JOINT MULTITASK LEARNING FOR IMAGE SEGMENTATION AND SALIENT OBJECT DETECTION IN HYPERSPECTRAL IMAGERY

Koushikey Chhapariya, Ientilucci Emmett J., A Benoit, Buddhiraju Usa Krishna Mohan, Kumar India Anil

JOINT MULTITASK LEARNING FOR IMAGE SEGMENTATION AND SALIENT OBJECT DETECTION IN HYPERSPECTRAL IMAGERY

Koushikey Chhapariya[1], Emmett J. Ientilucci[2], Alexandre Benoit[3], Krishna Mohan Buddhiraju[1], Anil Kumar[4]

[1]Centre of Studies in Resources Engineering, Indian Institute of Technology Bombay, India
[2]Digital Imaging and Remote Sensing Lab, Rochester Institute of Technology, New York, USA
[3] LISTIC, Université Savoie Mont-Blanc, Annecy, France
[4]Indian Institute of Remote Sensing, ISRO, Dehradun, India

ABSTRACT

With technological advancements, combining information from various tasks has become increasingly important. However, most feature learning approaches still focus on single-task learning. To address this, we propose a multitask learning-based model that simultaneously performs segmentation and saliency estimation. Our model is evaluated on two hyperspectral datasets: HS-SOD for computer vision and Pavia University (PU) for remote sensing, which demonstrate strong generalization capabilities. By utilizing the additional spectral dimension in hyperspectral data, the model improves its ability to distinguish between materials and objects, leading to higher accuracy. The architecture features a shared encoder-decoder structure for efficiency, with an attention block enhancing segmentation by capturing key spectral-spatial features and a dense ASPP block improving salient object detection through multi-scale context. Extensive testing shows our model outperforms single-task approaches and state-of-the-art methods, proving its effectiveness and efficiency.

***Index Terms*—** Hyperspectral Image Classification, Multitask Learning, Remote Sensing, Image Segmentation, Salient Object Detection, Shared-encoder decoder, ASPP, Attention Network

1. INTRODUCTION

Hyperspectral imaging is a vital technique in remote sensing and image processing, capturing detailed spectral information across numerous bands for precise material identification [1]. Unlike traditional RGB images, hyperspectral images capture extensive spectral data, enabling diverse applications in agriculture, environmental monitoring, medical imaging, and many more [2, 3]. With the rise of large datasets, advanced methods like machine learning and deep learning have shown significant progress in analyzing such data [4]. However, the computational cost increases when addressing tasks separately, as each requires substantial resources and training time. To address this, multitask learning has emerged as an effective approach, enhancing both efficiency and accuracy by simultaneously tackling multiple objectives and utilizing shared representations in areas such as computer vision and remote sensing.

Multitask learning (MTL) has evolved significantly, from its introduction to neural networks [5] to recent advancements in adaptive optimization for reinforcement learning [6]. Cipolla et al. [7] introduced the algorithm for handling multitask loss by computing uncertainty. Many researchers used a multitask learning approach with deep learning algorithms for applications such as building extraction, height and depth estimation of a building, and so on [8, 9]. Ji et al. [10] introduced a few-shot multitask learning approach for the scene classification of optical remote-sensing images. Adding to this, a semi-supervised few-shot multitask learning network for iterative label correction was introduced [11]. Further, Zhu et al. [12]proposed a novel semantic-guided image–text retrieval framework with segmentation. However, there is still room to explore multitasking for uncorrelated tasks, which could explore new applications for different objectives. Thus, our study focuses on applying multitasking to two distinct domains: saliency estimation and image segmentation.

In our study, we explore multitask learning with hyperspectral data, focusing on image segmentation and salient feature detection through saliency maps. The motivation arises from the need to enhance efficiency and accuracy in processing complex hyperspectral datasets, which present significant challenges for conventional single-task approaches due to their high dimensionality. By integrating image segmentation and salient object detection into a unified framework, we aim to exploit shared representations, improving overall model performance and reducing computational costs. This approach enhances segmentation accuracy while highlighting critical features, enabling more effective interpretations of hyperspectral imagery for applications like environmental monitoring and disaster management. Multitask learning is particularly crucial in scenarios requiring real-time processing, such as autonomous vehicles that need to segment roads and detect pedestrians simultaneously. Our task set consists of image

segmentation and saliency estimation, which are challenging to optimize jointly due to their low relatedness.

2. PROPOSED METHOD

In this section, we elaborate on the proposed architecture and parameters used in detail. The proposed network aims for joint image segmentation and salient object detection that helps in better scene understanding. The overview of the proposed network architecture is shown in Figure 1

2.1. Framework of the Proposed Architecture

For the proposed work, Principal Component Analysis (PCA) has been utilized to address the data dimensionality. We have selected the principal components that contribute to at least 98% of the total explained variance in the data as the reduced feature set. Further, to extract shared representations, an encoder-decoder approach is considered as it is highly flexible and can be adapted to various types of data and tasks. The proposed architecture adopts an encoder-decoder approach, employing ResNet-50 [13] as a shared encoder. ResNet is a good choice for a shared encoder in multitask learning due to its ability to effectively learn deep representations without the vanishing gradient problem, owing to its residual connections. These connections allow ResNet to train large deep networks, capturing complex features that are beneficial for multiple tasks. At the decoder end, we have incorporated an Attention block to enhance image segmentation and employed Dense ASPP (Atrous Spatial Pyramid Pooling) [14] to generate the saliency map.

2.1.1. Segmentation Block

The segmentation block consists of an attention block and a 4-block upsampling decoder. An attention block in a neural network architecture is designed to selectively emphasize or suppress certain parts of the input feature map, enabling the model to focus on more relevant information. The attention mechanism generates attention weights for each spatial location in the input, which are applied to modulate the feature map.

Let X be the input feature map, and W represents the set of learnable parameters in the attention block. The attention mechanism can be mathematically expressed as follows:

$$M = \sigma\left(Conv\left(X, W\right)\right) * X \tag{1}$$

$$\alpha = Softmax\left(Conv\left(M, W\right)\right) \tag{2}$$

$$\gamma = \alpha * X \tag{3}$$

where, $Conv(\cdot)$ denotes the convolutional layer, W is the learnable parameter, $\sigma(\cdot)$ is a non-linear activation function, $*$ denotes the elementwise multiplication, α represents the attention weights, and γ is the attention output.

2.1.2. Saliency Detection Block

Atrous Convolution handles the multi-scale feature extraction in salient object detection to achieve a large receptive field without increasing the computation cost. DeepLabV3 introduces multiple atrous convolution layer parallel called ASPP (Atrous Spatial Pyramid Pooling). The Dense ASPP module is generally represented as a concatenation of dilated convolutions with varying dilation rates. The mathematical formulation for a Dense ASPP module with dilation rates r_1, r_2,.....r_n can be represented as:

$$Y = Concat(Conv_{r_1}(X, W_{r_1}), ...Conv_{r_n}(X, W_{r_n})) \tag{4}$$

Here $Conv_{r_1}(X, W_{r_1})$ denotes the dilated convolution with dilation rate r_n applied to the input feature map X using the corresponding set of filters W_{r_i}. The $Concat(\cdots)$ operation combines the feature maps obtained from these dilated convolutions along the channel dimension. The dilation rate considered for the proposed work is 6, 12, and 18. Dense ASPP can be adapted for salient object detection, where the goal is to identify and highlight the most visually distinct objects or regions in an image.

2.2. Multitask Loss

This research uses multitask learning by considering the saliency detection and segmentation tasks together. Let $\chi = (X_i)_{i=1}^{N_1}$ represent number of bands in the hyperspectral image where X being input feature map with height H and width W for every image. The corresponding pixel-wise ground truth (GT) map for segmentation is given as

$$(GT_{i,j,k} \mid GT_{i,j,k} \epsilon \{1, 2, ..., C\})_{N_1 \times H \times W} \tag{5}$$

The salient object, S, represented as $(S_i)_{i=1}^{N_2}$ with the ground-truth binary map for all the bands of the hyperspectral image denoted as $\rho = (P_i)_{i=1}^{N_2}$, where P is the image pixel. Further, all the parameters in the shared encoder are represented as θ_{SH}; parameters of the salient detection block as θ_{SD} and the parameters of the segmentation block as θ_{SE}. The minimization of the cost function is given as

$$\begin{aligned} \jmath_1\left(\chi; \theta_{SH}, \theta_{SE}\right) &= -\frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbb{I}\left\{GT_{ijk} = C\right\} \\ &\cdot \log\left(h_{cjk}\left(X_i; \theta_{SH}, \theta_{SE}\right)\right) \end{aligned} \tag{6}$$

$$\jmath_2\left(\rho; \theta_{SH}, \theta_{SD}\right) = \frac{1}{N_2} \sum_{i=1}^{N_2} S_i - \int \left(P_i; \theta_{SH}, \theta_{SD}\right) \tag{7}$$

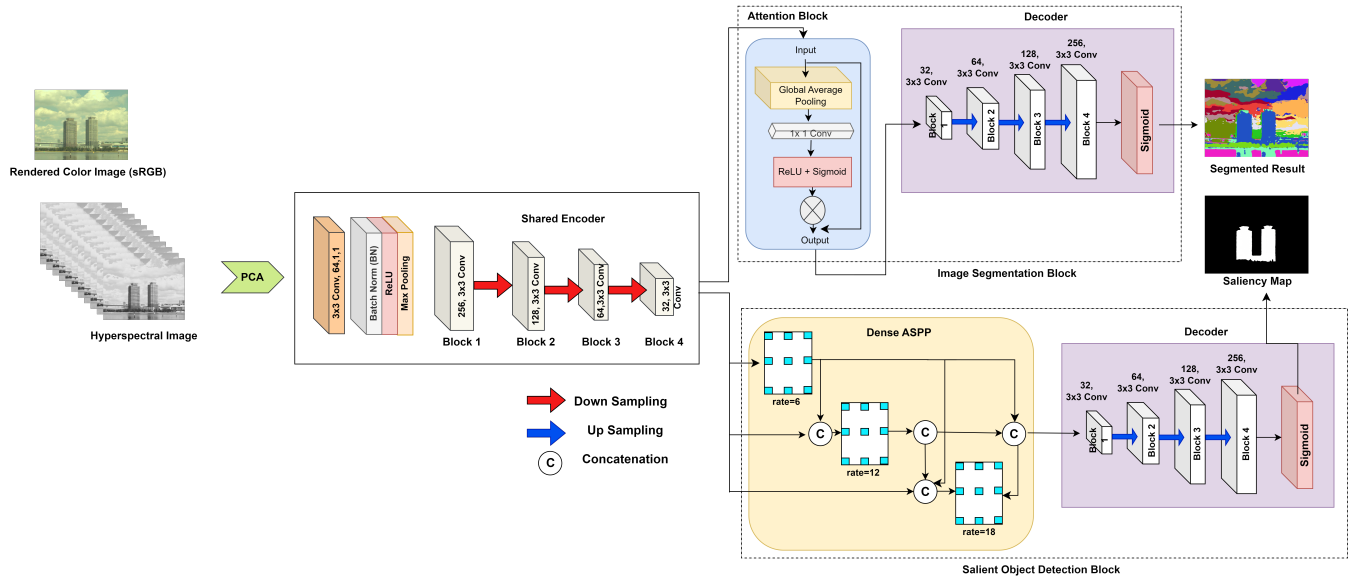

Fig. 1: Framework of the proposed joint multitask learning model for segmentation and saliency estimation for hyperspectral imagery.

Here, $\mathbb{I}$ is the indicator function, C represents the probability of the segmentation map, and h is the segmentation function. h_{cjk} is the $(j,k)^{th}$ element of the c^{th} probabilistic segmentation map and integral $\int$ is the output function for saliency map. Equation (6) represents a cross-entropy loss for the image segmentation task, and Equation (7) is the squared Euclidean loss for the salient object detection task.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section presents the dataset considered for this research work in Section 4.1 and the experimental settings in Section 4.2. The detailed experiment results are listed in Section 4.3 as quantitative analysis, and qualitative analysis is listed in Section 4.4.

3.1. Dataset

The efficacy of this study is evaluated using two datasets from different domains: the Hyperspectral Salient Object Detection (HS-SOD) is a computer vision hyperspectral dataset, and the Pavia University dataset (PU) is a remote sensing dataset.

The HS-SOD dataset includes 60 hyperspectral images captured in Japan's public parks, featuring 81 spectral bands in the 380–780 nm range, with a resolution of 1024×768 pixels. These images, acquired using an NH-AIK model hyperspectral camera, provide both ground truth and sRGB-rendered images, allowing for diverse variations in object size, position, and contrast. The PU dataset, acquired in 2001 by the ROSIS sensor over Northern Italy, consists of 103 spectral bands with a spatial resolution of 1.3m and an image size of 610×340 pixels covering nine land use/land cover classes,

3.2. Experimental Setup

The experiment was conducted on a system with an Intel Xeon Silver 4214R CPU (2.40 GHz), 64 GB RAM, and an NVIDIA GeForce RTX A6000 GPU (51 GB RAM), using Ubuntu 14.04 and PyTorch. The key hyperparameter includes a learning rate of 0.001 with stochastic gradient descent, momentum of 0.7, and weight decay of 0.0005. The network was trained for 100 epochs on full-resolution images with a mini-batch size of 128. For the HS-SOD dataset, 25% was used for training and 75% for testing, while the PU dataset was split into 10% training, 10% validation, and 80% testing. Random initialization was applied for kernel weights, and RMSProp was used for optimization.

3.3. Quantitative Analysis

The proposed multitask learning model demonstrated clear improvements over the single-task approach for both image segmentation and saliency estimation tasks on the HS-SOD and Pavia University (PU) datasets, as shown in Tables 1 and 2.

For image segmentation, the multitask model achieved higher IoU, precision, and recall compared to the single-task model. This suggests that sharing information between segmentation and saliency estimation tasks helped the model better capture spatial-spectral relationships in hyperspectral data, leading to more accurate and consistent segmentation results. IoU was chosen to measure the overlap between predicted and true regions. At the same time, precision and recall were used to assess the model's ability to minimize false positives and false negatives, which is crucial for accurate segmentation. For

Table 1: Comparison of the proposed method as single-task and multitask for image segmentation on both HS-SOD and Pavia University datasets.

Image Segmentation				
Metric	HS-SOD		PU	
	Single-task	Multitask	Single-task	Multitask
IoU (%)	94.89	**97.72**	91.98	95.98
Precision	0.892	**0.948**	0.828	0.912
Recall	0.887	**0.951**	0.813	0.929

Table 2: Comparison of the proposed method as single-task and multitask for saliency estimation on both HS-SOD and Pavia University datasets.

Salient Object Detection				
Metric	HS-SOD		PU	
	Single-task	Multitask	Single-task	Multitask
F-measure	0.907	**0.961**	0.892	0.942
AUC	0.912	0.943	0.893	**0.947**
MAE	0.081	**0.032**	0.097	0.041

saliency estimation, the multitask model also outperformed the single-task model in terms of F-measure, AUC, and MAE. The use of these metrics is crucial: F-measure captures the balance between precision and recall, AUC evaluates the model's performance across different threshold levels, and MAE reflects the model's precision in identifying salient objects. The lower MAE and higher AUC indicate more precise and reliable saliency detection, likely because the segmentation task helps better in defining object boundaries.

3.4. Qualitative Analysis

The visual representation of the proposed method on two different datasets, viz, HS-SOD and PU, is depicted in Figure 2 and Figure 3, respectively. The HS-SOD dataset categorizes objects with varying shapes, sizes, and complexities, offering a comprehensive assessment of the proposed model's performance across diverse scenarios. Additionally, overhead remote-sensing hyperspectral images featuring objects at multiple scales, such as those in the PU dataset, were analyzed. For saliency estimation, the focus is on highlighting specific objects or features—in this case, a building in the PU dataset. Our model successfully identifies and emphasizes the building, closely aligning with the ground truth, demonstrating its robustness and accuracy. The visualization of predicted results shows that the model excels in both image segmentation and saliency map generation, effectively handling diverse scenarios, including varying object scales. The multitasking approach consistently captures salient features and delivers precise segmentation.

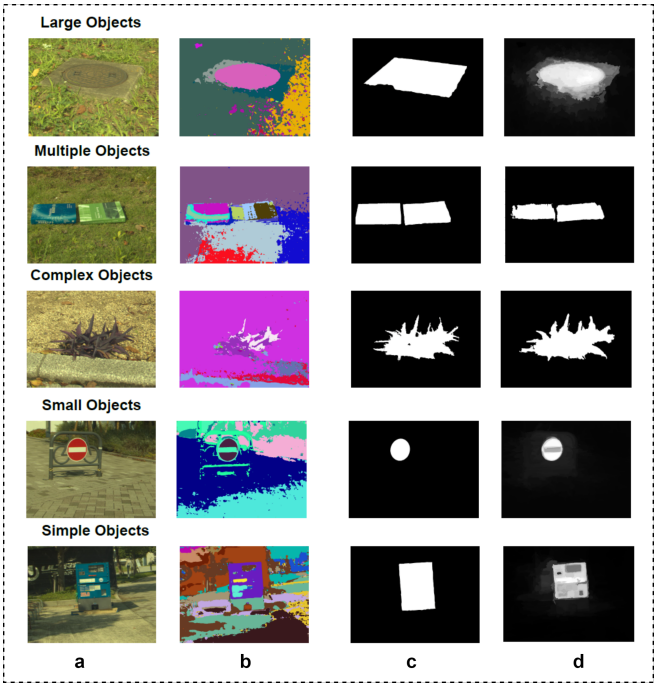

Fig. 2: Visual comparison of segmented results and saliency map on sample images from the HS-SOD dataset. Images are categorized to demonstrate the efficacy of the proposed methodology across various scenes. (a) RGB (b) segmented result (c) ground truth saliency map (d) predicted saliency map.

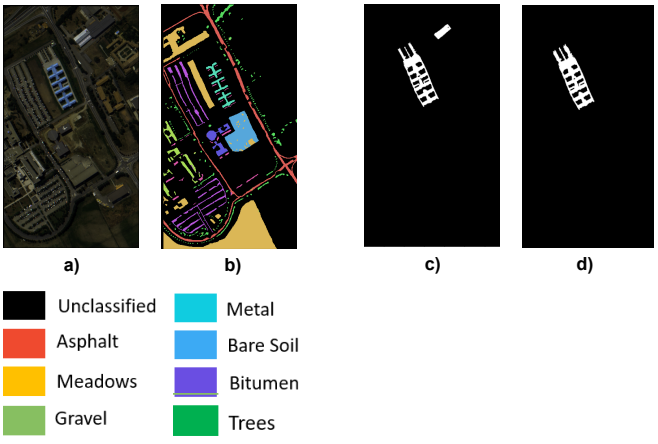

Fig. 3: Visual comparison of segmented results and saliency map on a remotely sensed Pavia University dataset. Legend depicting the segmented result and class assignments to demonstrate the efficacy of the proposed methodology across various scenes. (a) RGB (b) segmented result (c) ground truth saliency map (d) predicted saliency map.

4. CONCLUSION

In this research, we proposed an efficient learning-based multitasking model for joint image segmentation and saliency estimation using hyperspectral data. Our model integrates both spectral and spatial features through a shared encoder and separate decoders for segmentation and detection blocks, resulting in a lightweight architecture due to parameter sharing. We evaluated the model on two distinct datasets, HS-SOD (for computer vision) and Pavia University (for remote sensing), to assess its generalization capabilities. Visual comparisons of the generated saliency maps demonstrated improved results with sharper boundary edges compared to other methods. However, we faced difficulties in detecting edges for objects with faint contrasts and complex features, likely caused by inadequate parameter tuning and optimization. Notably, our approach outperformed single-task methods across various evaluation metrics, demonstrating its potential for extension to a wide range of complex multitasks in both computer vision and remote sensing. Future work involves refining the model through parameter tuning and optimization strategies to improve its ability to handle complex edge identification, with the potential integration of other data modalities to further enhance performance.

5. REFERENCES

[1] Chein-I Chang, *Hyperspectral imaging: techniques for spectral detection and classification*, vol. 1, Springer Science & Business Media, 2003.

[2] José M Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[3] Koushikey Chhapariya, Krishna Mohan Buddhiraju, and Anil Kumar, "Cnn-based salient object detection on hyperspectral images using extended morphology," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[4] Koushikey Chhapariya, Krishna Mohan Buddhiraju, and Anil Kumar, "A deep spectral-spatial residual attention network for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[5] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, New York, NY, USA, 2008, ICML '08, p. 160–167, Association for Computing Machinery.

[6] Aritz D. Martinez, Javier Del Ser, Eneko Osaba, and Francisco Herrera, "Adaptive multifactorial evolutionary optimization for multitask reinforcement learning," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 2, pp. 233–247, 2022.

[7] Roberto Cipolla, Yarin Gal, and Alex Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[8] Koushikey Chhapariya, Alexandre Benoit, Krishna Mohan Buddhiraju, and Anil Kumar, "A multitask deep learning model for classification and regression of hyperspectral images: Application to the large-scale dataset," *arXiv preprint arXiv:2407.16384*, 2024.

[9] Koushikey Chhapariya, Alexandre Benoit, Krishna Mohan Buddhiraju, and Anil Kumar, "A deep learning-based multitasking model for hyperspectral image analysis using novel taiga dataset," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 7883–7887.

[10] Hong Ji, Zhi Gao, Yongjun Zhang, Yu Wan, Can Li, and Tiancan Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[11] Hong Ji, Zhi Gao, Yao Lu, Ziyao Li, Boan Chen, Yanzhang Li, Jun Zhu, Chao Wang, and Zhicheng Shi, "Semi-supervised few-shot classification with multitask learning and iterative label correction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[12] Zicong Zhu, Jian Kang, Wenhui Diao, Yingchao Feng, Junxi Li, and Jingen Ni, "Sirs: Multitask joint learning for remote sensing foreground-entity image–text retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang, "Denseaspp for semantic segmentation in street scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.