Attn}(o^L_i; p^L_j)) \quad (10)$$

where X-Attn(X, Y) remains the same as Eq. (1). After that, we adopt a Local Image-to-text Alignment (LIA) loss L_{LIA} to pull p^L_i close to its cross-modal text embedding z^L_j but push away from other cross-modal text embeddings as follows

Model	Backbone	Accuracy(%)	Precision(%)	Avg Precision(%)
ViT-B/16	-	96.2	91.4	96.4
ViT-B/32	-	95.5	92.1	94.5
ResNet50	-	96.1	92.4	96.3
EfficientNetV2	-	96.1	90.1	95.1
VGG16	-	95.5	91.7	94.3
SwinV2-S	-	97.0	**93.1**	**96.7**
Swin-S	-	96.5	92.7	96.3
DiCLS-DeiT(ours)	DeiT-B/16	96.5	92.3	96.4
DiCLS-Swin(ours)	Swin Transformer-S	96.9	92.5	**96.7**
DiCLS-SwinV2(ours)	Swin Transformer V2-S	**97.2**	92.8	**96.7**
DiCLS-Res(ours)	ResNet50	96.2	92.2	96.1

Table 1. **Comparative Experiment Results** In the experiments, we selected Accuracy and Precision as the primary metrics, and the results indicated that the efforts of DiCLS in cross-modal alignment yielded little improvements. We believe that the main performance of the model still relies on the visual backbones.

$$L_{ITA} = -\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{i=1}^{n}\omega_i^j(L_1 + L_2) \tag{11}$$

$$L_1 = \text{infoNCE}(o^L, z^L),\ L_2 = \text{infoNCE}(z^L, o^L) \tag{12}$$

With the same manner and construct a Local Text-to-image alignment (LTA) loss L_{ITA}, The final objective of our CTA module is the combination of LIA and LTA losses:

$$L_{CTA} = \frac{1}{2}(L_{LIA} + L_{LTA}) \tag{13}$$

Cross-modal Prototype Alignment To achieve global semantic alignment between the entire image and complete sentence at a holistic level, we employ the Cross-modal Prototype Alignment (CPA) module from MGCA. This module harnesses the cross-modal inter-subject correspondences between images and text, allowing for effective alignment between the two modalities.

This module uses the Sinkhorn-Knopp algorithm to cluster the global image feature o_0^L and the sentence context feature p_0^L, resulting in q_o and $q_p \in \mathbb{R}^K$. Additionally, a predefined trainable prototype weight matrix $W_P = \{c_0, c_1, \cdots, c_K\}$ is established, where $c_k \in \mathbb{R}^d$ represents a prototype. The similarity between $o_0^L - c$ and $p_0^L - c$ is calculated using the following formula:

$$h_{o,k} = \text{softmax}(o_0^L W_P^T)^T,\ h_{p,k} = \text{softmax}(p_0^L W_P^T)^T \tag{14}$$

Next, the cross-entropy loss is computed between the clustering results and this similarity:

$$L(o_0^L, W_P) = \sum_{k=1}^{K} q_{o,k}\log h_{o,k} \tag{15}$$

$$L(p_0^L, W_P) = \sum_{k=1}^{K} q_{p,k}\log h_{p,k} \tag{16}$$

Finally, the overall CPA loss is the average of two prediction losses:

$$L_{CPA} = \frac{1}{2}(L(o_0^L, W_P) + L(p_0^L, W_P)) \tag{17}$$

4. Experiments

We conducted an evaluation of the performance of our model and other baseline models on the provided dataset for classifying diseased leaves. This dataset comprises 3k training images, 600 validation images, and 600 testing images. We typically converted this 12-class classification into a 6-class multi-label classification task. After training, the evaluation results were obtrained by assessing the models on the testing set through p^v or S_{align}. In the following sections, we delve into the dynamics of DiCLS by undertaking a comparative analysis of different backbones in Sec. 4.1, conducting rigorous ablation experiments in Sec. 4.2, and exploring the limits of our model in Sec. 4.3.

4.1. Fine-tuning backbones

We fine-tuned the pre-trained weights of different image classification base models, such as ViT, Swin, ResNet etc., as accuracy baselines and further evaluated the classification results when using these models as the backbone for DiCLS. The results, as shown in Tab. 1, indicate that the proposed model improves the overall performance of classification. However, no significant improvement was observed on the best backbone.

Setup In the experiments, the loss function for all base models pre-trained on ImageNet was set to cross-entropy

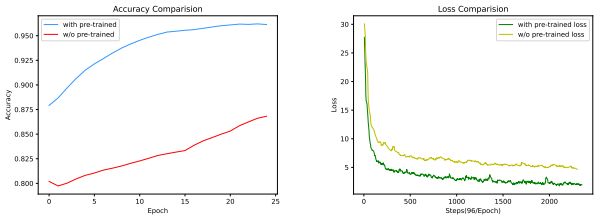
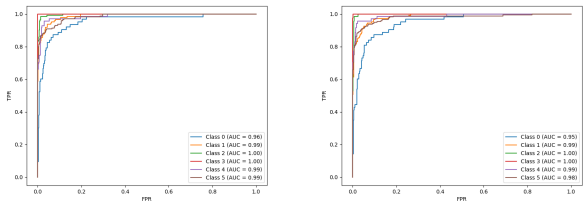


Figure 3. **Accuracy/Loss comparision between pre-trained and not pretrained** The performance of fine-tuning on pre-trained backbones significantly outperforms training from scratch. When backbones are not pre-trained, it usually requires more than 50 epochs to converge, and even then, the optimal performance still falls short of the fine-tuning scenario.



(a) Ours (b) Swint

Figure 4. **ROC curves** The fine-tuning of DiCLS (left) and plain Swin Transformer (right) yields similar ROC curves, indicating that Class 0 (complex) is a key category that influences performance.

loss, with a learning rate of 0.001. To achieve higher optimization performance, the same data augmentation methods were applied, including random rotation, random flipping, Mixup, and label smoothing. Additionally, Adam optimizer was used, along with the CosineAnnealingLR dynamic learning rate strategy. For DiCLS, a more complex strategy was employed in which is almost the same as above. To ensure fair comparison of experimental results, the performance of all models was evaluated on the test set after 10 epochs on one piece of 2080Ti GPU. It is worth noting that for DiCLS, we use p^v to evaluate benchmark metrics only.

Fine-tune on our dataset On the given dataset for classifying diseased leaves, we explored various options including frozen backbones, frozen stages of visual backbone layers, and full parameter fine-tuning. Ultimately, we opted for the frozen early visual backbone layers approach, which yielded the best performance. Through this method, we were able to achieve over 95% accuracy and 94% average precision across all base models. The results indicate that we outperformed the best baseline model(CLIP) in terms of accuracy and average precision. However, this improvement is limited, particularly in the aspect of cross-modal alignment, where it did not effectively contribute. We will provide further details on this matter in Sec. 4.2.

In summary, transferring knowledge from pre-trained weights through fine-tuning tends to be a better choice. Particularly in this case, we chose DeiT, Swin Transformer and ResNet. In such case, we found that models often struggle to converge within a predictable timeframe when pre-trained weights were not used, as shown in Fig. 3. This is because pre-trained models have already learned prior knowledge on large datasets, which enables them to converge to a position that aligns better with the data distribution in the physical world. Fine-tuning with a small amount of data further improves their performance on specific tasks.

4.2. Ablation studies

Backbones In the experiments, we standardized the deep fusion layers to two and employed the identical structure. Furthermore, we utilized the same loss configuration as described in the article, which will be specifically outlined in Sec. 4.2. Additionally, we incorporated the Arcface Loss to enhance the results of fine-grained classification, while kept the input text prompt the same.

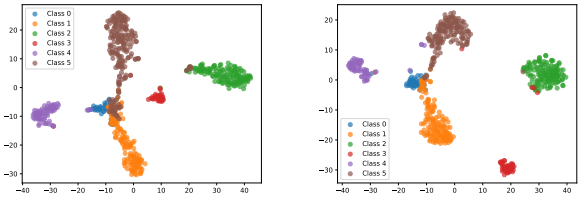
When testing DiCLS on different visual and language backbones with the same pre-training and fine-tuning settings, the model exhibited slightly varying performance, as shown in Tab. 2. The results indicate that using Swin Transformer V2 as the backbone yielded optimal results with a 97.2% Accuracy. Notably, aligning the language and visual backbones on the diseased leaf classification dataset proves to be quite challenging. As depicted in Fig. 6, the proposed Alignment Loss decreases more slowly compared to other loss functions when using the Swin Transformer+BERT combination.

Many studies have consistently demonstrated that pre-training on large-scale datasets is typically more beneficial than making minor improvements to network architectures. This observation is particularly evident in our exploration of cross-modal alignment. Backbones that have not been trained on image-text alignment do not exhibit significant performance improvements under the contrastive learning training paradigm. This is because achieving such alignment requires a vast amount of data for training (like CLIP).

Loss functions To better understand the impact of different loss functions on DiCLS, we investigated the model's performance under various loss configurations. We provided five optional losses for combination: (1) Arcface Loss L_{ARC} [3], which enhances the model's performance on fine-grained classification tasks. (2) Alignment Loss L_{ALN}, which aligns the image and word features and is represented by Eq. (6). (3) Token Loss L_{CTA}, which aligns the image patch token features and text token features and is described by Eq. (13). (4) Prototype Loss L_{CPA}, which

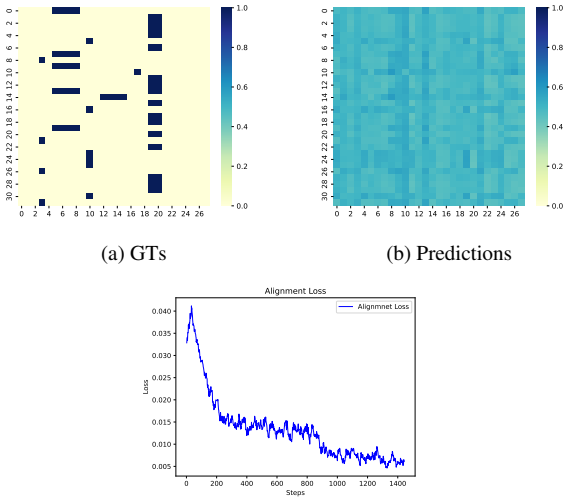
Model	Accuracy(%)	Precision(%)	Avg Precision(%)
Swin-S+BERT	96.2	92.4	96.4
Swin-S+GPT2	95.5	92.1	94.5
SwinV2-S+BERT	**96.9**	**93.1**	**96.8**
SwinV2-S+GPT2	96.4	92.7	96.1
DeiT+BERT	95.5	91.7	94.3
DeiT+GPT2	95.3	92.1	94.7
ResNet+BERT	95.6	92.7	95.3
ResNet+GPT2	94.9	92.3	95.6

Table 2. **Results with different backbones** In our experiments, we chose Accuracy and Precision as the primary metrics, and the results indicated that the SwinV2-S+BERT composition performed the best. But this is probably caused by prompt designing given that GPT2 and BERT has different prompt preferences.



(a) Ours (b) Swint

Figure 5. **The visualization of visual features based on t-SNE [21]** It can be observed that the visual-text fusion features obtained by DiCLS (left) exhibit better clustering performance in the feature space compared to the plain Swin Transformer (right). It is worth noting that we simultaneously observe that Class 0 (complex) and Class 5 (scab) are crucial factors affecting the model's performance.



(a) GTs (b) Predictions

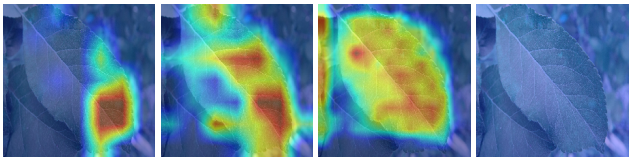
(c) Alignment Loss

Figure 6. **Image-word alignment scores** In the depicted figure, it is evident that the ground truth (left) and the predictions (right) are clearly mismatched, indicating that the cross-modal alignment did not achieve the desired outcome, even though the alignment loss decreased slowly (below).

aligns the image and text global features in a higher-level semantic space and is given by Eq. (17). (5) Focal Loss L_{FCE} [12], an improved version of BCE Loss in Eq. (3), which supervises the model's classification head for image classification as depicted in Fig. 1.

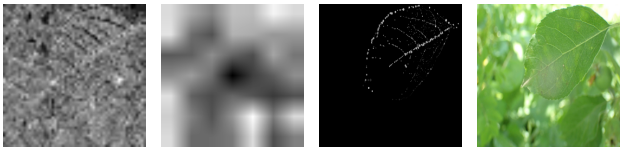
We assigned different weights to each loss and individually tested the results when (1)-(4) were applied separately. It is worth noting that we set (5) as **a fixed loss component**. This decision was made because in our experiments, we observed that using Alignment Loss alone hindered the model's convergence, which led to a Accuracy of less than 90%. In fact, we probably built a model that facilitates vision backbones only(see Fig. 5 and Fig. 6). for a comprehensive understanding of the experimental subject, an ablation experiment targeting all different loss functions has been conducted. Due to limitations in space and time, we only present the model's performance under the same set of loss weights.

Overall, we have found that all the proposed cross-modal alignment methods in the article have yielded limited improvements for DiCLS. Among these methods, we found: (1) The Focal Loss, which is a variant of traditional cross-entropy, remains the primary dynamic mechanism for supervising the model to learn the data distribution. (2) The Token Loss resulted in a marginal improvement of 0.1 percentage points on average. However, further research is needed to ascertain the generalizability of this enhancement across other baseline models. (3) Despite the decline observed in the Prototype Loss, it is, in fact, inconsequential. We believe that this decrease is a result of the backbone's improved ability to discriminate image features rather than any meaningful contribution from the Prototype Loss itself. (4) The Alignment Loss has had minimal impact and has not significantly influenced the performance of the network.



(a) Backbone (b) $L = 1$ (c) $L = 2$ (d) $L = 3$

Figure 7. **Visualization of features based on GradCAM [17]** It becomes apparent that as the model progresses, the deeper layers increasingly prioritize global information. However, we must acknowledge that an issue arises when $L \geq 3$, as we observe the occurrence of gradient vanishing. This particular phenomenon can possibly be traced back to certain design deficiencies that require further attention and investigation.



(a) Block 1 (b) Block 12 (c) SIFT (d) Original

Figure 8. **Visualization of feature maps** The results exhibit the feature maps of different blocks in Swin Transformer, revealing the varying levels of image features captured by the model at different layers. For instance, the earlier layers primarily extract edges and fine details, which is similar to SIFT, while the deeper layers focus on shape and semantic features.

(5) When using ViT as the backbone, the Arcface Loss resulted in a slight improvement in the model's performance, approximately 0.3%. By comparing their ROC curves, as illustrated in Fig. 4, we observed an enhancement in the model's performance on Class 0.

Deep Fusion variants In this experiment, our objective was to modify the settings of the Deep Fusion module and investigate their inherent properties. We conducted an analysis by varying the number of layers in the deep fusion module and observed that the model experienced a slight improvement when $L <= 2$. However, this improvement paused when $L > 2$. Through a visual examination of the attention weights of each fusion layer, as depicted in Fig. 7, we discovered that the model reached a saturation point that unable to acquire further intricate details when $L > 2$. Initially, our hypothesis suggested that this phenomenon was due to the loss of fine-grained image features, as the model primarily retained semantic information in its deeper layers. To address this, we attempted to integrate the features extracted from early layers into the Deep Fusion Block through skip connections. Regrettably, this approach did not yield substantial enhancements. Our analysis suggests three potential reasons for this outcome: (1) The capability of our feature extractor to capture these finer details remains insufficient, given the dataset's minimal inter-class differences but significant intra-class variations. (2) The final layers of the Deep Fusion module are unable to learn higher-level semantic information, which may require larger-scale models and corresponding pre-training. (3) The dataset itself contains some noise, making it challenging for the deeper layers to capture accurate higher-level semantic information.

Additionally, we attempted to use concatenated image-text features as input for the classifier, but no significant changes were observed. We believe this is because the fused image features alone possess sufficient representational capacity to encompass the information from the text features.

During the process of replacing the Attention mechanism in the Deep Fusion Block, we observed minimal improvements. This indicates that the traditional Attention mechanism is already capable of learning rich semantic information. In fact, recent research has also suggested that many new attention methods focus on reducing model computational complexity rather than improving performance.

Prompt engineerings It is intriguing to note that different prompt designs have a significant impact on the performance of the model [11]. This observation appears to deviate from the conclusion we drew in the Loss section (Sec. 4.2), and we will make every effort to provide an explanation. In our prompt design, we primarily divide the prompt into three parts: suffix, body, and prefix

$$\text{Prompt} = \text{suffix} + \text{body} + \text{prefix} \tag{18}$$

As demonstrated in Table 3, a body with richer semantics and shorter class names usually leads to significant performance improvements. Descriptive suffixes and prefixes also contribute to further enhancements. However, suffixes and prefixes often require specific designs. We have found that even prompts that align with intuition can sometimes result in a substantial decline in model performance. Due to the lack of further validation, we tentatively attribute this phenomenon to the preferences of the language backbone during pre-training. However, further research and verification are required to delve deeper into this issue.

4.3. Limitation analysis

Based on our experimental findings, we believe that these works in cross-modal alignment still have several limitations, primarily stemming from misconceptions about image-text alignment itself and the lack of in-depth research methods.

Firstly, the proposed DiCLS model failed to surpass the accuracy baseline on multiple evaluation metrics while increasing GPU memory consumption by approximately 1.5 times. Furthermore, the proposed model did not achieve the desired effectiveness in cross-modal alignment. Additionally, the study appears to have mistakenly judged the

effect of textual information on improving the discriminative power of the visual backbone, even though this requires further in-depth validation and research.

Moreover, the persuasiveness of the research needs further improvement, as the completeness of the conducted experiments does not adequately support the research findings.

5. Conclusion

In this paper, the DiCLS model is proposed to accomplish image classification tasks on a given dataset. The proposed method effectively addresses the issue of fine-grained classification in the dataset, while also conducting preliminary research and in-depth exploration of cross-modal alignment.

In conclusion, as a emerging field, cross-modal alignment holds profound dynamics that warrant further investigation to help us better understand the principles underlying cross-modal alignment work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[4] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[11] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 8

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 1

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[19] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[21] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[23] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. 2

[24] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1

[25] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 2

[26] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1