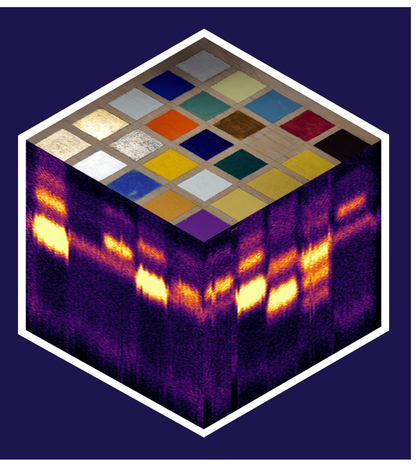
Vision Transformers for Enhanced Analysis of X-ray Fluorescence Spectra in Cultural Heritage

Eryk Zarębski
AGH University of Kraków



0. TL;DR

- We shoot paintings with high-energy radiation, collect data and map elemental distributions.
- Accurate deconvolution of characteristic elemental lines is essential. Machine learning methods, such as **NMF** or **PCA**, have been applied for this purpose [1]. Recently, deep learning approaches have also been explored [2]. However, to date, **Vision Transformers (ViT)** [3] have not yet been used for this task.

1. Physics

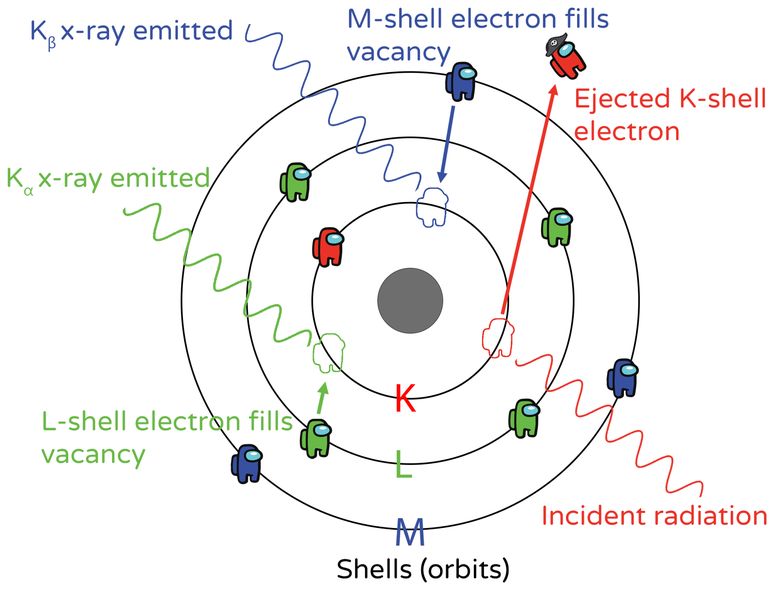


Figure 1: **X-ray fluorescence (XRF)** is the emission of characteristic (or fluorescent) X-rays from a material that has been excited by being bombarded with high-energy photons.

2. Problem

- **Full-Field X-Ray Fluorescence (FF-XRF)** [1] imaging technique is used, which captures an entire area (e.g., 100×100 pixels) at once, making it much faster than widely adapted **Micro-XRF**, which scans point-by-point.
- However, the **Gas Electron Multiplier (GEM)** detector used suffers from **low energy resolution** and **multiple artifacts** (see Figure 2), which makes XRF spectra difficult to deconvolve.

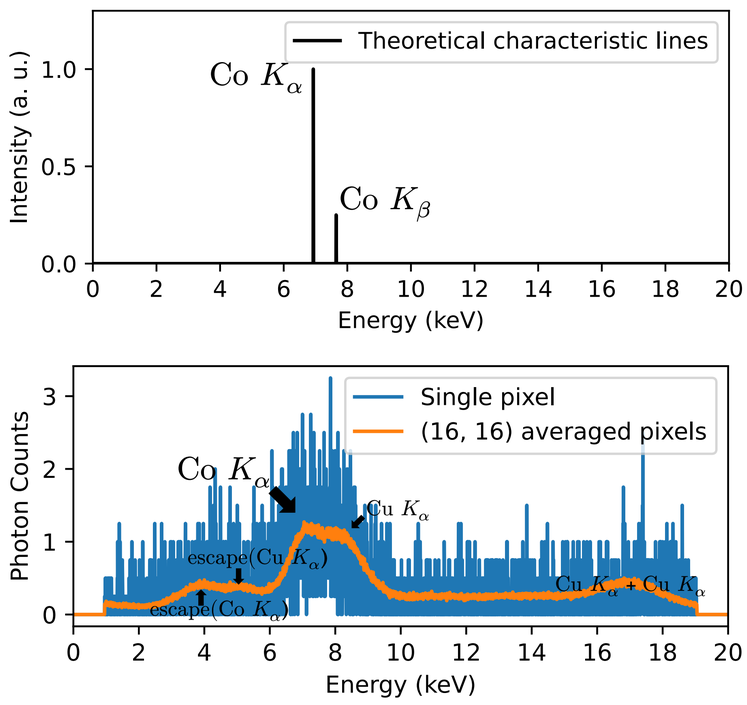


Figure 2: Theoretical characteristic lines of pure cobalt (top). Spectra measured from a cobalt-based pigment sample, showing multiple artifacts and high noise (bottom). The spectra were acquired from the object shown in the upper-right corner of the poster (2D projection of a hyperspectral cube).

3. Artificial Data Generation

- XRF spectra can be modeled as **mixtures of Gaussians** (see Figure 3).

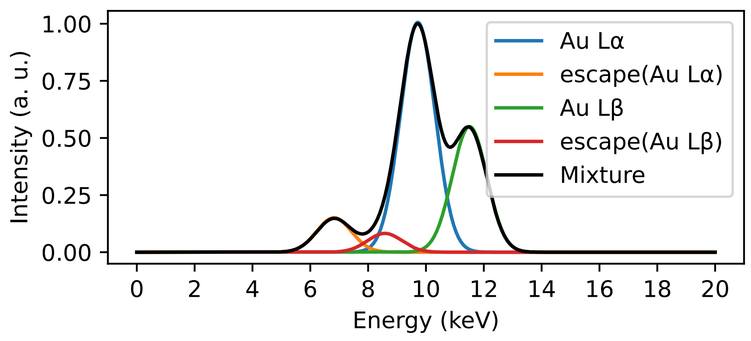


Figure 3: An artificial XRF spectrum of gold (*Au*) modeled as a mixture of Gaussian peaks, comprising two **theoretical** emission lines and their respective **escape** peaks at $E_{\text{peak}} - 2.96$ keV, originating from the GEM detector underlying mechanism.

- Complex, **augmented** multi-element mixtures can be generated along their target vectors (see Figure 4).

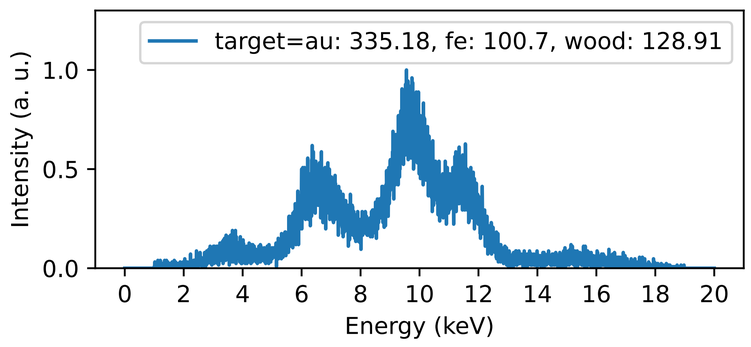


Figure 4: Sample consisting of signals from iron, gold, and wood (model of substrate background).

4. Deep Learning Models

- A **CNN** from [2], representing earlier deep learning approaches to XRF analysis, served as a benchmark for comparison with ViT models.
- Along with the original (Vanilla) ViT model [3] adapted for 1D data, another architecture was tested - the **Label Token Vision Transformer (LT-ViT)** [4]. It uses **auxiliary**, learnable [**LBL**] tokens that attend to **data** tokens instead of the usual [**CLS**] token. The idea behind [**LBL**] tokens is that the representations of **all true labels** in [**CLS**] token are not directly separable.
- Models were trained for both **multi-label classification** and **multivariate regression** task.

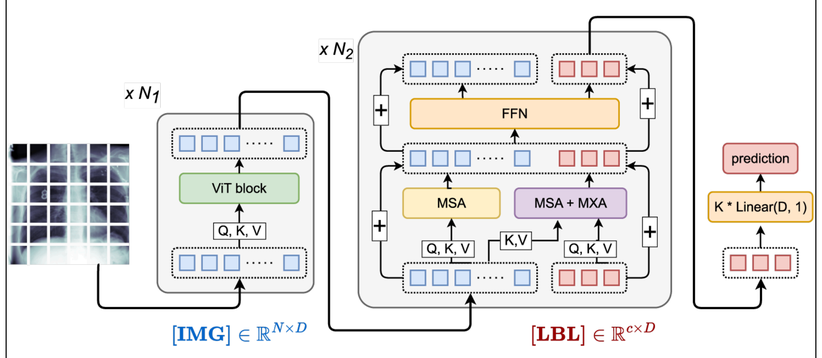


Figure 5: LT-ViT architecture. Adapted from [4].

5. Training Results

- LT-ViT model achieved the highest R^2 score after extensive hyperparameter optimization using **Optuna** framework on artificial dataset.

	LT-ViT	(Vanilla) ViT	CNN
R^2	**0.8409**	0.8353	0.7779

6. Elemental Maps

- Elemental distributions acquired using the best LT-ViT model are presented on Figure 6 and Figure 7.

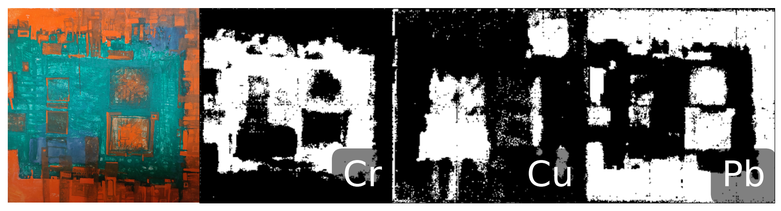


Figure 6: Photo and elemental classification maps of the prepared phantom painting. Three pigments were used in the painting: chrome green (chromium), azurite (copper) for the blue color, and minium (lead) for the orange.

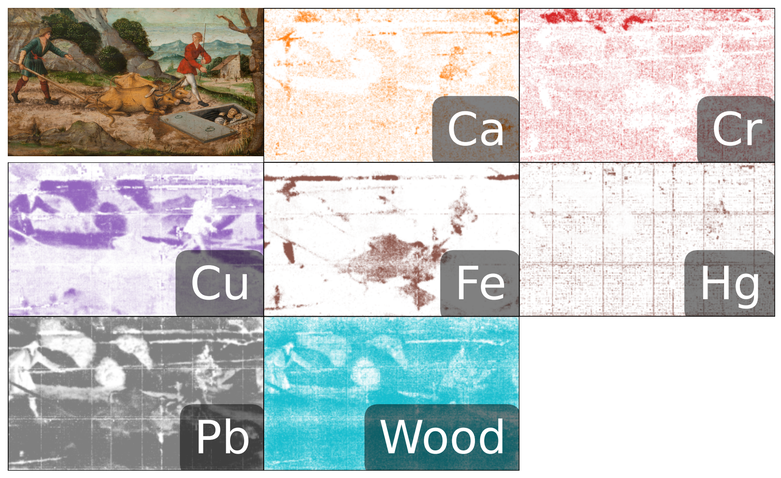


Figure 7: Photo of 16th-century oil painting *Discovering the Grave*, by unknown author and its elemental regression maps. Unfortunately, the peasant red clothing painted using vermilion pigment (mercury) should be more noticeable.

7. Conclusion

- The ViT architecture demonstrates superior performance in deconvolving noisy XRF spectra, outperforming the CNN model proposed in [2].
- Despite this improvement, poor data quality remains a limiting factor.

Bibliography

[1] B. Łach, "Rozwój systemu detekcyjnego do obrazowania przestrzennego rozkładu pierwiastków metodą fluorescencji rentgenowskiej," AGH University of Krakow, 2022.

[2] Z. Preisler, R. Andolina, A. Busacca, C. Caliri, C. Miliani, and F. P. Romano, "Deep learning for enhanced spectral analysis of MA-XRF datasets of paintings," *Science Advances*, vol. 10, no. 39, 2024, doi: 10.1126/sciadv.adp6234.

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." [Online]. Available: https://arxiv.org/abs/2010.11929

[4] U. Marikkar, S. Atito, M. Awais, and A. Mahdi, "LT-ViT: A Vision Transformer for multi-label Chest X-ray classification," 2023, doi: 10.48550/ARXIV.2311.07263.

Misc