BiMAE - A Bimodal Masked Autoencoder Architecture for Single-Label Hyperspectral Image Classification

Maksim Kukushkin
Leipzig University
Augustusplatz 10, 04109 Leipzig
kukushkin@informatik.uni-leipzig.de

Martin Bogdan
Leipzig University
Augustusplatz 10, 04109 Leipzig
bogdan@informatik.uni-leipzig.de

Thomas Schmid
Martin Luther University Halle-Wittenberg
Universitätsplatz 10, 06108 Halle (Saale)
thomas.schmid@medizin.uni-halle.de

Abstract

Hyperspectral imaging offers manifold opportunities for applications that may not, or only partially, be achieved within the visual spectrum. Our paper presents a novel approach for Single-Label Hyperspectral Image Classification, demonstrated through the example of a key challenge faced by agricultural seed producers: seed purity testing. We employ Self-Supervised Learning and Masked Image Modeling techniques to tackle this task. Recognizing the challenges and costs associated with acquiring hyperspectral data, we aim to develop a versatile method capable of working with visible, arbitrary combinations of spectral bands (multispectral data) and hyperspectral sensor data. By integrating RGB and hyperspectral data, we leverage the detailed spatial information from RGB images and the rich spectral information from hyperspectral data to enhance the accuracy of seed classification. Through evaluations in various real-life scenarios, we demonstrate the flexibility, scalability, and efficiency of our approach.

1. Introduction

Hyperspectral imaging offers wide variety of applications from remote sensing to analyzing level of plants in agricultural field. One of the main task of hyperspectral data is classification, which can be categorized into main two types: (i) Single-Label Hyperspectral Image Classification: This involves assigning a single label to hyperspectral image; (ii) Multi-Class Hyperspectral Image Classification: Here, the goal is to classify pixels of a hyperspectral image into multiple classes, allowing more detailed analysis and interpretation of the scene. This, for instance, applies for remote sensing dataset like Indian Pines, University of Pavia and etc.

While Multi-Class Hyperspectral Classification is extensively studied, Single-Label Hyperspectral Classification remains less explored, primarily due to limited data availability. In our study, we focus on Single-Label Hyperspectral Image Classification, made possible through collaboration with an industry partner. We specifically examine seed purity testing as a case study, which fits within the framework of Single-Label Hyperspectral Classification. However, we believe our research extends beyond this application, contributing to broader advancements in Single-Label Hyperspectral Image Classification.

In agricultural seed production, ensuring seed quality presents a significant challenge. Not only as customer expectations need to be met, but in many countries also mandatory goverment regulations [7, 33, 56]. For instance, the European Union (EU) implements rigorous quality control measures, such as certifying seed lots before they can be sold [57]. Thus, seed producers must regularly analyze and classify harvested seeds to comply with these regulations, which typically involves trained human analysts.

To address this challenge, researchers have explored the use of deep learning techniques based on RGB data. However, relying solely on color information may not be sufficient for accurately distinguishing between different types of seeds. While RGB analysis provides valuable insights into visual properties, it cannot capture information about the chemical composition of seeds beyond the visible spectrum, leading to limitations like metameric colors and difficulty in discriminating between species.

As an alternative, some studies have proposed using hyperspectral imaging, which captures a wider range of spectral information [17, 19]. Yet, this approach also has its drawbacks [29], including lower spatial resolution and challenges in model generalization due to variations in image acquisition conditions. Moreover, acquiring hyperspectral data is costly and time-consuming, often limiting its practical application in industries where speed is crucial. Thus, developers often prefer to use multispectral data, which involves selecting a subset of spectral bands for analysis.

Our study aims to streamline the spectral band selection process for model training, making it more flexible by necessitating only finetuning on chosen modalities and spectral bands. Through the utilization of self-supervised learning, which operates without the need for labeled data, and masked image modeling, we have developed a classification model capable of accommodating any number and combination of spectral bands. This model can seamlessly operate with either RGB, multispectral or hyperspectral modalities, offering versatility in its application.

2. Related Work

2.1. Computer Vision for Seed Analysis

Automated seed sorting, distinguishing desired from undesired seeds, has been explored extensively, with computer vision playing a pivotal role [22, 47]. A prevalent method involves classifying seed images based on labeled datasets, with various studies focusing on different seed types such as rice [31, 44], cottonseeds [30], sunflower [6, 39], tomato seeds [50], corn seeds [2, 51], wheat [1, 59], plum kernels [48], maize [8] and Canola seed [42].

Most studies in this domain employ machine learning (ML) techniques, with a recent surge in the adoption of deep learning methods. Transfer learning, for example, has been utilized for classifying various seed species [25, 26] and wheat varieties [59]. Additionally, Swin transformers have been employed for maize variety classification [8], while AlexNet has shown promise in classifying sunflower seeds [6]. An emerging trend involves the use of hyperspectral imaging, which offers richer spectral information than traditional RGB images [17–19, 34]. Furthermore, combining hyperspectral and RGB data has shown enhanced performance [35], leveraging the strengths of both modalities.

2.2. Self-Supervised Learning

In the field of self-supervised representation learning (SSL), models are trained on a pretext task where supervision comes directly from the input data itself, eliminating the need for labeled data. SSL can be divided into two main types: (i) Contrastive learning and (ii) generative modeling.

Contrastive learning involves learning representations by comparing positive and negative samples. Noteworthy methods in this area include SimCLR [12], MoCo v1-v3 [13, 14, 27], BYOL [23], DINO [10], and DINO v2 [41].

On the other hand, generative modeling attempts to construct a generative model capable of encapsulating the underlying data distribution. The VAE/GAN model, introduced by Larsen et al. (2016), combines the strengths of variational autoencoders (VAEs) and generative adversarial networks (GANs) to produce disentangled data representations. Meanwhile, PixelCNN [54] and PixelVAE [24] generate images incrementally, pixel by pixel, while taking into account the contextual information of previously generated pixels. In the field of generative modeling, there is a significant subgroup called masked modeling, which will be discussed in the following subsection.

2.3. Masked Image Modeling: Masked Autoencoders

Masked Autoencoders (MAEs) [5, 28] have emerged as a significant development in SSL, particularly within computer vision, drawing inspiration from successful approaches in natural language processing (NLP) such as BERT [15] and GPT [9, 45, 46]. Here we consider only four important to our work aspects of MAE: (i) Multimodal MAE, (ii) Cross-Attention, (iii) Scaling MAE and (iv) MAE in Hyperspectral Imaging. For other MAE aspects as well its applications could be found in [5, 60, 61].

Multimodal MAE. Multimodal masked autoencoders (Multimodal MAE) are an extension of unimodal masked autoencoders, allowing them to handle multiple types of data, making them useful for various tasks. Recent research, such as the work by Multi-MAE[4], has demonstrated the effectiveness of multimodal MAE in learning predictive coding across different data types. Yan et al.[58] utilized bimodal MAE for depth completion tasks, incorporating both RGB and depth data. Mizrahi et al.[40] further extended on Multimodal MAE by incorporating non-visual modalities like text, images, geometry, and semantics through discrete tokenization.

Cross-Attention. Cross-Attention is a type of attention mechanism widely used in deep learning. It enables the combination of sequences of different modalities, such as text, image, or sound. Unlike Self-Attention, Cross-Attention is more cost-effective in pooling information from a large set of visible tokens due to the asymmetric combination of two separate embedding sequences. The technique can be seen as a parametric form of pooling, where different features are weighted learnably [21]. In the context of Multimodal MAE, cross-attention is used in each decoder to integrate information from encoded tokens of other modalities, as demonstrated by Bachmann et al. [40] and Mizrahi et al. [40].

Scaling MAE. Despite vanilla MAE model exhibits efficiency in its asymmetric encoder-decoder design, it face challenges when handling volumetric data such as video or hyperspectral images due to the need for significant compu-

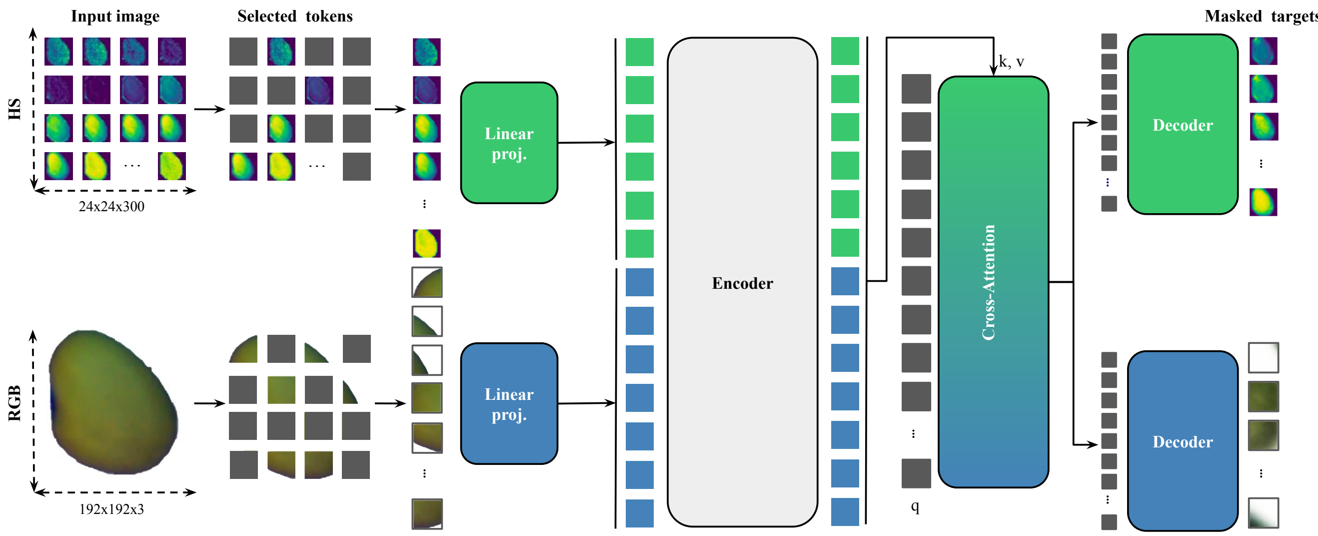

Figure 1. Overview of our Bimodal Masked Autoencoder (BiMAE) architecture.

tational resources. Proposed strategies, such as the "double-masking strategy" [55] (also known as "input and target masking" [40]) and Local Masked Reconstruction [11], aim to address this issue. Recent advancements [21] introduce a decoder architecture that utilizes cross-attention between masked and visible tokens, enhancing efficiency without sacrificing performance.

MAE in Hyperspectral Imaging. MAEs have recieved attention in hyperspectral imaging applications as well. SpectralMAE, which solves spectral reconstruction, are specifically designed to handle arbitrary combinations of spectral bands as inputs, making it versatile across different spectral sensors [62]. Furthermore, masked spatial-spectral autoencoders (MSSA) have been introduced to enhance hyperspectral image (HSI) analysis systems against adversarial attacks [43]. Noteworthy applications of MAEs in hyperspectral imaging include hyperspectral image classification (HSIC) [32, 49], few-shot classification [20], and multi-label classification [36].

It is worth mentioning that utilizing MAEs for hyperspectral data allows for much higher masking ratios, such as 0.9, compared to the RGB modality (0.75), without encountering performance degradation [62].

3. Bimodal Masked Autoencoding

In this section, we introduce the design of our primary contribution, the Bimodal Masked Autoencoder (BiMAE) architecture and analyze the key differences of our approach compared to recent masked pretraining approaches, such as MAE [28], Multi-MAE [4], and Cross-MAE [21]. Figure 1 gives a schematic overview of the BiMAE architecture.

Flexibility. To enhance the versatility of our approach for hyperspectral imaging, we incorporate two key architectural decisions:

1. BiMAE is based on the design of Vision Transformers (ViT) [16], allowing it to process a flexible number of input tokens, even partial input. Due to its computational efficiency compared to the larger *ViT*-Base and *ViT*-Large, we use the *ViT*-Small version of ViT. Detailed configuration of employed ViT available in the Supplemental Material (refer to Table 5).
2. We redefine the token specifically for the hyperspectral data, treating each spectral band of the hyperspectral image as a token (24x24x1).

Together, these two factors allow BiMAE to adapt to a flexible number of spectral bands. This enhanced versatility greatly expands the potential applications of the model in the processing of hyperspectral data.

Scalability and Efficiency. We apply and adopt following several techniques to improve the scalability and efficiency of BiMAE:

- We adopt a "double masking strategy" inspired by Wang et al. [55]. This strategy involves sending only a part of masked tokens to the decoders. This adjustment enables BiMAE to effectively handle volumetric data, like hyperspectral data with up to 300 bands. On the other hand, Multi-MAE[4] does not utilize this strategy. As a result, the imbalance between the low number of input tokens and the much higher number of target tokens can lead to significant computational costs in the decoder. This makes the Multi-MAE model less computationally efficient and scalable in a multi-modal context.
- We integrate Cross-Attention immediately after the encoder, following the approach proposed by Fu et al. (2024) [21]. Unlike conventional methods where a con-

catenation of mask and visible tokens is passed to self-attention decoders [28], BiMAE utilizes mask tokens to query visible tokens in a single cross-attention layer positioned before the decoders. This configuration allows mask tokens to gather information from visible tokens across different modalities without interacting with other mask tokens, thereby reducing the sequence length for the decoders and lowering computational costs. Moreover, by locating the cross-attention layer before the decoders, as in our BiMAE, instead of within the decoders as proposed in the Multi-MAE [4], we can utilize this layer as input for all decoders, not limited to just one. This means that unlike Multi-MAE, which employs separate cross-attention layers in each decoder, we utilize a single cross-attention layer for all decoders, further reducing computational costs.

- We employ two shallow MLP decoders for decoding each modality, which add little to the overall computational cost, and as He et al. [28] show, they perform similarly to deeper decoders on ImageNet-1K finetuning.

Together, these modifications enhance the capabilities and performance of the BiMAE model, making it particularly effective in leveraging both hyperspectral and RGB data.

Species (Class)	Train	Val.	Test	$\sum$
A. arvensis	4439	851	928	6218
A. lappa	4442	997	984	6423
A. myosuroides	3685	813	742	5240
B. napus	4054	1000	861	5750
B. officinalis	3949	855	784	5588
C. cyanus	3726	798	835	5359
E. crus-galli	3816	820	873	5509
G. aparine	3616	815	796	5227
G. dissectum	5328	1089	1094	7511
G. pratense	3831	810	794	5435
G. robertianum	4799	1021	995	6815
G. tetrahit	3850	850	848	5548
L. communis	5301	1091	1134	7526
P. aviculare	3802	825	820	5447
P. convolvulus	4477	989	980	6446
R. crispus	4263	925	942	6130
S. arvensis	3516	782	777	5075
S. media	3841	865	849	5555
T. pratense	5465	1150	1151	7766
$\sum$	80200	17185	17187	114572

Table 1. Sample Allocation in Training, Validation and Test Sets per each (RGB/MS/HS) modality

4. Experiments

4.1. Data

In order to assess the effectiveness of the proposed BiMAE architecture, we carrid out a series of experiments that employed a comprehensive bimodal dataset consisting of 114,572 RGB images paired with their corresponding hyperspectral counterparts, collected from 19 distinct species. The dataset was divided into training, validation, and test sets using a 70%/15%/15% split, as illustrated in Figure 2 and summarized in Table 1.

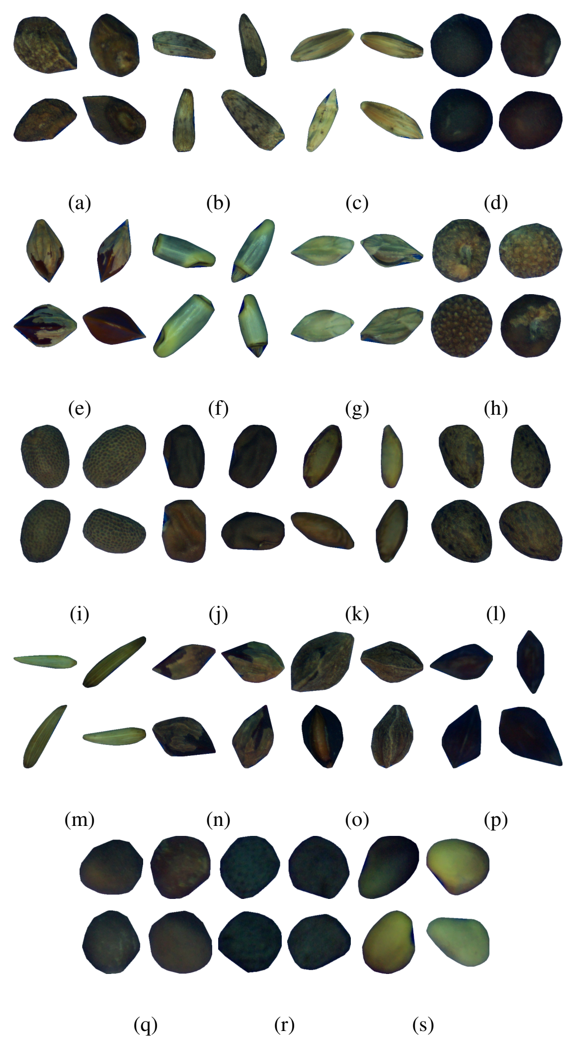

Figure 2. Examples from the test dataset. Images of (a-s) *A. arvensis* L., *A. lappa* L., *A. myosuroides* L., *B. napus* L., *B. officinalis* L., *C. cyanus* L., *E. crus-galli* L., *G. aparine* L., *G. dissectum* L., *G. pratense* L., *G. robertianum* L., *G. tetrahit* L., *L. communis* L., *P. aviculare* L., *P. convolvulus* L., *R. crispus* L., *S. arvensis* L., *S. media* L. and *T. pratense* L.

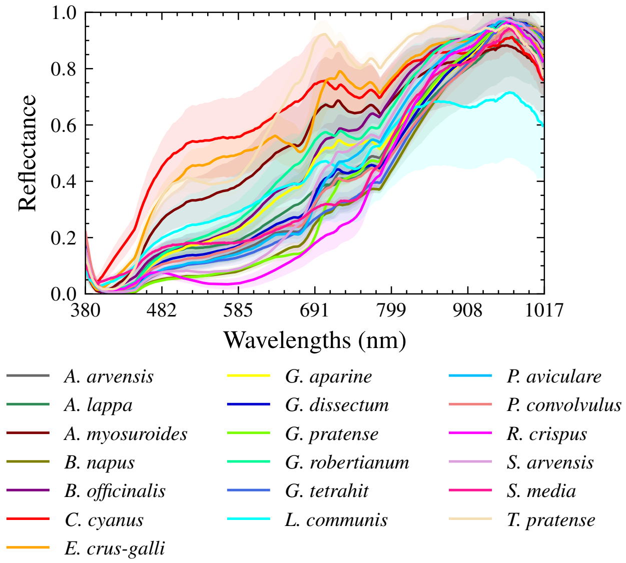

Figure 3. Spectra of the Region of Interest (ROI) for each seed species in the test dataset, illustrating the mean and standard deviation of a 8x8x300 dimensional ROI along last dimension

The RGB images were resized to a size of 192x192 pixels, while the hyperspectral images were resized to 24x24 pixels with a depth of 300 spectral bands. These bands were captured using the Resonon (USA) Pika L 100121-220 model, covering wavelengths ranging from 380 nm to 1000 nm in the visible and near-infrared (VNIR) region of the electromagnetic spectrum, with a spectral resolution of 5 nm (see Figure 3 for the mean spectra visualization).

The RGB images were acquired using the Sony (Japan) IMX477 model.

4.2. Pretraining

In all experiments, we employ the *ViT*-Small architecture [16] with a patch size of 24x24. For the hyperspectral (HS) modality, each token x_{hs} represents one of the 300 spectral bands, yielding 300 tokens per hyperspectral image. For RGB images that have a size of 192x192x3, each token x_{rgb} corresponds to a spatial patch (24x24x3) of the image. This results in a total of 64 tokens, ensuring compatibility between tokens from both modalities and facilitating their processing by the encoder.

We apply different masking ratios r to each modality: $r_{hs}(x_{hs}) = 0.9$ for hyperspectral and $r_{rgb}(x_{rgb}) = 0.75$ for RGB images. As mentioned earlier in Section 3, we utilize cross-attention for mask tokens to query visible tokens for further reconstruction in modality-specific decoders the masked patches. With the "double masking strategy" combination, we reconstruct only a subset $s_{hs}(x_{hs})$ and $s_{rgb}(x_{rgb})$ of masked tokens. For computational efficiency, we set $s_{hs}(x_{hs}) = 0.2$ and $s_{rgb}(x_{rgb}) = 0.5$.

Finally, we initialize our BiMAE and pretrain it for 300 epochs using the aforementioned dataset (see Section 4.1). We utilize the AdamW optimizer [38] with a base learning rate set to 1e-4 and weight decay of 0.05. The training process begins with a warm-up phase (30 epochs), starting with a learning rate of 1e-6, and gradually decays to 0 during training using cosine decay [37]. The training is conducted on Nvidia RTX A 6000 GPU with a batch size of 512. Data augmentation techniques, including random horizontal and vertical flips, are applied to both modalities with a probability of 0.5.

4.3. Finetuning (FT)

To evaluate the performance of the pretrained model, we extensively test its capabilities in single-label image classification as the downstream task (see Fig. 4). We replace the decoders with an average pooling operation over all encoded tokens, followed by LayerNorm [3] and a dense layer with softmax activation.

For end-to-end finetuning (FT), we utilize the supervised version of the dataset (cf. Section 4.1), training over 50 epochs on the entire training split containing 80,200 bimodal samples. We report the top-1 test accuracy and test loss. Similar to the pretraining phase, we employ the AdamW optimizer with a base learning rate set to 5e-4, weight decay of 0.05, a warmup phase lasting 5 epochs, and a warmup learning rate of 1e-6. We utilize cosine decay and maintain a batch size of 512. Data augmentation techniques used during pretraining are also applied in this phase.

Additionally, to simulate real-life scenarios where only a

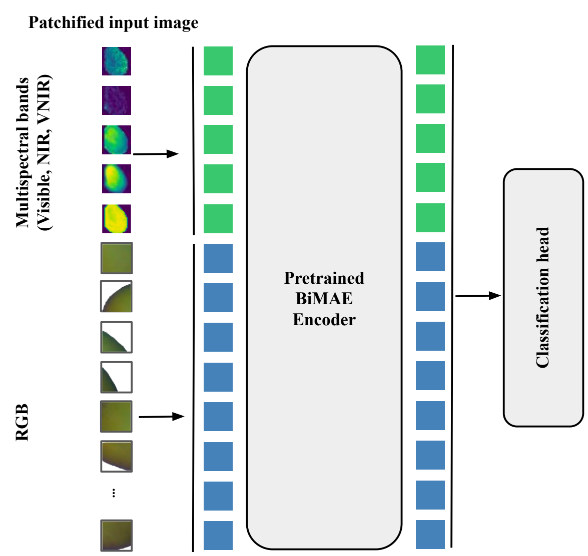

Figure 4. Single-Label Image Classification with BiMAE using bimodal data. For multispectral data, spectral bands along with their spectral indices should be provided. For unimodal data only corresponding part of pretrained BiMAE encoder is initialized.

limited number of spectral bands are available (multispectral rather than hyperspectral data), we reduce the number of spectral bands in the hyperspectral modality to n, creating a multispectral modality (MS) for downstream classification.

Overall, we examined how the pretrained BiMAE model transfers knowledge under following real-life scenarios:

(i) only RGB data is available;

(ii) only multispectral data is available;

(iii) both RGB and multispectral data are available.

To select spectral bands for multispectral modality we followed two band selection strategies (BSS):

(a) *Step60* - Sparse selection of every 60th spectral band from the hyperspectral data, representing the Visible and NIR spectrum (VNIR);

(b) *Step30* - Sparse selection of every 30th spectral band, representing VNIR;

An ablation study was conducted to analyze further how the number and selection of spectral bands for multispectral data affects performance on the downstream task.

4.4. Training from scratch (TFS)

In order to assess the effectiveness of transfer learning with our approach, we performed a thorough comparison by training BiMAE from scratch (TFS) on a classification task. This was carried out across all defined scenarios, using identical training settings as those employed for finetuning (see Section 4.3).

4.5. Results

4.5.1 Unimodal transfers

Examining BiMAE's performance in classification tasks, particularly when utilizing different modalities, reveals the effectiveness of models trained on multispectral data exclusively. Comparing models trained from scratch (TFS) with those fine-tuned (FT), it's evident that the latter outperforms the former. For instance, when fine-tuned on RGB data, BiMAE achieved an accuracy of 98.27%, while the model trained from scratch reached 97.73%. Utilizing multispectral data with 5 bands (*Step60*), BiMAE achieved an accuracy of 98.50% with finetuning and 97.45% with TFS. Increasing the bands to 10 in multispectral data (*Step30*) yielded even higher accuracy, reaching 99.06% with finetuning and 98.32% with TFS. This trend persists on hyperspectral modality as well, with BiMAE achieving an accuracy of 98.41% with finetuning and 97.93% with TFS.

4.5.2 Bimodal transfers

BiMAE, finetuned on bimodal data, shows significantly better results than using unimodal data only. Thus, BiMAE trained on RGB and multispectral data reaches the highest accuracy of 99.55%.

Mode	Modality	BSS	Loss ↓	Acc. (%) ↑
FT	RGB	-	0.074	98.27
TFS	RGB	-	0.093	97.73
FT	MS	Step60	0.079	98.50
TFS	MS	Step60	0.137	97.45
FT	MS	Step30	**0.039**	**99.06**
TFS	MS	Step30	0.085	98.32
FT	HS	-	0.057	98.41
TFS	HS	-	0.084	97.93

Table 2. Comparison of finetuning of BiMAE (FT) with training from scratch (TFS) performance using *single modality* only

Mode	Modalities	Loss ↓	Acc. (%) ↑
FT	RGB+MS	**0.018**	**99.55**
TFS	RGB+MS	0.030	99.28

Table 3. Comparison of finetuning (FT) with training from scratch (TFS) performance of BiMAE using *two modalities*

5. Discussion

Our BiMAE model demonstrates remarkable flexibility and effectiveness in accurately classifying various seed species samples, as evidenced by the results obtained during both training and testing phases.

5.1. Measuring the influence of spectral band selection for multispectral modality

The selection of specific spectral bands within the electromagnetic spectrum (EM) can significantly affect classification accuracy. To evaluate this influence, we conducted additional experiments in a bimodal setting by adding four more band selection strategies to early introduced *Step60* and *Step30*:

(i) *Top5* - Selection of the first 5 spectral bands, representing the Visible spectrum;

(ii) *Top10* - Selection of the first 10 spectral bands, representing the Visible spectrum;

(iii) *Bottom5* - Selection of the last 5 spectral bands, representing the Near-Infrared (NIR) spectrum;

(iv) *Bottom10* - Selection of the last 10 spectral bands, representing the NIR spectrum;

Analyzing the results in Table 4, we can see, that BiMAE finetuned on multispectral data with *Step30* strategy performed the best, reaching the accuracy of 99.55%. Using *Step60* strategy reduced the accuracy of the BiMAE only marginally (99.49%). Selecting the spectral bands from NIR part of the spectrum only, leads to lower accuracy of 98.74% when using 5 bands (*Bottom5*) and a bit higher ac-

Spectrum	Nb. bands	BSS	Loss ↓	Acc. (%) ↑
Visible	5	Top5	0.050	98.92
Visible	10	Top10	0.091	97.99
NIR	5	Bottom5	0.058	98.74
NIR	10	Bottom10	0.046	98.95
VNIR	5	Step60	**0.018**	**99.55**
VNIR	10	Step30	0.024	99.49

Table 4. Finetuning (FT) performance of BiMAE using *two modalities* (RGB and MS) using various band selection strategies (BSS)

curacy of 98.95% when using 10 bands (*Bottom10*). The lowest accuracy of 97.99% is reached, by selecting 10 bands of Visible part of spectrum (*Top10*).

5.2. Model Comparison

When comparing the performance of BiMAE trained on unimodal data, it becomes evident that finetuned models consistently outperform those trained from scratch. Furthermore, a comparison of the modalities on which BiMAE was trained (FT or TFS) reveals that those trained on multispectral modality achieve the best results. Additionally, the number of bands in multispectral data proves to be crucial; for example, bands selected using the *Step60* strategy outperform those chosen with the *Step30* strategy.

Another notable finding is the inability of unimodal BiMAE, finetuned on hyperspectral data, to surpass the performance of unimodal BiMAE, finetuned on multispectral data. This might be attributed to the limited training time for FT and TFS (50 epochs), suggesting that FT and TFS may require more computational time, especially for hyperspectral data. Additionally, the Hughes phenomenon or curse of dimensionality of data [52, 53] might be involved here.

In the comparison of BiMAE models trained on unimodal and bimodal data, it is evident that finetuned models consistently outperform those trained from scratch. Overall, bimodal models exhibit superior performance compared to those trained on unimodal data, indicating that training on more diverse data enhances the model's classification capabilities. This principle extends to the diversity of spectral bands in the multispectral modality, as demonstrated by the results presented in Table 4, which show that selecting spectral bands from various parts of the VNIR spectrum can significantly enhance classification accuracy.

The variety of experiments conducted in Section 4.1, utilizing different combinations of modalities in both unimodal and bimodal settings, underscores the versatile potential applications of the model.

6. Conclusion and Outlook

Modern data-driven AI has demonstrated potential to greatly contribute to sustainable agriculture. By automating tasks and reducing errors, AI can simplify the work of farmers. The focus of our current work is on enhancing seed purity, traditionally a human task. In particular, we have proposed our BiMAE architecture for bimodal single-label classification that allows to enhance the efficiency and accuracy of seed purity, thus promoting sustainability in agriculture. In our study, we have showcased the effectiveness, adaptability, and scalability of our BiMAE model in classifying various seed species using RGB, multispectral, and hyperspectral images. These findings underscore the potential of our approach to streamline and expedite seed production in agriculture.

Looking forward, our future research will concentrate on refining single label classification techniques in agriculture. For instance, we plan to explore additional applications of BiMAE beyond classification, such as seed segmentation, which could enable deeper seed analysis. Additionally, we aim to evaluate the model's performance on unseen species using zero-shot or few-shot learning techniques. Furthermore, incorporating more modalities and investigating their synergies could offer further insights. Lastly, we aim to identify the key wavelengths crucial for distinguishing between different seed types. This insight could simplify classification, enhance efficiency, and potentially reduce costs in agricultural processes.

7. Acknowledgments

We express our gratitude to NPZ Innovation GmbH for generously providing the dataset. Moreover, we would like to thank Matthias Enders and Simon Goertz for discussions on various aspects of agricultural seed purity testing.

This project is supported by funds from the German Federal Ministry of Food and Agriculture (BMEL), based on a decision of the Parliament of the Federal Republic of Germany. The German Federal Office for Agriculture and Food (BLE) provides coordinating support for artificial intelligence (AI) in agriculture as the funding organization, grant number 28DK116C20.

References

[1] Diwakar Agarwal, P Bachan, et al. Machine learning approach for the classification of wheat grains. *Smart Agricultural Technology*, 3:100136, 2023. 2

[2] Aqib Ali, Salman Qadri, Wali Khan Mashwani, Samir Brahim Belhaouari, Samreen Naeem, Sidra Rafique, Farrukh Jamal, Christophe Chesneau, and Sania Anam. Machine learning approach for the classification of corn seed using hybrid features. *International Journal of Food Properties*, 23(1):1110–1124, 2020. 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 2, 3, 4

[5] Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024. 2

[6] Mikel Barrio-Conde, Marco Antonio Zanella, Javier Manuel Aguiar-Perez, Ruben Ruiz-Gonzalez, and Jaime Gomez-Gil. A deep learning image system for classifying high oleic sunflower seed varieties. *Sensors*, 23(5):2471, 2023. 2

[7] Lin Batten, Maria José Plana Casado, and Josephine van Zeben. Decoding seed quality: a comparative analysis of seed marketing law in the eu and the united states. *Agronomy*, 11(10):2038, 2021. 1

[8] Chunguang Bi, Nan Hu, Yiqiang Zou, Shuo Zhang, Suzhen Xu, and Helong Yu. Development of deep learning methodology for maize seed variety recognition based on improved swin transformer. *Agronomy*, 12(8):1843, 2022. 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[11] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022. 3

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[16] 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, 5

[17] Gamal ElMasry, Nasser Mandour, Salim Al-Rejaie, Etienne Belin, and David Rousseau. Recent applications of multispectral imaging in seed phenotyping and quality monitoring—an overview. *Sensors*, 19(5):1090, 2019. 2

[18] Samson Damilola Fabiyi, Hai Vu, Christos Tachtatzis, Paul Murray, David Harle, Trung Kien Dao, Ivan Andonovic, Jinchang Ren, and Stephen Marshall. Varietal classification of rice seeds using RGB and hyperspectral images. *IEEE Access*, 8:22493–22505, 2020.

[19] Lei Feng, Susu Zhu, Fei Liu, Yong He, Yidan Bao, and Chu Zhang. Hyperspectral imaging for seed quality and safety inspection: A review. *Plant methods*, 15(1):1–25, 2019. 2

[20] Pengming Feng, Kaihan Wang, Jian Guan, Guangjun He, and Shichao Jin. Spectral masked autoencoder for few-shot hyperspectral image classification. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023. 3

[21] Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, Xudong Wang, Adam Yala, Trevor Darrell, Alexei A. Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *arXiv preprint arXiv:2401.14391*, 2024. 2, 3

[22] Zhaoyong Gong, Fang Cheng, Zihao Liu, Xiaoling Yang, Bujin Zhai, and Zhaohong You. Recent developments of seeds quality inspection and grading based on machine vision. In *2015 ASABE Annual International Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2015. 2

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[24] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016. 2

[25] Yonis Gulzar, Yasir Hamid, Arjumand Bano Soomro, Ali A Alwan, and Ludovic Journaux. A convolution neural network-based seed classification system. *Symmetry*, 12(12): 2018, 2020. 2

[26] Yasir Hamid, Sharyar Wani, Arjumand Bano Soomro, Ali A Alwan, and Yonis Gulzar. Smart seed classification system based on mobilenetv2 architecture. In *2022 2nd International Conference on Computing and Information Technology (ICCIT)*, pages 217–222. IEEE, 2022. 2

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 4

[29] Maryam Imani and Hassan Ghassemian. An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges. *Information fusion*, 59:59–83, 2020. 2

[30] KS Jamuna, S Karpagavalli, MS Vijaya, P Revathi, S Gokilavani, and E Madhiya. Classification of seed cotton yield based on the growth stages of cotton crop using machine learning techniques. In *2010 International Conference on Advances in Computer Engineering*, pages 312–315. IEEE, 2010. 2

[31] Kantip Kiratiratanapruk, Pitchayagan Temniranrat, Wasin Sinthupinyo, Panintorn Prempree, Kosom Chaitavon, Supanit Porntheeraphat, and Anchalee Prasertsak. Development of paddy rice seed classification process using machine learning techniques for automatic grading machine. *Journal of Sensors*, 2020, 2020. 2

[32] Weili Kong, Baisen Liu, and Xiaojun Bi. Instructional mask autoencoder: A powerful pretrained model for hyperspectral image classification. 2023. 3

[33] Katrin Kuhlmann and Bhramar Dey. Using regulatory flexibility to address market informality in seed systems: A global study. *Agronomy*, 11(2):377, 2021. 1

[34] Maksim Kukushkin, Matthias Enders, Reinhard Kaschuba, Martin Bogdan, and Thomas Schmid. Canola seed or not? autoencoder-based anomaly detection in agricultural seed production. In *INFORMATIK 2023 - Designing Futures: Zukünfte gestalten*, pages 1645–1652. Gesellschaft für Informatik e.V., Bonn, 2023. 2

[35] Maksim Kukushkin, Martin Bogdan, and Thomas Schmid. BiCAE - A Bimodal Convolutional Autoencoder for Seed Purity Testing. 2024. 3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE). 2

[36] Junyan Lin, Feng Gao, Xiaochen Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 3

[37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5, 1

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5, 1

[39] Zhengguang Luan, Chunlei Li, Shumin Ding, Miaomiao Wei, and Yan Yang. Sunflower seed sorting based on convolutional neural network. In *Eleventh International Conference on Graphics and Image Processing (ICGIP 2019)*, pages 428–434. SPIE, 2020. 2

[40] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[41] 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. 2

[42] Salman Qadri, Syed Furqan Qadri, Abdul Razzaq, Muzammil Ul Rehman, Nazir Ahmad, Syed Ali Nawaz, Najia Saher, Nadeem Akhtar, and Dost Muhammad Khan. Classification of canola seed varieties based on multi-feature analysis using computer vision approach. *International Journal of Food Properties*, 24(1):493–504, 2021. 2

[43] Jiahao Qi, Zhiqiang Gong, Xingyue Liu, Chen Chen, and Ping Zhong. Masked spatial–spectral autoencoders are excellent hyperspectral defenders. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–15, 2024. 3

[44] Zhengjun Qiu, Jian Chen, Yiying Zhao, Susu Zhu, Yong He, and Chu Zhang. Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences*, 8(2):212, 2018. 2

[45] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[46] 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. 2

[47] Anisur Rahman and Byoung-Kwan Cho. Assessment of seed quality using non-destructive measurement techniques: a review. *Seed Science Research*, 26(4):285–305, 2016. 2

[48] Ewa Ropelewska, Xiang Cai, Zhan Zhang, Kadir Sabanci, and Muhammet Fatih Aslan. Benchmarking machine learning approaches to evaluate the cultivar differentiation of plum (prunus domestica l.) kernels. *Agriculture*, 12(2):285, 2022. 2

[49] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2175, 2023. 3

[50] U Škrubej, Č Rozman, D Stajnko, et al. Assessment of germination rate of the tomato seeds using image processing and machine learning. *European Journal of Horticultural Science*, 80(2):68–75, 2015. 2

[51] Jared Taylor, Chien-Ping Chiou, and Leonard J. Bond. A methodology for sorting haploid and diploid corn seed using terahertz time domain spectroscopy and machine learning. In *AIP Conference Proceedings*. Author(s), 2019. 2

[52] Prasad S Thenkabail and John G. Lyon, editors. *Hyperspectral Remote Sensing of Vegetation*. CRC Press, 2016. 7

[53] Prasad S Thenkabail, Isabella Mariotto, Murali Krishna Gumma, Elizabeth M Middleton, David R Landis, and K Fred Huemmrich. Selection of hyperspectral narrowbands (hnbs) and composition of hyperspectral twoband vegetation indices (hvis) for biophysical characterization and discrimination of crop types using field reflectance and hyperion/eo-1

data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2):427–439, 2013. 7

[54] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[55] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3

[56] Tamara Wattnem. Seed laws, certification and standardization: outlawing informal seed systems in the global south. *The Journal of Peasant Studies*, 43(4):850–867, 2016. 1

[57] Tone Winge. Seed legislation in europe and crop genetic diversity. *Sustainable Agriculture Reviews: Volume 15*, pages 1–64, 2015. 1

[58] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*, pages 378–395. Springer, 2022. 2

[59] Ali Yasar. Benchmarking analysis of cnn models for bread wheat varieties. *European Food Research and Technology*, 249(3):749–758, 2023. 2

[60] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022. 2

[61] Zexian Zhou and Xiaojing Liu. Masked autoencoders in computer vision: A comprehensive survey. *IEEE Access*, 11:113560–113579, 2023. 2

[62] Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23 (7):3728, 2023. 3