# SWIR ハイパースペクトル画像における廃棄物セグメンテーションのための直交性制約付き光学フィルタ学習

WANG Songchao　【 画像情報工学研究室 】

# Learning Optical Filters with Orthogonality Constraints for SWIR-Based Waste Segmentation in Hyperspectral Images

WANG Songchao　【 Image Processing and Informatics Lab. 】

## 1 Introduction

Automated waste sorting requires material-level recognition because objects with similar appearances can consist of different materials. Hyperspectral imaging provides dense spectral information that supports material discrimination, particularly in the shortwave infrared (SWIR) range. However, directly acquiring and processing hundreds of bands increases sensor complexity, memory use, computation, and the amount of training data required.

This study learns a small set of optical transmission responses from hyperspectral data. The objective is to compress a 224-band SWIR patch into a compact representation while preserving information useful for semantic segmentation. Unlike principal component analysis (PCA) [3], which maximizes input variance, the proposed filters are optimized using material labels. Their transmission curves are also encouraged to be non-negative, smooth, and mutually complementary.

## 2 Proposed Method

The proposed framework has two stages. In Stage 1, 2048 pixel spectra are sampled from each hyperspectral patch. A set of $K$ optical filters projects each 224-dimensional spectrum to $K$ responses, which are classified by a lightweight multilayer perceptron. The raw filter parameters are clamped to non-negative values and normalized to a maximum transmission of one. For spectrum $\mathbf{x}$, filter $k$ produces

$$r_k = \sum_{c=1}^{224} w_{k,c} x_c. \tag{1}$$

The filters and classifier are jointly optimized by

$$\mathcal{L}_{\mathrm{S1}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{smooth}} \mathcal{L}_{\mathrm{smooth}} + \lambda_{\mathrm{ortho}} \mathcal{L}_{\mathrm{ortho}}, \tag{2}$$

where second-order smoothness suppresses rapidly oscillating responses. After $L_2$ normalization of each realized transmission curve, the orthogonality term is

$$\mathcal{L}_{\mathrm{ortho}} = \sum_{i \neq j} \left( \widehat{\mathbf{w}}_i^{\mathsf{T}} \widehat{\mathbf{w}}_j \right)^2, \qquad \widehat{\mathbf{w}}_i = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}. \tag{3}$$

This soft penalty encourages complementary spectral responses; it does not impose exact orthogonality as a hard equality constraint.

In Stage 2, the learned filters are fixed and applied densely to each hyperspectral patch to obtain $K$ filter-response maps. SegFormer-B0 [2] then learns spatial context and predicts the segmentation mask. This separation provides direct material supervision to the spectral projection before spatial segmentation is learned.

## 3 Experimental Setup

Experiments use the SpectralWaste dataset [1], which contains 224-band hyperspectral patches spanning approximately 900–1700 nm. The task includes background and six foreground classes: film, basket, cardboard, video tape, filament, and bag. Stage 2 comparisons use the same SegFormer-B0 backbone, augmentation, optimizer, learning-rate schedule, class weighting, and model-selection rule. Performance is evaluated using foreground mean intersection over union (mIoU), precision, recall, F1, and pixel accuracy.

Three principal input representations are compared: the raw 224-band patch (Full-band), a fixed three-component PCA projection computed from the training split (PCA-3), and three learned filter responses (Ours). Additional experiments compare $K = 2, 3, 4$, and 5, remove orthogonality regularization, and evaluate Stage 1 with grouped cross-validation.

## 4 Results and Discussion

Table 1 summarizes the Stage 2 test results. The proposed three-filter representation achieves the highest foreground mIoU, precision, F1, and pixel accu-

Table 1　Stage 2 test results under matched settings.

| Input | mIoU | Precision | Recall | F1 | Pixel Acc. |
|---|---|---|---|---|---|
| Full-band | 0.500 | 0.549 | **0.829** | 0.644 | 0.866 |
| PCA-3 | 0.532 | 0.594 | 0.818 | 0.671 | 0.884 |
| Ours ($K = 3$) | **0.561** | **0.634** | 0.817 | **0.699** | **0.887** |

racy. Its mIoU is 0.561, compared with 0.500 for Full-band and 0.532 for PCA-3. Although PCA-3 preserves 99.78% of sampled training variance, its components are not optimized for material labels. This indicates that preserving global variance is not equivalent to preserving task-discriminative information. Per-class IoU also shows that the learned responses obtain the best result for five of the six foreground categories, with the largest gain over PCA-3 observed for filament.

The three learned transmission curves differ clearly from the PCA-3 projection coefficients. The learned curves are non-negative and smooth, whereas PCA coefficients contain positive and negative values and have arbitrary signs. More importantly, the proposed curves are optimized with class supervision rather than only spectral variance.

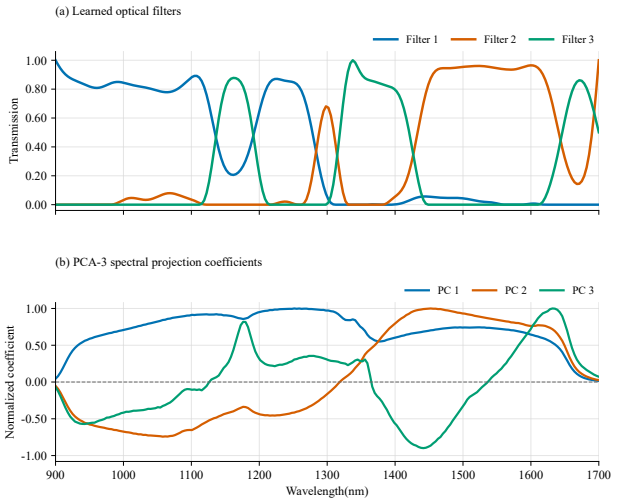



Fig. 1　Learned transmission curves and normalized PCA-3 projection coefficients.

Under one shared hyperparameter setting, $K = 3$ gives the best Stage 2 performance (Table 2). The result does not establish that three filters are universally optimal because the regularization coefficients were not independently tuned for every $K$. It shows that $K = 3$ provides the best balance of compactness and performance in the tested configuration.

Removing orthogonality changes Stage 2 mIoU only from 0.561 to 0.557. Its main effect is instead structural: the mean pairwise cosine similarity between learned curves decreases from 0.596 without the term to 0.097 with it. Thus, orthogonality regularization substantially reduces spectral redundancy, although its direct accuracy gain is small.

Table 2　Effect of the number of filters.

| | Stage 1 validation | | Stage 2 test | |
|---|---|---|---|---|
| $K$ | mIoU | F1 | mIoU | F1 |
| 2 | 0.281 | 0.405 | 0.481 | 0.626 |
| 3 | 0.342 | 0.487 | **0.561** | **0.699** |
| 4 | 0.346 | 0.486 | 0.465 | 0.611 |
| 5 | **0.347** | **0.489** | 0.504 | 0.644 |

Five-fold grouped cross-validation for $K = 3$ gives Stage 1 foreground mIoU $0.320 \pm 0.025$ and F1 $0.460 \pm 0.029$. After permutation alignment, the mean correlation between corresponding filter shapes is 0.761. The filters repeatedly emphasize related spectral regions, although exact peaks vary with the acquisition groups. Three-fold experiments for $K = 4$ and $K = 5$ also show stable Stage 1 metrics but lower shape consistency, supporting the compact $K = 3$ configuration when repeatability is considered.

## 5　Conclusion

A two-stage method was developed to learn compact optical responses for SWIR-based waste segmentation. Three learned filters reduce 224 spectral bands to three response maps and outperform both Full-band and PCA-3 inputs under matched Stage 2 settings. Orthogonality mainly improves complementarity among the learned curves, while grouped validation shows recurring response structure across data partitions. Future work will investigate optical fabrication, calibration, and evaluation with a physical SWIR imaging system.

## 参考文献

[1] S. Casao et al., "SpectralWaste Dataset: Multimodal Data for Waste Sorting Automation," *Proc. IEEE/RSJ IROS*, pp. 5852–5858, 2024.

[2] E. Xie et al., "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *Advances in Neural Information Processing Systems*, 2021.

[3] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.