Supplementary Material

Kilian Declercq[1,2] Abderrahmane Rahiche[1] Mohamed Cheriet[1]
[1]Synchromedia Laboratory, École de Technologie Supérieure (ÉTS), Montréal, Canada
[2]ECAM LaSalle, Lyon, France
kilian.declercq@ecam.fr, {kilian.declercq.1, abderrahmane.rahiche.1}@ens.etsmtl.ca,
Mohamed.Cheriet@etsmtl.ca, Code : https://github.com/Kilian-Declercq/PRISM

1. Metrics definition

Here, we present the definitions of all metrics used in this paper: F-measure (FM), Negative Rate Metric (NRM), Distance Reciprocal Distortion (DRD), and Peak Signal to Noise Ratio (PSNR).

F-measure (FM) or F1-score: measures prediction accuracy:

$$FM = \frac{2 \times Recall \times Precision}{Recall + Precision}, \tag{1}$$

where $Recall = \frac{TP}{TP+FN}$, $Precision = \frac{TP}{TP+FP}$, and TP, FP, FN denote True Positive, False Positive, and False Negative predictions, respectively.

Negative Rate Metric (NRM): quantifies pixel-level mismatches between predicted and ground truth (GT) images:

$$NRM = \frac{NR_{FN} + NR_{FP}}{2}, \tag{2}$$

where $NR_{FN} = \frac{FN}{FN+TP}$ and $NR_{FP} = \frac{FP}{FP+TN}$.

Distance Reciprocal Distortion (DRD): measures distortion between two binary images:

$$DRD = \frac{\sum_{k=1}^{N} DRD_k}{NUBN}, \tag{3}$$

where DRD_k is the distortion of the weighted sum of pixels in a 5×5 block of the GT image and the predicted image. NUBN is the number of non-uniform 8×8 blocks in GT.

Peak Signal to Noise Ratio (PSNR): ratio between the maximum possible pixel value I_{max} and the MSE (Mean Square Error) between two images:

$$PSNR = 10 \log_{10} \left(\frac{I_{max}^2}{MSE} \right). \tag{4}$$

Root Mean Square Error (RMSE): square root of the MSE between estimated and true values:

$$\text{RMSE} = \sqrt{\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (Y_{hw} - \hat{Y}_{hw})^2} \tag{5}$$

2. Loss function

The total loss function of our model combines three main terms, each addressing a different aspect:

$$\mathcal{L}_{Total} = \mathcal{L}_{SAD} + \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{orth}, \tag{6}$$

where λ_1, λ_2 are two penalty parameters selected empirically (See Section 5 for detailed hyperparameters values).

We assess the reconstruction quality of the MS images using the Mean Square Error (MSE), which measures the pixel-wise reconstruction quality as follows:

$$\mathcal{L}_{MSE} = \frac{1}{hwb} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{l=1}^{b} (\mathbf{Y}_{ijl} - \hat{\mathbf{Y}}_{ijl})^2, \tag{7}$$

where $\mathbf{Y}_{ijl}$ is a pixel from the input MS cube, and $\hat{\mathbf{Y}}_{ijl}$ is the corresponding pixel from the reconstructed output.

Following [9], we complement MSE by the Spectral Angle Distance (SAD), related to the cosine similarity, to measure the spectral similarity between the input and output spectral vectors:

$$\mathcal{L}_{SAD}(\mathbf{Y}_{ij.}, \hat{\mathbf{Y}}_{ij.}) = \text{acos}\left(\frac{\langle \mathbf{Y}_{ij.}, \hat{\mathbf{Y}}_{ij.} \rangle}{\|\mathbf{Y}_{ij.}\| \|\hat{\mathbf{Y}}_{ij.}\|}\right), \tag{8}$$

where acos denotes the arccos function, $\mathbf{Y}_{ij.}$ is the input spectral vector of the pixel at position i, j, $\hat{\mathbf{Y}}_{ij.}$ is the corresponding output spectral vector, and $\langle , \rangle$ is the inner product.

To promote the extraction of independent abundance maps, we enforce orthogonality between the rows of the abundance matrix $\mathbf{A}$ as follows:

$$\mathcal{L}_{orth} = \|\mathbf{A}\mathbf{A}^T - \mathbf{I}_r\|_1, \tag{9}$$

where $\|.\|_1$ is the L_1 norm, $\mathbf{I}_r$ is an $r \times r$ identity matrix, and r is the rank, which consists of the remaining abundance maps after pruning. This reduces linear correlation between extracted abundances, so that a single element does not appear on several of them.

This multi-term loss function allows for a balanced optimization that considers pixel-wise accuracy, spectral fidelity, and abundance map independence.

3. Algorithm Pseudo-code

The complete algorithmic workflow of the PRISM framework is formalized in the following algorithm :

Algorithm 1: PRISM: Progressive Pruning with MDL-based Rank Selection

Input: MS cube $\mathbf{Y}$;
Initial number of components r_{init};
Minimum number of components to evaluate r_{min};
Hyperparameters $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_{SAD'}$;
Smoothing kernel K (for spatial convolution);
Output: Optimal Abundance Maps $\mathbf{A}_{opt}$, Optimal Endmembers $\mathbf{E}_{opt}$, Optimal Rank r

Initialize PRISM model with r_{init} components (all active);
active_map_indices $\leftarrow \{1, 2, \ldots, r_{init}\}$;
best_overall_mdl_cost $\leftarrow \infty$;
best_model_config $\leftarrow$ null;
best_rank $\leftarrow r_{init}$;
for *k from r_{init} down to r_{min}* **do**
  Train PRISM model using only maps in active_map_indices until early stopping;
  Let trained_model$_k$ be the converged model parameters;
  `// Evaluate current model and calculate MDL cost`
  $\mathbf{A}_k, \mathbf{E}_k, \mathbf{S}_k, \hat{\mathbf{Y}} \leftarrow$ trained_model$_k(\mathbf{Y})$;
  $L_{recon} \leftarrow \mathcal{L}_{SAD}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_1 \mathcal{L}_{MSE}(\mathbf{Y}, \hat{\mathbf{Y}})$ `// Reconstruction error` $L(D|H)$
  $L_{struct} \leftarrow \lambda_2 \mathcal{L}_{orth}(\mathbf{A}_k)$ `// Structural complexity part of` $L(H)$
  $L_{rank\_penalty} \leftarrow \lambda_3 \cdot k$ `// Rank penalty part of` $L(H)$
  current_mdl_cost $\leftarrow L_{recon} + L_{struct} + L_{rank\_penalty}$;
  `// Update best configuration if current MDL cost is lower`
  if *current_mdl_cost $<$ best_overall_mdl_cost* **then**
    best_overall_mdl_cost $\leftarrow$ current_mdl_cost;
    best_model_config $\leftarrow$ trained_model$_k$;
    best_rank $\leftarrow k$;
  `// If more pruning is needed, select one map to prune for the next iteration`
  if $k > r_{min}$ **then**
    Let $\mathbf{A}_{active}$ be the set of abundance maps corresponding to active_map_indices from $\mathbf{A}_k$;
    Let $\mathbf{E}_{active}$ be the set of endmembers corresponding to active_map_indices from $\mathbf{E}_k$;
    for *each map A_m in $\mathbf{A}_{active}$* **do**
      $\mathcal{A}_m \leftarrow K * A_m$ `// Spatial convolution (smoothing)`
    for *each pair of distinct maps $(\mathcal{A}_i, \mathcal{A}_j)$ (and corresponding E_i, E_j from $\mathbf{E}_{active}$)* **do**
      Compute similarity S_{ij} using $\mathcal{A}_i, \mathcal{A}_j, E_i, E_j$ (using Eq. (11))
    $(i^*, j^*) \leftarrow$ pair with maximum similarity;
    map_to_prune $\leftarrow \arg\min_{m \in \{i^*, j^*\}} \|A_m\|_F$ (using A_{i^*}, A_{j^*} from $\mathbf{A}_{active}$);
    Remove map_to_prune from active_map_indices;
    Prune connections producing map_to_prune in the pre-abundance convolution layer and corresponding endmembers weights in decoder;
Load PRISM model with parameters from best_model_config;
$\mathbf{A}_{opt}, \mathbf{E}_{opt} \leftarrow$ Loaded_Best_Model$(\mathbf{Y})$;
$r \leftarrow$ best_rank;
return $\mathbf{A}_{opt}, \mathbf{E}_{opt}, r$

4. Attention block

Figure 1 illustrates our Attention block. Point-wise convolutions are indicated in green, dilated convolution in red and classical depthwise convolution in blue.

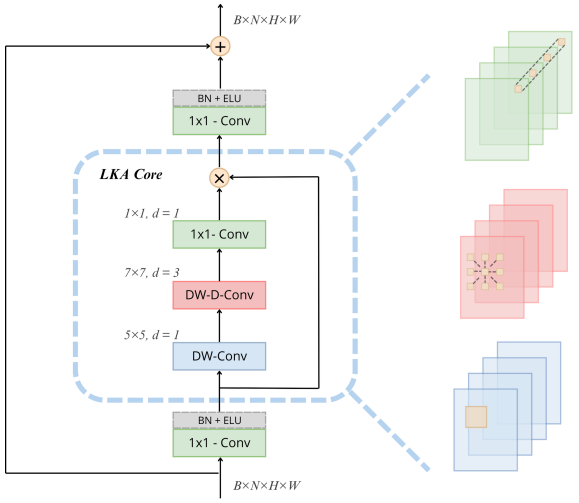


Figure 1. Illustration of the Large Kernel Attention block.

We display some attention maps to visualize how the model emphasizes relevant regions of the images. In Fig. 2, we observe that the incorporated attention helps highlight the regions of interest in the images.

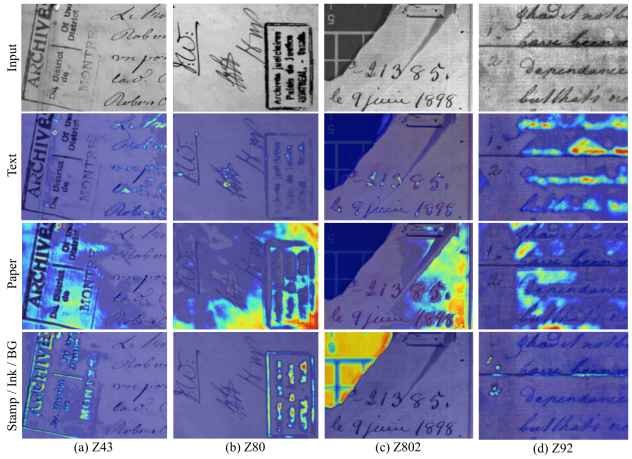


Figure 2. Visualization of the attention maps at the output of the attention block for different input images from MSTEx-2. The 1st row shows the input images. The following rows show characteristics attended at different iterations.

5. Implementation details

Our model is implemented using PyTorch 2.3.1 with CUDA 11.2 on Python 3.10.12. The model architecture incorporates non-negativity constraints using PyTorch's parametrize module, while pruning is implemented via the nn.utils.prune module. DW convolutions in the encoder use 3×3 kernels and are followed by BatchNorm2d and Elu.

For the MSTex 1 & 2 and Urban datasets, we employ the Adam optimizer with a learning rate of 1e-4 for a maximum training epoch of 200, subject to early stopping (patience=10). For the MSBin dataset, which contains larger images and thus requires more computational resources, we adjust the parameters to 150 epochs with a learning rate of 0.001, maintaining the same early stopping criterion. Our experiments revealed that the initial state has a significant impact on model convergence, particularly given the complexity of degradation in Historical Document Images. Input MS images are normalized to the range [0,1] using a MinMax scheme.

Based on our experimental, we set $\lambda_1 = 5 \times 10^{-2}$ for the mean squared error (MSE), $\lambda_2 = 5 \times 10^{-2}$ for the orthogonality constraint in Eq. (6) and $\lambda_3 = 5 \times 10^{-2}$. The initial number of abundances r_{init} was set to 12, corresponding to the number of spectral bands available in MSBin. This choice ensures an overdetermined or at least well-determined unmixing problem, as having at least as many spectral bands as endmembers is generally required for reliable spectral separation. We fixed $r_{min} = 2$, representing at least the background and text components. Both ASC and ANC were enforced using a softmax function with a temperature parameter $T = 0.5$, ensuring that abundances remain non-negative and sum to unity for each pixel. For the Urban dataset, which exhibits more overlapping abundance mixing, the orthogonality constraint λ_2 is relaxed to 5×10^{-3} while the temperature T is augmented to 2.

$$\text{softmax}(x_i) = \frac{\exp(x_i/T)}{\sum_{j=1}^{K} \exp(x_j/T)} \tag{10}$$

6. Similarity Measure

We recall the criterion we developed to perform a similarity measure between the i^{th} and j^{th} abundance maps and their corresponding endmembers as follows :

$$\mathcal{S}_{i,j} = \frac{1}{|\langle \mathcal{A}_i - \mathcal{A}_j, \mathbf{1} \rangle| + \epsilon} \frac{\langle \mathcal{A}_i, \mathcal{A}_j \rangle}{\|A_i\| \|A_j\|} + \lambda_{\text{SAD}'} \mathcal{L}_{\text{SAD}'}(E_i, E_j), \tag{11}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product, $\mathcal{A}_i$ is the i^{th} abundance map, which represents the smoothed matrix form of the i^{th} row of the abundance matrix $\mathbf{A}$, $\mathcal{L}_{\text{SAD}'}(\mathbf{E}_i, \mathbf{E}_j) = \frac{\pi}{2} - \text{acos}(\frac{\langle \mathbf{E}_i, \mathbf{E}_j \rangle}{\|\mathbf{E}_i\| \|\mathbf{E}_j\|})$ is the complementary of the spectral angular distance between the i^{th} and j^{th} columns of the endmembers matrix $\mathbf{E}$, $\| \cdot \|$ represents the Euclidean norm, $| \cdot |$ the absolute value, $\lambda_{\text{SAD}'} = 5 \times 10^{-3}$ is a hyperparameter added to scale the SAD′ metric, and ϵ is a small positive constant added to avoid division per zero. The smoothing kernel used is defined as :

$$K = \begin{bmatrix} 0.0369 & 0.0392 & 0.0400 & 0.0392 & 0.0369 \\ 0.0392 & 0.0416 & 0.0424 & 0.0416 & 0.0392 \\ 0.0400 & 0.0424 & 1.0000 & 0.0424 & 0.0400 \\ 0.0392 & 0.0416 & 0.0424 & 0.0416 & 0.0392 \\ 0.0369 & 0.0392 & 0.0400 & 0.0392 & 0.0369 \end{bmatrix}$$

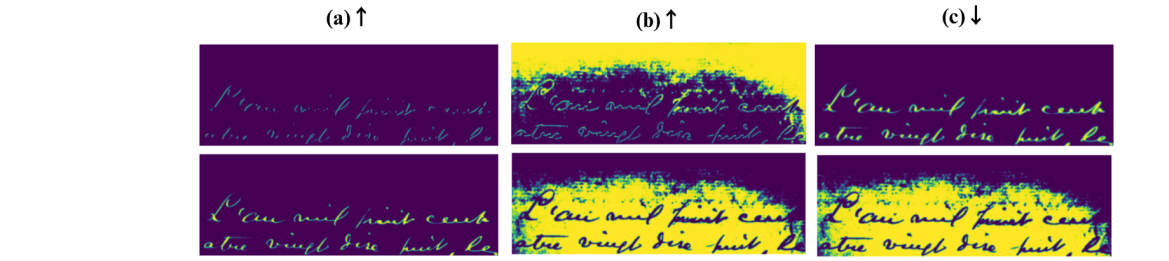


Figure 3. Ablation of the spatial component in the pairwise similarity measure. The figure shows three abundance map pairs: (a) and (b) represent cases where high similarity is desired, while (c) represents a case where low similarity is expected. The spatial component achieves this discrimination through: (1) Inverse scaling by the absolute of the activation sum differences between pairs $|\langle \mathcal{A}_i - \mathcal{A}_j, \mathbf{1} \rangle|$, which penalizes pairs with very different abundance distributions, and (2) Spatial smoothing of abundance by convolution with kernel K.

The proposed similarity criterion comprises two complementary components: spectral similarity for identifying materials with similar spectral signatures, and spatial similarity for comparing abundance distributions. As illustrated in Fig. 3, pure correlation fails to identify similarity in pair (a), which represents identical text regions, while yielding low similarity values for pair (b), despite both regions corresponding to the same background material. This limitation arises because these regions share minimal spatial overlap, with contact occurring only at boundary pixels. To address these deficiencies, we introduced the inverse scaling factor based on the absolute difference between abundance sums. This modification enhances the similarity measure for both pairs (a) and (b), while maintaining low similarity for pair (c), which represents genuinely dissimilar materials (text & paper). Furthermore, the incorporation of spatial smoothing mitigates the fragmentation artifacts observed in pair (a), resulting in appropriately high similarity scores for semantically related regions. This multi-component approach thus provides a more robust measure of material similarity that accounts for both spectral characteristics and spatial distribution patterns.

7. Additional Experimental Results

7.1. Decomposition for enhanced binarization

Full decomposition of the image scene has a significant impact on traditional tasks, such as image binarization. Typically, a well-segmented text image yields superior binarization performance. To demonstrate the impact of decomposition on the binarization task, we have included this ablation study. Specifically, we apply binarization directly to one of the set images for each MS cube using one of the SoTA binarization methods (Howe [2]), and we compare these results with those obtained after our decomposition process. This comparison allows us to illustrate the advantages of our approach in enhancing the performance of subsequent image-processing tasks. Due to the orthogonality regularization, the resulting abundance maps exhibit near-binary characteristics, enabling straightforward binarization through a winner-takes-all strategy where each pixel is assigned to the class with the maximum abundance value.

Table 1. Binarization difference on MSTEx 1 & 2, with and without decomposition.

Method	FM↑	DRD↓	NRM↓	PSNR↑
Binarization only (Howe) [2]	76.96	6.63	8.94	14.94
Decomposition + binarization **(Ours)**	**86.47**	**3.11**	**7.30**	**17.24**
Difference	**+9.51**	**−3.52**	**−1.64**	**+2.30**

Table 2. Binarization difference on MSBin BT.

Method	FM↑	DRD↓	NRM↓	PSNR↑
Binarization only (Howe) [2]	62.48	28.73	20.76	11.70
Decomposition + binarization **(Ours)**	**92.35**	**7.90**	**4.68**	**15.95**
Difference	**+29.87**	**−20.83**	**−16.08**	**+4.25**

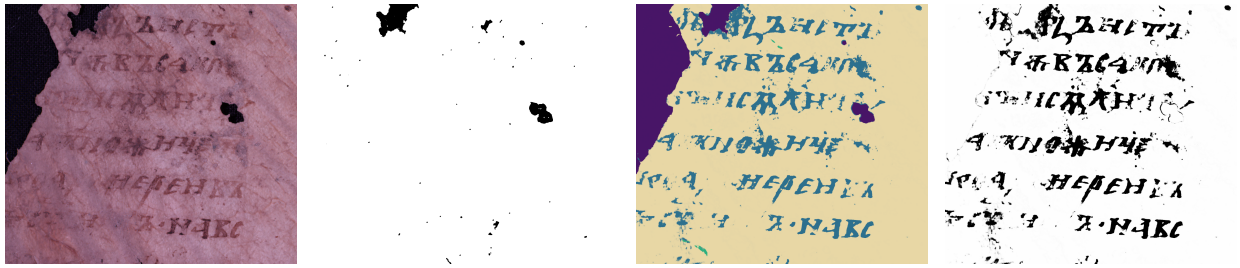
Table 3. Binarization difference on MSBin EA.

Method	FM↑	DRD↓	NRM↓	PSNR↑
Binarization only (Howe) [2]	14.06	51.69	45.02	8.88
Decomposition + binarization **(Ours)**	**69.77**	**32.84**	**18.01**	**10.76**
Difference	**+55.71**	**−18.85**	**−27.02**	**+1.89**

As seen on Tabs. 1 to 3, without decomposition, the binarization achieves lower performance compared to the approach that incorporates decomposition, especially for the highly degraded images such as in MSBin EA book. Moreover, the binarization-only approach involves a manual intervention to select the best suitable image for binarization.

7.2. Extended Visual Results for Text Extraction

Figure 4 shows the highly degraded sample EA58 from MSBin, where the Howe method fails to extract text content and instead misclassifies darker background areas as text regions. Despite some imperfections, PRISM succeeds in extracting it.



(a) Pseudo color Image (b) Howe [2] **(c) Ours (Decomposition)** **(d) Ours (Extracted text)**

Figure 4. Example of a highly degraded sample from MSBin EA book where the Howe method fails to identify any text.

The extended analysis of different methods on the BT56 sample demonstrates that most approaches face challenges in separating background from text content, especially in the degraded region visible in the top left corner. Furthermore, distinguishing between different ink colors (red & brown) proves problematic for the majority of methods.

(a) MS Cube (b) Pseudo color image

(c) OSPAEU [1] (d) Howe [2] (e) MS-Former [3]

(f) SKKHM [4] (g) Endnet [5] (h) DAEU [6]

(i) MTAEU [7] (j) CNNAEU [8] **(k) Ours**

Figure 5. Extended version of Fig. 1 (from paper) on sample BT56 from MSBin dataset. Rank was fixed to 4 for the different methods, except for Howe, which only extracts text and Ours, which uses iterative pruning.

7.3. Extended Results of Iterative Pruning on Urban Dataset

Following the method from [9], the RMSE between the generated abundances and the corresponding GT was computed, selecting the best result from 10 independent runs to mitigate initialization effects for the three scenarios. In this setup, for our method, the minimum rank possible r_{min} was fixed to four for each run, abundance maps corresponding to each GT scenario being saved. Results are displayed below in Tabs. 4 to 6.

Table 4. Quantitative comparison with HS unmixing methods on the Urban dataset for SIX elements. Best results are in **bold** and the second best in **blue**.

Metric	Element	CNNAEU [8]	Endnet [5]	DAEU [6]	OSPAEU [1]	MTAEU [7]	Ours
RMSE ↓	Asphalt	0.2270	0.1528	**0.1322**	0.2994	**0.1517**	0.1786
	Grass	0.3622	0.2141	0.2352	**0.1782**	**0.1862**	0.2080
	Tree	0.1972	**0.0939**	0.1492	**0.1358**	0.1152	0.1492
	Roof	0.1252	0.1060	**0.0915**	0.1460	0.1152	**0.0852**
	Soil	0.2096	0.2017	0.2195	0.2354	**0.1395**	**0.1732**
	Metal	0.1710	0.2508	**0.1606**	**0.0844**	0.1808	0.1881
	Average	0.2154	0.1699	0.1647	0.1847	**0.1515**	**0.1637**

Table 5. Quantitative comparison with HS unmixing methods on the Urban dataset for FIVE elements. Best results are in **bold** and the second best in **blue**.

Metric	Element	CNNAEU [8]	Endnet [5]	DAEU [6]	OSPAEU [1]	MTAEU [7]	Ours
RMSE ↓	Asphalt	0.2499	**0.1102**	**0.1266**	0.3159	0.1295	0.2221
	Grass	0.2563	**0.1688**	0.1856	0.2064	**0.1620**	0.2113
	Tree	0.2022	**0.1082**	0.1169	0.1644	**0.1105**	0.1197
	Roof	0.1212	0.0870	0.1022	0.1563	**0.0693**	**0.0591**
	Soil	0.2641	**0.1534**	0.1815	0.2410	**0.1157**	0.2326
	Average	0.2187	**0.1255**	0.1426	0.2168	**0.1117**	0.1689

Table 6. Quantitative comparison with HS unmixing methods on the Urban dataset for FOUR elements. Best results are in **bold** and the second best in **blue**.

Metric	Element	CNNAEU [8]	Endnet [5]	DAEU [6]	OSPAEU [1]	MTAEU [7]	Ours
RMSE ↓	Asphalt	0.2369	**0.1084**	0.1703	0.3028	0.1426	**0.0948**
	Grass	0.2756	0.1660	0.1678	0.2688	**0.1346**	**0.1562**
	Tree	0.2070	0.1019	**0.0762**	0.2134	**0.0951**	0.1252
	Roof	0.1876	**0.0845**	0.0867	0.2876	0.0904	**0.0560**
	Average	0.2268	**0.1152**	0.1253	0.2682	0.1157	**0.1081**

PRISM demonstrates competitive performance compared to state-of-the-art hyperspectral unmixing methods on this dataset, with particularly strong results for the roof abundance maps. While PRISM does not achieve the lowest RMSE across all material classes, it consistently outperforms the CNNAEU baseline method that served as its architectural foundation. Higher RMSE, especially for the Asphalt and Soil components, can be attributed to similarities between those two materials, PRISM classifying some off-roads as Asphalt rather than Soil. This confusion is mitigated when these components are combined into a single material in the last GT scenario, with PRISM achieving its best quantitative results.

7.4. Extended Ablation on MDL Cost

An extended ablation study highlights the effectiveness of the Minimum Description Length (MDL) principle. As illustrated in Fig. 6, relying solely on the loss function leads to an overcomplete model, with the best results clustering around suboptimal Ranks 4, 5, and 6. Qualitatively, these higher-rank solutions yield components with fragmented text materials, indicating a less meaningful decomposition. In contrast, incorporating the total MDL cost not only improves the convergence properties but also guides the model to a more robust solution. The MDL cost correctly identifies a stable and acceptable range of solutions around Ranks 2, 3, and 4, with a clear global minimum at Rank 3. This confirms that the MDL cost is crucial for preventing the model from overfitting and for correctly determining the true number of latent components in the data.

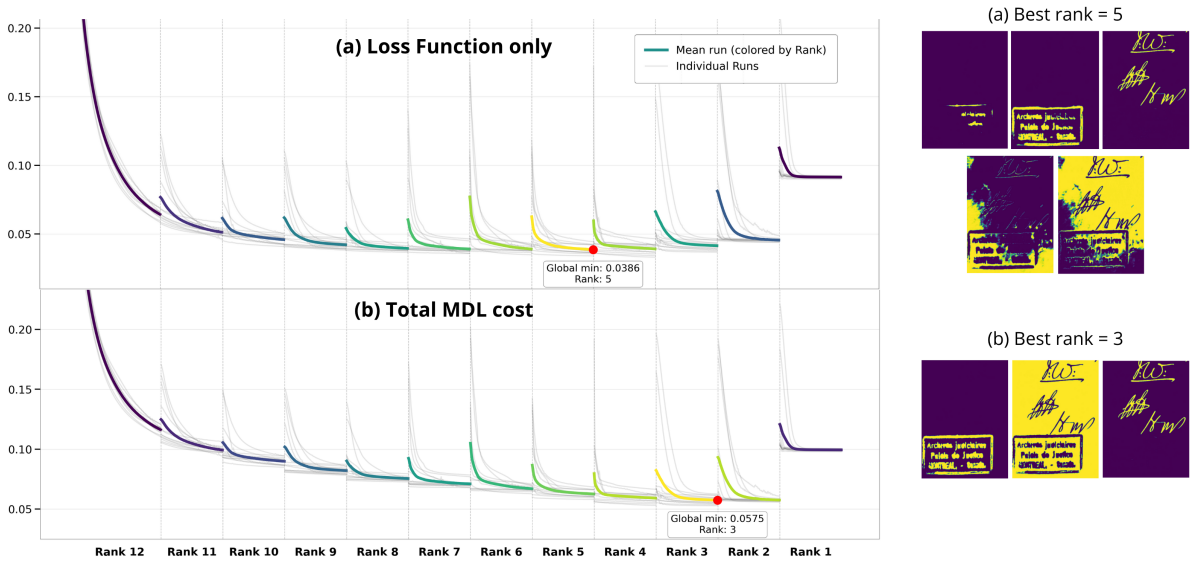


Figure 6. Ablation on rank selection with and without the MDL Cost.

References

[1] Zeyang Dou, Kun Gao, Xiaodian Zhang, Hong Wang, and Junwei Wang. Hyperspectral unmixing using orthogonal sparse prior-based autoencoder with hyper-laplacian loss and data-driven outlier detection. *IEEE TGRS*, 58(9):6550–6564, 2020. 5, 6

[2] Nicholas R Howe. Document binarization with automatic parameter tuning. *International journal on document analysis and recognition (IJDAR)*, 16:247–258, 2013. 4, 5

[3] Sanaz Karimijafarbigloo, Reza Azad, Amirhossein Kazerouni, and Dorit Merhof. MS-former: Multi-scale self-guided transformer for medical image segmentation. In *Medical Imaging with Deep Learning*, 2023. 5

[4] Qi Li, Nikolaos Mitianoudis, and Tania Stathaki. Spatial kernel k-harmonic means clustering for multi-spectral image segmentation. *IET Image Processing*, 1(2):156–167, 2007. 5

[5] Savas Ozkan, Berk Kaya, and Gozde Bozdagi Akar. Endnet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing. *IEEE TGRS*, 57(1):482–496, 2019. 5, 6

[6] Burkni Palsson, Jakob Sigurdsson, Johannes R. Sveinsson, and Magnus O. Ulfarsson. Hyperspectral unmixing using a neural network autoencoder. *IEEE Access*, 6:25646–25656, 2018. 5, 6

[7] Burkni Palsson, Johannes R. Sveinsson, and Magnus O. Ulfarsson. Spectral-spatial hyperspectral unmixing using multitask learning. *IEEE Access*, 7:148861–148872, 2019. 5, 6

[8] Burkni Palsson, Magnus O Ulfarsson, and Johannes R Sveinsson. Convolutional autoencoder for spatial-spectral hyperspectral unmixing. In *IEEE IGARSS*, pages 357–360, 2019. 5, 6

[9] Burkni Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. Blind hyperspectral unmixing using autoencoders: A critical comparison. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1340–1372, 2022. 1, 6