Learning with Limited Labeled Data for Image and Video Understanding

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Razieh Kaviani Baghbaderani

August 2022

This dissertation is dedicated to my husband, my parents, and my sister who constantly encouraged me to pursue my dreams and finish my dissertation

Acknowledgements

I am forever grateful to everyone who has directly or indirectly supported, encouraged and guided me in my quest for knowledge.

First and foremost, I would like to express my profound appreciation to my advisor, Dr. Hairong Qi, for her timely and powerful mentorship, unwavering support, and continuous encouragement during the past five years. Her immense knowledge, rigorous research attitude, and invaluable advice have greatly helped shape my research direction and future career. She has set an example of excellence as a researcher, mentor, and a successful woman in balancing life and work. I could not have imagined having a better advisor and mentor for my Ph.D. study and life.

Besides, I would like to thank my committee members, Dr. Jens Gregor, Dr. Amir Sadovnik, and Dr. Chuanren Liu, for their valuable advice and guidance in my research. I appreciate their time and constructive suggestions to my dissertation.

To my co-authors, Dr. Hairong Qi, Dr. Ying Qu, Dr. Shuangquan Wang, Dr. Yuanxin Li, Dr. Zhifei Zhang, Dr. Yang Song, Fanqi Wang, and Craig Stutts, thank you for your valuable knowledge and insights in my research. I also want to thank all my labmates for their continuous support and memorable experiences.

My journey would not have been possible without my constant source of support and encouragement, my husband, Ramin Arvin, during the challenges of graduate school and life, my parents, sister, and my beloved family for their unconditional love and support throughout these years and for helping me to reach this stage in my life.

Abstract

Deep learning-based algorithms have remarkably improved the performance in many computer vision tasks. However, deep networks often demand a large-scale and carefully annotated dataset and sufficient sample coverage of every training categories. However, it is not practical in many real-world applications where only a few examples may be available or the data annotation is costly and require expert knowledge. To mitigate this issue, learning with limited data has gained considerable attention and is investigated thorough different learning methods, including few-shot learning, weakly/semi supervised learning, open-set learning, etc.

In this work, the classification problem is investigated under an open-world assumption to handle unpredictable categories in a long-tail distribution dataset. For the open-set recognition, a representative-discriminative multi-task learning framework is presented which is able to characterize the categories more effectively for known vs. unknown category recognition. Experiments on multiple satellite benchmarks and RGB image datasets demonstrate significant improvement over state-of-the-art open-set recognition algorithms.

In addition, the generalization capability of a recognition system is studied where the goal is to build a model which has a reasonable performance across diverse datasets for a certain task. To bridge the gap between different but related datasets, an unmixing-based domain adaptation approach is proposed which projects the datasets to a domain-invariant space to align the domains. The experimental results show stable deployment performance across multiple satellite datasets.

To extend this work to higher dimensional data, we also study semi-supervised learning for video segmentation task as only certain frames are often annotated in a pixel level. To take advantage of labeled frames more efficiently, a bidirectional framework is proposed to segment the unlabeled frames with the help of segmented ones. Besides, an occlusion estimation approach is introduced to improve the segmentation performance by effectively fusing the bidirectional propagated semantic labels. Extensive experiments on driving video benchmarks demonstrate superiority of the proposed method on segmentation accuracy, temporal consistency, and computation cost compared to the state-of-the-art methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

Deep learning has demonstrated remarkable success in a variety of computer vision tasks such as image classification [8; 9; 10], object detection [11; 12], semantic segmentation [13; 14], video scene understanding [15; 16], etc. However, these methods heavily rely on large quantities of labeled datasets with sufficient examples of every training categories (different illumination conditions, viewing angles, etc.) to achieve high performance. Data collection and annotation for such datasets is time-consuming, costly, and labor intensive. This issue becomes more sever when annotation requires expert knowledge (e.g. medical imaging), processing wide range of data (e.g. satellite imagery), or is based on classes that rarely occur.

To tackle the issue of availability of large-scale datasets, researchers have explored approaches that learn from limited labeled data. These methods include few-shot learning [17; 18; 19], weakly-supervised learning [20; 21; 22], semi supervised learning [23; 24], open-set learning [25; 26; 27], domain adaptation [28; 29; 30], etc.

1.1.1 Open-set Recognition

A vast majority of supervised learning algorithms are based on the close-set assumption, that is, the training dataset is assumed to contain all the classes that might appear in the testing phase. However, this assumption can be violated when the algorithm is deployed in real-world applications where there are many more categories compared to the number of categories present in any available dataset [25; 26]. As a result, the algorithm trained on the closed-set assumption is forced to classify a given unknown class sample as one of the known classes, and this affects the performance and the usability of the system. However, a desirable classifier/recognizer should be able to know when it does not know. This behavior is critical for safety applications such as autonomous driving and medical diagnosis system. Having such a wise system avoids malfunctions by triggering additional sensors or asking a human operator or doctor for intervention and further analysis.

To tackle this issue, open-set recognition (OSR) was introduced to extend the closed-set classification problem into a more realistic scenario. For this purpose, instead of assuming that the dataset is comprehensive, the data is assumed to be incomplete during training to be able to detect samples belonging to none of the training classes [25]. The goal of an open-set recognition system is to correctly identify test samples as either known or unknown while maintaining the classification performance on known classes as shown in Fig. 1.1.

1.1.2 Domain Adaptation

Machine learning algorithms assume that the training and test data come from the same distributions and aim to learn a model which is generically useful across diverse datasets for a certain application. However, this assumption may not hold true in real-world applications since test data can originate from a different distribution due to various reasons, e.g. different acquisition conditions or changes in the statistical properties of a domain over time. Hence, existence of domain-shift between training

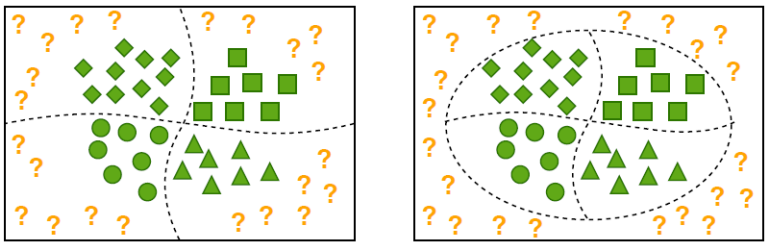


Figure 1.1: Comparison between closed set classification and open set recognition. Green symbols and orange question marks indicate known and unknown class samples, respectively. Dashed lines show the decision boundaries [1].

and test data will likely degrade the performance of a well-trained model when applying on the test data [31; 32; 2].

To address the above problem, a new research area called domain adaptation is developed in the machine learning community. Given the training and test data as the source and the target domains, respectively, the goal is to learn a model from a source labeled data that is able to generalized well to the target domain, as shown in Fig. 1.2 [31]. Note that the typical solution of fine-tuning the model on the target data is unacceptable since it is often prohibitively expensive and time-consuming to collect enough labeled samples in the target domain.

1.1.3 Semi-supervised Learning

Semi-supervised learning is a type of machine learning that lies somewhere between supervised learning and unsupervised learning. It is used in the scenario where relatively small amount of labeled data is available, but a large amount of unlabeled data is provided [23; 24]. One of such scenarios occurs for video segmentation where labeling every frame in the video requires more effort and cost. Hence, certain number of frames are annotated [33] and the task is performed under the semi-supervised learning strategy.

Semi-supervised video segmentation takes advantage of temporal information existed in the video and propagates the labels in key-frames to the rest of the video frames [34; 35; 23; 36]. These methods often utilize optical flow [37] for label propagation and in some extent track the objects throughout the video. The label propagation is illustrated in Fig. 1.3. Recent works has employed different approaches, including flow-based random field propagation model [38], segmentation on bilateral space [3], patch-stem based propagation strategy [39], etc.

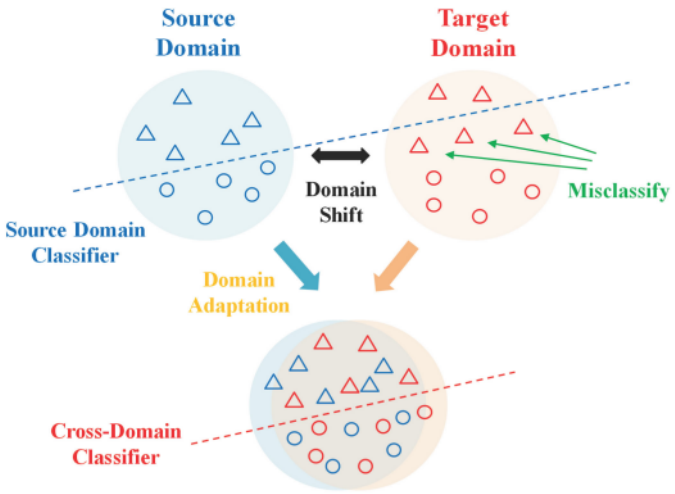


Figure 1.2: Illustration of the domain-shift between source and target domains and the effect of domain adaptation strategy [2].

Figure 1.3: Video segmentation utilizes temporal information to propagate the labels [3].

1.2 Motivation and Contribution

This work focuses on learning with limited labeled data for image recognition and video segmentation tasks. Three approaches to handle the availability of limited annotated data, including open-set recognition, domain adaptation, and semi-supervised video segmentation are studied.

Land cover classification of satellite imagery is one of the active research topics that plays an important role in many remote sensing applications, including ecology management, urban development, etc [40]. Although deep learning-based approaches have achieved very promising results, they suffer from two challenges: 1) a considerable amount of labeled data is required to prevent the model from over-fitting or misclassification issues, 2) the model tend to perform poorly when the training and test images are from different domains [41]. The former comes from the fact that the vast area covered by the satellite imagery makes the task of generating representative training samples almost impossible, as there are a large variety of materials existed on the Earth's surface, especially those not well-exploited regions. The latter is an inevitable issue in remote sensing area where spectral variability is prevalent due to different acquisition conditions, i.e. illumination and atmospheric conditions.

One of the most essential capabilities of land cover classification is to be able to automatically identify which test image and which area or pixel location of the image, has a higher probability of hosting new classes of materials. We propose a representative-discriminative framework where both the representative and discriminative aspects of data are exploited in order to best characterize the differences between known and unknown classes. The proposed representative and discriminative learning framework operates among three spaces, including 1) the transformation from the raw image space to the embedding feature space, and 2) the transformation from the embedding feature space to a so-called abundance space. The first transformation creates a more discernible input to be fed into the

subsequent open-set learning network. The second transformation provides a finer-scale representation which is useful for the unknown detection part.

To bridge the gap between source and target domains, we present an unmixing-based domain adaptation framework that performs classification in a domain-invariant embedding space. It decreases the domain discrepancy by projecting both the source and target data to a so-called abundance space which is regularized by some physical characteristics. In addition, the two domains are further aligned in the abundance space to minimize the remaining domain-shift between the distributions. In this way, the model trained on the source domain would be able to perform well on the target dataset without extra data annotation efforts.

Video semantic segmentation has received significant attention due to the rise of applications in autonomous vehicles and video analysis. A naïve frame-by-frame approach applying an image semantic segmentation method on each frame has raised concern about the unaffordable computational cost and less temporal consistency. Many works exploit the temporal information in the video to balance the trade-off between quality and speed. They utilize the content continuity between frames and propagate the high-level features extracted at the key frame to other time steps. The propagation process is challenging and causes distortions due to the misalignment caused by the motion between frames.

We introduce an occlusion-aware bidirectional framework consisting of two main components: bidirectional feature propagation and occlusion-based feature correction. Unlike the previous works, it propagates the features of the key frames toward the non-key frames in forward and backward directions and estimates the distorted regions based on the flows predicted between the current frame and the previous and next frames. The proposed method is able to highlight the occluded regions corresponding to the forward and backward directions in the video and compensate the features in the occluded areas with the intact propagated features from the opposite direction and the current frame.

1.3 Dissertation Outline

The first chapter serves as an introduction and describes the problem to be solved, as well as the major contributions of this dissertation. Chapter 2 provides a survey of the approaches for the two prevalent perception tasks such as classification problem and segmentation problem, and reviews various open-set recognition and domain adaptation methods in the literature. The next three chapters (Chapter 3-5) are the core parts of the dissertation, which describes the developed open-set learning framework, domain adaptation method, and novel video semantic segmentation approach. In each chapter, the methodology is described, comprehensive experimental evaluations are conducted, and comparative analysis with the state-of-the-art methods are illustrated. The last chapter summarizes the contributions of this dissertation and discusses several future directions.

Chapter 2

Literature Survey

The advent of deep learning has accelerated the progress in many computer vision tasks. The classification and segmentation problems are the two fundamental perception algorithms which are utilized in various real-world applications such as satellite imagery processing, self-driving cars, etc. In this chapter, a review of the classification and segmentation approaches are provided. Then, the extended versions of the classification problem which are called open-set recognition and domain adaptation are described and a survey of the approaches are presented.

2.1 Classification Problem

2.1.1 Image Classification

The classification problem plays an important role in many applications. In computer vision area, image classification require two steps: features extraction and classification. Conventional classification methods including Nearest neighbor or SVM [42] reach promising results. Recently, the advent of the deep learning networks alongside large-scale datasets such as ImageNet [8] has remarkably improved the performance. To name a few, ResNet [9], VGGNet [10], and InceptionNet [43] have demonstrated promising results on different images datasets.

2.1.2 Land Cover Classification

Land cover classification or material classification is one of the building blocks of satellite image analysis, providing essential inputs to a series of subsequent tasks including object segmentation, 3D reconstruction and modeling, as well as texture mapping. Supervised land cover classification involves categorization of multispectral or hyperspectral image pixels into predefined material classes, e.g., asphalt, tree, concrete, water, metal, soil, etc. Note that both multispectral and hyperspectral satellite images (MSI and HSI) try to provide additional spectral information, beyond the visible spectra, to reveal extra details and compensate for the coarse spatial resolution of these images.

The land cover classification methods mainly employ a traditional classifier on spectral information, including support vector machine (SVM) [44; 45], k-nearest-neighbor [46], logistic regression [47], maximum likelihood criterion [48], where its discriminative power is further enhanced through feature engineering algorithms such as minimum noise fraction (MNF) [49], principal component analysis (PCA) [50], linear discriminative analysis (LDA) [51], independent component analysis (ICA) [52], morphological profiles [53], and spectral unmixing [54; 55; 56].

The advent of deep learning has enabled the extraction of hierarchical features automatically and achieved unprecedented performance. Spectral-based methods [57; 58] are developed based on 1D convolutional neural network (CNN) architecture to take into account the correlation between adjacent spectral bands. Spatial-based approaches use a patch surrounding the desired pixels by adopting a 2D-CNN structure [59; 60] to incorporate the spatial correlation as well. More recently, integration of both spectral and spatial domains using 3D-CNN structures have been employed to further improve the classification accuracy, including 3D-CNN-based [61], Spec-Spat [62], HSI-CNN [63], SSRN [64], HybridSN [65], HT-CNN [66].

Although each approach has its own merit, all the existing land cover classification approaches work under the closed-set assumption where the training and testing sets

share the same label set. In addition, the state-of-the-art methods assume that the training and the test data are drawn from the same domain which may be feasible in a real-world scenario.

2.2 Segmentation Problem

2.2.1 Image Semantic Segmentation

The success of deep learning for object classification [10; 9; 67] has been motivating researchers to exploit it for semantic segmentation. As a seminal work, the fully convolutional network (FCN) [13] replaces fully-connected layers with convolutional layers to achieve dense prediction. Follow-up models [68; 69] have extended the FCN network with explicit encoder-decoder architectures to get high-resolution output. Methods of [70; 71] utilize the dilated convolution to enlarge the receptive fields with minimal computational cost. Different from these works, the spatial pyramid pooling module [72] presented in PSPNet [73] aggregates multi-scale contextual information obtained with different filters and pooling operations. DeeplabV3+ [6] combines the advantages from both encoder-decoder structure and atrous spatial pyramid pooling [14; 74]. Recently, HRNet [75] fuses various resolution representations in parallel to enhance the final prediction.

2.2.2 Video Semantic Segmentation

Unlike static images, videos contain useful temporal information that can be exploited for video semantic segmentation. A number of works leverage cross-frame relations by applying the same image semantic segmentation network to each video frame and aggregating the features over time. NetWarp [76] combines the intermediate features warped from the previous time-step with the ones extracted from the current frame. STFCN [77] and GRFP [78] incorporate a spatial-temporal LSTM and a gated recurrent unit, respectively, to temporally aggregate semantic labels. TDNet

[79] combines features extracted from several shallower sub-networks by an attention-based propagation module. Although these methods boost the segmentation accuracy, they suffer from high computation burden as all features are recalculated at each frame.

Another group of works aiming at efficient video semantic segmentation take advantage of temporal continuity to reuse the deep features extracted at only sparse key frames. Clockwork Net [80] directly reuses the features of preceding key frame and updates them at the current frame according to their semantic stability. DFF [81] adopts optical flow to propagate high-level features to non-key frames. Further, Accel [82] updates the propagated features with shallow features obtained from the current frame. DAVSS [7] proposes to correct the propagated features only on the distorted regions which is estimated by a light-weight network. As indicated in [81; 82; 7], although the overall computational cost is decreased compared to the single-frame approaches, there is a drop in the segmentation accuracy due to potential error in optical flows, large motions, and occlusions-disocclusion.

2.3 Open-set Recognition

Open-set recognition has gained considerable attention due to its handling of unknown class samples based on incomplete knowledge of the data during model training.

Researchers have explored the modeling of open-set recognition from two perspective, the discriminative and generative perspectives [27]. From the discriminative model perspective, early efforts were made to adapt traditional machine learning algorithms, including Support Vector Machine (SVM) [42], Nearest Neighbor, and Sparse representation [83], etc., to the open-set recognition scenario. The open-set version of Nearest Neighbor was developed based on the distance of the testing samples to the known samples [84]. The SVM-based approaches employed different regularization terms or kernels to detect unknown samples [85; 86]. In [87], the

residuals from the Sparse Representation-based Classification (SRC) algorithm were used as the score for unknown class detection.

Recently, a great number of works dealt with open-set recognition by improving the feature space of deep neural networks (DNNs) through calibrated layers [26], enhanced objective function [88], activation function of each layer [89], etc. [26] employed a statistical model to calibrate the SoftMax scores and produced an alternative layer, called OpenMax. [88] improved upon the OpenMax layer approach by maximizing the inter-class distance and minimizing the intra-class distance on the penultimate layer. The work of [89] proposed a k-sigmoid activation-based loss function for training a neural network to be able to find an operating threshold on the final activation layer. [90] incorporated the latent representation for reconstruction along with the discriminative features obtained from a classification model to enhance the feature vector used for open-set detection.

In recent years, the generative models has gained considerable attention, which are capable of learning the probabilistic distribution over the data from its samples and making up new instances similar to real seen data. A myriad of works have taken advantage of the adversarial learning (AL) [91] to generate artificial unknown samples from the known class training data to fine-tune the model [92; 93; 94]. However, the likelihood-based generative models learn the density using the tractable likelihood approximation and researchers, as an intuitive approach, utilize likelihood as a metric to identify open-set samples [95; 96; 97].

[98] utilized reconstruction error obtained from a multi-task learning framework as a detection score. Recently, [99] proposed using self-supervision and augmented the input image to learn richer features to improve separation between classes.

Ge et al. [92] extended OpenMax [26] by synthesizing the unknown samples using a Generative Adversarial Network (GAN) based framework. Along the same line, [93] proposed the counterfactual image generation (OSRCI) framework which employs a GAN to generate samples placing between decision boundaries that can be treated as unknown examples. [100] proposed class-conditioned auto-encoder (C2AE) algorithm

where conditional reconstruction helps learning of both known and unknown score distributions.

It should be noted that there are related problems in the literature including outlier detection [101; 102] and anomaly detection [103; 104] that have some overlap with open-set recognition and can be treated as a relaxed version of open-set recognition. These problems assume the availability of one abnormal class during training. However, general open-set recognition problems usually do not provide information about the type or the number of the unknown classes in advance.

2.4 Domain Adaptation

Domain Adaptation is a sub-discipline of transfer learning [105], where the former refers to the situation that only data domains change for a certain task while the latter indicates a scenario which either the tasks and/or domains may differ. Domain Adaptation aims to align the disparity between domain distributions such that the trained model in the source domain can be generalized well into the target domain. Existing domain adaptation methods can be broadly categorized into three mains categories: 1) discrepancy-based methods, 2) reconstruction-based methods, and 3) adversarial-based methods.

Discrepancy-based approaches tend to minimize the domain shift between domain distributions by means of some distance metrics. The popular distance measures in domain adaptations are maximum mean discrepancy (MMD) [106], Kullback-Leibler (KL) divergence [107], correlation alignment (CORAL) [108], and contrastive domain discrepancy (CDD) [109]. The transfer component analysis [110] decreases the marginal distributions with the help of MMD criterion. The advent of deep learning has motivated researchers to utilize deep neural networks (DNNs) for learning transferable features across domains. Deep adaptation network (DAN) [111] align the marginal distributions using multiple adaptation layers in the DNN through multiple kernel variant of MMD (MK-MMD). Deep transfer network (DTN) [112] matches

conditional distributions in addition to the marginal distributions by minimizing both the marginal and conditional MMD based objective function.

Another category of domain adaptation methods aligns the data distributions by learning an invariant representation across domains. They mostly utilize autoencoder network and train parameters of the encoder using the data from the source domain and learn the decoder to reconstruct samples of the target domain. For instance, [113] employs stacked autoencoders (SDA) [114] to extract high-level representation from both the source and the target domains by minimizing the reconstruction error. Marginalized SDAs [115] addresses the limitations of SDAs by the use of linear denoisers to learn parameters without the need for stochastic gradient descent (SGD) [116]. To tackle domain adaptation in object recognition, an encoder-decoder network, called deep reconstruction-classification network (DRCN) [117] was proposed where the encoder is trained to predict the source labels while the decoder generate the samples in the target domain.

Adversarial domain adaptation methods learns domain-invariant features and minimizes the cross-domain discrepancy through an adversarial learning objective function. [118] incorporated an effective Gradient Reversal Layer (GRL) to the model to align the domain distributions. Adversarial domain adaptation using one discriminator matches only marginal distributions and leads to discrepancy in the classes across domains. Multi-adversarial domain adaptation (MADA) [119] employed multiple class-wise domain discriminators to reduces the shift between class-conditional distributions. From another perspective, [120; 121] proposed minimax optimization based adversarial domain adaptation to align marginal and conditional distributions by combining several objective functions.

Domain adaptation is crucially necessary in remote sensing areas since domain shift is inevitable in satellite imagery acquisition due to equipment differences and environment changes [122]. Many approaches have been proposed to tackle the challenging domain shift existed in hyperspectral image classification. A group of works rely on labeled samples in the target domain to transfer knowledge across data

sets [123; 124; 125; 126]. For instance, [127] used multiple geodesic flows obtained from source and target domain as features to perform classification with an SVM. [128] and [129] presented a tensor alignment technique and a heterogeneous method for adapting the domains with a few labeled samples, respectively. HT-CNN [66] transferred features learned from low-level and mid-level layers of VGG [10] network through an attention mechanism to the target dataset.

Another group of works develop unsupervised domain adaptation methods for cross domain knowledge transferring. [130] presented a principal component based approach to align the domain in a tensor feature space. [131] proposed a discriminative transfer joint matching framework to match the distributions. [132] developed a framework to adapt both the classifier and the feature extraction to the target domain. [133] proposed an unsupervised domain adaptation by transforming data into a shared space. Our method falls into the second category as it does not need any labeled samples from the target domain.

Chapter 3

Representative-Discriminative Learning for Open-set Recognition

Recent advancements in computer vision, especially the advent of Convolutional Neural Networks (CNN), have significantly improved the performance of image classification [8; 9; 10], detection [11; 12], and segmentation [13; 14] tasks, enabling their deployment in many different fields. However, the vast majority of existing approaches work under the "static closed world" assumption, meaning that the training and testing sets are drawn from the same label set. As a result, a system observing any unknown class is forced to misclassify it as one of the known classes, thus, weakening the recognition performance. Therefore, a more realistic scenario is to work in a non-stationary and open environment that not all categories are known *a priori* and testing samples from unseen classes can emerge unexpectedly. Recognition of known and unknown samples in a given dataset and correctly classifying known samples is defined as "open-set Recognition".

In this chapter, we study the problem of open-set recognition that identifies the samples belonging to unknown classes during testing phase, while maintaining performance on known classes. This work [134] is motivated from satellite imagery processing (described in Sec. 3.1) which is although inherently a classification problem,

both representative and discriminative aspects of data need to be exploited in order to better distinguish unknown classes from known ones. We propose a representative-discriminative open-set recognition (RDOSR) framework (in Sec. 3.2), which 1) projects data from the raw image space to the embedding feature space that facilitates differentiating similar classes, and further 2) enhances both the representative and discriminative capacity through transformation to a so-called abundance space. Experimental results (Sec. 3.3) on multiple satellite benchmarks demonstrate effectiveness of the proposed method. We also show the generality of the proposed approach by achieving promising results on open-set classification tasks using RGB images.

3.1 Motivation

Land cover classification which aims to assigns a semantic label to each individual pixel in satellite imagery faces a unique challenge: the vast area covered by the satellite imagery makes the task of generating representative training samples almost impossible, as there are a large variety of materials existed on the Earth's surface, especially those not well-exploited regions. Therefore, one of the most essential capabilities of land cover classification is to be able to automatically identify which test image and which area or pixel location of the image, has a higher probability of hosting new classes of materials. This would provide essential guideline to human operators in collecting training samples for the new classes.

Existing land cover classification models assume a closed-set setting where both the training and testing classes belong to the same label set. However, due to the unique characteristics of satellite imagery with extremely vast area of versatile cover materials, the training data are bound to be non-representative. Hence, a more realistic scenario that is able to work in a dynamic and unpredictable environment is needed. Recognition of known and unknown pixels in a given satellite image and correctly classifying known pixels is defined as "open-set land cover classification".

Fig. 3.1 explains this process using a real-world satellite image. It illustrates that there might be some regions in the satellite image with unknown categories that need to be identified.

Motivated by this real-world problem, we present a multi-tasking representative-discriminative open-set recognition (RDOSR) framework to address the challenging land cover classification problem, where both the representative and discriminative aspects of data are exploited in order to best characterize the differences between known and unknown classes. We propose the representative and discriminative learning among three spaces, as shown in Fig. 3.2, including 1) the transformation from the raw image space to the embedding feature space, and 2) the transformation from the embedding feature space to a so-called abundance space. See Supplement A for an illustration of the effect in different spaces.

The benefits of the proposed framework can be summarized from four aspects. First, unlike other open-set recognition methods applied on the raw image space directly, we propose to first learn a classification network that would transform from the raw image space to an embedding feature space such that a more discernible input is fed into the subsequent open-set learning network. Second, we propose to use the so-called Dirichlet-net to transform data from the embedding feature space to the abundance space. Due to the resolution issue, each pixel in a satellite image covers a large area with more than one constituent material, resulting in "mixed pixel". The mixtures are generally assumed to be a linear combination of a few spectral bases, with the corresponding mixing coefficients (or abundances). This way, instead of looking at the mixed pixel, we study the mixing coefficients of each spectral basis in making up the mixture. Thus the abundance space provides a finer-scale representation.

Third, to the best of our knowledge, this work is the first attempt to address the critical open-set land cover classification problem essential for analyzing the Earth's surface. Fourth, while the proposed method was motivated by satellite imagery analysis, it is generalizable to RGB images and achieves promising results.

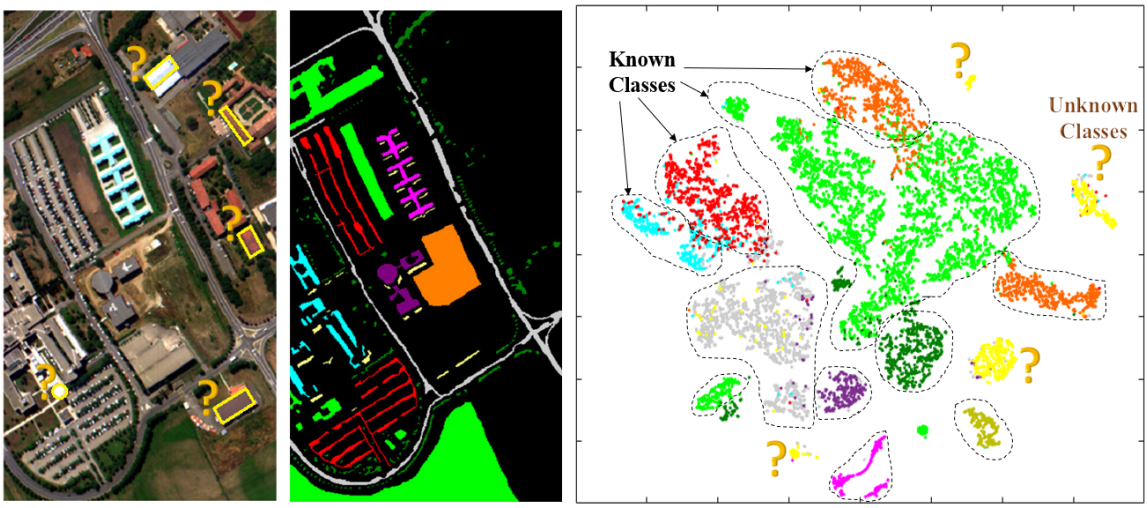


Figure 3.1: Open-set land-cover classification: Data samples corresponding to ground truth categories are from the known class set (**K**). It is likely that some categories are not known during training and will be encountered at testing, *i.e.*, samples from unknown class set (**U**). The goal is to identify pixels coming from (**U**), while correctly classify any pixel belonging to (**K**). From left to right, a satellite image from the Pavia University dataset [4] showing unknown materials surfaces with yellow bounding boxes, the ground truth labels, and visualization of the feature space for both known and unknown classes using tSNE [5].

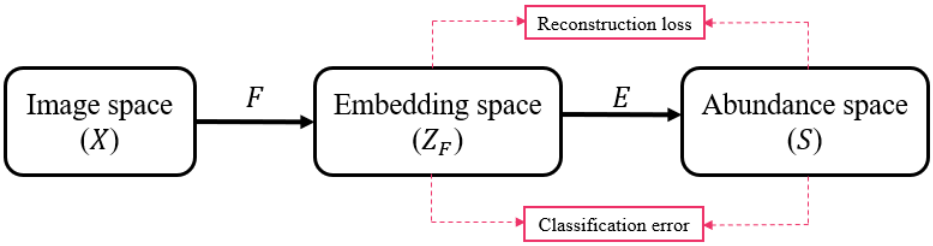


Figure 3.2: Representative-discriminative learning through the transformation among 3 spaces: the raw image space, the embedding space, and the abundance space.

3.2 Approach

We propose a representative-discriminative open-set recognition (RDOSR) structure, as shown in Fig. 3.3. The network mainly consists of two components, 1) a closed-set embedding component to project the data from the original image domain to the embedding domain, such that different classes with similar spectral characteristics are more distinguishable, and 2) a multi-task representative-discriminative learning component to learn a better representation scheme at a finer scale in the abundance space, such that unknown classes can be better differentiated from known classes.

3.2.1 Network Architecture

One challenging issue of the open-set satellite land cover classification problem is that different classes may possess similar spectral characteristics. Thus, it is likely that an unknown class, whose spectral profile is close to that of a known class, may be misclassified as the known class. To address this issue, instead of detecting unknown classes on the image domain, we detect them on the embedding domain projected by a closed-set embedding layer, as shown in Fig. 3.3.i. The closed-set embedding layer increases the discriminative power of network to a large extent, such that unknown classes can be better recognized even if their spectra are similar to those of the known classes. The weights of the closed-set embedding layer are trained with a classifier F, which is further elaborated in Sec. 3.2.2.

To recognize unknown classes in the embedding domain, we propose a multi-task representative-discriminative feature learning framework to boost both the representative and discriminative power of the extracted feature vector, such that it is more informative and effective to recognize unknown samples. This is shown in Fig. 3.3.ii. The network consists of an encoder-decoder architecture with the representative features S extracted using a sparse Dirichlet encoder E, and a decoder formed by the bases shared among known classes. A classifier C applied on S is also included to further increase its discriminative capability. In this way, the data from

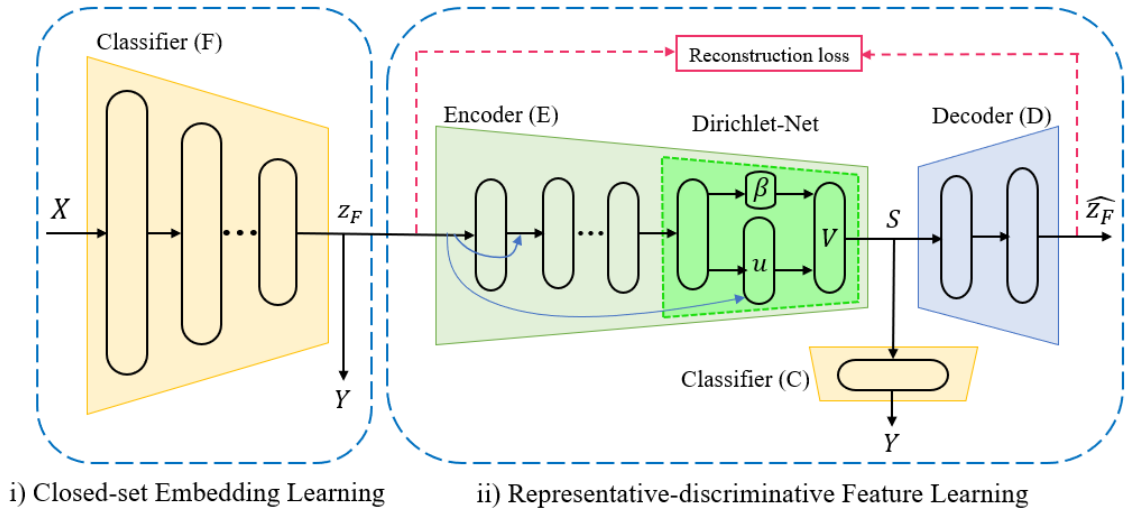


Figure 3.3: An overview of the proposed framework: i) Closed-set embedding learning: the classifier F is trained on the spectral domain X to produce latent discriminative embedding $\mathbf{z_F}$. ii) Representative-discriminative feature learning: the encoder E takes the embedding feature $\mathbf{z_F}$ and derives the representative features S using a Dirichlet-Net. The classifier C applied on S enhances the discriminative aspect of S, and the reconstruction error between the decoder output ($\hat{\mathbf{z}}_\mathbf{F}$) and input to encoder ($\mathbf{z_F}$) enhances the representative aspect of S.

unknown classes fed into the network would produce higher reconstruction error, thus can be detected accordingly. The details of network design are further elaborated in Sec. 3.2.3.

3.2.2 Closed-set Embedding Learning

Given the set of sample pixels $X_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_k}\}$ from the known classes with each pixel, $\mathbf{x}_i$, being a high-dimensional vector recording the reflectance readings of different spectral bands in the hyperspectral image, the corresponding labels are denoted with $Y_k = \{y_1, y_2, \ldots, y_{N_k}\}$, where N_k is the number of known pixels and $\forall y_i \in \{1, 2, \ldots, L\}$, where L is the number of known classes. To distinguish classes with similar spectral distributions, we project the input data X_k from the image domain to the embedding domain Z_F. The projection is learned through a classifier, F, with parameters Θ_F and the embedded features, $\mathbf{z_F}$ is forced to be discriminative through the cross-entropy loss,

$$\mathcal{L}_f(\Theta_F) = -\frac{1}{N_k} \sum_{i=1}^{N_k} y_i \log[F(\mathbf{x}_i)], \tag{3.1}$$

where y_i is a one-hot encoded label and $F(\mathbf{x}_i)$ denotes the vector carrying the predicted probability score of the i^{th} known sample. Such vector is generated by applying a softmax function on the features $\mathbf{z_F}$ in the embedding domain.

This general structure is sufficient for a common classification problem where the encountered classes are known. However, our goal is to increase the discriminative power of the features from classes with similar spectral characteristics. Therefore, we further increase the discriminative capacity of the embedded features with the l_1-norm sparse constraint defined by

$$\mathcal{L}_z(\Theta_F) = \frac{1}{N_k} \sum_{i=1}^{N_k} \|\mathbf{z_F}_i\|, \tag{3.2}$$

where $\mathbf{z}_{\mathbf{F}i}$ is the embedding feature vector learned by the classifier F. With such constraint, the embedded features from samples of different classes are more discriminative, even if their spectra are similar in the image domain.

3.2.3 Representative-Discriminative Feature Learning

With the proposed closed-set embedding layer, the samples are projected from the image domain to the embedding domain possessing more distinguishable features. In order to better identify unknown samples, both discriminative and representative nature of the samples need to be exploited. Previous approaches [98; 100] usually train a general auto-encoder to reconstruct samples from known classes. When the samples from unknown classes are fed into the network, the reconstruction error is expected to be larger than that of the known classes in the ideal case, since the network weights are optimized during the training procedure using known samples. However, the challenge lies in the scenario when the unknown classes especially the ones close to the known classes may potentially contribute to small reconstruction error too which would lead to failure in detection.

In this work, instead of adopting a general-purpose auto-encoder, we propose a multi-task representative-discriminative feature learning framework to improve the detection accuracy. The purpose of this network is to decrease the reconstruction error of the known classes while increasing the reconstruction error of unknown classes intensively.

Due to the resolution issue, each pixel in a satellite image usually covers a large geographical area or footprint (e.g., 30×30m for Landsat-8), resulting in the so-called "mixed pixel" (i.e., each pixel tends to cover more than one constituent materials). These mixtures are generally assumed to be a linear combination of a few spectral bases with the corresponding mixing coefficients (or abundance). The proposed method is designed based on this assumption as shown in Eq. 3.3. Assume that the feature vector of a sample from known classes $\mathbf{z}_{\mathbf{F}}$ is a linear combination of a few

bases B, and such bases are shared among features of the known classes. Thus, each sample of known classes can be decomposed with

$$\mathbf{z_F} = \mathbf{s}B, \tag{3.3}$$

where $\mathbf{s}$ denotes the proportional coefficients of the shared bases, which serves as a form of "representation" of the embedding feature, that we refer to as the abundance. The abundance vector, or representation, should satisfy two physical constraints, i.e., non-negative and sum-to-one. The samples from unknown classes are also able to be decomposed by Eq. 3.3 using shared bases of the known classes, B. However, since B does not include the bases of the unknown classes, the distribution of its representations $\mathbf{s}$ should deviate from that of the known classes. Therefore, we design a network following the model of Eq. 3.3, which enforces $\mathbf{s}$ from known classes to follow a certain distribution. And if the network can extract $\mathbf{s}$ from unknown classes with similar distributions, then we expect they have high reconstruction errors.

The flowchart of the proposed multi-task representative-discriminative feature learning is detailed in Fig. 3.4. The network performs both the reconstruction task and the classification task. The reconstruction branch consists of a sparse Dirichlet-based encoder E with weights Θ_E and a decoder D with weights Θ_D. The encoder and decoder can be defined by the functions $E : Z_F \to S$ and $D : S \to Z_F$, respectively, where Z_F is the embedding space obtained by the closed-set classifier F, and S is the abundance space of latent representations projected by the encoder E. The representations $\mathbf{s}$ in the latent space S is enforced to follow a Dirichlet distribution. And a sparse constraint is introduced to enhance the representativeness of $\mathbf{s}$. More details and justifications are provided below. In addition, S is also enforced to be discriminative by the classifier C, which can be defined by the function $C : S \to Y$ with weights Θ_C, where Y is the space of known labels.

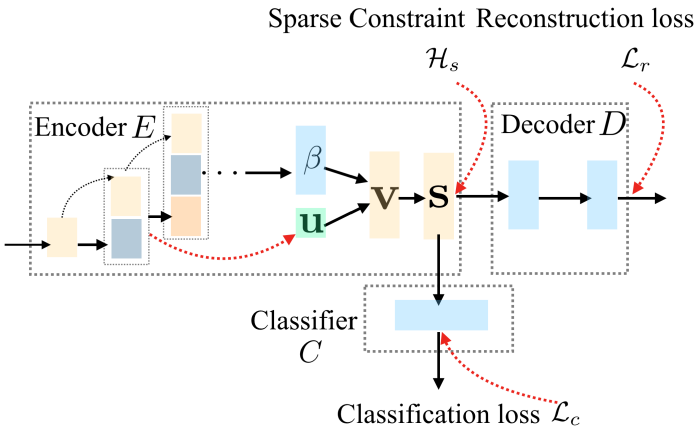


Figure 3.4: The flowchart of the multi-task representative-discriminative feature learning framework.

Representative Feature Learning with Reconstruction

The reconstruction branch is constructed according to Eq. 3.3, where the shared bases are embedded in the decoder D of the network and the corresponding representation $\mathbf{s}$ is extracted with the encoder E. As mentioned in Sec. 3.2.3, $\mathbf{s}$, the proportional coefficients of the bases, needs to satisfy the non-negativity and sum-to-one constraints. Since the Dirichlet distribution is able to provide this property as its output is a distribution in which the sum of the variables always equals 1, we enforce $\mathbf{s}$ to follow a Dirichlet distribution. Following the work of [135; 136; 137], we adopt the stick-breaking structure in the encoder to enforce the representations $\mathbf{s}$ to follow the Dirichlet distribution.

Assuming that the representation $\mathbf{s}$ is denoted as $\mathbf{s} = (s_1, \cdots, s_j, \cdots, s_c)$, the stick-breaking structure states that a single element s_j in $\mathbf{s}$ can be expressed by

$$s_j = \begin{cases} v_1 & \text{for} \quad j = 1 \\ v_j \prod_{o<j}(1 - v_o) & \text{for} \quad j > 1, \end{cases} \tag{3.4}$$

where v_o is drawn from a Beta distribution, i.e., $v_j \sim \text{Beta}(u, \alpha, \beta)$. Since it is not easy to get samples from the Beta distribution directly, the inverse transform of Kumaraswamy distribution, as shown in Eq. 3.5, is utilized for sample generation. In particular, the inverse transform of Kumaraswamy distribution with $\alpha = 1$ or $\beta = 1$ is equivalent to Beta distribution.

$$\text{Kuma}(u, \alpha, \beta) = \alpha\beta u^{\alpha-1}(1 - u^{\alpha})^{\beta-1} \tag{3.5}$$

where $\alpha > 0$, $\beta > 0$, and $u \in (0, 1)$. The characteristics of Kumaraswamy include having a closed-form CDF and the inverse transform as follows.

$$v_j \sim (1 - (1 - u^{\frac{1}{\beta}})^{\frac{1}{\alpha}}). \tag{3.6}$$

We set $\alpha = 1$ and the two parameters u and β are learned as the hidden layers in the encoder of the network as illustrated in Fig 3.4. A softplus activation function is adopted on the layer β due to its non-negative property, and a sigmoid is used to map u into the $(0, 1)$ range at the layer u. More details of the stick-breaking structure can be found in [136] and [137].

In addition, the entropy function [138] is adopted to reinforce the sparsity of the representation layer. Let $\hat{s}_j = \frac{|s_j|}{\|\mathbf{s}\|}$, for each pixel, the entropy function is defined as,

$$\mathcal{H}_s(\Theta_E) = -\sum_{j=1}^{c} \hat{s}_j \log \hat{s}_j. \tag{3.7}$$

where c is the dimension of the representation $\mathbf{s}$. The reconstruction loss $\mathcal{L}_r$ is adopted to reduce the reconstruction error of the known classes. It is defined by,

$$\mathcal{L}_r(\{\Theta_E, \Theta_D\}) = \frac{1}{N_k} \sum_{i=1}^{N_k} \|\mathbf{z}_{\mathbf{F}i} - \hat{\mathbf{z}}_{\mathbf{F}}\|_2, \tag{3.8}$$

where $\mathbf{z}_{\mathbf{F}i}$ is the embedding feature vector fed into the encoder E, and $\hat{\mathbf{z}}_{\mathbf{F}}$ is the reconstructed $\mathbf{z}_{\mathbf{F}i}$ obtained from the decoder D.

Discriminative Feature Learning with Classification Branch.

To further increase the discriminative capacity of the representations, a classifier is adopted on the representations $\mathbf{s}$ with the classification loss $\mathcal{L}_c$ defined as,

$$\mathcal{L}_c(\{\Theta_E, \Theta_C\}) = \frac{1}{N_k} \sum_{i=1}^{N_k} y_i \log[E(\mathbf{z}_{\mathbf{F}i})], \tag{3.9}$$

where y_i is the ground truth label and $E(\mathbf{z}_{\mathbf{F}i})$ denotes the representative feature vector of the i^{th} known sample. Note that the weights of both the reconstruction branch and classifier C are updated together, such that the learned representations can be both representative and discriminative.

3.2.4 Training Procedure and Network Settings

We first learn the embedding projection by optimizing the weights Θ_F of the classifier F with the loss function,

$$\min_{\Theta_F} \lambda_f \mathcal{L}_f + \lambda_z \mathcal{L}_z, \tag{3.10}$$

where λ_f and λ_z are two parameters to balance the trade-off between the cross-entropy loss and the sparsity loss.

Then, having the learned embedding layer, the multi-task representative-discriminative feature learning network is trained to minimize both the reconstruction loss and the classification error of the known classes with loss function,

$$\min_{\Theta_E,\Theta_D,\Theta_C} \lambda_r \mathcal{L}_r + \lambda_s \mathcal{H}_s + \lambda_c \mathcal{L}_c, \tag{3.11}$$

where λ_r, λ_s, and λ_c are parameters balancing the trade-off between the reconstruction loss, the sparsity loss, and the classification loss.

The structures of the F, E, D, and C networks are as [512,1024,512,32,L], [3,3,3,3, 10], [10,10,L], and [L], respectively, which each number represents the number of nodes in the fully connected layer and L is the number of known classes.

3.3 Experiments and Results

In this section, the effectiveness of the proposed RDOSR method is evaluated on several widely used benchmark hyperspectral image datasets. In addition, we demonstrate the generalization capacity of the proposed approach on RGB image datasets. Furthermore, the contribution from each component of the proposed framework is analyzed through ablation study.

3.3.1 Implementation Details

We train the network, described in Sec. 3.2.1, using an Adam optimizer [139], with a learning rate of 10^{-3}. The classifier F and the joint structure of encoder-decoder-classifier (E-D-C) are trained separately for a total number of 15K epochs. However, the other methods which do not have two separate components were trained for 6K epochs.

For training the classifier F, λ_f and λ_z are set equal to 1 and 0.1, respectively. The weights for the reconstruction λ_r, sparsity λ_s, and classification λ_c losses in training the E-D-C structure are set equal to 0.5, 10^{-3}, and 0.5, respectively. The sparsity weight λ_s is decayed with a weight decay of 0.9977. The classifier F is trained until its accuracy reaches 0.9988. It should be noted that all input data of the datasets are normalized to their mean value and unit variance. In addition, the feature vector obtained from the classifier F is divided by 10 to avoid divergence.

One of the factors that affects the performance of the open-set recognition algorithm is Openness [25] of the problem, defined as,

$$Openness = 1 - \sqrt{\frac{2 \times N_{train}}{N_{test} + N_{target}}}, \tag{3.12}$$

where N_{train}, N_{test} and N_{target} are the number of classes known during training, the number of classes given during testing, and the number of classes that need to be recognized correctly during testing phase, respectively. In the experiments, classes of each dataset is partitioned into known and unknown sets according to the Openness.

The code is written in TensorFlow, and all the experiments are performed on a desktop computer having GeForce GPU of 10 GB Memory.

3.3.2 Metrics

To compare performance of different methods, there are several metrics including overall accuracy or F-score on a combination of known and unknown classes, and

Receiver Operating Characteristic (ROC). The first two metrics do not characterize the performance of the model well due to their sensitivity not only to the performance of the model in classifying the known classes, but also an arbitrary operating threshold for detecting unknown samples.

On the other hand, the ROC curve would illustrate the ability of a binary classification system (here, known vs. unknown detection) as a discrimination threshold is varied from the minimum to the maximum value of the given detection measure (here, reconstruction error). Thus, it provides a measure free from calibration. To have a quantitative comparison, the area under the ROC (AUC) is computed in the experiments.

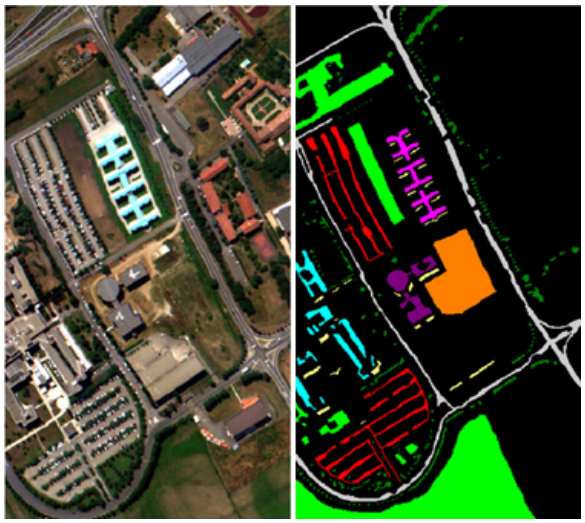
3.3.3 Open-set Recognition for Hyperspectral Data

The experiments are conducted on three hyperspectral image datasets pictured in Fig. 3.5. Note that the datasets are composed of individual pixels as the representation learning in the encoder-decoder networks and the classification in the classifier are performed in pixel-wise manner.

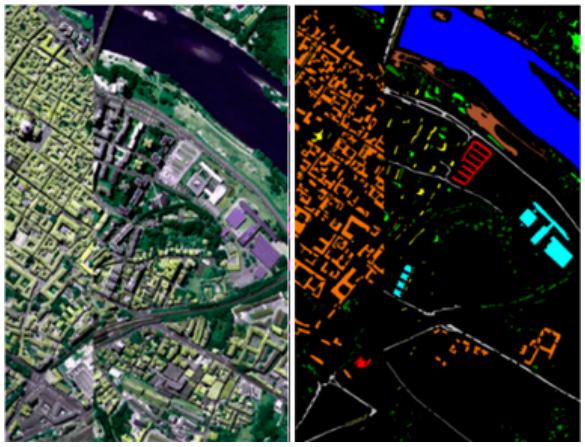
Pavia University (PU) and Pavia Center (PC). Both PU and PC datasets were gathered over Northern Italy in 2011 by the Reflective Optics Systems Imaging Spectrometer which has a resolution of 1.3 m. The dimension of the PU dataset is 1096×715 pixels with 103 spectral bands, ranging from 430 to 860 nm. The PC dataset has 610×340 pixels with 102 spectral bands. The PU and PC datasets both include nine land cover categories.

Indian Pines (IN). The IN dataset was collected over Northwest Indiana in 1912 by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). It has a dimension of 145×145 with a resolution of 20 m by pixel and 200 spectral bands. Its ground truth consists of 16 land cover classes.

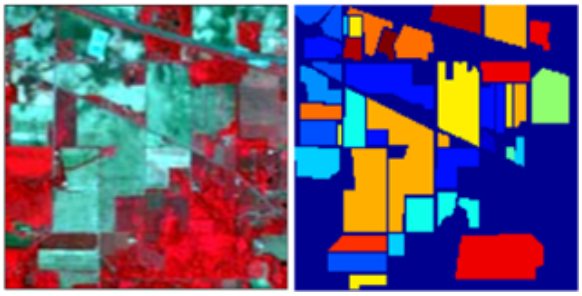
We compare the performance of the proposed approach with three methods described as follows.



(a)



(b)



(c)

Figure 3.5: The three hyperspectral benchmarks used in this chapter, including the satellite images and the corresponding ground truths with color indexes. The datasets are: a) Pavia University, b) Pavia Center, c) Indian Pines.

SoftMax: In a neural network classifier, a common confidence-based approach to detect open-set examples is thresholding the SoftMax scores. We use network structure of the classifier F without considering sparsity constraint.

OpenMax [26]: This approach calibrates the SoftMax scores in a classifier and augments them with a $N_k + 1$ class for an unknown category. The replaced SoftMax layer with an OpenMax layer is used for open-set recognition. We adopt the classifier mentioned earlier in the SoftMax method, and use the Weibull fitting approach with parameter $Weibull\,tail\,size = 10$ to generate OpenMax layer values.

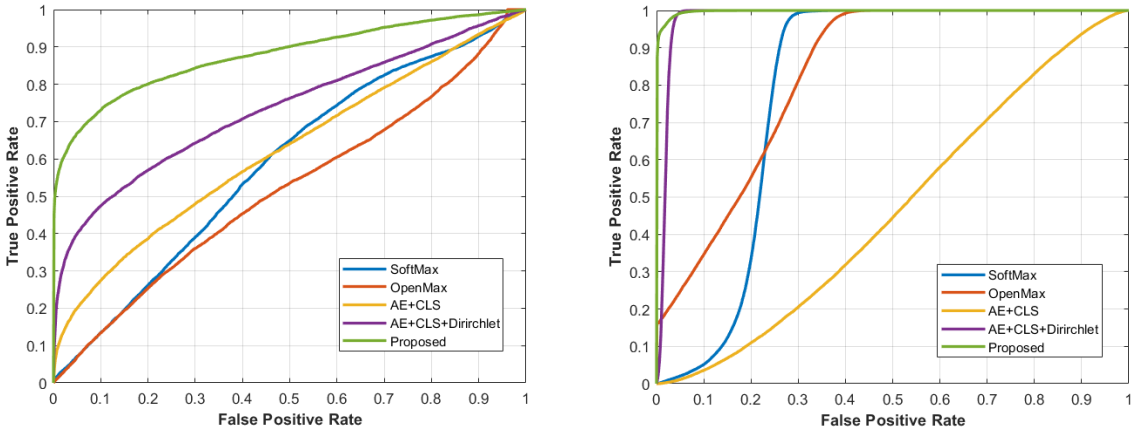
AE+CLS: Fully connected version of MLOSR [98] which utilizes a multi-task learning framework, composed of a classifier and a decoder with a shared feature extraction part, to detect open-set examples. To have a fair comparison with our approach, the encoder, decoder, and classifier are designed as our E (without the Dirichlet-Net), D, C, and trained with $\mathcal{L}_r$ and $\mathcal{L}_c$ loss, with weights of 0.5.

First, each of the L classes is assumed to be unknown which equates to an openness of 2.99%, 2.99%, and 1.63% for PU, PC, IN, respectively. The AUC values corresponding to choosing each of the L labels as unknown are averaged and reported for each method in Table 3.1. It can be observed that the proposed method outperforms other methods on all three datasets. See Supplement B for detailed comparison on the PU dataset. The minor improvement on the PC dataset can be justified by its differentiated spectrum of different classes which diminishes the effect of the classifier F.

Second, for the Openness equal to 6.46%, the ROC curves of different methods for the PU and IN datasets are illustrated in Fig. 3.6. As seen from the results on both datasets, the *AE+CLS+Dirichlet* method which adopts the Dirichlet net to the *AE+CLS* framework and the proposed method lie above all other methods. It should be noticed that our proposed method is able to detect unknown classes with 60% and above 90% accuracy and almost zero false detection for the PU and PC datasets, respectively.

Table 3.1: Area under the ROC curve for open-set detection. Results are averaged over L partitioning of the selected dataset to $L-1$ known and 1 unknown classes.

Method	**PU**	**PC**	**IN**
SoftMax	0.385	0.816	0.555
OpenMax [26]	0.441	0.884	0.415
AE+CLS [98]	0.586	0.757	0.669
AE+CLS+Dirichlet	0.714	0.927	0.681
RDOSR (Ours)	**0.773**	**0.963**	**0.802**



(a) Open-set detection on PU dataset **(b)** Open-set detection on PC dataset

Figure 3.6: Receiver Operating Curve curves for open-set recognition for PU and PC datasets, for $L = 7$ (openness=6.46%).

Third, the histograms of reconstruction error for both known and unknown sets with Openness= 2.99% are shown in Fig. 3.7. It can be observed that the reconstruction errors corresponding to the known set have small values. However, the unknown set produces larger error due to mismatches in terms of representative and discriminative features learned from the known classes examples.

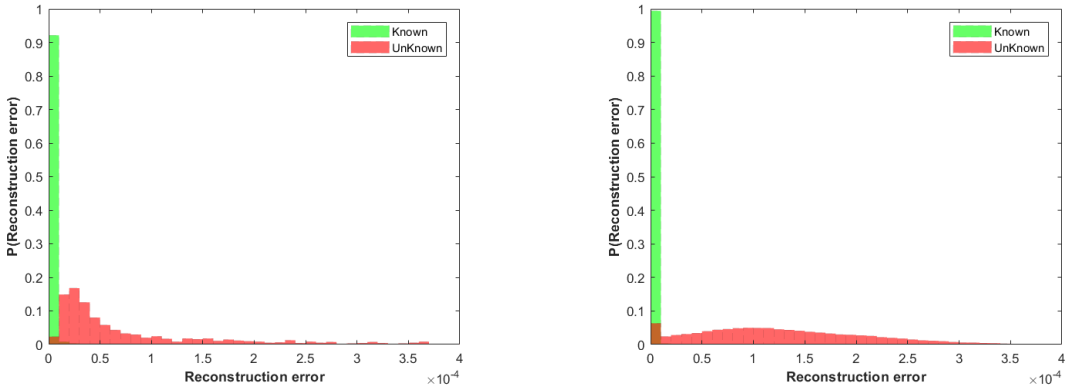
3.3.4 Open-set Recognition for RGB Images

To show the generalization capacity of the proposed method, we evaluate the performance of the proposed approach on two RGB datasets and compare with several state-of-the-art methods. For this purpose, the classifier F performing pixel-wise classification is substituted with a DenseNet structure which takes a 2D image as input.

Following the protocol in [93], we sample 4 known classes from CIFAR10 [140] to have Openness=13.39% and 20 known classes out of 200 categories of TinyImageNet [141] resulting in an Openness of 57.35%. Table 3.2 summarizes the results where the values other than the proposed RDOSR are taken from [99]. It can be observed that the proposed method has better performance over the compared methods, except for GDOSR [99], on CIFAR10. However, it achieves significant improvement on TinyImageNet. It may be due to the similarity between the classes in TinyImageNet which hinders detecting unknown samples in the image space while RDOSR addresses this issue by operating in the embedding space.

3.3.5 Ablation Study

Starting with a baseline, *AE+CLS*, each component is gradually added to the framework to show its effectiveness. The results corresponding to the ablation study are shown in Fig. 3.8. It can be seen that employing the baseline structure applied on the spectra domain has the worst performance. However, adding the Dirichlet-based



(a) PU dataset **(b)** PC dataset

Figure 3.7: Reconstruction error distribution of known and unknown classes using the proposed method for PU and PC datasets, for $L = 8$.

Table 3.2: Area under the ROC curve for Open-set recognition.

Method	**CIFAR10**	**TinyImageNet**
SoftMax	0.677	0.577
OpenMax [26]	0.695	0.576
OSRCI [93]	0.699	0.586
C2AE [100]	0.711	0.581
GDOSR [99]	**0.807**	0.608
RDOSR (Ours)	0.744	**0.752**

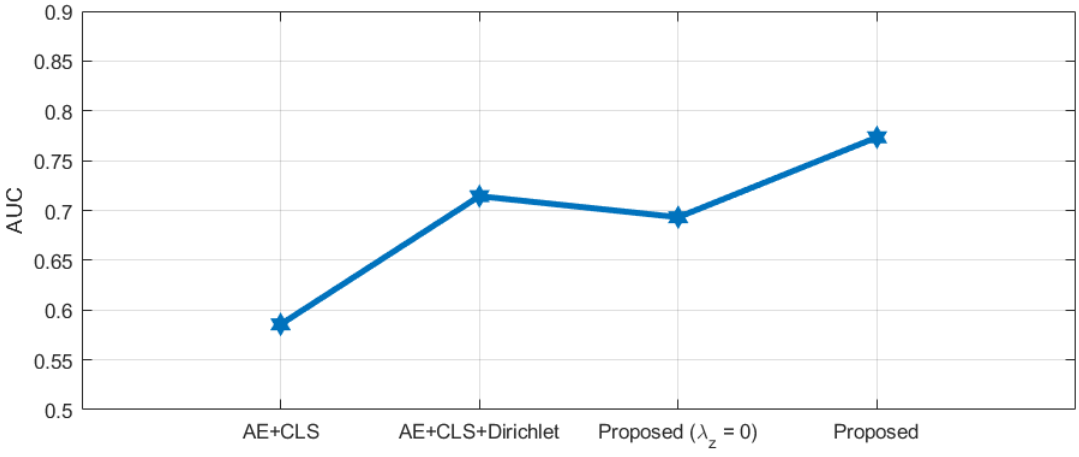


Figure 3.8: Ablation study of the proposed method on PU dataset

network makes a major improvement due to applying physical constraints on the latent space learned by the encoder E. Directly performing open-set recognition on an embedding space causes instability problem which is confirmed by a performance drop compared to the *AE+CLS+Dirichlet* method. Our proposed method addresses the instability issue by adopting a sparsity constraint on the embedding feature vector $\mathbf{z_F}$. As seen from Fig. 3.8, our proposed method achieves the highest AUC value compared to three other baseline methods.

3.4 Summary

We studied the challenging problem of open-set land cover recognition in satellite images. Although inherently a classification problem, both representative and discriminative features need to be learned in order to best characterize the difference between known and unknown classes. We presented the transformation among three spaces, that is, the original image space, the embedding feature space, and the abundance space, where features with both representative and discriminative capacity can be learned to maximize success rate. The proposed multi-tasking representative-discriminative learning structure was evaluated on three hyperspectral and two RGB image datasets and exhibited significant improvement over state-of-the-art open-set recognition algorithms.

Chapter 4

Hyperspectral Image Domain Adaptation Through Unmixing Abundance Maps

Most machine learning-based algorithms assume that training and test data share the same distribution. Hence, it is expected that a model trained on the training set would perform well on the test data. However, in real-world applications this assumption may not work since the training and the test sets may come from different distributions due to many reasons, including various data acquisition conditions, variation in the data itself over time, etc. The discrepancy between the training and test data is referred to as the domain shift [105]. As a result, directly deploying the trained model on the test data may lead to degradation in the performance. Therefore, a more realistic approach is to align the two domain distributions such that the data shift between the training and test domain is minimized [110; 142].

In this chapter, we study the problem of domain adaptation for the classification task to align the distributions across the training and the test domains [143]. This work is motivated from satellite imagery analysis (described in Sec. 4.1) where domain shift is inevitable leading to importance of domain adaptation in this field. We

propose an unsupervised domain adaptation approach (in Sec. 4.2) for hyperspectral land cover classification. The uniqueness of this approach is that instead of aligning the distribution in either the training domain or the testing domain, we propose to first project both training and test data on a common embedding space where domain alignment can be done in a more efficient manner. The key is how to find this embedding space. Experimental results (Sec. 4.3) on several hyperspectral benchmarks demonstrate superiority of the proposed method.

4.1 Motivation

Despite the great success of land cover classification in satellite imagery with the development of convolutional neural network [60; 61; 62; 63; 64], it suffers from two challenges: 1) a considerable amount of labeled samples are required for training a deep learning-based approach, 2) it assumes that both the training and test data share similar distributions leading to poor performance when it is deployed into images acquired from regions with different atmospheric, illumination, or environmental conditions. Therefore, a non-negligible time and effort needs to be invested to generate a representative training dataset containing all the variations of the spectra for different categories existed in the dataset.

The main reason for the problem mentioned above is spectral variability phenomenon where spectral of the pixels belonging to a certain category vary across hyperspectral images due to different acquisition conditions. The spectral variability is an inevitable characteristic of satellite imagery which hinders the deployment of the land cover classification methods to new regions as it requires collecting new samples which is laborious and time-consuming. Hence, a more preferable scenario is to take advantage of the existing annotated datasets and adapt the trained model such that it performs well on the images taken from new domains. This setting is called hyperspectral domain adaptation, where the dataset with labeled samples and

the new dataset collected under different acquisition condition are termed as source and target domains, respectively.

To tackle this problem, some methods work based on knowledge transferring from the source data to the target data, such that the pretrained model is often finetuned in the target domain. For instance, [144] trains a multi-kernel classifier on the source data and the acquired samples from the target dataset, sequentially. [145] examines the correlation between domains using a deep network with some salients samples extracted from each domain, and finetunes it for the classification task. [146] learns spectral-spatial features through a hierarchical stacked sparse autoencoder and transfer them with help of few samples labeled from the target data. HT-CNN [66] transfers features learned from low-level and mid-level layers of a VGG [10] network through an attention mechanism to the target dataset. [147] minimizes the data disparity among the datasets and enhances feature embedding space with few samples of the target data which have been labeled. Despite their competitive performance on the target domain, the finetuning step requires a large set of annotated data ifrom the target domain which may not be feasible for some applications.

Unlike the works mentioned above, we present an unsupervised domain adaptation method which does not require any annotated samples from the test dataset. The proposed multi-tasking unmixing-based representation learning aligns the representations across the domains, while the discriminative aspect of the data is preserved. In particular, it transfers the data from the image space to a latent space, which is called abundance space, where it is easier to classify pixels (regardless which domains they are from) in the abundance space.

The benefits of the proposed approach is three-folds. First, unlike other domain adaptation methods that blindly minimize the domain discrepancy, we propose an unmixing-based approach which transforms the data from the image domain to a physically constrained space where the classification has a good performance on both the source and target domains. Second, the representation of the source and target

data in the latent abundance space are further aligned by penalizing their marginal distribution distance which helps to minimize the discrepancy between distributions.

4.2 Approach

We present a multi-tasking domain adaptation approach that reduces the disparity between source and target domains through mapping them into an unmixing-based representation space, as illustrated in Fig. 4.1. The proposed framework composed of three main components, 1) an unmixing-based representation learning component to map the pixel spectra from both source and target image spaces to a latent space which is physically constrained, 2) a distribution alignment approach to further minimize the domain-shift across domains, and 3) a 3D-CNN-based classifier applied on the learned abundance representation to predict the category taking into consideration both spectral and spatial information to enhance discriminative power.

4.2.1 Problem Setting

Domain adaptation involves two datasets: a source data $G_s = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_s}$ with labels $y_i \in 1, 2, ..., l$ where l denotes the number of categories, and a target data $G_t = \{(\mathbf{x}_j)\}_{j=1}^{N_t}$ without labels, where N denotes the number of samples. Note that there is a domain shift between the source set G_s and the target set G_t since they are drawn from different distributions.

To be specific, the representation learning structure consists of two encoder-decoder networks, each of which operating on a specific domain. The encoder E with shared weights among domains transforms the source data $\mathbf{x}_s$ and the target data $\mathbf{x}_t$ to the shared abundance space as $\mathbf{a}_s$ and $\mathbf{a}_t$, respectively. The distribution of the abundance representation corresponding to the source and the target data are further matched using a metric-based distribution alignment method. To perform classification, a 3D-CNN-based classifier is applied on the abundance representation

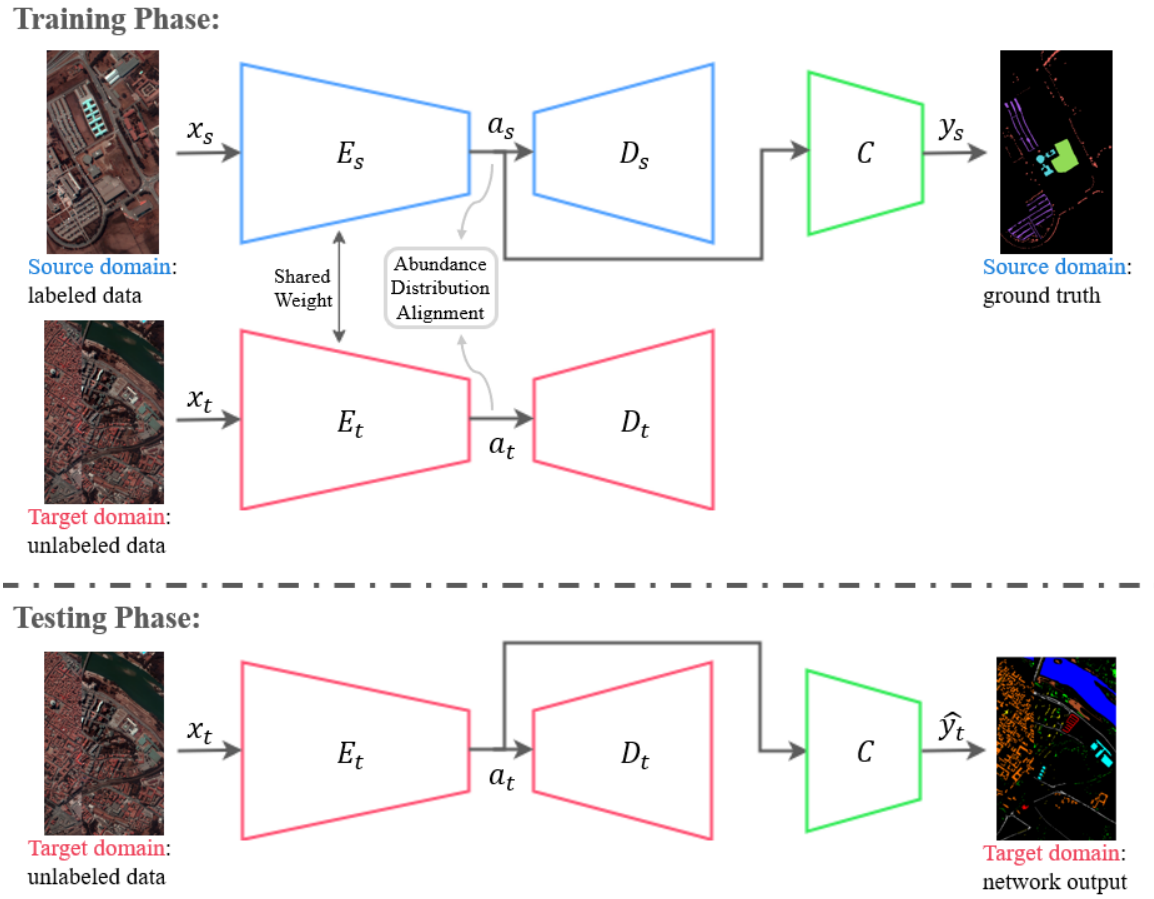


Figure 4.1: The overall framework of the proposed method: E_s, D_s and E_t, D_t are encoder-decoder networks operating on the source and target domains, respectively, where the weights between the encoder networks are shared. C is the 3D-CNN-based classifier applying on the abundance representation a. Note that source and target are represented as "s" and "t", respectively.

extracted from the source $\mathbf{a}_s$ and the target data $\mathbf{a}_t$ to predict the corresponding labels, $\hat{y}_s$ and $\hat{y}_t$, respectively.

In the following, we elaborate on the unmixing-based representation learning in Sec. 4.2.2, describe the abundance distribution alignment approach in Sec. 4.2.3, and present the 3D-CNN-based classifier in Sec. 4.2.4.

4.2.2 Unmixing-based Representation Learning

Due to trade-off between spectral and spatial acquisition, satellite imagery usually have low spatial resolution. Hence, each pixel covering a large area of the scene contains more than one building materials. According to unmimxing concept, the mixed pixels can be decomposed into a series of spectral bases and their corresponding mixing coefficients (i.e. abundance), such that the mixing coefficients should be non-negative and sum to one.

Inspired by this unmixing procedure, we propose to project the data from the image space into the latent abundance space where the discrepancy across domains can be minimized. We hypothesize that the pixels belonging to the same categories would have similar mixing coefficients even though the spectral bases are different due to domain shifting. Hence, the abundance feature that have a physical intuition can be considered as a finer representation of the pixel spectra and can be aligned better between different data domains.

Assume the spectra of the pixels in the source and the target domains are represented as $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively. Each of the spectra can be decomposed into the domain-specific bases and their corresponding mixing coefficients as follows.

$$\mathbf{x}_s = \mathbf{a}_s B_s \tag{4.1}$$

$$\mathbf{x}_t = \mathbf{a}_t B_t \tag{4.2}$$

where $\mathbf{a}_s \in \mathbb{R}^{1\times c}$ and $\mathbf{a}_t \in \mathbb{R}^{1\times c}$ denote the proportional coefficients of their corresponding spectral bases $B_s \in \mathbb{R}^{c\times l}$ and $B_t \in \mathbb{R}^{c\times l}$ for the source and target domains, respectively. The number of spectral bands of the given pixel $\mathbf{x}$ and the number of spectral bases are denoted as l and c, respectively. Note that the proportional coefficients $\mathbf{a}$ is equivalent to $\mathbf{s}$ in Eq. 3.3. However, we represent the proportional coefficients with $\mathbf{a}$ to avoid confusion with the subscript "s" associated with source domain in the domain adaptation concept.

Each reconstruction branch is constructed with an encoder E and a decoder D, with the corresponding weights as θ_E, θ_D, respectively, as illustrated in Fig. 4.2. The bases B are embedded in the decoder and the mixing coefficients $\mathbf{a}$ are estimated using the encoder. In particular, these two networks can be defined as $E : X \to A$ and $D : A \to X$, respectively, where X is the image space and A is the abundance space. Since the abundance vector $\mathbf{a}$ represents the mixing coefficients of the bases for a mixed pixel, they should satisfy two physical criteria: non-negativity and sum-to-one. For this purpose, we enforce them to follow Dirichlet distribution. The Dirichlet-based encoder is built by adopting the stick-breaking structure [135; 136; 137] in the encoder.

Given that the abundance feature vector $\mathbf{a}$ is equivalent to $\mathbf{s}$ in Eq. 3.3, it can be extracted with the encoder E following the stick-breaking structure described in Eq. 3.4, Eq. 3.5, and Eq. 3.6.

In addition, according to the fact that each mixed pixel is usually constructed by a few spectral bases, the mixing coefficient vector $\mathbf{a}$ is encouraged to be sparse using the entropy function [138]. Assuming $\hat{a}_k = \frac{|a_k|}{\|\mathbf{a}\|}$, the entropy function is expressed as,

$$\mathcal{H}(\theta_E) = -\sum_{i=1}^{N_s}\sum_{k=1}^{c} \hat{\mathbf{a}}_{\mathbf{s_{i,k}}} \log \hat{\mathbf{a}}_{\mathbf{s_{i,k}}} - \sum_{j=1}^{N_t}\sum_{k=1}^{c} \hat{\mathbf{a}}_{\mathbf{t_{j,k}}} \log \hat{\mathbf{a}}_{\mathbf{t_{j,k}}}. \tag{4.3}$$

where c denotes the number of bases. The reconstruction loss penalizing the encoder-decoder networks for reconstructing the source and target samples is defined as,

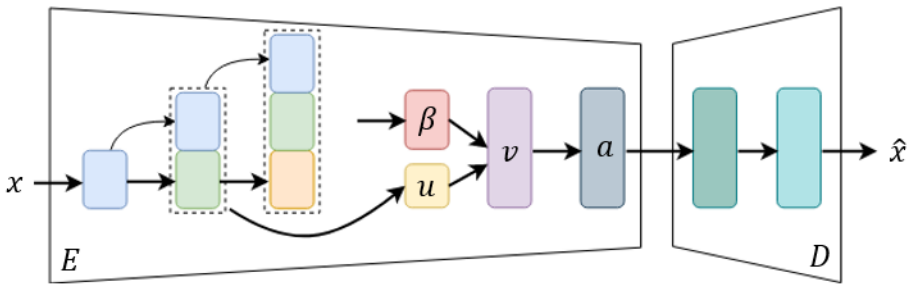


Figure 4.2: The structure of the encoder E and decoder D. The features are concatenated to the features of the subsequent layers to enhance the representative power of the encoder.

$$\mathcal{L}_r(\{\theta_E, \theta_{D_s}, \theta_{D_t}\}) = \frac{1}{N_s}\sum_{i=1}^{N_s} \|\mathbf{x_{s_i}} - \hat{\mathbf{x}}_{\mathbf{s_i}}\|_2 + \frac{1}{N_t}\sum_{j=1}^{N_t} \left\|\mathbf{x_{t_j}} - \hat{\mathbf{x}}_{\mathbf{t_j}}\right\|_2, \tag{4.4}$$

where $\hat{\mathbf{x}}$ denotes the reconstructed sample through encoder-decoder networks.

4.2.3 Distributions Alignment in Abundance Space

To further reduce the distribution discrepancy in the abundance space, a distance called maximum mean discrepancy (MMD) [106] is adopted. Given two domains G_s and G_t, the marginal distribution discrepancy between the two distributions can be obtained as

$$\mathcal{L}_a(\{\theta_E, \theta_{D_s}, \theta_{D_t}\}) = d(G_s, G_t) = \left\| \frac{1}{N_s}\sum_{x_i \in G_s} E(x_i) - \frac{1}{N_t}\sum_{x_j \in G_t} E(x_j) \right\|_H^2 \tag{4.5}$$

where H represents the reproducing kernel Hilbert space (RKHS) and $\|.\|_H$ denotes the RKHS norm. For more details, the reader can refer to [148; 149; 150]. Note that E represents the encoder and the MMD distance is calculated on the representation space obtained from the encoder network. By minimizing the the objective function mentioned in Eq. 4.5, the marginal abundance distributions from source and target data can be drawn closer to match each other.

4.2.4 3D-CNN-based Classifier

To perform classification taking into consideration both the spectral and spatial information, a 3D-CNN-based classifier C [133] with weights θ_C is employed, as illustrated in Fig. 4.3. The classifier network comprises of several blocks and a fully-connected at the end. Each of the blocks consists of a 3D convolutional layer, a bachnorm layer [151], and a ReLU activation function [152]. In addition, the output of each block is densely concatenated to the input of its subsequent blocks to strengthen the discriminative power of learned feature. The classifier takes abundance maps

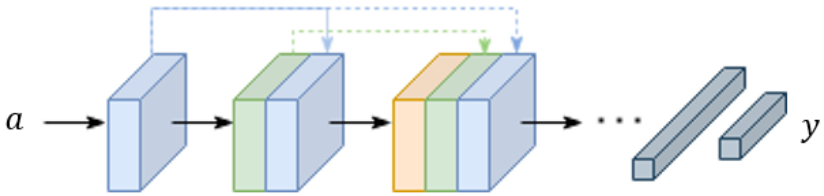


Figure 4.3: The structure of the 3D-CNN-based classifier C. The features are concatenated to the features of the subsequent layers to enhance the discriminative power of the classifier.

extracted from the input image patch using the Dirichlet-based encoder E and predicts a semantic label corresponding to the pixel in the center of the image patch. The classifier is trained with a cross-entropy objective function using annotated samples in the source domain as follows.

$$\mathcal{L}_c(\{\theta_E, \theta_C\}) = -\sum_{i=1}^{N_s} y_i log(C(E(\mathbf{x}_{s_i}))) \tag{4.6}$$

where y_i denotes the true label of the ith sample of the source dataset, $\mathbf{x}_{s_i}$. The E and C represent the encoder and the classifier of the proposed framework, respectively.

4.2.5 Training Procedure

Training the proposed framework has two stages. First, we train the reconstruction branches comprising of the encoder E, the decoders D_s and D_t, with the corresponding weights θ_E, θ_{D_s}, θ_{D_t}, respectively, with the following optimization function.

$$\min_{\theta_E, \theta_{D_s}, \theta_{D_t}} \mathcal{L}_r + \lambda_h \mathcal{H} + \lambda_a \mathcal{L}_a, \tag{4.7}$$

where λ_h and λ_a are the parameters determining the importance of the sparsity and alignment losses, respectively. Note that the λ_a is gradually increased using $\lambda_a = 2/(1 + \exp(-10 \times (e_i/e_n))) - 1$ to emphasize the impact of domain matching throughout the training process, where e_i and e_n indicate epoch at iteration i and the total number of epochs, respectively.

In the second stage, the abundance representations are further aligned across domains for the classification task with optimizing the loss function as follows,

$$\min_{\theta_E, \theta_{D_s}, \theta_{D_t}, \theta_C} \mathcal{L}_r + \lambda_h \mathcal{H} + \lambda_a \mathcal{L}_a + \mathcal{L}_c, \tag{4.8}$$

where λ_h and λ_a are parameters that balance the trade-off between different losses. Since reconstruction loss and classification loss are the main objective losses and have the same importance, their coefficients are set to 1.

4.3 Experiments and Results

In this section, we evaluate the performance of the proposed method on several hyperspectral datasets and compare it with the some state-of-the-art approaches on both source and target data using several evaluation metrics.

4.3.1 Datasets

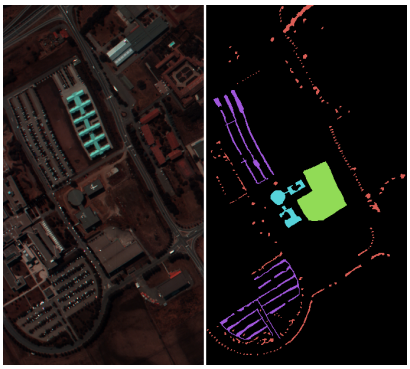
For the experiments, two pairs of source-target datasets are used. For each pair, the common categories between the two datasets are selected for the evaluation. The overlapped bands are kept for the datasets with different number of bands. Note that the datasets are composed of individual pixels as the representation learning in the unmixing-based encoder-decoder framework is performed pixel-wise. Hence, each dataset has a size of $N \times l$ where N and l are the number of pixels and the number of bands, respectively. For the 3D-CNN-based classifier, a $p \times p$ patch surrounding each pixel is used and the input data has a size of $N \times p \times p \times l$.

Pavia University (PU) and Pavia Center (PC). The PU and PC datasets, pictured in Fig. 4.4, were captured by Reflective Optics Systems Imaging Spectrometer (ROSIS) sensor over Northern Italy in 2011 with a resolution of 1.3m. The PU and PC have the size of 1096×715 and 610×340, respectively, with 102 spectral bands and common categories as Trees, Bare soil, Bitumen, and Bricks.

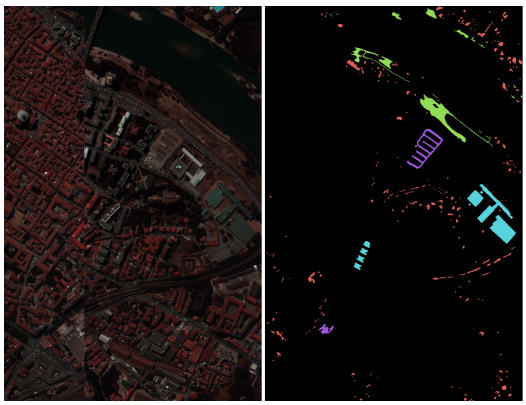
Houston 2013 (H13) and Houston 2018 (H18). The H13 and H18 datasets captured a large region of the University of Houston with a resolution of 2.5m which are illustrated in Fig. 4.5. The overlapped area is resized to the size of 209×955 containing seven common categories, including, Grass healthy, Grass stressed, Trees, Water, Residual buildings, Non-residual buildings, and Road.

4.3.2 Implementation Details

The encoder E and decoder D are constructed of 6 and 3 fully connected layers, each

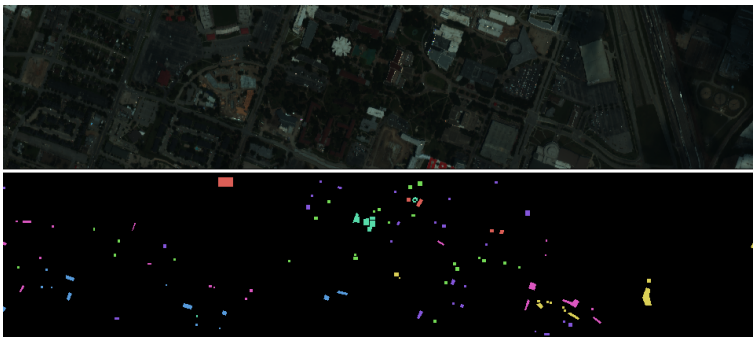


(a) Pavia University

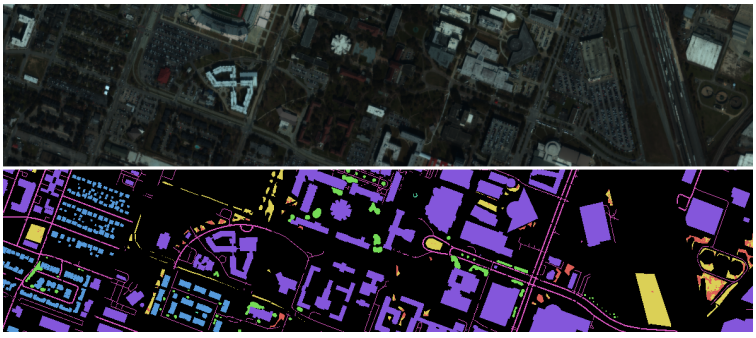


(b) Pavia Center

Figure 4.4: The satellite images and the corresponding ground truths with color indexes for, a) Pavia University (source), b) Pavia Center (target).



(a) Houston 2013



(b) Houston 2018

Figure 4.5: The satellite images and the corresponding ground truths with color indexes for, a) Houston 2013 (source), b) Houston 2018 (target).

of which containing 3 and 11 nodes, respectively. The classifier C is composed of 5 3D-CNN blocks with kernel of $7 \times 7 \times 3$, where the layers are as [12,32,12,12,30,l].

The encoder and the decoders are trained using the objective function Eq. 4.7 for 5k epochs and an Adam optimizer [139] with a learning rate of 10^{-3} which is decayed with a coefficients of 0.9997. The balancing parameter λ_h is set to 0.001.

Then, the whole framework is trained for 300 epochs minimizing the objective function Eq. 4.8. It is worth mentioning that all input data are normalized to their mean values and unit variance. The code is written in TensorFlow and experiments are performed on a desktop computer having GeForce GPU of 10 GB Memory.

4.3.3 Metrics

To compare the performance of different approaches, three popular metrics are employed, including overall accuracy (OA), average accuracy (AA), and Kappa (κ) [41]. The Kappa is a statistical metric that compares the obtained accuracy with an expected accuracy (random chance). In fact, it indicates the degree of reliability for a classification task where the value 1 is the best and 0 is the worst.

4.3.4 Classification Results

We compare the proposed method with different land-cover classification approaches, including 1D-CNN-based classifier [58], 3D-CNN [61], Spec-Spat [62], HSI-CNN [63], SSRN [64], HybridSN [65], HT-CNN [66], and PCTL [133].

The classification results on the PU-PC and H13-H18 pair datasets are reported in Table 4.1 and Table 4.2. To evaluate generalization performance, the experiments are conducted in two scenarios: 1) the model is trained and tested on the same region, i.e. source data, 2) the model is trained on the source data and tested on the target data. It can be observed that the proposed method outperforms other methods on the challenging target datasets while achieving comparable results on the source datasets.

Table 4.1: Evaluation on the Pavia University (PU) and Pavia Center (PC) pair datasets, as the source and the target domains, respectively.

Method	PU			PC		
	OA ↑	AA ↑	κ ↑	OA ↑	AA ↑	κ ↑
1D-CNN [58]	97.1	95.2	95.1	90.5	80.0	86.8
3D-CNN [61]	88.5	78.7	83.5	88.0	89.0	84.9
Spec-Spat [62]	95.6	94.9	93.8	91.0	90.0	87.8
HSI-CNN [63]	93.2	91.1	90.4	88.9	86.9	84.9
SSRN [64]	99.8	99.8	99.7	74.9	70.1	66.1
HybridSN [65]	99.4	99.2	99.1	17.1	16.5	10.8
HT-CNN [66]	97.9	97.6	97.1	81.7	73.2	74.4
PCTL [133]	98.0	98.6	97.2	90.1	85.2	86.3
Ours	99.1	99.0	98.7	**92.7**	**88.9**	**89.9**

Table 4.2: Evaluation on the Houston 2013 (H13) and Houston 2018 (H18) pair datasets, as the source and the target domains, respectively.

Method	H13			H18		
	OA ↑	AA ↑	κ ↑	OA ↑	AA ↑	κ ↑
1D-CNN [58]	84.1	82.6	81.3	67.1	65.1	46.9
3D-CNN [61]	25.6	24.5	13.1	29.9	21.9	8.5
Spec-Spat [62]	31.7	30.4	18.9	51.3	26.8	13.6
HSI-CNN [63]	57.4	57.8	50.1	55.4	47.1	33.0
SSRN [64]	85.0	86.1	82.5	46.6	56.3	31.6
HybridSN [65]	89.5	90.0	87.7	33.2	39.8	17.9
HT-CNN [66]	62.5	59.6	55.6	34.4	32.1	20.7
PCTL [133]	89.1	89.8	87.3	57.9	44.9	40.1
Ours	**90.3**	**91.5**	**88.7**	**59.9**	**46.3**	**41.4**

It demonstrates that our method bridges the gap between the two different domain aiming to reach a balance of performance on both domains.

4.3.5 Ablation Study

To demonstrate the effectiveness of each component and parameter setting in the proposed method, we conduct an ablation study and the results on the PU-PC pair datasets are shown in Table 4.3. Starting with a baseline, each component is gradually added to the network to examine its effectiveness. It can be observed that employing encoder-decoder structure without any proposed alignment techniques obtains the worst performance both on the source and the target domains. Although adding Dirichlet-based structure improves the performance, sparsity constraint complements it by making the abundance embedding space more representative and enhances the performance. It can be seen that adopting the MMD alignment tool without on unregularized abundance space deviates the distribution leading to poor results. Finally, the full proposed method including all three components achieves the best results on the target domain and comparable performance on the source dataset.

4.4 Summary

We studied the challenging generalization problem for a recognition system, where the model trained on one dataset should be adaptable to perform well on another related dataset. We presented the unmixing-based representation learning framework that project the data to a so-called abundance space which physically regularized. The source and the target data projected the abundance space are further aligned to minimize the discrepancy between them. The proposed approach was evaluated on two hyperspectral benchmark and demonstrated the superior performance compared to other methods.

Table 4.3: Ablation study. Effect of different components in our proposed method on the Pavia University (PU) and Pavia Center (PC) pair datasets, as the source and the target domains, respectively.

Dirichlet	Sparsity	Alignment	PU			PC		
			OA	AA	κ	OA	AA	κ
			98.6	98.2	98.00	87.1	79.6	82.1
✓			98.8	98.6	98.3	89.4	84.7	85.5
✓	✓		99.3	99.4	99.0	90.3	85.8	86.6
✓		✓	94.4	96.5	91.9	87.5	83.0	82.8
✓	✓	✓	99.1	99.0	98.7	92.7	88.9	89.9

Chapter 5

Occlusion-aware Bidirectional Video Semantic Segmentation

Video semantic segmentation, as a semi-supervised learning for video data, aims to associate each pixel of sequential video frames with a semantic label. However, conducting per-frame image segmentation in a video is computationally prohibitive and is often prone to temporal inconsistency. The fact that a video, unlike an image, contains temporal information has motivated researchers in this field to explore the inherent content continuity in the consecutive frames for video segmentation. These methods are mostly based on the flow-based feature propagation. However, the flow-based segmentation results tend to suffer from distortions due to the inevitable erroneous optical flow and the challenging occlusion phenomenon.

In this chapter, we explore a new research direction in video semantic segmentation and propose a bidirectional feature propagation and an occlusion-aware feature rectification method (Sec. 5.2) with the goal of improving performance of video semantic segmentation at a low computation cost [153]. We refer to it as BiVSS. Our main idea is that the features propagated in both forward and backward directions are complementary such that the ambiguous regions caused by occlusions can be compensated by the features propagated from the reverse direction. In particular,

we design an occlusion network to estimate the potentially distorted areas based on bidirectional optical flows, and utilize them as cues for correcting and fusing the propagated features. Extensive experiments (Sec. 5.3) on Cityscapes and CamVid datasets demonstrate that our proposed method achieves superior performance in terms of segmentation accuracy and temporal consistency, striking a great balance between performance and computational cost.

5.1 Introduction

Semantic segmentation is a critical visual recognition task that aims to assign a semantic label to each pixel. Image semantic segmentation has witnessed a tremendous progress benefited from the invent of deep convolutional neural network [13] combined with the proliferation of datasets (e.g. Coco [154], Pascal VOC [155], Cityscapes [156], CamVid[157]). In recent years, the use of video semantic segmentation in real-world applications such as autonomous vehicles [158], robotics [159], video surveillance [160] has drawn significant attention in research community. These challenging applications require a video analysis method with high accuracy and low latency.

As a naïve approach for video semantic segmentation, an image semantic segmentation method can be applied on each frame separately and independently. But this strategy mostly is not acceptable in real-world applications due to the unaffordable computational cost and poor temporal consistency. Actually, a single-frame approach treats the video as a collection of uncorrelated images and ignores the fact that the consecutive frames of a video typically have a substantial semantic similarity with each other [161]. To leverage the temporal continuity in a video, an intuitive idea is to reuse the high-level features extracted at sparse key frames when segmenting other frames [81; 162; 163].

The observation that high-level features change more slowly compared to shallower features extracted from video frames [80] has motivated prior works to relegate the

expensive feature extraction component in a video recognition architecture and warp the high-level representations based on optical flow [81]. Although feature warping reduces the overall computational cost, the accuracy of optical flow estimation affects the propagated features and hence, determines the performance of semantic segmentation.

Despite the significant progress in optical flow estimation [164; 165], the spatial misalignment between key frames and other frames may result in inaccurate optical flow. The situation gets even worse when scene motion causes occlusion or the entire scene moves relative to the camera, making noticeable warping errors. To compensate for fast scene evolution, features can be updated by features computed at the current frame [82]. However, this method treats every pixel equally – even the regions where the features are correctly propagated may be enforced to be rectified, causing the false correction. To rectify the propagated features in an optimal way, distorted regions can be estimated, thereby the feature correction is performed only on the distorted regions as preserving propagated features for other regions [7].

While approaches based on feature correction appear effective, their performance are constrained by the update branch that computes features of the current frame. In particular, these methods may not be able to rectify all the distortions using a light-weight network due to its weak feature extraction architecture. Note that utilizing a strong deep network as the update network contradicts with the goal of keyframe-based approaches. In addition, identifying the distortions is challenging since they are accumulated through propagation over multiple frames.

To address the challenges of efficient video segmentation, we propose a bidirectional video semantic segmentation framework (BiVSS), which leverages the high quality features extracted at the key frames using an expensive image semantic segmentation network, and propagates them toward the non-key frames in both forward and backward directions using flow fields. The distorted regions due to object motion or a moving observer are corrected at the current frame following the guidance from a learned occlusion map. The map essentially highlights the potentially

occluded regions by the use of bidirectional optical flows embodying forward-backward movement information, which are then used as cues to compensate for the distortions from the features warped in the reverse direction. As can be observed in Fig. 5.1, the warped features from two key frames are complementary, and the proposed BiVSS is capable of merging them effectively.

5.2 Approach

We propose to perform semantic segmentation in a bidirectional framework. This framework satisfies two goals: 1) leveraging the high-level representations of two key frames for segmenting the non-key frames, and 2) estimating occlusions which is crucial for correcting the distortions occurred through feature propagation via flow fields. A visual illustration of our bidirectional video semantic segmentation model is presented in Fig. 5.2.

We start with a pair of consecutive key frames, i.e. I_k and I_{k+D}, and perform semantic segmentation using an off-the-shelf image segmentation network $SegNet_k$, which is deep and expensive. The high-level features obtained from the key frames are then warped along the optical flow to the subsequent and previous time steps, respectively, in a frame-by-frame fashion to align with the non-key frames. The Occlusion network, $OccNet$, on the other hand, learns regions where occlusion and disocclusion may happen during propagation in both forward and backward directions. Next, we rectify the propagated features under the cues of the predicted occlusion maps. The remaining slight distortions is corrected with the help of the features extracted from the current frame using a shallow image segmentation network, $SegNet_{nk}$.

In the following, we elaborate on the bidirectional feature propagation framework in Sec. 5.2.1, describe the occlusion prediction network in Sec. 5.2.2, and present the effective feature correction in Sec. 5.2.3.

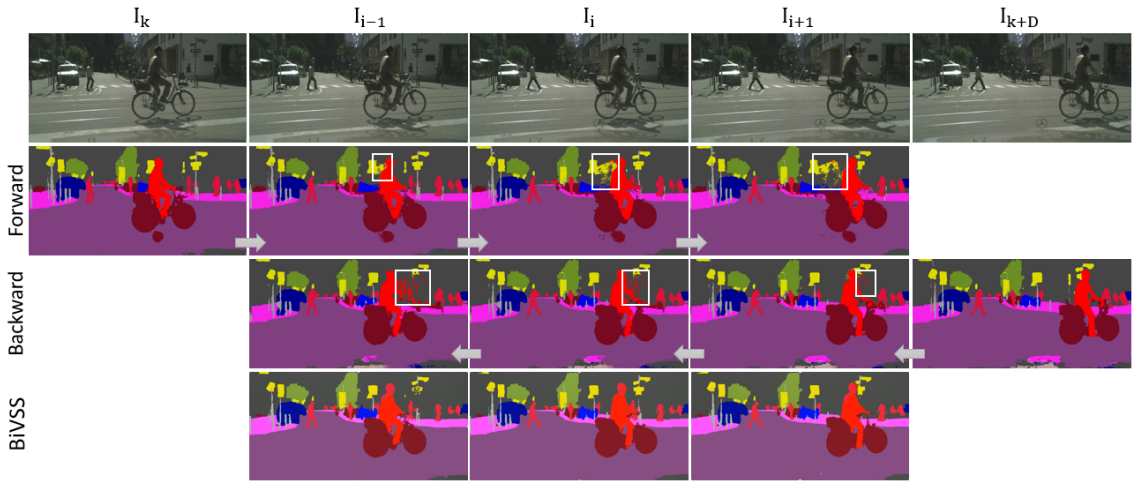


Figure 5.1: Illustration of how distortions due to occlusion-disocclusion (highlighted with white bounding boxes) can be mitigated through the proposed bidirectional framework, BiVSS.

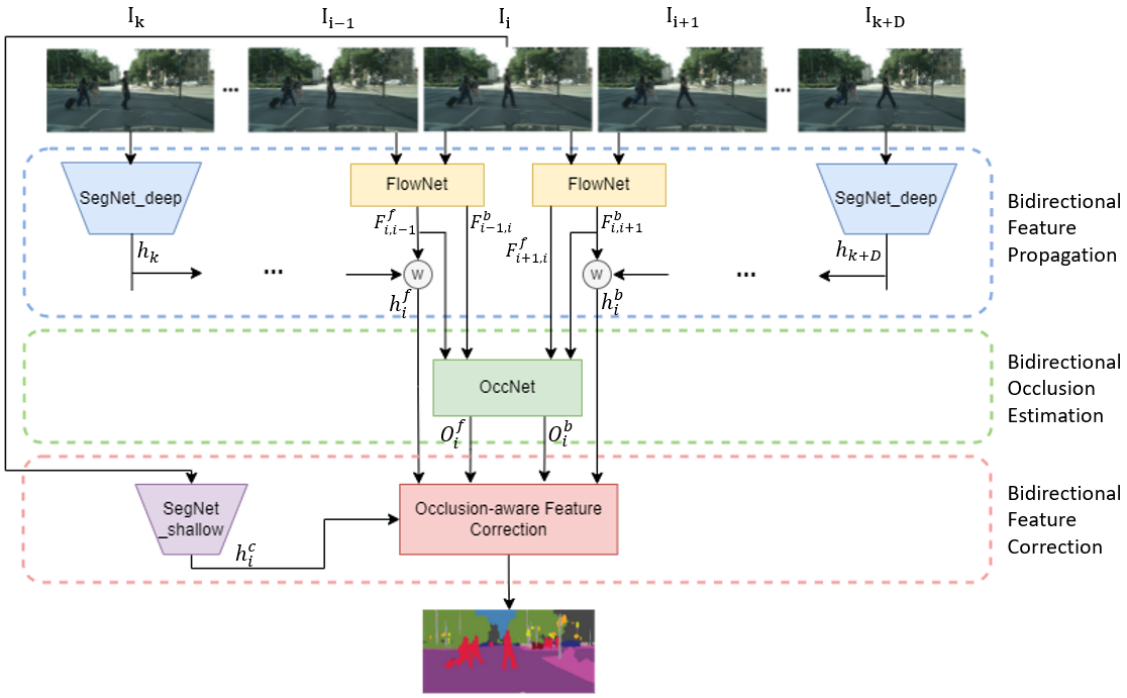


Figure 5.2: An overview of the proposed BiVSS framework that comprises of a bidirectional feature propagation component (Sec. 5.2.1), a bidirectional occlusion estimation component (Sec. 5.2.2), and an occlusion-aware feature correction component (Sec. 5.2.3). It leverages the high quality deep features extracted from two key frames (I_k and I_{k+D}) using a static deep segmentation network ($SegNet_{deep}$) and propagates them toward the non-key frame (I_i) using optical flows. The potential ambiguities or uncertainties in the propagated features are rectified with the help of occlusion maps estimated by an Occlusion network ($OccNet$). The ambiguities are further corrected by a shallow segmentation network ($SegNet_{shallow}$) executed on the current frame to save computation cost.

5.2.1 Bidirectional Feature Propagation

The technique bidirectional optical flow was initially proposed in [166]. It has been used for video frame interpolation [166; 167; 168; 169; 170; 171; 172], motion detection [173], and robust optical flow estimation [174; 175; 176]. In this work, the bidirectional optical flow is employed for feature propagation of two consecutive key frames toward a non-key frame.

Assuming a fixed distance D between the successive key frames, a deep image segmentation network $SegNet_k$, where k denotes key frames, is executed only on selected key frames. The extracted key frame feature, $h_k = SegNet_k(I_k)$ is forward-propagated to subsequent frames using a computed optical flow, $F^f_{i,i-1} = FlowNet(I_{i-1}, I_i)$ where $FlowNet$ represents the flow estimation network. Similarly, the feature obtained from the incoming key frame, $h_{k+D} = SegNet_k(I_{k+D})$, is warped backward to previous frames along a flow field, $F^b_{i,i+1} = FlowNet(I_{i+1}, I_i)$. During the propagation process, the features are warped along optical flow using bilinear interpolation as the warping operator W. The feature warping in forward and backward directions at the current frame, I_i, result in $h^f_i = W(h^f_{i-1}, F^f_{i,i-1})$ and $h^b_i = W(h^b_{i+1}, F^b_{i,i+1})$, respectively. These two features are later fused with the guidance of occlusion maps which will be elaborated in Secs. 5.2.2 and 5.2.3.

5.2.2 Forward-Backward Occlusion Estimation

Occlusion estimation plays a critical role in many vision tasks, including image instance/panoptic segmentation [177; 178], video frame interpolation [172; 170], object tracking [179; 180; 181], and optical flow estimation [174; 175; 176]. In video segmentation, occlusion handling is critical as object motion causes ambiguous regions in the feature propagation process.

As discussed in Sec. 5.1, the spatial misalignment between key frames and other frames may result in inaccurate optical flow, hence feature correction is a necessary step toward high-fidelity segmentation result. To largely reduce the amount of false

correction and only rectify the propagated features in distorted regions but preserve those for other regions, occlusion estimation is essential since the distorted regions in video semantic segmentation are mostly the occlusion-disocclusion areas occurred due to object motion or a moving observer in videos.

Let I_{i-1} and I_i be two consecutive non-key frames. Classical optical flow estimation methods are based on the observation that a pixel in frame I_{i-1} should be similar to pixel which is mapped by optical flow in frame I_i [182]. However, this observation does not hold true for pixels which become occluded. Inspired by occlusion reasoning based on bidirectional flow consistency [174; 175], we estimate occlusion maps in forward and backward directions.

The forward optical flow, F^f maps I_{i-1} to I_i. The backward flow F^b mapping I_i to I_{i-1} is acquired by interchanging the inputs to the optical flow estimation model. The intuition behind our occlusion detection model is built upon the forward-backward consistency assumption [183], stating that the forward flow F^f negates the backward flow F^b on the pixels where there is no occlusion, hence, the pixels where there is a large mismatch between these two flows will be marked as occluded pixels [175]. Consequently, the occlusion in the forward direction is obtained whenever the following constraint is violated [175].

$$\begin{aligned} &|F^f(x) + F^b(x + F^f(x))|^2 \\ &< \alpha_1(|F^f(x))|^2 + |F^b(x + F^f(x))|^2) + \alpha_2 \end{aligned} \tag{5.1}$$

where x is the pixel location, and α_1 and α_2 are some scalar parameters. Similarly, the occlusion in the backward direction is calculated when F^f and F^b are exchanged.

Despite the promising results of the occlusion detection using Eq. 5.1, it only estimates the occluded regions between two consecutive frames which may not be adequate when the distance between two key frames is large. Flow-based feature warping aggregates the distortions throughout the frames and thus certain distortions would remain in such pair-wise occlusion caused by the previous frames. To detect all

the aggregated occlusion areas, an occlusion-based network is designed to highlight the occlusion areas in both forward and backward directions. The structure of the occlusion network is depicted in Fig. 5.3.

The occlusion network has a U-Net shape structure that takes four optical flows from three consecutive frame (I_{i-1}, I_i, I_{i+1}), $[F^f_{i-1,i}, F^b_{i,i-1}, F^f_{i,i+1}, F^b_{i+1,i}]$, as input and estimates the occlusion maps. To be specific, from these complementary optical flows, it predicts occlusion maps in forward direction, O^f_i, where a region of frame I_{i-1} becomes occluded in frame I_i, and in backward direction, O^b_i, where a region of frame I_{i+1} becomes occluded in frame I_i.

5.2.3 Occlusion-Aware Feature Correction

Through utilizing forward-backward propagation, our model is able to estimate the occlusion-disocclusion areas using optical flow in both forward and backward directions, and rectify the features on these ambiguous regions with the help of features in its complementary direction.

Let h^f_i and h^b_i denote the propagated features from the preceding key frame I_k and the incoming key frame I_{k+D} to the current frame I_i, respectively, and h^c_i be the extracted feature from the current frame using the shallow image segmentation network $SegNet_{nk}$. The feature correction module adopts the weighted sum to perform feature rectification. Hence we have:

$$h_i = h^f_i \times O^f_i + h^b_i \times O^b_i + h^c_i \times (1 - O^f_i - O^b_i) \tag{5.2}$$

where $\times$ represents the spatially element-wise multiplication. Later, the Softmax operation is applied on the corrected features to output the final segmentation result.

5.3 Experiments and Results

In order to evaluate the effectiveness of the proposed BiVSS, we use two challenging

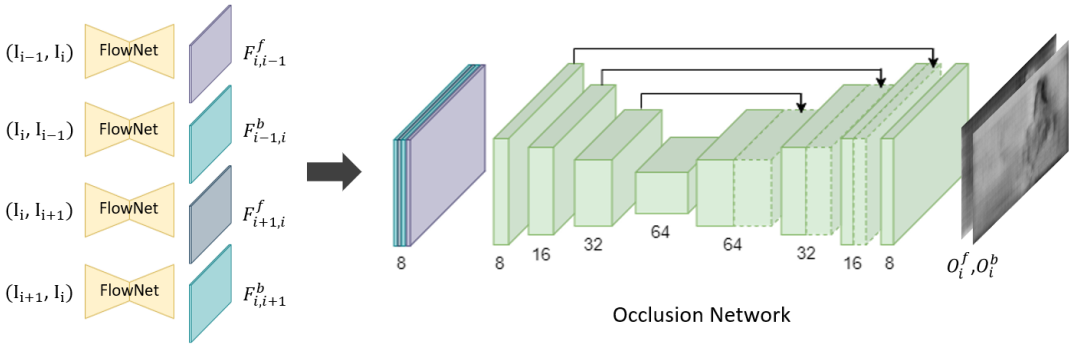


Figure 5.3: Occlusion network for occlusion map estimation.

video semantic segmentation datasets, namely, Cityscapes [156] and CamVid [157] pictured in Fig. 5.4, and compare its performance with the state-of-the-art works from the perspectives of accuracy, temporal consistency, and computational cost.

5.3.1 Experimental Setup

Datasets. Cityscapes [156] consists of 5000 snippets recorded at 17 fps of street scenes from 50 different cities. The dataset is divided into 2975/500/1525 sets corresponding to training/validation/test sets. Each frame has size of 1024×2048, and the 20th frame of each snippet is carefully annotated with 19 semantic labels. CamVid [157] contains 4 video clips captured at 30 fps with resolution of 720×960. The 11-class high quality dense pixel annotation is provided for each 30th frame. Following [7], the trainval and test sets have 468 and 233 samples, respectively.

Evaluation Metrics. Three performance metrics are used in the experiments. To evaluate segmentation accuracy of different video semantic segmentation methods, we use the mean intersection over union (mIoU) [184]. The temporal consistency of the video semantic segmentation results is measured by tracing semantic labels of trajectories thoughout frames of a video [183]. It calculates the ratio of the label-consistent trajectories and is reported as mIoU based temporal consistency (mTC) [185]. In addition, the amount of floating point operations (FLOPs) [186; 187] is adopted to measure computation cost.

Implementation Details. There are four key components in our proposed framework, including the deep image semantic segmentation network $SegNet_k$ executed on the key frames, the optical flow estimation model $FlowNet$, the occlusion network $OccNet$, and the shallow image segmentation network $SegNet_{nk}$ applied on the non-key frames. We conduct experiments using DeepLabv3+ [6] and HRNetV2 [75] as $SegNet_k$, due to their superior performance in terms of accuracy and efficiency in segmenting static images. They are pretrained on ImageNet [188] dataset and then finetuned on target segmentation dataset. We adopt the modified FlowNet2-S [165]

Figure 5.4: The two video semantic segmentation datasets used in this chapter, including an example frame of the datasets images and the corresponding ground truths with color indexes. The datasets are: a) Cityscapes, b) CamVid.

as optical flow estimation network which has less complexity compared to the original one. $FlowNet$ is pretrained on the synthetic Flying Chairs dataset [164] and then fine-tuned on the particular dataset during the bidirectional feature propagation learning. Both $SegNet_{nk}$ and $OccNet$ have a U-Net shape structure comprising of a few convolutional layers, interlaced with batchnorm and Leaky-ReLU layers in the encoder and a few deconvolutional layers interlaced with Leaky-ReLU in the decoder.

We train our model in two phases. In phase one, the bidirectional feature propagation framework only comprises of $SegNet_k$ and $FlowNet$. It fine-tunes $FlowNet$ while the weights of $SegNet_k$ are kept fixed. In parallel, the shallow static segmentation network $SegNet_{nk}$ is separately trained on the particular dataset (e.g. Cityscapes). In phase two, the $OccNet$ in the full framework is trained from scratch while the weights of other networks are kept frozen.

The bidirectional framework is trained with a batch of three frames $[I_{g-p}, I_g, I_{g+q}]$, where the image I_g has semantic annotation. The index of the images in the batch is generated based on the relation as $q = D - p + 1$ where $1 \leq p \leq D$. It enables the framework to propagate the features with various propagation distances while the distance between the key frames is constant as D.

In training steps, the Adam optimizer [139] is adopted with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The optimizations are performed for 100 epochs with learning rate of 10^{-4}. Key frames are selected at regular intervals, starting with the first frame in the video. The proposed method uses key frame interval as 5, unless otherwise stated.

5.3.2 Comparison with State-of-the-Arts

We compare the proposed BiVSS with different video semantic segmentation approaches. We use DeeplabV3+ [6] and HRNetV2 [75], respectively, as backend to our model. The quantitative comparisons on the Cityscapes dataset validation subset are shown in Table 5.1. Note that for the keyframe-based methods like DFF [81], Accel [82], DAVSS [7], and our method, the GFLOPs of the non-key frames are reported.

Table 5.1: Evaluation on the Cityscapes and CamVid datasets. Key and Non-key indicate keyframe and non-keyframe based methods, respectively. We adopt HRNetV2 as baseline mainly for non keyframe based approaches like GRFP and TDNet due to its relatively low computation cost. We adopt DeeplabV3+ as baseline mainly for keyframe-based approaches like DFF, Accel, DAVSS, and the proposed BiVSS for its relatively high accuracy.

	Method	Cityscapes			CamVid		
		mIoU ↑	mTC ↑	GFLOPs ↓	mIoU ↑	mTC ↑	GFLOPs ↓
Non-key	DeeplabV3+ [6]	76.6	76.6	820	72.0	83.2	270
	HRNetV2 [75]	75.9	81.0	156	75.0	83.9	52
	GRFP [78]	76.6	83.8	468	74.6	87.2	156
	TDNet [79]	76.5	81.6	161	72.6	84.7	54
Key	DFF [81]	68.7	-	180	66.0	-	60
	Accel [82]	72.1	-	510	66.7	-	170
	DAVSS [7]	75.4	84.5	212	71.1	85.0	72
	Ours (DeeplabV3+)	**76.5**	84.8	231	71.8	85.5	78
	Ours (HRNetV2)	75.7	**86.5**	231	**74.4**	**88.4**	78

The computation of key frames has GFLOPs equal to the one reported for the baseline model. The GRFP and TDNet models are re-implemented with HRNetV2 network as backbone. It can be observed that the proposed method based on the DeepLabV3+ outperforms the state-of-the-art keyframe-based video segmentation methods in terms of segmentation accuracy, and has comparable mTC value. Although there is usually a trade-off between segmentation accuracy and temporal consistency, our model is able to achieve a promising balance in them with a reasonable computation cost. Comparing our model based on HRNetV2 with two other non keyframe-based methods, i.e. GRFP and TDNet, confirms that non keyframe methods achieve better segmentation mIoU since they aggregate the information over multiple frames. However, they may suffer from high computation cost or low temporal consistency as the results of GRFP and TDNet indicate, respectively.

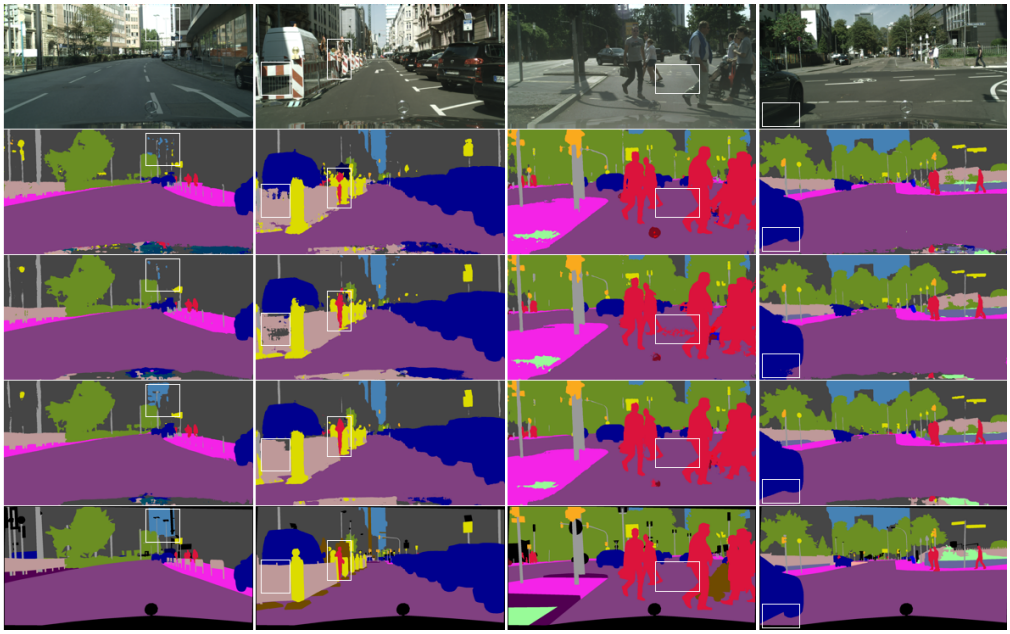
It is worth mentioning that the proposed method gets a bit lower mIoU compared to the baseline models but outperforms them in terms of temporal consistency. The improvement of 7.2% and 5.5% in mTC over DeeplabV3+ and HRNetV2, respectively, with comparable mIoU demonstrate that our model is able to reach a better balance between accuracy, temporal consistency, and computation cost. The 70% less computation burden with DeeplabV3+ being the baseline also leads to a noticeable speed-up. It also shows the segmentation results on CamVid dataset. It can be observed that our proposed method outperforms the state-of-the-art methods, and has comparable performance compared to the corresponding baseline frame-by-frame segmentation networks.

In order to compare the results in more detail, the class-wise IoUs are presented in Table 5.2. One can see that most of our gains come from large objects, like sidewalk, wall, truck, bus, and train. However, it is common in keyframe-based approaches that slender objects, like pole or traffic sign, get deteriorated due to the flow-based propagation.

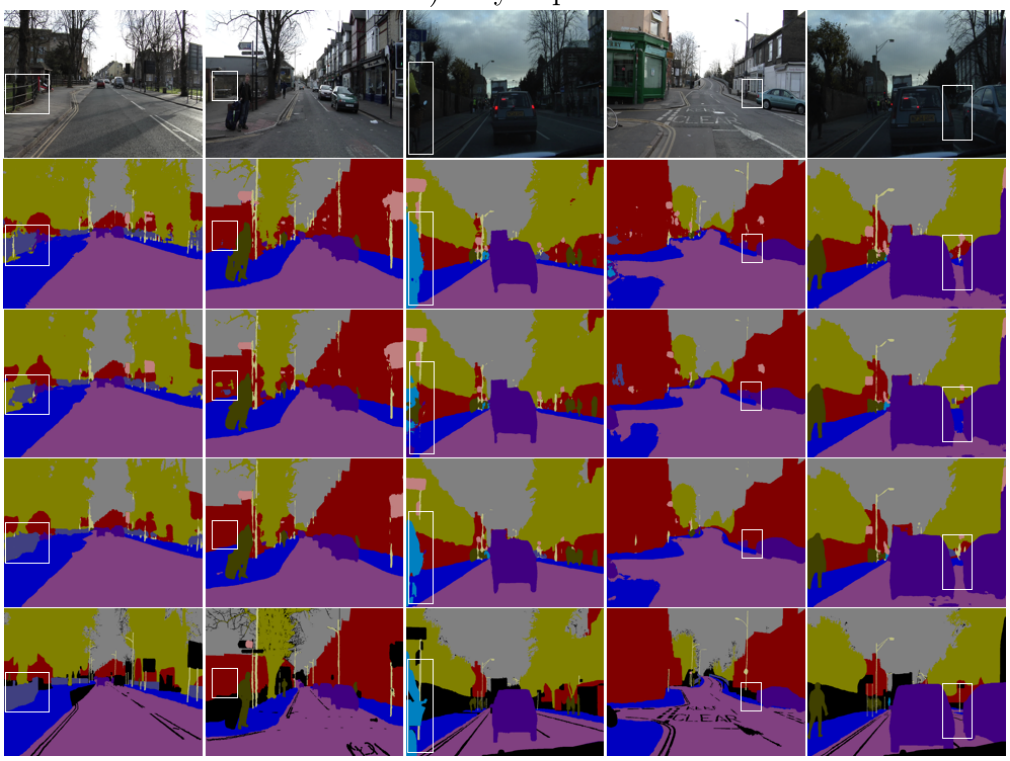
Visual Comparison. The qualitative comparisons of our method and the single-frame DeeplabV3+ [6] are shown in Fig. 5.5 on the two datasets used. In particular

Table 5.2: Class-wise IoUs on the Cityscapes validation set with HRNetV2 and DeeplabV3+ as baselines.

Class	HRNetV2	Ours	DeeplabV3+	Ours
Road	98.2	98.1	98.1	98.1
Sidewalk	85.3	84.7	84.6	84.5
Building	92.4	91.9	92.0	91.7
Wall	56.0	58.0	59.0	58.6
Fence	60.2	60.9	60.8	59.9
Pole	64.6	57.4	58.8	55.0
Traffic light	69.9	67.9	64.9	63.8
Traffic sign	77.2	76.3	75.0	73.9
Vegetation	92.5	91.9	92.1	92.0
Terrain	64.2	64.4	64.7	65.2
Sky	94.3	94.0	94.5	94.5
Person	81.2	78.4	79.6	78.8
Rider	59.4	58.5	59.0	58.7
Car	94.6	94.3	94.6	94.6
Truck	67.6	75.7	83.8	84.8
Bus	81.4	81.1	85.5	88.5
Train	67.6	69.2	70.5	73.1
Motorcycle	59.7	62.0	63.7	64.9
Bicycle	75.9	74.1	74.6	74.0
Avg.	75.9	75.7	76.6	76.5



a) Cityscapes



b) CamVid

Figure 5.5: Semantic segmentation results on two datasets: a) Cityscapes and b) CamVid. From top to bottom: the image, image segmentation by DeeplabV3+ [6], video segmentation by DAVSS [7], video segmentation by our framework, and the ground truth.

our approach is able to better segment the large regions where ambiguities exist.

Effect of Propagation Interval. To evaluate how the segmentation accuracy changes with respect to key frame interval, D, the proposed method with DeeplabV3+ backbone is compared with other keyframe-based methods for various propagation distances ranging from 1 to 9. Fig. 5.6 depicts the results of different models on Cityscapes validation set. It is in line with intuition that as the key frame interval increases, the segmentation accuracy decreases due to inaccuracy in optical flow estimation and the distortions caused by occlusion-disocclusion. Meanwhile, our model is able to achieve stable performance which demonstrates the important role of bidirectional framework on rectifying the distortions aggregated over the frames.

5.3.3 Method Analysis

Ablation Study. The contribution of each component in the proposed framework is investigated by removing them one by one. Table 5.3 illustrates the results on the Cityscapes dataset with DeeplabV3+ and forward propagation as baseline (see 1st row in Table 5.3). It is observed that incorporating the backward propagation to the baseline improves the segmentation accuracy by 3.9% in mIoU while introducing 25 ms of latency per frame (see 3rd row in Table 5.3). Adding the features extracted from the current frame slightly improves the baseline (see 2nd row in Table 5.3), but deteriorate the performance of the forward-backward framework (see 4th row in Table 5.3). This is because of the fact that the feature correction is applied to all pixels of the frames, even those which have correctly propagated to the current frame. Adding OccNet to the bidirectional framework further enhances the segmentation accuracy by rectifying only the distorted areas indicated in the occlusion maps (see 5th row in Table 5.3). Our full framework outperforms the other variants and achieves 76.7% mIoU with a reasonable overhead computation cost.

Effect of the Feature Correction Method. The impact of the occlusion-aware feature correction is compared with other fusion methods including directly adding

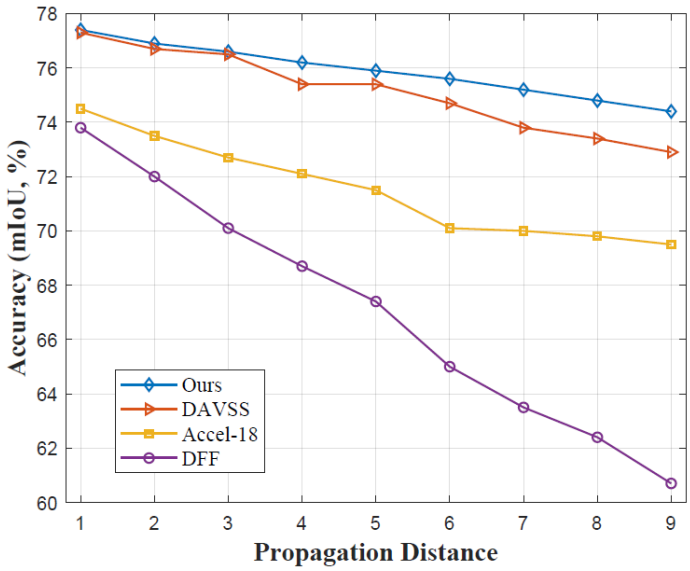


Figure 5.6: Accuracy vs. propagation distance on Cityscapes validation set. The keyframe-based method are compared over different propagation distances.

Table 5.3: Contribution of different components to BiVSS, including backward propagation (Bkwd), Occlusion network (OccNet), and the use of update branch at the current frame (Curr). The mIoU scores and runtimes on a non-key frame for the Cityscapes dataset are provided.

Bkwd	OccNet	Curr	mIoU (%)	Time (ms/f)
			72.0	171
		✓	72.1	184
✓			75.9	196
✓		✓	75.2	213
✓	✓		76.3	253
✓	✓	✓	76.5	265

feature maps (Add) and 1×1 convolutional layer (Conv1×1). As shown in Table 5.4, concatenating the three features with a simple convolutional layer leads to the worst accuracy. Our "OccNet" module outperforms the "Add" approach which does not take into account the spatial misalignment.

Occlusion Visualization. The intermediate results including the propagated feature in forward and backward directions, feature extracted from the current frame, and the occlusion maps are illustrated in Fig. 5.7. It can be observed that the proposed method is able to estimate the occlusion maps for both forward and backward directions, and highlight the regions where both propagated features are uncertain about the predictions which is used for extra refinement with the help of shallow features extracted from the current frame.

5.4 Summary

We proposed a novel bidirectional framework for video semantic segmentation that consists of two main components, the bidirectional feature propagation module and the occlusion-aware feature correction module. The former helps to propagate the features of the key frames toward the non-key frames. The latter works on fusing the propagated features and the ones obtained from the current frame with the guidance of the estimated occlusion maps. The experiments on Cityscapes and CamVid demonstrate the superiority of the proposed method with a better balance between accuracy, temporal consistency, and computation cost.

Table 5.4: Effect of different feature correction modules in our framework on the Cityscapes dataset. "OccNet" represents the proposed feature correction based on occlusion maps. "Add" and "Conv1×1" denote adding features maps and fusion using a 1×1 convolution layer, respectively.

Method	mIoU (%)	Time (ms/f)
OccNet	76.5	265
Add	75.2	213
Conv1×1	61.4	244

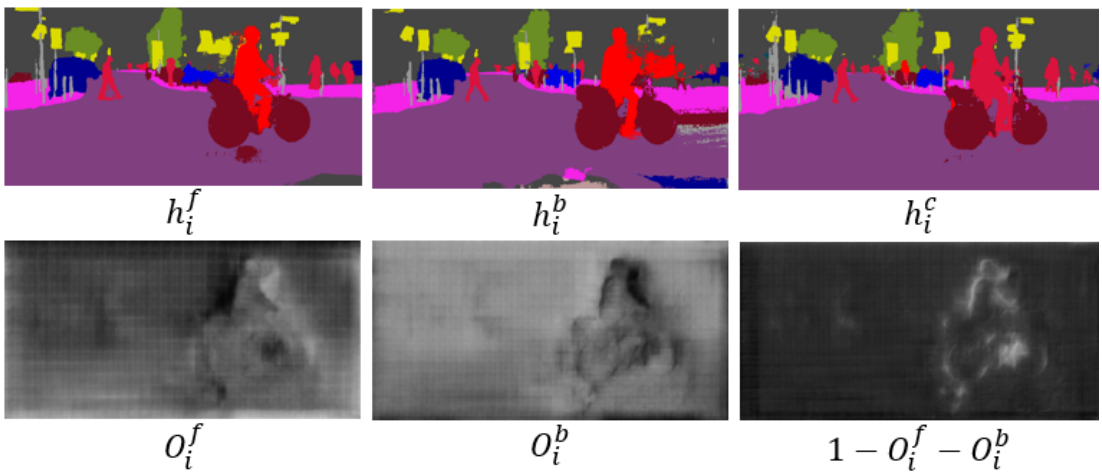


Figure 5.7: Example intermediate results from BiVSS on Cityscapes. Top row, from left to right: the propagated features in forward and backward directions, and the one obtained from the current frame. Bottom row, from left to right: the occlusion maps in forward and backward directions, and the remaining distortion map.

Chapter 6

Conclusions and Future Works

This dissertation focuses on learning with limited data and how to deal with it for image and video processing. We studied open-set recognition in Chapter 3 to learn a recognition system algorithm under a more realistic open-world assumption where there could be incomplete knowledge of the world during training and the system may encounter unknown categories in deployment phase. We presented a representative-discriminative framework in order to best characterize the differences between known and unknown categories. It was shown that transforming the data to an embedding feature space facilitates differentiating similar classes, and the transformation from the embedding feature space to a so-called abundance space enhances the representative capacity of the recognition model.

In Chapter 4, we investigated the generalization capability of a recognition system under a real-world scenario where there may be some domain-shift between the data that the model is being deployed on compared to the training data. We presented a physic-based representation learning approach to perform classification in a domain-invariant embedding space. It was demonstrated that projecting the data to a physic-regularized space reduces discrepancy between the source and the target domains, which is further enhanced by a distribution alignment method.

In Chapter 5, we explored video semantic segmentation in a semi-supervised learning strategy where certain frames of a video have been labeled. We proposed an occlusion-aware bidirectional framework to take advantage of two labeled frames instead of one. It was demonstrated that the features propagated in both forward and backward directions are complementary specially for the occluded areas. We designed an occlusion network to estimate the potentially distorted areas based on bidirectional optical flows, and utilized them as cues for correcting and fusing the propagated features.

The future works of this dissertation lie in three aspects. First, as the semantic segmentation is an important perception tool in safety critical applications like autonomous driving system, it needs to be robust to unknown objects that may occur. The fact that the segmentation problem is inherently a classification problem which assigns each pixel of an image to a semantic label, it could be learned under an open-world assumption. In the future work, the proposed open-set recognition work will be extended for the segmentation task.

Second, the proposed domain-adaptation method works in an unsupervised setting where there are no labeled data available from the target domain. As future research, we will investigate how much improvement we could achieve if few samples of the target domain are annotated, i.e. semi-supervised learning, as conditional distribution can be aligned more effectively with having those cues.

Third, the presented video semantic segmentation approach assumes fixed interval between keyframes which may be hard to tune for each video separately. In the future, we will study some keyframe selection approaches to adaptively choose a frame as keyframe and perform bidirectional video semantic propagation accordingly.

Bibliography

Bibliography

[1] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 102–117. xii, 3

[2] X. Li, W. Zhang, and Q. Ding, "Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 7, pp. 5525–5534, 2018. xii, 4, 5

[3] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 743–751. xii, 4, 5

[4] http://lesun.weebly.com/hyperspectral-data-set.html. xii, 20

[5] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008. xii, 20

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818. xv, 11, 66, 68, 69, 70, 72

[7] J. Zhuang, Z. Wang, and B. Wang, "Video semantic segmentation with distortion-aware feature correction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. xv, 12, 58, 66, 68, 69, 72

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 1, 9, 17

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 1, 9, 11, 17

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 1, 9, 11, 16, 17, 41

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 1, 17

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 1, 17

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. 1, 11, 17, 57

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 1, 11, 17

[15] Y. Cheng, Y. Yang, H.-B. Chen, N. Wong, and H. Yu, "S3-net: a fast and lightweight video scene understanding network by single-shot segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3329–3337. 1

[16] W. Wang, T. Zhou, F. Porikli, D. Crandall, and L. Van Gool, "A survey on deep learning technique for video segmentation," *arXiv preprint arXiv:2107.01153*, 2021. 1

[17] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020. 1

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017. 1

[19] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208. 1

[20] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1742–1750. 1

[21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 685–694. 1

[22] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 642–651. 1

[23] V. Badrinarayanan, I. Budvytis, and R. Cipolla, "Semi-supervised video segmentation using tree structured graphical models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2751–2764, 2013. 1, 4

[24] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 1, 4

[25] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012. 1, 2, 30

[26] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572. 1, 2, 13, 33, 34, 36

[27] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3614–3631, 2020. 1, 12

[28] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305. 1

[29] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526. 1

[30] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2019, pp. 6936–6945. 1

[31] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in Data Science and Information Engineering*, pp. 877–894, 2021. 4

[32] S. Zhu, B. Du, L. Zhang, and X. Li, "Attention-based multiscale residual adaptation network for cross-scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021. 4

[33] F. Galasso, N. S. Nagaraja, T. J. Cárdenas, T. Brox, and B. Schiele, "A unified video segmentation benchmark: Annotation, metrics and analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3527–3534. 4

[34] V. Badrinarayanan, F. Galasso, and R. Cipolla, "Label propagation in video sequences," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3265–3272. 4

[35] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label mrf optimization," *International journal of computer vision*, vol. 100, no. 2, pp. 190–202, 2012. 4

[36] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 985–998, 2018. 4

[37] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3899–3908. 4

[38] S. Vijayanarasimhan and K. Grauman, "Active frame selection for label propagation in videos," in *European conference on computer vision*. Springer, 2012, pp. 496–509. 4

[39] S. Avinash Ramakanth and R. Venkatesh Babu, "Seamseg: Video object segmentation using patch seams," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 376–383. 4

[40] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, 2017. 6

[41] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 279–317, 2019. 6, 52

[42] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 9, 12

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9. 9

[44] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004. 10

[45] H. Yu, L. Gao, W. Liao, B. Zhang, A. Pižurica, and W. Philips, "Multiscale superpixel-level subspace-based support vector machines for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2142–2146, 2017. 10

[46] L. Samaniego, A. Bárdossy, and K. Schulz, "Supervised classification of remotely sensed imagery using a modified k-nn technique," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2112–2125, 2008. 10

[47] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4085–4098, 2010. 10

[48] J. Ediriwickrema and S. Khorram, "Hierarchical maximum-likelihood classification for improved accuracies," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 810–816, 1997. 10

[49] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Transactions on geoscience and remote sensing*, vol. 26, no. 1, pp. 65–74, 1988. 10

[50] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear pca for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 3, pp. 447–451, 2011. 10

[51] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862–873, 2009. 10

[52] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994. 10

[53] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended

morphological transformations," *IEEE Transactions on Geoscience and remote sensing*, vol. 43, no. 3, pp. 466–479, 2005. 10

[54] B. Luo and J. Chanussot, "Unsupervised classification of hyperspectral images by using linear unmixing algorithm," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 2877–2880. 10

[55] I. Dópido, M. Zortea, A. Villa, A. Plaza, and P. Gamba, "Unmixing prior to supervised classification of remotely sensed hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 760–764, 2011. 10

[56] R. K. Baghbaderani, F. Wang, C. Stutts, Y. Qu, and H. Qi, "Hybrid spectral unmixing in land-cover classification," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2019. 10

[57] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, 2015. 10

[58] Y. Song, Z. Zhang, R. K. Baghbaderani, F. Wang, C. Stutts, and H. Qi, "Land cover classification for satellite images through 1d cnn," in *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2019. 10, 52, 53

[59] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with markov random fields and a convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2354–2367, 2018. 10

[60] V. Sharma, A. Diba, T. Tuytelaars, and L. Van Gool, "Hyperspectral cnn for image classification & band selection, with application to face recognition," *Technical report KUL/ESAT/PSI/1604, KU Leuven, ESAT, Leuven, Belgium*, 2016. 10, 40

[61] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-d deep learning approach for remote sensing image classification," *IEEE Transactions on geoscience and remote sensing*, vol. 56, no. 8, pp. 4420–4434, 2018. 10, 40, 52, 53

[62] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sensing*, vol. 9, no. 1, p. 67, 2017. 10, 40, 52, 53

[63] Y. Luo, J. Zou, C. Yao, X. Zhao, T. Li, and G. Bai, "Hsi-cnn: a novel convolution neural network for hyperspectral image," in *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE, 2018, pp. 464–469. 10, 40, 52, 53

[64] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847–858, 2017. 10, 40, 52, 53

[65] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2019. 10, 52, 53

[66] X. He, Y. Chen, and P. Ghamisi, "Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3246–3263, 2019. 10, 16, 41, 52, 53

[67] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. 11

[68] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. 11

[69] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934. 11

[70] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. 11

[71] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480. 11

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015. 11

[73] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890. 11

[74] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 11

[75] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020. 11, 66, 68, 69

[76] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video cnns through representation warping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4453–4462. 11

[77] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, F. Huang, and R. Klette, "Stfcn: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 493–509. 11

[78] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6819–6828. 11, 69

[79] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8818–8827. 12, 69

[80] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 852–868. 12, 57

[81] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2349–2358. 12, 57, 58, 68, 69

[82] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8866–8875. 12, 58, 68, 69

[83] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008. 12

[84] P. R. M. Júnior, R. M. De Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017. 12

[85] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014. 12

[86] P. R. M. Júnior, T. E. Boult, J. Wainer, and A. Rocha, "Specialized support vector machines for open-set recognition," *arXiv preprint arXiv:1606.03802*, 2016. 12

[87] H. Zhang and V. M. Patel, "Sparse representation-based open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1690–1696, 2016. 12

[88] M. Hassen and P. K. Chan, "Learning a neural-network-based representation for open set recognition," *arXiv preprint arXiv:1802.04365*, 2018. 13

[89] L. Shu, H. Xu, and B. Liu, "Doc: Deep open classification of text documents," *arXiv preprint arXiv:1709.08716*, 2017. 13

[90] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4016–4025. 13

[91] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016. 13

[92] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," *arXiv preprint arXiv:1707.07418*, 2017. 13

[93] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 613–628. 13, 35, 36

[94] I. Jo, J. Kim, H. Kang, Y.-D. Kim, and S. Choi, "Open set recognition by regularising classifier with fake data generated by generative adversarial networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 2686–2690. 13

[95] H. Choi, E. Jang, and A. A. Alemi, "Waic, but why? generative ensembles for robust anomaly detection," *arXiv preprint arXiv:1810.01392*, 2018. 13

[96] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 13

[97] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," *arXiv preprint arXiv:1909.11480*, 2019. 13

[98] P. Oza and V. M. Patel, "Deep cnn-based multi-task learning for open-set recognition," *arXiv preprint arXiv:1903.03161*, 2019. 13, 24, 33, 34

[99] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, and V. M. Patel, "Generative-discriminative feature representations for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 814–11 823. 13, 35, 36

[100] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2307–2316. 13, 24, 36

[101] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1511–1519. 14

[102] C. You, D. P. Robinson, and R. Vidal, "Provable self-representation based outlier detection in a union of subspaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3395–3404. 14

[103] R. Chalapathy, A. K. Menon, and S. Chawla, "Robust, deep and inductive anomaly detection," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 36–51. 14

[104] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Advances in Neural Information Processing Systems*, 2018, pp. 9758–9769. 14

[105] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016. 14, 39

[106] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," *Advances in neural information processing systems*, vol. 19, 2006. 14, 47

[107] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951. 14

[108] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*. Springer, 2017, pp. 153–171. 14

[109] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2019, pp. 4893–4902. 14

[110] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE transactions on neural networks*, vol. 22, no. 2, pp. 199–210, 2010. 14, 39

[111] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105. 14

[112] X. Zhang, F. X. Yu, S.-F. Chang, and S. Wang, "Deep transfer network: Unsupervised domain adaptation," *arXiv preprint arXiv:1503.00591*, 2015. 14

[113] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011. 15

[114] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, vol. 11, no. 12, 2010. 15

[115] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," *arXiv preprint arXiv:1206.4683*, 2012. 15

[116] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 421–436. 15

[117] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 597–613. 15

[118] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189. 15

[119] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Thirty-second AAAI conference on artificial intelligence*, 2018. 15

[120] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4068–4076. 15

[121] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," *arXiv preprint arXiv:1412.4446*, 2014. 15

[122] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE geoscience and remote sensing magazine*, vol. 4, no. 2, pp. 41–57, 2016. 15

[123] X. Zhou and S. Prasad, "Deep feature alignment neural networks for domain adaptation of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5863–5872, 2018. 16

[124] J. Shen, X. Cao, Y. Li, and D. Xu, "Feature adaptation and augmentation for cross-scene hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 622–626, 2018. 16

[125] Z. Wang, B. Du, Q. Shi, and W. Tu, "Domain adaptation with discriminative distribution and manifold embedding for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1155–1159, 2019. 16

[126] X. Li, L. Zhang, B. Du, and L. Zhang, "On gleaning knowledge from cross domains by sparse subspace correlation analysis for hyperspectral image

classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3204–3220, 2018. 16

[127] T. Liu, X. Zhang, and Y. Gu, "Unsupervised cross-temporal classification of hyperspectral images with multiple geodesic flow kernel learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9688–9701, 2019. 16

[128] Y. Qin, L. Bruzzone, and B. Li, "Tensor alignment based domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9290–9307, 2019. 16

[129] Y. Qin, L. Bruzzone, B. Li, and Y. Ye, "Cross-domain collaborative learning via cluster canonical correlation analysis and random walker for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3952–3966, 2019. 16

[130] G. Gao and Y. Gu, "Tensorized principal component alignment: A unified framework for multimodal high-resolution images classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 46–61, 2018. 16

[131] J. Peng, W. Sun, L. Ma, and Q. Du, "Discriminative transfer joint matching for domain adaptation in hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 972–976, 2019. 16

[132] W. Wei, W. Li, L. Zhang, C. Wang, P. Zhang, and Y. Zhang, "Robust hyperspectral image domain adaptation with noisy labels," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1135–1139, 2019. 16

[133] Y. Qu, R. K. Baghbaderani, W. Li, L. Gao, Y. Zhang, and H. Qi, "Physically constrained transfer learning through shared abundance space for hyperspectral

image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 455–10 472, 2021. 16, 47, 52, 53

[134] R. Kaviani Baghbaderani, Y. Qu, H. Qi, and C. Stutts, "Representative-discriminative learning for open-set land cover classification of satellite imagery," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17. 17

[135] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, pp. 639–650, 1994. 27, 45

[136] E. Nalisnick and P. Smyth, "Stick-breaking variational autoencoders," *International Conference on Learning Representations (ICLR)*, 2017. 27, 28, 45

[137] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse dirichlet-net for hyperspectral image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2511–2520, 2018. 27, 28, 45

[138] S. Huang and T. D. Tran, "Sparse signal recovery via generalized entropy functions minimization," *IEEE Transactions on Signal Processing*, vol. 67, no. 5, pp. 1322–1337, 2018. 28, 45

[139] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 30, 52, 68

[140] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. 35

[141] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, 2015. 35

[142] L. Zhang, S. Wang, G.-B. Huang, W. Zuo, J. Yang, and D. Zhang, "Manifold criterion guided transfer learning via intermediate domain generation," *IEEE*

transactions on neural networks and learning systems, vol. 30, no. 12, pp. 3759–3773, 2019. 39

[143] R. Kaviani Baghbaderani, Y. Qu, and H. Qi, "Hyperspectral image domain adaptation through unmixing abundance maps," 2022. 39

[144] C. Deng, X. Liu, C. Li, and D. Tao, "Active multi-kernel domain adaptation for hyperspectral image classification," *Pattern Recognition*, vol. 77, pp. 306–315, 2018. 41

[145] J. Lin, L. Zhao, S. Li, R. Ward, and Z. J. Wang, "Active-learning-incorporated deep transfer learning for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4048–4062, 2018. 41

[146] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1741–1754, 2018. 41

[147] Z. Li, X. Tang, W. Li, C. Wang, C. Liu, and J. He, "A two-stage deep domain adaptation method for hyperspectral image classification," *Remote Sensing*, vol. 12, no. 7, p. 1054, 2020. 41

[148] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 47

[149] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space embedding for distributions," in *International Conference on Algorithmic Learning Theory*. Springer, 2007, pp. 13–31. 47

[150] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010. 47

[151] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456. 47

[152] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010. 47

[153] R. Kaviani Baghbaderani, Y. Li, S. Wang, and H. Qi, "Occlusion-aware bidirectional video semantic segmentation," 2022. 56

[154] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 57

[155] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. 57

[156] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. 57, 66

[157] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009. 57, 66

[158] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1013–1020. 57

[159] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, 2015. 57

[160] S. Liu, C. Wang, R. Qian, H. Yu, R. Bao, and Y. Sun, "Surveillance video parsing with single frame supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 413–421. 57

[161] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, "Dynamic video segmentation network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6556–6565. 57

[162] B. Mahasseni, S. Todorovic, and A. Fern, "Budget-aware deep semantic video segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1029–1038. 57

[163] Y. Li, J. Shi, and D. Lin, "Low-latency video semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5997–6005. 57

[164] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766. 58, 68

[165] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470. 58, 66

[166] A. Alshin, E. Alshina, and T. Lee, "Bi-directional optical flow for improving motion compensation," in *28th Picture Coding Symposium*. IEEE, 2010, pp. 422–425. 62

[167] A. Alshin and E. Alshina, "Bi-directional pptical flow for future video codec," in *2016 Data Compression Conference (DCC)*. IEEE, 2016, pp. 83–90. 62

[168] E. Herbst, S. Seitz, and S. Baker, "Occlusion reasoning for temporal interpolation using optical flow," *Department of Computer Science and Engineering, University of Washington, Tech. Rep. UW-CSE-09-08-01*, 2009. 62

[169] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International journal of computer vision*, vol. 92, no. 1, pp. 1–31, 2011. 62

[170] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1701–1710. 62

[171] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008. 62

[172] D. Gu, Z. Wen, W. Cui, R. Wang, F. Jiang, and S. Liu, "Continuous bidirectional optical flow for video frame sequence interpolation," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1768–1773. 62

[173] S. S. Sengar and S. Mukhopadhyay, "Motion detection using block based bi-directional optical flow method," *Journal of Visual Communication and Image Representation*, vol. 49, pp. 89–103, 2017. 62

[174] J. Hur and S. Roth, "Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 312–321. 62, 63

[175] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 62, 63

[176] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4884–4893. 62

[177] Y. Chen, G. Lin, S. Li, O. Bourahla, Y. Wu, F. Wang, J. Feng, M. Xu, and X. Li, "Banet: Bidirectional aggregation network with occlusion handling for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3793–3802. 62

[178] Z. Wu, F. Li, R. Sukthankar, and J. M. Rehg, "Robust video segment proposals with painless occlusion handling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4194–4203. 62

[179] Y. Huang and I. Essa, "Tracking multiple objects through occlusions," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 1051–1058. 62

[180] A. Yilmaz, X. Li, and M. Shah, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1531–1536, 2004. 62

[181] Y. Zhou and H. Tao, "A background layer model for object tracking through occlusion," in *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 1079–1085. 62

[182] D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of mathematical models in computer vision*. Springer, 2006, pp. 237–257. 63

[183] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *European conference on computer vision*. Springer, 2010, pp. 438–451. 63, 66

[184] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015. 66

[185] A. Kundu, V. Vineet, and V. Koltun, "Feature space optimization for semantic video segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3168–3175. 66

[186] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016. 66

[187] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. 66

[188] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 66

Appendix

A Visual comparison among spaces X, Z_f, and S

Figures A1, A2, and A3 illustrate the mean of samples belonging to different classes in spaces X, Z_f, and S, respectively, using the PU dataset. The feature vectors learned through F and E, shown in Figs. A2 and A3, respectively, are sparse due to the sparsity constraint.

It can be seen from Fig. A1 that the spectrum of samples belonging to class 3 and 8 are close. However, the feature vectors, learned through F, corresponding to class 3 and 8 are more discriminative. Further, the discriminative and representative characteristics of the features are enhanced through encoder E, as illustrated in Fig. A3.

B Comparison of open-set recognition performing on space X and Z_f

To better compare the effectiveness of performing open-set recognition in spaces X and Z_f, we show the results of performing in each space separately using the PU dataset in Table A1.

Comparing the results of $AE + CLS$ and $AE + CLS + Dirichlet$ approaches performed on spaces X and Z_f, it can be seen that discriminative features learned through the classifier F contribute to a substantial improvement. In addition, the Dirichlet network plays more critical role when performing open-set recognition in space X as compared to space Z_f.

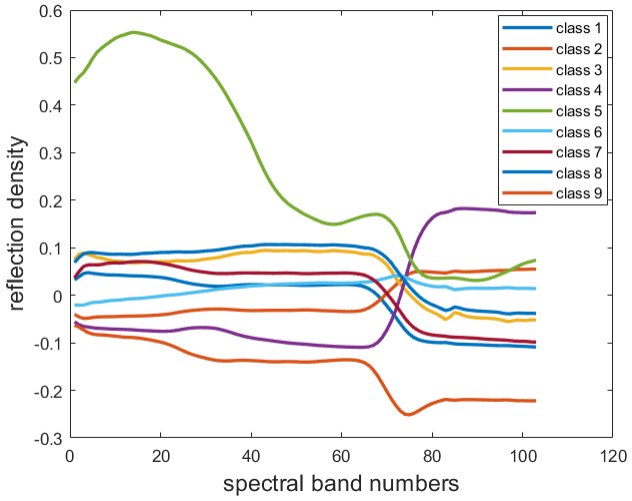


Figure A1: Mean of samples in space X, belonging to classes 1 to 9, using the PU dataset

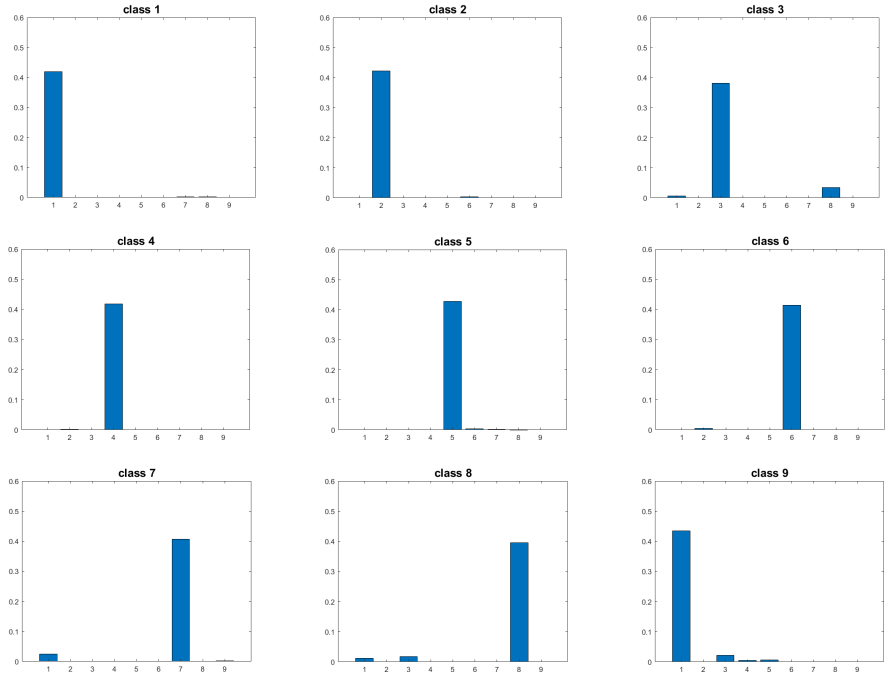


Figure A2: Mean of samples in space Z_f, belonging to classes 1 to 9, using the PU dataset

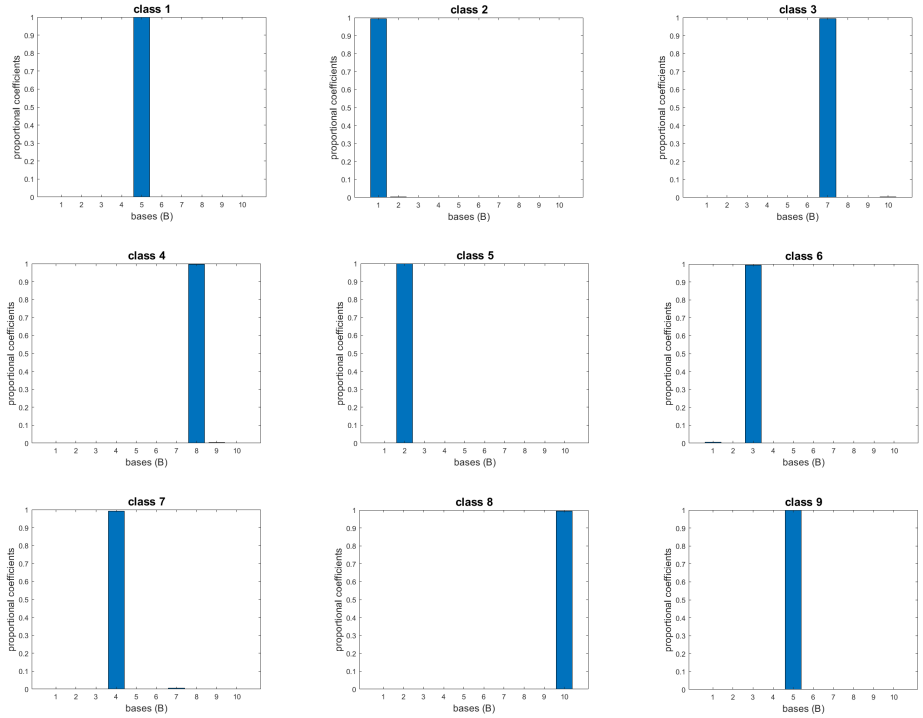


Figure A3: Mean of samples in space S, belonging to classes 1 to 9, using the PU dataset

Table A1: Area under the ROC curve for open-set recognition. Results are from partitioning PU dataset to $L-1$ known and the mentioned unknown classes, (openness=2.99%). Note that L denotes the number of classes in the original PU dataset

Space	**Method**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Avg.**
X	SoftMax	0.54	0.52	0.51	0.42	0.14	0.38	0.23	0.64	0.09	0.39
	OpenMax [26]	0.67	0.37	0.45	0.40	0.99	0.35	0.12	0.57	0.04	0.44
	AE+CLS [30]	0.51	0.53	0.54	0.83	1.0	0.46	0.48	0.46	0.46	0.59
	AE+CLS+Dirichlet	0.79	0.69	0.47	0.90	1.0	0.64	0.47	0.48	0.97	0.71
Z_f	AE+CLS	0.91	0.70	0.68	0.72	1.0	0.62	0.46	0.66	0.94	0.74
	AE+CLS+Dirichlet	0.91	0.70	0.71	0.72	1.0	0.68	0.51	0.80	0.93	**0.77**

Vita

In 2008, Razieh Kaviani Baghbaderani entered the Isfahan University of Technology (IUT), Isfahan, Iran, as an undergraduate student in the Department of Electrical Engineering. In 2012, she received the Bachelor's degree as an outsatnding student ranked top 5%, and then she entered the Master's program in Sharif University of Technology (SUT), Tehran, Iran.

In the Fall of 2018, she began to pursue the doctoral degree at the University of Tennessee - Knoxville, TN. She received the Tennessee's Top 100 Fellowship in 2018 and Excellence in Graduate Research in 2020. In the four-year study period supervised by Dr. Hairong Qi in the Department of Electrical Engineering and Computer Science, she focused on approaches to address the challenges in learning with limited labeled data for image and video understanding. During the doctoral program, she had internships with Samsung Semiconductor Inc, San Diego, CA, and American Family Insurance, Madison, WI. In June 2022, she successfully passed the Ph.D. defense and will return to Samsung Semiconductor Inc. as a full-time senior research scientist.