SpectralWaste Dataset: Multimodal Data for Waste Sorting Automation

Sara Casao[1,*] Fernando Peña[1,*] Alberto Sabater[1]
Rosa Castillón[2] Darío Suárez[1] Eduardo Montijano[1] Ana C. Murillo[1]

***Abstract*— The increase in non-biodegradable waste is a worldwide concern. Recycling facilities play a crucial role, but their automation is hindered by the complex characteristics of waste recycling lines like clutter or object deformation. In addition, the lack of publicly available labeled data for these environments makes developing robust perception systems challenging. Our work explores the benefits of multimodal perception for object segmentation in real waste management scenarios. First, we present SpectralWaste, the first dataset collected from an operational plastic waste sorting facility that provides synchronized hyperspectral and conventional RGB images. This dataset contains labels for several categories of objects that commonly appear in sorting plants and need to be detected and separated from the main trash flow for several reasons, such as security in the management line or reuse. Additionally, we propose a pipeline employing different object segmentation architectures and evaluate the alternatives on our dataset, conducting an extensive analysis for both multimodal and unimodal alternatives. Our evaluation pays special attention to efficiency and suitability for real-time processing and demonstrates how hyperspectral imaging can bring a boost to RGB-only perception in these realistic industrial settings without much computational overhead.**

I. Introduction

The global issue of waste production intensifies as societies grow and consumption rises. The sheer volume of waste generated, particularly non-biodegradable waste such as plastics, has reached concerning proportions. Recycling and reuse are key strategies to lessen the environmental burden of waste. Hence, automating waste management facilities not only increases the volume of properly processed waste but also safeguards worker health and comfort. To achieve automated manipulation of relevant elements in these real industrial environments, the first step is to improve and adapt existing perception systems to this domain.

Despite the great advances in automated visual recognition tasks in recent years, real industrial settings often present recurring problems that hinder real-world applicability, such as the lack of annotated data to achieve precise domain adaptation or high computational requirements [1]. The most common method to identify and localize elements of interest in automated tasks is through the segmentation of RGB images [2]. However, accurate detection based only on visual features is extremely challenging in waste management scenarios with severe clutter, high materials diversity, deformable or broken objects, and translucent elements (see sample images in Figure 2). To overcome these issues, the use of more complex sensing modalities like hyperspectral imaging (HSI), commonly used for raw material classification [3], can provide insights beyond the visual appearance of objects. While conventional RGB cameras capture the visible spectrum, hyperspectral cameras are able to acquire light across a wide range of wavelengths. Thus, leveraging the combined information from both modalities improves the performance of complex perception tasks such as segmentation of buildings [4] or terrain classification [5]. Unfortunately, the advantages derived from using hyperspectral information along with RGB images remain unexplored in object recognition for automatic waste sorting, where the task is mostly approached using RGB information [6].

This work demonstrates the benefits of using multimodal segmentation approaches in waste management and contributes to mitigating two key challenges that currently hinder their adoption: the scarcity of public multimodal waste datasets and the high computational demands associated with HSI data. Specifically, our main contributions are twofold: (1) We introduce SpectralWaste[1], the first multimodal dataset obtained from an operational waste sorting facility, featuring in-the-wild industrial data from both hyperspectral and RGB cameras (Figure 1). This dataset addresses the identification of critical objects that frequently appear in real trash flows and impact sorting efficiency by either clogging machinery if not removed or holding value if recovered. (2) We present a comprehensive object segmentation analysis that underscores the boost in performance when combining both modalities and, for the first time, explores the suitability of using HSI for object segmentation in waste sorting scenarios. The proposed pipeline places particular emphasis on employing efficient architectures. Furthermore, to ensure consistency in annotated masks between modalities and reduce the labeling effort required, we propose a novel label transfer algorithm that automatically adapts RGB-annotated masks to HSI without any calibration needed.

II. Related Work

This section summarizes existing works regarding waste datasets for object identification, and image segmentation methods using both hyperspectral imaging exclusively and multimodal information.

*Authors contributed equally to this work.
[1]I3A, Universidad de Zaragoza, [2]ATRIA Innovation

This work was supported by DGA projects T45_23R and T58_23R, and by MCIN/AEI/ERDF/European Union NextGenerationEU/PRTR projects PID2021-125514NB-I00 and PID2022-136454NB-C22.

[1]Dataset website: https://sites.google.com/unizar.es/spectralwaste

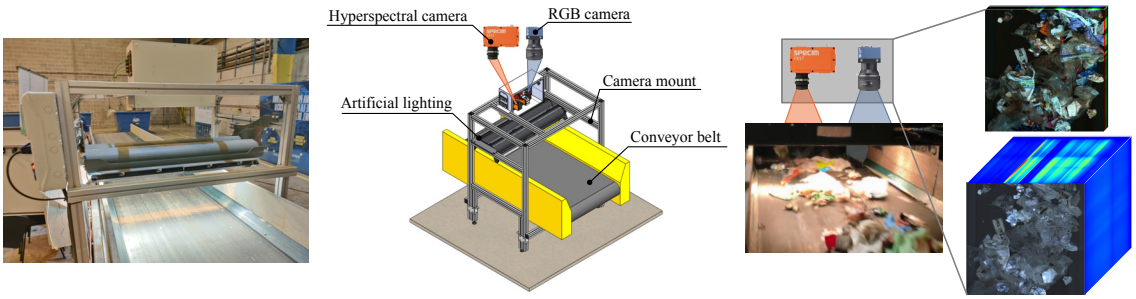


Fig. 1: The setup in the waste sorting facility contains two synchronized line-scan cameras (RGB and hyperspectral). *Left:* Real prototype installed in the facility. *Middle:* Diagram of the setup for data capture. *Right:* Example of a scene captured by both cameras (HSI with false-color representation).

A. Waste Object Identification Datasets

Numerous datasets spanning diverse domains have focused on collecting data for waste identification. Stanford TrashNet [7] consists of classifying single objects on an empty white background. Aiming to tackle the localization problem in addition to classification, Trash Annotations in Context (TACO) [8] presents a dataset in open and real environments, such as streets, lakes, or beaches where world litter is shown. Following the idea of reducing waste in natural environments, Floating Waste (FloW) [9] is dedicated to the efficient cleaning of inland water areas with autonomous boats. Due to the demands of this task, a multimodal subdataset, FloW-RI, is included, providing millimeter wave radar data synchronized with the images. Similarly, other works aim to tackle challenging tasks by providing complementary information alongside RGB images. For instance, [10] introduces a multimodal dataset comprising RGB-D, thermal infrared and object poses to address the issue of transparent object identification.

Closer to our work is the ZeroWaste dataset [6] and its extension ZeroWaste-v2, proposed in the VisDA2022 challenge [11]. In these papers, they provide a dataset comprised of conventional RGB images for industrial waste object segmentation that have been collected from a real sorting plant. In contrast, our dataset includes hyperspectral data synchronized with the RGB images. Thus, this spectral information combined with data from the visible spectrum holds significant potential for enhancing object identification tasks within the recycling processes.

B. Identification with Hyperspectral Data

A wide variety of techniques have been explored to leverage hyperspectral sensors for raw material identification [12], [13]. However, the limited amount of data in the existing HSI datasets poses a challenge for training data-based models. Common methods involve pixel-wise training and testing on a reduced set of images [14], leading to information leakage and suboptimal generalization capabilities [15].

Final applications with HSI typically focus on fields where the information captured by RGB cameras lacks sufficient detail, e.g., environmental monitoring, agriculture, medical imaging, or remote sensing [16]. More specifically, the use of hyperspectral sensors for automatic plastic sorting in recycling facilities is a widespread technique. For example, the structure of plastic material is analyzed by sparse pixels [3], or hyperspectral imaging is used to densely label images based on per-pixels classifications [17]. Unlike these works focused on raw material identification, we study the use of HSI for object segmentation, which can potentially overcome the limitations of RGB data in waste sorting tasks by leveraging spectral properties for material differentiation.

Regarding multimodal sensors for segmentation tasks, multiple works have addressed this problem through sensor fusion [18]. In particular, the combination of HSI with RGB information has been used in multiple fields. For instance, combining HSI with RGB among other modalities is exploited in environmental monitoring with UAV's [19] and autonomous terrain classification [5]. Moreover, perception works combine both modalities for different tasks like classification of building materials [4] or rise seeds inspection [20]. However, the exploration of multimodal segmentation in real-world industrial waste sorting scenarios remains an open area of research.

III. SpectralWaste Dataset

This section describes the data acquisition and the annotation process of the novel multimodal SpectralWaste dataset.

A. Data Acquisition

The dataset was collected in a real waste sorting industry specialized in plastics, cartons and cans, with a true-to-life prototype of the conveyor belt installed on the waste separation line (see Figure 1). This prototype closely mimics the real installation, ensuring that the captured waste streams accurately mirror those arriving at the facility for separation.

The setup involved two synchronized cameras for multimodal data capture: a line-scan RGB camera (Teledyne DALSA Linea) and line-scan hyperspectral sensor (Specim FX17) that captures 224 contiguous spectral bands in a range from 900 to 1700 nm. Both cameras were housed in an industrial enclosure situated at a height of 1.7 m. This installation was supplemented with a set of LED illuminators and infrared halogen lighting to ensure a suitable image capture in the spectral domain. RGB images were stored with a resolution of 1200×1184 pixels and 8-bit color depth, while HSI images were stored as data cubes of size $600 \times 640 \times 224$ with 16-bit precision.

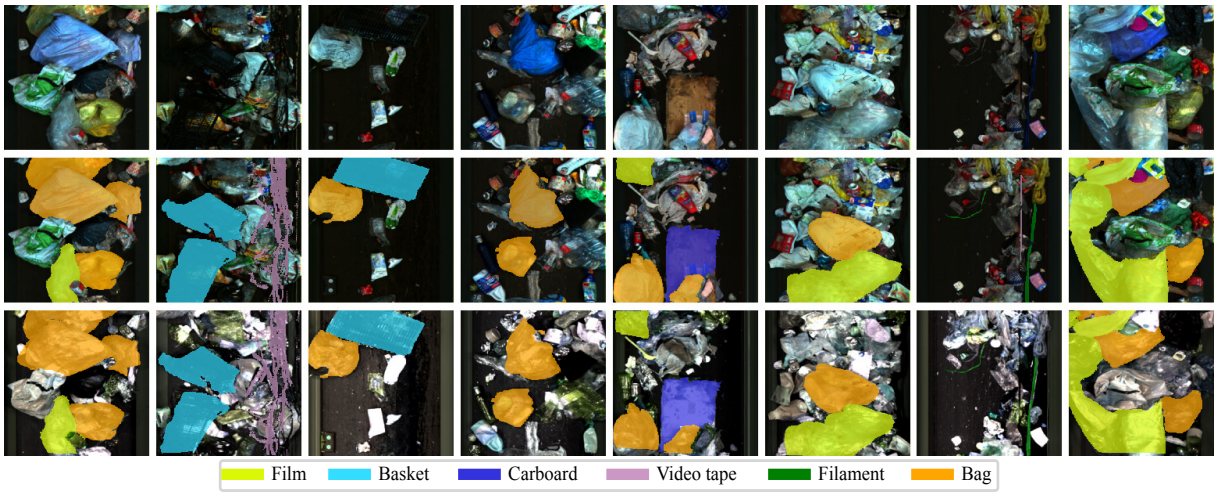


Fig. 2: Examples of images included in the dataset. *First row*: RGB images. *Second row*: ground-truth RGB annotations. *Third row*: hyperspectral images annotated with labels obtained with the proposed label-transfer algorithm (three hyperspectral bands have been manually selected for visualization).

B. Data Annotation

The classes chosen for annotation in the presented dataset were selected according to the requirements of the facility. Labeled objects represent elements that commonly cause operational problems in recycling lines, impacting the efficiency of the sorting process. Among these problems, machinery jams pose a significant issue, causing a complete stoppage of the waste separation until the obstructing object is removed. Thus, the selected objects for automatic identification include *film* and *basket*, large objects that can clog the conveyor belts as they are not easily breakable; *video tape* and *filament*, representing long objects prone to entangle in waste separation zones and requiring manual intervention; *trash bag*, which encompasses closed bags containing waste that need to be mechanically opened for further processing; and *cardboard*, paper objects received at the facility whose recovery adds value sending it to another recycling process.

To streamline the time-consuming labeling process, we developed an interactive segmentation tool leveraging the point-prompt feature from the Segment Anything Model (SAM) [21]. In essence, the user can select points belonging to an object which generates a new mask displayed over the image for further refinements or saving. In this work, we used our tool to manually create the ground truth masks of the defined objects in the RGB image set.

C. Dataset Content

The result of the entire data acquisition and annotation process is encompassed in the SpectralWaste dataset. The dataset provides annotations for six object classes: *film, basket, video tape, filaments, trash bags*, and *cardboard*, totaling 2059 annotated instances across a set of 852 non-overlapping images. Table I presents the overview of the annotations, while Figure 2 and 6 illustrate sample images.

In addition to the labeled set, SpectralWaste contains 6803 unlabeled multimodal images (RGB-HSI). We believe that releasing these unlabeled images is valuable for the community, enabling the researchers to explore the advantage of hyperspectral or multimodal object identification in an industrial waste facility through further study of different techniques. These techniques may involve refining labeled segmentation with semi-supervised or self-supervised methods and exploring unsupervised segmentation approaches.

TABLE I: Summary of the instances annotated in SpectralWaste.

	Instances per class					
Total	*Film*	*Basket*	*Card.*	*Tape*	*Filam.*	*Bag*
2059	339	300	68	287	111	954

IV. Waste Segmentation

This section describes the proposed pipeline for waste object segmentation using RGB and HSI images. Figure 3 summarizes the key steps of the process where we consider different configurations that can take one or both modalities. A detailed description of the steps involved in the baselines and the adapted architectures is provided in the following.

A. Data Preprocessing

General preprocessing: The pipeline includes a series of common preprocessing steps. We first crop the mismatched areas captured by each sensor to align the space shown in each image, resize the RGB and HSI images to 256×256 pixels.

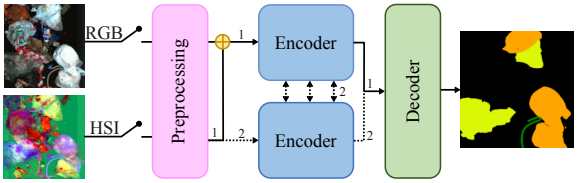


Fig. 3: Pipeline of the segmentation process. *Path 1*: data flow for the architectures designed for unimodal RGB segmentation (MiniNet-v2 and SegFormer). *Path 2*: HSI data flow in the multimodal CMX architecture.

HSI channel reduction: Hyperspectral imaging imposes significant computational demands due to the large amount of information contained in each pixel. Aiming to explore efficient solutions within our pipeline, we implement dimensionality reduction across the spectral channels using principal component analysis (PCA) to select the three main components (HYPER3). This PCA analysis is conducted on the pixel values from the training set of the dataset and then applied to the validation and testing sets to obtain their reduced version. The reduction process maintains 99.7% of the explained variance over the training set. By reducing the HSI channels to three components, we not only obtain a compressed representation of the information but also provide a balanced input when combining HSI with RGB images in the multimodal configurations.

B. Segmentation Architectures

Considering the application addressed in this work of segmenting objects in an industrial setting, we compare three different architectures within our pipeline paying special attention to alternatives suitable for real-time inference.

The first architecture is *MiniNet-v2* [22], a lightweight convolutional neural network designed for segmentation tasks. This model uses multi-dilation depthwise separable convolutions and two convolutional branches instead of skip connections to achieve a favorable trade-off between accuracy and computation. The second architecture is *SegFormer* [23], a well-known transformer-based segmentation network. Specifically, we opted for the version with the smallest encoder (SegFormer-B0) as it is reported to be suitable for real-time environments and closer to MiniNet-v2 in terms of computational requirements and number of parameters. Since both MiniNet-v2 and SegFormer were originally designed for processing only RGB images, we adapt them for multimodal segmentation by early-fusing both modalities. This early fusion involves concatenating the RGB and HSI images across the channel dimension before feeding them into the network (see path 1 in Figure 3). Finally, we evaluate *CMX* [24], a hybrid-fusion transformer based on SegFormer. This network receives RGB and HSI data separately, processing them with two encoders that share information throughout the network (see path 2 in Figure 3). This model also integrates feature rectification techniques to mitigate the effects of noisy measurements from different modalities, which is crucial in our case. It is noteworthy that, while CMX has been previously assessed with multiple modalities (depth, polarization, event and LiDAR), including multispectral ones (thermal and infrared) [25], it has not been evaluated on HSI data before.

In summary, our pipeline offers flexibility in handling different data types. MiniNet-v2 and SegFormer can be used for processing individual RGB or HSI images (unimodal inputs). Additionally, all three architectures, i.e., MiniNet-v2, SegFormer, and CMX, can be employed for multimodal segmentation combining RGB and HSI data.

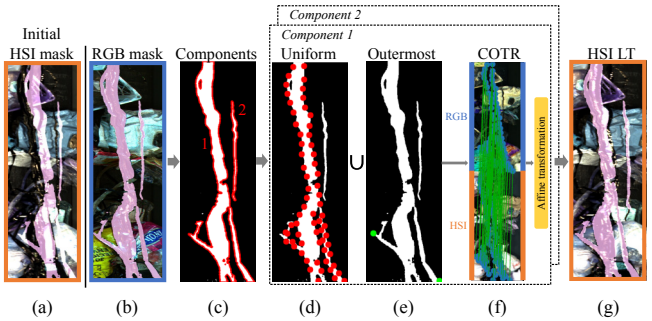


Fig. 4: Visualization of the main steps of the proposed label transfer method. (a) initial mask on HSI image (b) manually annotated mask on RGB image; (c) mask components; (d) uniformly-sampled points; (e) outermost points; (f) COTR matching for selected points in RGB to the HSI image used to compute an affine transformation for the mask; (g) resulting mask on HSI.

C. Label Transfer

To explore the unimodal HSI configurations of our pipeline, we address the challenge of data misalignment, i.e., the inaccurate alignment of manually RGB-annotated masks to the corresponding objects in HSI. This effect is due to the physical distance between the cameras and the heterogeneous internal settings of each sensor, which causes the perspective captured by each camera to differ. Since the employed cameras are non-conventional (line-scan cameras) and we only have two views with no additional spatial information such as depth, the well-established calibration methods for pinhole cameras are not applicable [26]. Therefore, to ensure consistency in annotated masks between modalities while improving labeling efficiency, we propose a novel label transfer algorithm. Our approach automatically adapts annotated segmentation masks from one camera (RGB) to another (HSI), relying exclusively on both images. The objective is to find affine transformations per mask that adapt the existing segmentation to the size and perspective of the corresponding object in the target image.

The different steps of the algorithm are visualized in Figure 4. First, we extract the contours of the segmentation mask and process each connected component independently. For example, in Figure 4(c) two different local affine transformations are computed, one per component. Then, we sample each contour to obtain a sparse representation of the shape, generated by uniformly-sampled points (Figure 4(d)) and always including the outermost points for better coverage (Figure 4(e)). The next step involves identifying the corresponding points in the target image. To accomplish this, we leverage the feature matching system COTR [27], which uses a transformer-based model to find the point on a target image that corresponds to a query point on a source image (Figure 4(f)). Finally, we obtain the affine transformation from both sets of points and apply it to the component. The final mask is obtained by combining all the transformed components. Figure 4(g) illustrates the resulting mask transferred with the proposed algorithm, which compared to the initial one (Figure 4(a)), shows a significant improvement in fitting the object.

V. Experiments

This section presents several experiments to evaluate the waste segmentation pipeline (Section IV) in the Spectral-Waste dataset (Section III).

A. Experimental Settings

Training configuration: The training of our pipeline follows the recommendations from the original works [22], [23], [28]. MiniNet-v2 is trained using Adam with an initial rate of 1×10^{-3}, a polynomial schedule with the power set to 0.9 and a weight decay of 1×10^{-4}. SegFormer models are trained using AdamW with an initial learning rate of 1×10^{-3} and a polynomial scheduler with the power set to 0.1. The selected CMX architecture is based on SegFormer and is trained using AdamW with an initial learning rate of 1×10^{-3} and a polynomial schedule with the power set to 0.9. In all cases, the loss is calculated using the cross-entropy function and the models are trained with batch size 12 for 200 epochs. We weight the classes in the loss calculation using median pixel-level frequency balancing to account for class imbalance. Regarding data augmentation, the training set is augmented with random rotations of ± 30 degrees and random vertical and horizontal flips.

Metrics: To evaluate the implemented architectures, preprocessing steps, and fusion methods, we compute the intersection over union (IoU) per class. As the final metric, we report the mean over all classes (mIoU).

B. Label Transfer Evaluation

To obtain a reliable evaluation of the proposed algorithm for transferring labels, we manually annotate 20 hyperspectral images with 81 object instances, ensuring that all the classes appear in the set.

Table II presents the results of the proposed label transfer approach (LT) in comparison to the base manual alignment (MA). The manual alignment process involves cropping the excess image captured by each camera to align the space shown and resizing them to the same shape. The evaluation of both methods is based on the intersection over union (IoU) of the resulting masks with the set of 81 instances manually labeled in the hyperspectral images. In the case of big objects, i.e., film, basket, cardboard, and trash bag, label transfer demonstrates an improvement on every class ranging between 9.7 and 17.2 percentage points compared to the manual alignment. On the more complex classes with thin objects, i.e., video tape and filaments, the IoU reaches 57.1% and 81.2% respectively, while the manual alignment baselines stay below 28%. Regarding the mean IoU over all classes, the proposed label transfer method achieves a mIoU of 79.6%, improving the baseline mIoU by 23.3 percentage points.

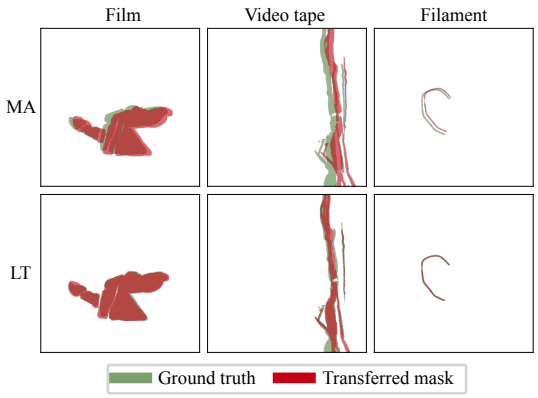


Fig. 5: Qualitative results of the label transfer evaluation. *First row*: manual alignment (MA). *Second row*: proposed label transfer (LT).

Figure 5 shows qualitative results from both algorithms for the challenging thin classes (video tape and filaments) and the largest class in the dataset (film). The ground truth is shown in green and the resulting masks in red. The first row corresponds to MA and the second row to LT alignment. The visualization shows how the masks transferred with the proposed label transfer algorithm match significantly better the ground truth than the manual alignment process.

In addition, we also evaluate the impact of training the unimodal models, MiniNet-v2 and SegFormer, with the annotations resulting from each of the alternative methods, MA and LT, as labels. Table III summarizes this study. The results demonstrate a higher mIoU using LT than MA in every configuration analyzed. Thus, we validate that our label transfer approach generates masks that better fit the objects in the target images, hyperspectral in our case, by reducing the noise of the segmentation annotations.

C. Segmentation Architectures Evaluation

Unimodal segmentation analysis: First, we analyze the potential of using hyperspectral data for the object segmentation task by evaluating MiniNet-v2 and SegFormer with hyperspectral or RGB data. The results, shown in Table IV, demonstrate the value of using HSI information (HYPER) for object segmentation, achieving a higher mean intersection over union (mIoU) than conventional RGB images in both architectures. Another relevant aspect to note arises when comparing the results using all hyperspectral bands directly as input (HYPER) with the PCA reduction to three channels (HYPER3). The good results achieved with the reduced input confirm the high variance covered with just three components. Figure 6 shows a qualitative comparison between the different architectures studied. For instance, note how the filament (fourth row and green mask) is barely identified by SegFormer using RGB data while the resulting mask using HSI is highly accurate for this challenging class.

Multimodal segmentation analysis: The assessment of the multimodal evaluation is summarized in Table IV, the CMX architecture with hybrid fusion outperforms the early-fusion baselines. Considering that the images from both modalities are not perfectly aligned, it is reasonable that hybrid feature fusion smooths out the noise and leverages the combination of data better than the early-fusion methods.

Efficiency analysis: This evaluation is conducted on a computer equipped with an AMD Ryzen 9 5950X CPU and a NVIDIA GeForce RTX 4090 GPU. Table IV examines the

TABLE II: Evaluation of the annotations alignment between modalities of manual alignment (MA) and automatic label transfer (LA). Results are computed with the IoU of the obtained masks with a set of 81 instances manually labeled in a subset of 20 hyperspectral images.

	IoU (%) ↑						
Alignment method	*Film*	*Basket*	*Cardboard*	*Video tape*	*Filament*	*Trash bag*	**mIoU (%) ↑**
Manual Alignment (MA)	69.0	61.1	81.2	27.4	25.1	74.0	56.3
Label Transfer (LT)	78.7	78.3	93.5	57.1	81.2	88.8	79.6

TABLE III: HSI segmentation with different supervision: labels resulting from manual alignment (MA) and the proposed automatic label transfer (LT).

			IoU (%) ↑						
Backbone	**Modality**	**Labels**	*Film*	*Basket*	*Cardboard*	*Video tape*	*Filament*	*Trash bag*	**mIoU (%) ↑**
MiniNet-v2	HYPER	MA	59.2	57.2	76.2	17.2	19.2	49.2	46.3
MiniNet-v2	HYPER	LT	61.2	61.0	78.8	28.8	30.5	56.3	52.8
MiniNet-v2	HYPER3	MA	56.8	52.8	62.9	19.2	6.9	45.9	40.7
MiniNet-v2	HYPER3	LT	58.8	61.9	69.4	**30.2**	23.0	50.5	49.0
SegFormer-B0	HYPER	MA	61.8	59.1	85.0	18.6	26.3	52.0	50.5
SegFormer-B0	HYPER	LT	**65.4**	**63.2**	85.2	21.9	33.1	**57.2**	**54.3**
SegFormer-B0	HYPER3	MA	56.6	55.4	84.7	14.4	27.5	46.7	47.5
SegFormer-B0	HYPER3	LT	60.4	58.4	**86.6**	22.6	**43.0**	49.9	53.5

TABLE IV: Object segmentation evaluation on SpectralWaste dataset with different architectures (MiniNet-v2, SegFormer and CMX) and different modalities (RGB, HYPER, HYPER3, RGB-HYPER and RGB-HYPER3).

			IoU (%) ↑									
Backbone	**Modality**	**Fusion**	*Film*	*Basket*	*Card.*	*Tape*	*Filam.*	*Bag*	**mIoU (%) ↑**	**Img./s ↑**	**Param. (M) ↓**	**GFLOPs ↓**
MiniNet-v2	RGB	-	63.1	58.9	55.4	30.6	10.0	49.2	44.5	126.7	0.522	1.343
MiniNet-v2	HYPER	-	61.2	61.0	78.8	28.8	30.5	56.3	52.8	125.8	0.585	3.429
MiniNet-v2	HYPER3	-	58.8	61.9	69.4	30.2	23.0	50.5	49.0	125.9	0.522	1.431
MiniNet-v2	RGB-HYPER	early	67.3	59.1	82.6	24.1	6.1	55.4	49.1	124.6	0.586	3.457
MiniNet-v2	RGB-HYPER3	early	57.9	53.3	69.6	13.5	6.5	49.6	41.7	125.0	0.523	1.459
SegFormer-B0	RGB	-	66.9	71.3	48.9	**33.6**	15.2	54.6	48.4	156.2	3.716	3.508
SegFormer-B0	HYPER	-	65.4	63.2	85.2	21.9	33.1	57.2	54.3	152.2	4.062	6.347
SegFormer-B0	HYPER3	-	60.4	58.4	86.6	22.6	**43.0**	49.9	53.5	152.9	3.717	3.596
SegFormer-B0	RGB-HYPER	early	71.3	62.9	**87.5**	21.2	22.0	56.9	53.6	155.9	4.067	6.385
SegFormer-B0	RGB-HYPER3	early	57.7	59.2	80.9	10.2	34.4	48.6	48.5	157.1	3.721	3.634
CMX-B0	RGB-HYPER	hybrid	**77.7**	**74.9**	80.2	31.1	20.7	**64.5**	**58.2**	54.7	11.539	8.365
CMX-B0	RGB-HYPER3	hybrid	71.7	71.6	71.7	27.8	37.7	59.4	56.6	55.1	11.193	5.615

computational load (GFLOPs) of each configuration when processing one image at a time. Measurements also include memory (number of parameters) and inference throughput (images/s). Note that batch inference can improve the processing time per image. When HYPER3 is used, we account additionally for the dimensionality reduction overhead. The throughput results showcase the real-time running capability of the implemented architectures while a significant increase in GFLOPs occurs when using the 224-channel hyperspectral images (HYPER). Employing HYPER data as input brings slight improvements in accuracy with respect to the reduced version HYPER3 in the analyzed configurations. However, the efficiency analysis clearly suggests that the reduced version of hyperspectral imaging is a better choice, offering the best trade-off between accuracy and computational load. Each of the architectures evaluated presents distinct trade-offs in performance and efficiency. MiniNet-v2 is designed to run efficiently on CPU, prioritizing low computational and memory demands (GFLOPs and number of parameters respectively). This design, optimized for CPU execution, results in slower processing speeds (images/s) on GPUs compared to SegFormer. Conversely, CMX requires significantly higher computational demands but surpasses both MiniNet-v2 and SegFormer in terms of mIoU, achieving the highest segmentation accuracy.

VI. Conclusions

This paper introduces SpectralWaste, the first multimodal dataset collected from a real waste sorting facility comprising RGB and HSI images. The presented dataset addresses the identification of critical objects that impact the efficiency of the sorting process. In the context of waste object segmentation, we also propose a pipeline that pays special attention to employing efficient architectures and exploiting the synergies between multiple sensing modalities. Our evaluation demonstrates the benefits of using RGB and HSI together for waste object segmentation. On the other hand, the low performance in segmenting some of the classes with current state-of-the-art architectures, remarks the open challenges and opportunities that SpectralWaste poses for future research in waste segmentation.

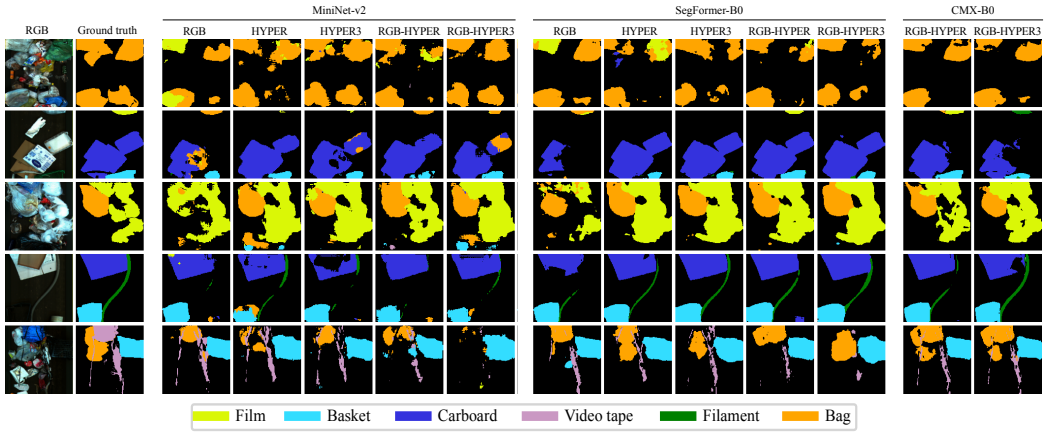


Fig. 6: Qualitative results of the implemented architectures MiniNet-v2, SegFormer-B0 and CMX using the RGB, HYPER and RGB-HYPER modalities.

REFERENCES

[1] T. Linder, N. Vaskevicius, R. Schirmer, and K. O. Arras, "Cross-modal analysis of human detection for robotics: An industrial case study," in *International Conference on Intelligent Robots and Systems*. IEEE, 2021, pp. 971–978.

[2] J. Lee, J. Hur, I. Hwang, and Y. M. Kim, "MasKGrasp: Mask-based grasping for scenes with multiple general real-world objects," in *International Conference on Intelligent Robots and Systems*. IEEE, 2022, pp. 3137–3144.

[3] M. L. Henriksen, C. B. Karlsen, P. Klarskov, and M. Hinge, "Plastic classification via in-line hyperspectral camera analysis and unsupervised machine learning," *Vibrational Spectroscopy*, vol. 118, 2022.

[4] N. Habili, E. Kwan, W. Li, C. Webers, J. Oorloff, M. A. Armin, and L. Petersson, "A hyperspectral and RGB dataset for building façade segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 258–267.

[5] S. Kodgule, A. Candela, and D. Wettergreen, "Non-myopic planetary exploration combining in situ and remote measurements," in *International Conference on Intelligent Robots and Systems*. IEEE, 2019, pp. 536–543.

[6] D. Bashkirova, M. Abdelfattah, Z. Zhu, J. Akl, F. Alladkani, P. Hu, V. Ablavsky, B. Calli, S. A. Bargal, and K. Saenko, "ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes," in *Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 147–21 157.

[7] M. Yang and G. Thung, "Classification of trash for recyclability status," *CS229 project report*, vol. 2016, no. 1, p. 3, 2016.

[8] P. F. Proença and P. Simões, "TACO: Trash annotations in context for litter detection," 2020, arXiv preprint arXiv:2003.06975.

[9] Y. Cheng, J. Zhu, M. Jiang, J. Fu, C. Pang, P. Wang, K. Sankaran, O. Onabola, Y. Liu, D. Liu, and Y. Bengio, "FloW: A dataset and benchmark for floating waste detection in inland waters," in *International Conference on Computer Vision*, 2021, pp. 10 933–10 942.

[10] J. Kim, M.-H. Jeon, S. Jung, W. Yang, M. Jung, J. Shin, and A. Kim, "TRansPose: Large-scale multispectral dataset for transparent object," *The International Journal of Robotics Research*, 2023.

[11] D. Bashkirova, S. Mishra, D. Lteif, P. Teterwak, D. Kim, F. Alladkani, J. Akl, B. Calli, S. A. Bargal, K. Saenko, D. Kim, M. Seo, Y. Jeon, D.-G. Choi, S. Ettedgui, R. Giryes, S. Abu-Hussein, B. Xie, and S. Li, "VisDA 2022 challenge: Domain adaptation for industrial waste sorting," in *NeurIPS 2022 Competition Track*, 2022, pp. 104–118.

[12] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz, "The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data," *Remote Sensing of Environment*, vol. 44, no. 2-3, pp. 145–163, 1993.

[13] S. Seidlitz, J. Sellner, J. Odenthal, B. Özdemir, A. Studier-Fischer, S. Knödler, L. Ayala, T. J. Adler, H. G. Kenngott, M. Tizabi, *et al.*, "Robust deep learning-based semantic organ segmentation in hyperspectral images," *Medical Image Analysis*, vol. 80, 2022.

[14] A. Wendel and J. Underwood, "Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging," in *International Conference on Robotics and Automation*. IEEE, 2016, pp. 5128–5135.

[15] J. Nalepa, M. Myller, and M. Kawulok, "Validating hyperspectral image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1264–1268, 2019.

[16] R. Grewal, S. S. Kasana, and G. Kasana, "Hyperspectral image segmentation: A comprehensive survey," *Multimedia Tools and Applications*, vol. 82, pp. 20 819–10 872, 2023.

[17] A. C. Karaca, A. Ertürk, M. K. Güllü, M. Elmas, and S. Ertürk, "Automatic waste sorting using shortwave infrared hyperspectral imaging system," in *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. IEEE, 2013.

[18] Z. Chen, T. R. Scott, S. Bearman, H. Anand, D. Keating, C. Scott, J. R. Arrowsmith, and J. Das, "Geomorphological analysis using unpiloted aircraft systems, structure from motion, and deep learning," in *International Conference on Intelligent Robots and Systems*. IEEE, 2020, pp. 1276–1283.

[19] H. Qin, W. Zhou, Y. Yao, and W. Wang, "Individual tree segmentation and tree species classification in subtropical broadleaf forests using UAV-based LiDAR, hyperspectral, and ultrahigh-resolution RGB data," *Remote Sensing of Environment*, vol. 280, 2022.

[20] S. D. Fabiyi, H. Vu, C. Tachtatzis, P. Murray, D. Harle, T. K. Dao, I. Andonovic, J. Ren, and S. Marshall, "Varietal classification of rice seeds using RGB and hyperspectral images," *IEEE Access*, vol. 8, pp. 22 493–22 505, 2020.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *International Conference on Computer Vision*, 2023, pp. 4015–4026.

[22] I. Alonso, L. Riazuelo, and A. C. Murillo, "MiniNet: An efficient semantic segmentation ConvNet for real-time robotic applications," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1340–1347, 2020.

[23] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[24] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, pp. 14 679–14 694, 2023.

[25] K. Chen, J. Liu, and H. Zhang, "IGT: Illumination-guided RGB-T object detection with transformers," *Knowledge-Based Systems*, vol. 268, 2023.

[26] J. Behmann, A.-K. Mahlein, S. Paulus, H. Kuhlmann, E.-C. Oerke, and L. Plümer, "Calibration of hyperspectral close-range pushbroom cameras for plant phenotyping," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 106, pp. 172–182, 2015.

[27] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "COTR: Correspondence transformer for matching across images," in *International Conference on Computer Vision*, 2021, pp. 6207–6217.

[28] Z. Wang, F. Colonnier, J. Zheng, J. Acharya, W. Jiang, and K. Huang, "TIRDet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation," in *ACM International Conference on Multimedia*. Association for Computing Machinery, 2023, p. 2663–2672.