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

Maksim Kukushkin[1], Martin Bogdan[1], Thomas Schmid[2,3,1]

[1]Leipzig University, [2]Martin Luther University Halle-Wittenberg, [3]Lancaster University in Leipzig

Summary

➤ Introduce a novel approach for **Single-Label Hyperspectral Image Classification** to the task of seed purity testing.

➤ Seed assessment [1, 2] in the industry is crucial but labor-intensive.

➤ Utilize **Self-Supervised Learning** and **Masked Image Modeling** techniques to enhance classification accuracy.

➤ **Goal**: Create a method adaptable to *visible (RGB)*, *multispectral (MS)*, and *hyperspectral (HS)* data, which:

- Integrates *RGB* and *HS* data to leverage spatial and spectral information
- Demonstrates flexibility, scalability, and efficiency in real-life scenario evaluations

1. Methodology

1.1 Dataset

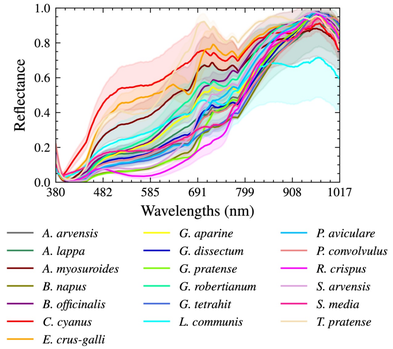

Figure 1. Mean Spectra of 8x8x300 Region of Interest

➤ Dataset contains bimodal samples of RGB and hyperspectral of seed images

➤ 19 seed species

➤ More than 114K bimodal samples

➤ 70/15/15 Train/val/test split

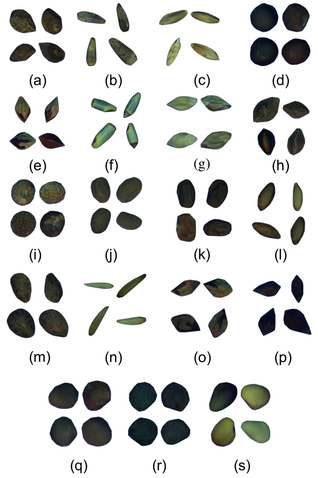

Figure 2. Examples from the test dataset. Images of (a-s) *A. arvensis* L., *A. lappa* L., *A. myosuroides* L.,*B. napus* L., *B. officinalis* L., *C. cyanus* L., *E. crus-galli* L., *G. aparine* L., *G. dissectum* L., *G. pratense* L., *G. robertianum* L., *G. tetrahit* L., *L. communis* L., *P. aviculare*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

1.2 Pretraining

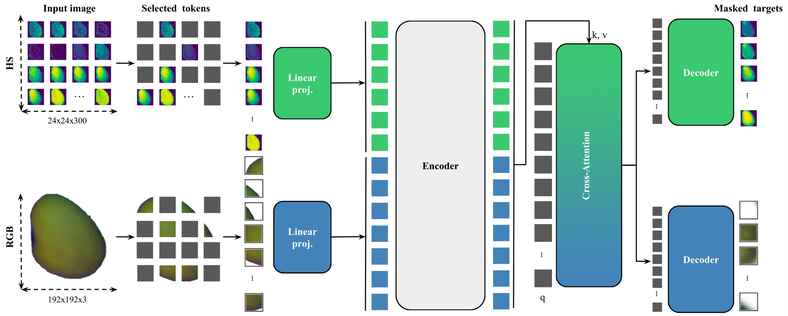

Figure 3. Overview of our Bimodal Masked Autoencoder (BiMAE) architecture

➤ Scalable & efficient ViT
- dual masking strategy [3, 4]
- shallow decoders [5]

➤ Multimodal ViT
- Cross-Attention for cross-modal token prediction

➤ ViT-Small backbone

➤ Epochs: 300

➤ Batch size: 512

2. Experiments & Results

2.1 Finetuning using two modalities

➤ Evaluate the performance of pretrained BiMAE, finetuning (FT) it on RGB and multispectral data

➤ Compare results with BiMAE trained from scratch (TFS) using both RGB and multispectral data

➤ For MS data select equidistant spectral band subset from HS with *Step30* and *Step60 (10 and 5 spectral bands respectively)*

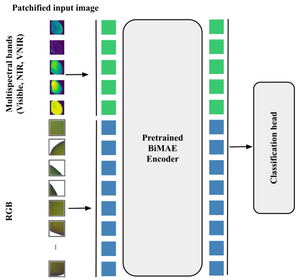

Figure 4. Single-Label Image Classification with BiMAE using bimodal data.

Mode	Modalities	Loss ↓	Acc. ↑(%)
FT	RGB+MS	**0,018**	**99.55**
TFS	RGB+MS	0,030	99.28

Table 1. Comparison of finetuning (FT) with training from scratch (TFS) performance of BiMAE, trained for 50 epochs using two modalities

2.2 Finetuning using one modality

➤ Evaluate the performance of pretrained BiMAE, finetuning (FT) it on one of RGB, multispectral or hyperspectral modalities

➤ Compare results with BiMAE trained from scratch (TFS)

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 (FT) with training from scratch (TFS) performance of BiMAE, trained for 50 epochs using one modality only

3. Conclusions

➤ Proposed novel self-supervised flexible approach for multiple modalities such as RGB, multispectral and hyperspectral data

➤ Demonstrated efficiency and versatility of BiMAE architecture on Single Lable Hyperspectral Image Classification

4. References

[1] Kukushkin, Maksim, et al. "BiCAE–A Bimodal Convolutional Autoencoder for Seed Purity Testing." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Cham: Springer Nature Lithuania, 2024 (accepted)

[2] Erichsmeier, Fabian, et al. "Automating the purity analysis of oilseed rape through usage of hyperspectral imaging." *Photonic Technologies in Plant and Agricultural Science*. Vol. 12879. SPIE, 2024.

[3] Wang, Limin, et al. "Videomae v2: Scaling video masked autoencoders with dual masking." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[4] Mizrahi, David, et al. "4M: Massively Multimodal Masked Modeling." *Advances in Neural Information Processing Systems* 36 (2024).

[5] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Contact:
Maksim Kukushkin
Leipzig University
kukushkin@informatik.uni-leipzig.de

KIRa Project:
Funding: 28DK116A20

With support from Federal Ministry of Food and Agriculture, by decision of the German Bundestag