SAR Image Classification Using Few-shot Cross-domain Transfer Learning

Mohammad Rostami
University of Pennsylvania,
Philadelphia, PA, 19104

Soheil Kolouri
HRL Laboratories, LLC
Malibu, CA, 90265

Eric Eaton
University of Pennsylvania,
Philadelphia, PA, 19104

Kyungnam Kim
HRL Laboratories, LLC
Malibu, CA, 90265

Abstract

Data-driven classification algorithms based on deep convolutional neural networks have reached human-level performance for many tasks within Electro-Optical (EO) computer vision. Despite being the prevailing visual sensory data, EO imaging is not effective in applications such as environmental monitoring at extended periods, where data collection at occluded weather is necessary. Synthetic Aperture Radar (SAR) is an effective imaging tool to circumvent these limitations and collect visual sensory information continually. However, replicating the success of deep learning on SAR domains is not straightforward. This is mainly because training deep networks requires huge labeled datasets and data labeling is a lot more challenging in SAR domains. We develop an algorithm to transfer knowledge from EO domains to SAR domains to eliminate the need for huge labeled data points in the SAR domains. Our idea is to learn a shared domain-invariant embedding for cross-domain knowledge transfer such that the embedding is discriminative for two related EO and SAR tasks, while the latent data distributions for both domains remain similar. As a result, a classifier learned using mostly EO data can generalize well on the related task for the SAR domain.

1. Introduction

Electro-Optical (EO) images are the dominant visual data that are collected and processed as input sensory data in computer vision applications for supervised learning. With the emergence of deep convolutional neural networks (CNNs), autonomous systems can now rely on classification and detection algorithms that process and learn from EO data with human-level performance. This success stems from the fact that deep nets can be trained in a data-driven scheme using a huge labeled dataset of images to automatically extract abstract and high-quality features for a given task. This possibility has helped to bypass feature engineering, which was a major bottleneck in vision applications.

Despite wide range of applicability of EO imaging, in applications such as continuous environmental monitoring and large-scale surveillance [18] and earth remote sensing [21] which require imaging at extended time periods, EO imaging is not feasible. In these applications, using SAR imaging is inevitable, since SAR imaging can provide high-resolution images using the radar signals that can propagate in occluded weather. While both the EO and the SAR domain images describe the common physical world, processing EO and SAR data and developing suitable learning algorithms can be quite different. In particular, as opposed to EO domains, training and using CNNs in SAR domains is more challenging. This is because training CNNs depends on the availability of huge labeled datasets to supervise learning. Generating such datasets can be challenging. This challenge is currently tackled through crowdsourcing labeling platforms such as Amazon Mechanical Turk for EO domain tasks, e.g. ImageNet [8]. In a crowdsourcing platform, EO data points, i.e. images, are presented to a pool of participants with common basic knowledge for labeling. Each participant selects a label for each given image. Upon collecting labels from the pool of applicants, collected labels are aggregated to increase labeling accuracy [29]. Despite being very effective for EO domains, crowdsourcing platforms are not suitable for SAR domains:

- Preparing devices for collecting SAR data, solely for generating training datasets is much more expensive compared to EO datasets [22].
- SAR images are often classified data, making access to SAR data heavily regulated and limited. This limits the number of participants who can be hired to help with processing and labeling.
- Labeling SAR images needs trained experts, as opposed to tasks within the EO domain images [31]. This

makes labeling SAR data more expensive.

- Continuous collection of SAR data is common in SAR applications. This can make the labeled data unrepresentative of the current distribution, obligating persistent labeling for model retraining [12].

As a result, generating labeled datasets for the SAR domain data is challenging. Additionally, training a CNN using most existing SAR datasets leads to overfit models that do not generalize well on test sets [3, 32]. In other words, we face situations in which the amount of accessible SAR data is not sufficient for training deep neural networks. Learning is these scenarios has been investigated within transfer learning [25]. Building upon prior works in the area of transfer learning, several recent works have used the idea of knowledge transfer to address challenges of SAR domains [12, 22, 35, 34, 19, 32]. The common idea in these works is to transfer knowledge from a secondary related domain, where labeled data is easy and cheap to obtain. Following this line of work, we focus on addressing scarcity of labeled data in SAR domains through cross-domain knowledge transfer from a related task in EO domains.

A common technique for cross-domain knowledge transfer is to map or encode data points of the two related domains to a domain-invariant embedding space such that knowledge can be transferred across the domains via the embedding space. Consider a classification task in two domains, e.g. SAR and EO, where we have sufficient labeled data points in the source domain, i.e. EO, but only few labeled data points in the target domain, i.e. SAR. This setting is called semi-supervised domain adaptation in the computer vision literature [24]. If we can train two deep encoders to map the data points from both domains into a shared embedding space such that both domains would have similar class-conditioned probability distributions in the embedding space, then a classifier trained using the source-domain data points in the shared embedding, would generalize well to the target domain [28]. This goal can be achieved by training the deep encoders such that the empirical distribution discrepancy between the two domains is minimized in the shared output of the deep encoders with respect to some probability distribution metric[33, 10].

In this paper, our contribution is to propose a novel semi-supervised domain adaptation algorithm to transfer knowledge from the EO domain to the SAR domain using the above explained procedure. We use the Sliced-Wasserstein Distance (SWD) [26] to measure and minimize the discrepancy between the source and the target domain distributions in order to supervise training of domain-specific encoders. SWD is an effective metric for the space of probability distributions that can be computed efficiently. More importantly, it is a differentiable metric with non-vanishing gradients, which make it a suitable objective function for training deep networks using gradient-based optimization [17, 28]. This is important as most optimization problems for training deep neural networks are solved using gradient-based methods, e.g. stochastic gradient descent (SGD). This strategy on its own might not succeed because distributions may not be aligned class-conditionally. We use the few accessible labeled data points in the SAR domain to align both distributions class-conditionally to tackle the class matching challenge [14]. We provide experimental results to validate our approach in the area of maritime domain awareness, where the goal is to understand activities that could impact the safety and the environment. Our results demonstrate our approach is effective and leads to SOA performance.

2. Related Work

Several prior works have applied the idea of transfer learning to the SAR domain. Huang et al. [12] address the problem of labeled data scarcity in the SAR domain via unsupervised learning. The idea is to use a large pool of unlabeled SAR data to train an autoencoder. As a result, the embedding space learned by the autoencoder is discriminative and can be used as a mean for better feature extraction to benefit from knowledge transfer. The trained encoder subnetwork can be concatenated with a classifier network and both would be fine-tuned using the labeled portion of data to map the data points to the label space. Hansen et al. [22] proposed to transfer knowledge using synthetic SAR images which are easy to generate. Their major novelty is to demonstrate how to generate a simulated dataset for a given SAR problem based on simulated object radar reflectivity. A CNN is then pretrained on the synthetic dataset and then used as an initialization for the real SAR domain problem. Due to the pretraining stage, the model can be fine-tuned using fewer real labeled data points. Zhang et al. [35] propose to transfer knowledge from a secondary source SAR task, where labeled data is available. Their idea is to pretrain a CNN on the task with labeled data and then fine-tune it on the target task. Lang et al. [19] use automatic identification system (AIS) as the secondary domain for knowledge transfer. AIS is a tracking system for monitoring movement of ships that can provide labeling information. Shang et al. [32] amend a CNN with an information recorder. The recorder is used to store spatial features of labeled samples and the recorded features are used to predict labels of unlabeled data points based on spatial similarity to increase the number of labeled samples. Finally, Weng et al. [34] use an approach more similar to our framework. Their proposal is to transfer knowledge using VGGNet as a feature extractor in the learning pipeline, which itself has been trained on a large EO dataset. Despite being novel, these past works mostly use a pretrained deep network that is trained using a secondary source of knowledge, which is then fine-tuned using few labeled data points on the target

SAR task. Hence, knowledge transfer occurs as a result of selecting a better initial point using the secondary source. We follow a different approach by recasting the problem as a domain adaptation (DA) problem [10], where the goal is to adapt a model trained on the source domain to generalize well in the target domain. Our contribution is to demonstrate how to transfer knowledge from EO imaging domain in order to train a deep network for the SAR domain. The idea is to train a deep network on a related EO problem with abundant labeled data and adapt the model using only few labeled SAR data points such that the distributions of both domains become similar within a mid-layer of the network.

Domain adaptation has been investigated in the computer vision literature for a broad range of scenarios. The goal is to learn a model on a source data distribution with sufficient data such that it generalizes well on a different, but related target data distribution with insufficient labeled data. Early DA algorithms either develop domain invariant and stable features which can be used on both domains [6] or learn a function to map the target data points into the source domain [30]. Despite being very different solutions, both approaches try to preprocess data such that the distributions of both domains become similar after preprocessing. As a result, a classifier trained using the source data, can also be used on the target domain. In this paper, we consider that two deep convolutional neural networks with a shared output space, i.e. deep encoders, preprocess data to enforce both EO and SAR domains data to have similar probability distributions in their shared output. This space can be considered as an intermediate embedding space between the input space from each domain and label space of a shared classifier network between the two domains. These deep encoders are trained such that the discrepancy between the source and the target domain distributions is minimized in the shared embedding space, while overall classification is supervised via the EO domain labeled data. This procedure has been done via both adversarial learning [9] and as an optimization problem with probability matching objective [4].

In order to minimize the distance between two probability distributions, we minimize with respect to a measure of distance between two empirical distributions. Early works in domain adaptation used the Maximum Mean Discrepancy (MMD) metric for this purpose [10]. MMD measures the distance between two distribution as the Euclidean distance between their means. However, MMD might not be an accurate measure when the distributions are multi-modal. Other common discrepancy measures such as KL divergence and Jensen-Shannon divergence can be used for a broader range of domain adaptation problems [7]. But these measures have vanishing gradients when the distributions are too distant, which makes them inappropriate for deep learning as deep networks are trained using gradient-based first-order optimization [16]. For this reason, recent works in deep learning use the Wasserstein Distance (WD) metric as an objective function to match distributions [33]. WD has non-vanishing gradient but it does not have a closed-form definition and is defined as a linear programming (LP) problem. Solving the LP problem can be computationally expensive for high-dimensional distributions. To circumvent this challenge, we use the Sliced Wasserstein Distance (SWD). SWD approximates WD as sum of multiple Wasserstein distances of one-dimensional distributions which possess a closed-form solution [26, 1, 2, 16].

3. Problem Formulation and Rationale

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the domain space of SAR data. Consider a multiclass SAR classification problem with k classes in this domain, where i.i.d data pairs are drawn from the joint probability distribution, i.e. $(\boldsymbol{x}_i^t, \boldsymbol{y}_i^t) \sim q_T(\boldsymbol{x}, \boldsymbol{y})$ which has the marginal distribution $p_T(\boldsymbol{x})$ over $\mathcal{X}$. Here, a label $\boldsymbol{y}_i^t$ identifies the class membership of the vectorized SAR image $\boldsymbol{x}_t^i$ to one of the k classes. We have access to $M \gg 1$ unlabeled images $\mathcal{D}_{\mathcal{T}} = (\boldsymbol{X}_{\mathcal{T}} = [\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_M^t]) \in \mathbb{R}^{d \times M}$ in this target domain. Additionally, we have have access to O labeled images $\mathcal{D}'_{\mathcal{T}} = (\boldsymbol{X}'_{\mathcal{T}}, \boldsymbol{Y}'_{\mathcal{T}})$, where $\boldsymbol{X}'_{\mathcal{S}} = [\boldsymbol{x}'^t_1, \ldots, \boldsymbol{x}'^t_O] \in \mathbb{R}^{d \times O}$ and $\boldsymbol{Y}'_{\mathcal{S}} = [\boldsymbol{y}'^t_1, \ldots, \boldsymbol{y}'^t_O] \subset \mathbb{R}^{k \times O}$ contains the corresponding one-hot labels. Our goal is to train a parameterized classifier $f_\theta : \mathbb{R}^d \rightarrow \mathcal{Y} \subset \mathcal{R}^k$, i.e. a deep neural network with weight parameters θ, on this domain. Given that we have access to only few labeled data points and considering model complexity of deep neural networks, training the deep network such that it generalizes well using solely the SAR labeled data is not feasible and would lead to overfitting on the few labeled data points such that the trained network would generalize poorly.

To tackle the problem of label scarcity, we consider a domain adaptation scenario, where we have access to sufficient labeled data points in a related source EO domain problem. Let $\mathcal{D}_{\mathcal{S}} = (\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{Y}_{\mathcal{S}})$ denote the dataset in the EO domain, with $\boldsymbol{X}_{\mathcal{S}} \in \mathcal{X} \subset \mathbb{R}^{d' \times N}$ and $\boldsymbol{Y}_{\mathcal{S}} \in \mathcal{Y} \subset \mathbb{R}^{k \times N}$ ($N \gg 1$). Note that we are considering the same classification problem in two domains, i.e. the same classes similar to the target domain exist in the source domain. We assume the source samples are drawn i.i.d. from the source joint probability distribution $q_{\mathcal{S}}(\boldsymbol{x}, \boldsymbol{y})$, which has the marginal distribution $p_{\mathcal{S}}$. Given that extensive research and investigation has been done in EO domains, we hypothesize that finding such a labeled dataset is likely feasible and is easier than labeling more SAR data points. Our goal is to use the similarity between the EO and the SAR domains to train a model for classifying SAR images using the knowledge that can be learned from the EO domain.

Since we have access to sufficient labeled data points in the EO domain, training a deep network in this domain is straightforward. Following the standard supervised learning setting, we can use empirical risk minimization (ERM) to learn the network optimal weight parameters:

$$\hat{\theta} = \arg\min_{\theta} \hat{e}_{\theta} = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_{\theta}(\boldsymbol{x}_i^s), \boldsymbol{y}_i^s) \ , \quad (1)$$

where $\mathcal{L}$ is a proper loss function (e.g., cross entropy loss). Given enough training data points, the empirical risk is a suitable surrogate for the real risk function:

$$e = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_S(\boldsymbol{x},\boldsymbol{y})}(\mathcal{L}(f_{\theta}(\boldsymbol{x}), \boldsymbol{y})) \ , \quad (2)$$

which is the objective function for Bayes optimal inference. This means that the learned classifier would generalize well on data points if they are drawn from p_S. A naive approach to transfer knowledge from the EO domain to the SAR domain is to use the classifier that is trained on the EO domain directly in the target domain. However, since distribution discrepancy exists between the two domains, i.e. $p_S \neq p_T$, the trained classifier on the source domain $f_{\hat{\theta}}$, might not generalize well on the target domain. Therefore, there is a need for adapting the training procedure for $f_{\hat{\theta}}$. The simplest approach which has been used in most prior works is to fine-tune the EO classifier using the few labeled target data points to employ the model in the target domain. But we want to use a more principled approach and take advantage of the unlabeled SAR data points which are accessible and provide additional information about the SAR domain marginal distribution. Additionally, fine tuning requires $d = d'$ which might not always be the case.

In our approach, we consider the EO deep network $f_{\theta}(\cdot)$ to be formed by a feature extractor $\phi_{\boldsymbol{v}}(\cdot)$, i.e. convolutional layers of the network, which is followed by a classifier subnetwork $h_{\boldsymbol{w}}(\cdot)$, i.e. fully connected layers of the network, that inputs the extracted feature and maps them to the label space. Here, $\boldsymbol{w}$ and $\boldsymbol{v}$ denote the corresponding learnable parameters for these sub-networks, i.e. $\theta = (\boldsymbol{w}, \boldsymbol{v})$. In other words, the feature extractor sub-network $\phi_{\boldsymbol{v}} : \mathcal{X} \rightarrow \mathcal{Z}$ maps the data points into a discriminative embedding space $\mathcal{Z} \subset \mathbb{R}^f$, where classification can be done easily by the classifier sub-network $h_{\boldsymbol{w}} : \mathcal{Z} \rightarrow \mathcal{Y}$. The success of deep learning stems from optimal feature extraction which converts the data distribution into a multimodal distribution which allows for class separation. Following the above, we can consider a second encoder network $\psi_{\boldsymbol{u}}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^f$, which maps the SAR data points to the same target embedding space at its output. The idea that we want to explore is based on training $\phi_{\boldsymbol{v}}$ and $\psi_{\boldsymbol{u}}$ such that the discrepancy between the source distribution $p_S(\phi(\boldsymbol{x}))$ and target distribution $p_T(\phi(\boldsymbol{x}))$ is minimized in the shared embedding space. As a result of matching the two distributions, the embedding space becomes invariant with respect to the domain. Consequently, even if we train the classifier sub-network using solely the source labeled data points, it will still generalize well when target data points are used for testing. The key question is how to train the encoder sub-networks such that the embedding space becomes invariant. Figure 1 presents a block diagram visualization of our framework. In the figure, we have visualized images from two related real world SAR and EO datasets. Notice that SAR images are confusing for the untrained human eye, compared to EO ship/no-ship images which suggests that as we discussed SAR labeling is more challenging and requires expertise.

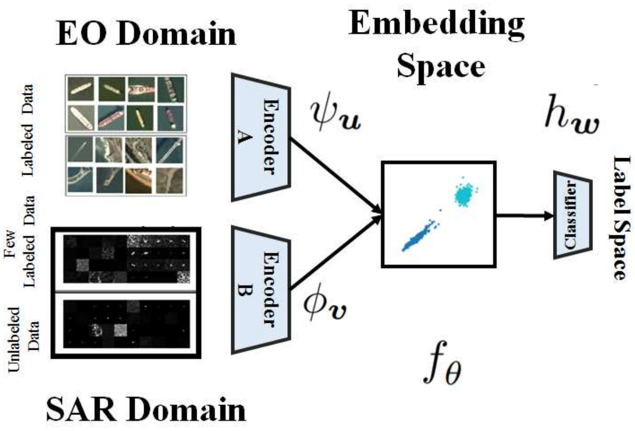


Figure 1: Block diagram architecture of the proposed framework for transferring knowledge from the EO to the SAR domain.

4. Proposed Optimization Solution

In our solution, the encoder sub-networks need to be learned such that the extracted features in the encoder output are discriminative. Only then, the classes become separable for the classifier sub-network (see Figure 1). This is a direct result of supervised learning for EO encoder. Additionally, the encoders should mix the SAR and the EO domains such that the embedding becomes domain-invariant. Hence, the SAR encoder indirectly is enforced to be discriminative for the SAR domain. Domain invariance can be enforced by minimizing the discrepancy between the distributions of both domains in the embedding space. Following the above, we can formulate the following optimization problem for computing optimal values for $\boldsymbol{v}, \boldsymbol{u}$ and $\boldsymbol{w}$:

$$\begin{aligned}\min_{\boldsymbol{v},\boldsymbol{u},\boldsymbol{w}} & \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\big(h_{\boldsymbol{w}}(\phi_{\boldsymbol{v}}(\boldsymbol{x}_i^s)), \boldsymbol{y}_i^s\big) \\ & + \frac{1}{O}\sum_{i=1}^{O} \mathcal{L}\big(h_{\boldsymbol{w}}(\psi_{\boldsymbol{u}}(\boldsymbol{x}_i^{\prime t})), \boldsymbol{y}_i^{\prime t}\big) \\ & + \lambda D\big(\phi_{\boldsymbol{v}}(p_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}})), \psi_{\boldsymbol{u}}(p_{\mathcal{T}}(\boldsymbol{X}_{\mathcal{T}}))\big) \\ & + \eta \sum_{j=1}^{k} D\big(\phi_{\boldsymbol{v}}(p_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}})|C_j), \psi_{\boldsymbol{u}}(p_{\mathcal{T}}(\boldsymbol{X}'_{\mathcal{T}})|C_j)\big) \ ,\end{aligned} \quad (3)$$

where $D(\cdot,\cdot)$ is a discrepancy measure between the probabilities and λ and η are trade-off parameters. The first two terms in Eq. (3) are empirical risks for classifying the EO and SAR labeled data points, respectively. The third term is the cross-domain unconditional probability matching loss. The matching loss is computed using all available data points from both domains to learn the learnable parameters of encoder sub-networks and the classifier sub-network is simultaneously learned using the labeled data from both domains. Finally, the last term is Eq. (3) is added to enforce semantic consistency between the two domains. This term is important for knowledge transfer. To clarify this point, note that the domains might be aligned such that their marginal distributions $\phi(p_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}))$ and $\psi(p_{\mathcal{T}}(\boldsymbol{X}_{\mathcal{T}}))$ have minimal discrepancy, while the distance between $\phi(q_{\mathcal{S}}(\cdot,\cdot))$ and $\psi(q_{\mathcal{T}}(\cdot,\cdot))$ is not minimized. This means that the classes may not have been aligned correctly, e.g. images belonging to a class in the target domain may be matched to a wrong class in the source domain or, even worse, images from multiple classes in the target domain may be matched to the cluster of another class of the source domain. In such cases, the classifier will not generalize well on the target domain as it has been trained to be consistent with spatial arrangement of the source domain in the embedding space. This means that if we merely minimize the distance between $\phi(p_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}))$ and $\psi(p_{\mathcal{T}}(\boldsymbol{X}_{\mathcal{T}}))$, the shared embedding space might not be a consistently discriminative space for both domains domain in terms of classes. The challenge of class-matching is a known problem in domain adaptation and several approaches have been developed to address this challenge [20]. In our framework, the few labeled data points in the target SAR domain can be used to match the classes consistently across both domains. We use these data points to computer the fourth term in Eq. (3). This term is added to match class-conditional probabilities of both domains in the embedding space, i.e. $\phi(p_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})|C_j) \approx \psi(p_{\mathcal{T}}(\boldsymbol{x}|C_j)$, where C_j denotes a particular class.

The final key question is to select a proper metric to compute $D(\cdot,\cdot)$ in the last two terms of Eq 1. KL divergence and Jensen-Shannon divergence have been used extensively to measure closeness of probability distributions as maximizing the log-likelihood is equivalent to minimizing the KL-divergence between two distributions but note that since we will use SGD to solve the optimization problem in Eq 1, they are not suitable. This is a major reason for success of adversarial learning [9]. Additionally, the distributions $\phi(p_{\mathcal{S}}(\boldsymbol{x})$ and $\psi(p_{\mathcal{T}}(\boldsymbol{x})$ are unknown and we can rely only on observed samples from these distributions. Therefore, we should be able to compute the discrepancy measure, $D(\cdot,\cdot)$ using only on the drawn samples. Optimal transport [33] is a suitable metric to deal with the above issues and due to be an effective metric, it has been used extensively in deep learning literature recently [5, 4, 15, 28]. In this paper, we use the Sliced Wasserstein Distance (SWD) [27] is a good approximate of optimal transport [17] and can be computed more efficiently.

Algorithm 1 FCS (L, η, λ)

```
Input: data
$\mathcal{D}_{\mathcal{S}} = (\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{Y}_{\mathcal{S}}); \mathcal{D}_{\mathcal{T}} = (\boldsymbol{X}_{\mathcal{T}},, \boldsymbol{Y}_{\mathcal{T}}), \mathcal{D}'_{\mathcal{T}} = (\boldsymbol{X}'_{\mathcal{T}}),$
Pre-training: initialization
    $\hat{\theta}_0 = (\boldsymbol{w}_0, \boldsymbol{v}_0) = \arg\min_\theta 1/N \sum_{i=1}^N \mathcal{L}(f_\theta(\boldsymbol{x}_i^s), \boldsymbol{y}_i^s)$
for $itr = 1, \ldots, ITR$ do
    Update encoder parameters using:
        $\hat{\boldsymbol{v}}, \hat{\boldsymbol{u}} = \lambda D\big(\phi_{\boldsymbol{v}}(p_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}})), \psi_{\boldsymbol{u}}(p_{\mathcal{T}}(\boldsymbol{X}_{\mathcal{T}}))\big)$
                $+\eta \sum_j D\big(\phi_{\boldsymbol{v}}(p_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}})|C_j), \psi_{\boldsymbol{v}}(p_{\mathcal{S}\mathcal{L}}(\boldsymbol{X}'_{\mathcal{T}})|C_j)\big)$
    Update entire parameters:
        $\hat{\boldsymbol{v}}, \hat{\boldsymbol{u}}, \hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w},\boldsymbol{v},\boldsymbol{u}} 1/N \sum_{i=1}^N \mathcal{L}\big(h_{\boldsymbol{w}}(\phi_{\boldsymbol{v}}(\boldsymbol{x}_i^s)), \boldsymbol{y}_i^s\big)$
                        $+1/O \sum_{i=1}^O \mathcal{L}\big(h_{\boldsymbol{w}}(\psi_{\boldsymbol{u}}(\boldsymbol{x}_i'^t)), \boldsymbol{y}_i'^t\big)$
end for
```

The Wasserstein distance is defined as the solution to a linear programming problem. However, for the case of one-dimensional probability distributions, this problem has a closed form solution which can be computed efficiently. The solution is equal to the ℓ_p-distance between the inverse of the cumulative distribution functions of the two distributions. SWD has been proposed to benefit from this property. The idea is to decompose a d-dimensional distributions into one-dimension marginal distributions by projecting the distribution along all possible hyperplanes that cover the space. This process is called slicing the high-dimensional distributions. For a distribution $p_{\mathcal{S}}$, a one-dimensional slice of the distribution along the projection direction γ is defined as:

$$\mathcal{R}p_{\mathcal{S}}(t;\gamma) = \int_{\mathcal{S}} p_{\mathcal{S}}(\boldsymbol{x})\delta(t - \langle\gamma, \boldsymbol{x}\rangle)d\boldsymbol{x} \quad , \tag{4}$$

where $\delta(\cdot)$ denotes the Kronecker delta function, $\langle\cdot,\cdot\rangle$ denotes the vector dot product, and $\mathbb{S}^{d-1}$ is the d-dimensional unit sphere. We can see that $\mathcal{R}p_{\mathcal{S}}(\cdot;\gamma)$ is computed via integrating $p_{\mathcal{S}}$ over the hyperplanes which are orthogonal to the projection directions γ that cover the space.

The SWD is computed by integrating the Wasserstein distance between sliced distributions over all γ:

$$SW(p_{\mathcal{S}}, p_{\mathcal{T}}) = \int_{\mathbb{S}^{d-1}} W(\mathcal{R}p_{\mathcal{S}}(\cdot;\gamma), \mathcal{R}p_{\mathcal{T}}(\cdot;\gamma))d\gamma \quad , \tag{5}$$

where $W(\cdot,\cdot)$ denotes the Wasserstein distance. Computing the above integral directly, is computationally expensive. But, we can approximate the integral in Eq. (5) using a Monte Carlo style integration by choosing L number of random projection directions γ and after computing the Wasserstein distance, average along the random directions.

In our problem, we have access only to samples from the two source and target distributions, so we approximate the one-dimensional Wasserstein distance as the ℓ_p-distance between the sorted samples, as the empirical commutative probability distributions. Following the above procedure, the SWD between f-dimensional samples $\{\phi(\mathbf{x}_i^{\mathcal{S}}) \in \mathbb{R}^f \sim p_{\mathcal{S}}\}_{i=1}^M$ and $\{\phi(\mathbf{x}_i^{\mathcal{T}}) \in \mathbb{R}^f \sim p_{\mathcal{T}}\}_{j=1}^M$ can be approximated as the following sum:

$$SW^2(p_{\mathcal{S}}, p_{\mathcal{T}}) \approx \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{M} |\langle \gamma_l, \phi(\mathbf{x}^{\mathcal{S}}_{s_l[i]})\rangle - \langle \gamma_l, \phi(\mathbf{x}^{\mathcal{T}}_{t_l[i]})\rangle|^2 \ , \tag{6}$$

where $\gamma_l \in \mathbb{S}^{f-1}$ is uniformly drawn random sample from the unit f-dimensional ball $\mathbb{S}^{f-1}$, and $s_l[i]$ and $t_l[i]$ are the sorted indices of $\{\gamma_l \cdot \phi(\mathbf{x}_i)\}_{i=1}^M$ for source and target domains, respectively. We utilize the SWD as the discrepancy measure between the probability distributions to match them in the embedding space. Our proposed algorithm for few-shot SAR image classification (FSC) using cross-domain knowledge transfer is summarized in Algorithm 1. Note that we have added a pretraining step which trains the EO encoder and the shared classifier sub-network solely on the EO domain for better initialization.

5. Experimental Validation

We demonstrated effectiveness of our method in the area of maritime domain awareness on SAR ship detection.

5.1. Ship detection dataset

We tested our approach in the binary problem of ship detection using SAR images [31]. This problem arises within maritime domain awareness (MDA) where the goal is monitoring large areas of ocean to decipher maritime activities that could impact the safety and the environment. Ships are important object in this application as the majority of important activities is related to ships. To reach this end, SAR imaging is highly effective because monitoring is done continually over extended time intervals. In order to automize the monitoring process, classic image processing techniques are used to determine regions of interest in aerial SAR images. First, land areas are removed and then ships, ship-like, and ocean regions are identified and then extracted as square image patches. These image patches are then fed into a classification algorithm to determine whether the region corresponds to a ship or not.

The dataset that we have used is obtained from aerial SAR images of the South African Exclusive Economic Zone. The dataset is preprocessed into 51×51 pixels subimages [31]. Each instance either contains ships (positive data points), or no-ship (negative data points). The dataset contains 1436 positive examples and 1436 negative subimages. The labels are provided by experts. We recast the problem as a few-shot learning problem. To solve this problem using knowledge transfer, we use the "EO Ships in Satellite Imagery" dataset [11]. The dataset is prepared to automate monitoring port activity levels and supply chain analysis and contains images extracted from Planet satellite imagery over the San Francisco Bay area with 4000 RGB 80×80 images. Again, each instance is either a ship image (a positive data point), or no-ship (a negative data point). The dataset is split evenly into positive and negative samples. Instances from both datasets are visualized in Figure 1 (left).

5.2. Methodology

We consider a deep CNN with 2 layers of convolutional filters for the SAR domain, with 16, and 32 filters in these layers respectively. We have used both maxpool and batch normalization layers in these convolutional layers. These layers are used as the SAR encoder sub-network in our framework, ϕ. These layers are followed by a flattening layer and a subsequent layer dense layer as the embedding space with dimension f which potentially can be tuned as a parameter. After the embedding space layer, we have used a shallow two-layer classifier based on Eq. (3). The EO encoder has similar structure with the exception of using three convolutional layers. We have used three layers as EO dataset seems to have more details and more complex model can be helpful. We used TensorFlow for implementation and the Adam optimizer [13].

For comparison purpose, we compared our results against the following learning settings:

1) Supervised training on the SAR domain (ST): we just trained a network directly in the SAR domain using the few labeled SAR data points to generate a lower-bound for approach to demonstrate that knowledge transfer is effective.

2) Direct transfer (DT): we just directly used the network that is trained on EO data directly in the SAR domain. In order to do this end, we resized the EO domain to 51×51 pixels so we can use the same shared encoder networks for both domains. As a result, potentially helpful details may be lost. This can be served as a second lower-bound to demonstrate that we can benefit from unlabeled data.

3) Fine tuning (FT): we used the no transfer network from previous method, and fine-tuned the network using the few available SAR data points. As discussed before, this is the main strategy that prior works have used in the literature to transfer knowledge from the EO to the SAR domain and is served to compare against previous methods.

In our experiments, we used a 90/10 % random split for training the model and testing performance. In our experiments, we report the performance on the SAR testing split to compare the methods. We use the classification accuracy rate to measure performance and whenever necessary, we used cross validation to tune the hyper parameters. We have repeated each experiment 20 times and have reported the average and the standard error bound to demonstrate statistical significance in the experiments.

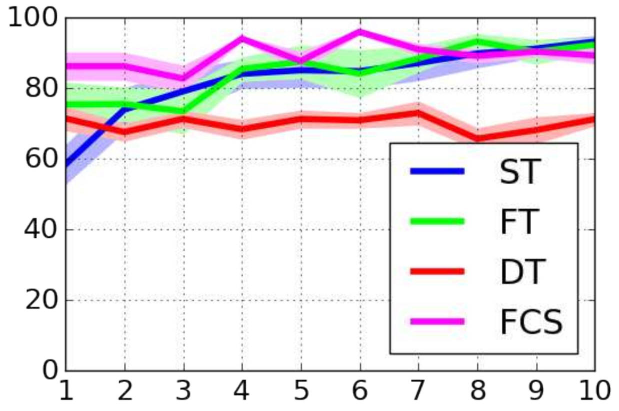
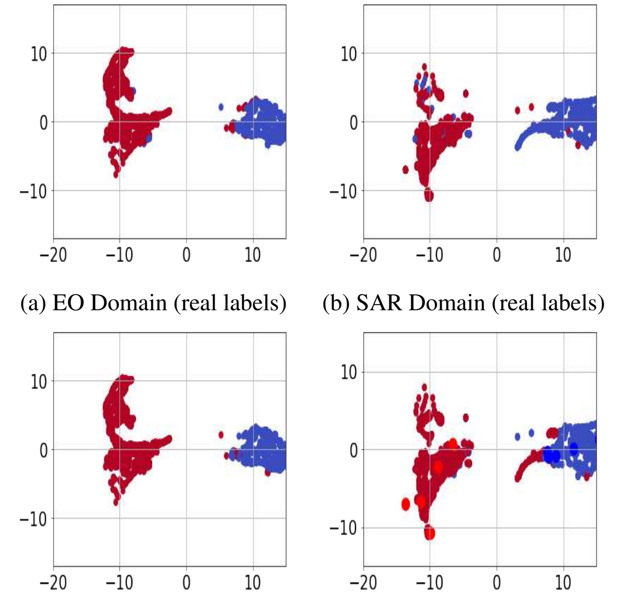


Figure 2: The SAR test performance versus the number of labeled data per class.



(a) EO Domain (real labels) (b) SAR Domain (real labels)

(c) EO Domain (predicted labels) (d) SAR Domain (labeled and unlabeled data)

Figure 3: Umap visualization of the EO versus the SAR dataset in the shared embedding space. (view in color.)

5.3. Results

O	1	2	3	4	5	6	7
ST	58.5	74.0	79.2	84.1	85.2	84.9	87.2
FT	75.5	75.6	73.5	85.5	87.6	84.2	88.5
DT	71.5	67.6	71.4	68.5	71.4	71.0	73.1
FCS	86.3	86.3	82.8	94.2	87.8	96.0	91.1

Table 1: Comparison results for the SAR test performance.

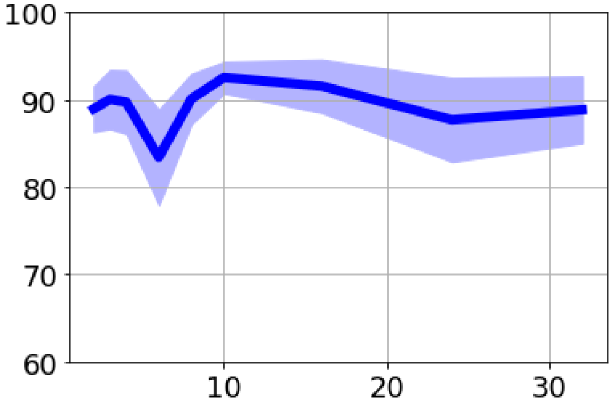
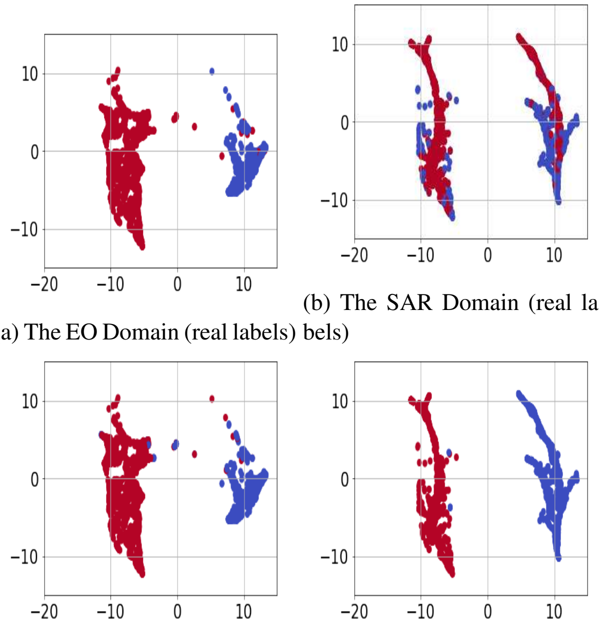


Figure 4: The test performance versus the dimension of the embedding space.

Figure 2 presents the performance results on the data test split for our method along with the three mentioned methods above, versus the number of labeled data points per class that has been used for the SAR domain. For each curve, the solid line denotes the average performance over all ten trials and the shaded region denotes the standard error deviation. These results accord with intuition. Supervised training on the SAR domain is not effective in few shot learning regime, i.e. its performance is close to chance. Direct transfer method boosts the performance in few-shot regime considerably (about 20%) but after 2-3 labeled samples per class, as expected supervised training overtakes direct transfer as task data is used. In other words, it only helps to start learning from a better initial point. Fine tuning can improve the DT performance, but only few-shot regime, and beyond few-shot learning regime the performance is similar to supervised training. Our method outperforms all methods as we have benefited from SAR unlabeled data points. As it can be seen, our approach is effective and leads to 30% boost from almost no-learning baseline, i.e. 50% performance, in few-shot learning regime. For a more clear quantitative comparison, we have presented data in Figure 2 in Table 1 for different number of labeled SAR data points per class (O).

For having better intuition, Figure 3 denotes the Umap visualization [23] of the EO and SAR data points in the learned embedding as the output of the feature extractor encoders. In this figure, we have used 5 labeled data points per class in the SAR domain. In Figure 3, each color corresponds to one of the classes. In Figures 3a and 3b, we have used real labels for visualization, and in Figures 3c and 3d, we have used the predicted labels by networks trained using our method for visualization. In Figure 3, the points with brighter red and darker blue colors are the SAR labeled data points that has been used in training. By comparing the top row with the bottom row, we see that the embedding is discriminative for both domains. Additionally, by comparing the left column with the right column, we see



(a) The EO Domain (real labels)

(b) The SAR Domain (real labels)

(c) The EO Domain (predicted labels)

(d) The SAR Domain (predicted labels)

Figure 5: Umap visualization of the EO versus the SAR dataset for ablation study. (view in color.)

that the domain distributions are matched in the embedding class conditionally, suggesting our framework formulated is Eq. (3) is effective. This result suggests that learning an invariant embedding space can be served as a helpful strategy for transferring knowledge. Additionally, we see that labeled data points are important to determine the boundary between two classes which suggests that why part of one of the classes (blue) is predicted mistakenly.

We also preformed a set of experiments to empirically study the effect of dimension of the embedding space on performance of our algorithm. Figure 4 presents performance on SAR testing set versus dimension of the embedding space when 10 SAR labeled data per class is used for training. The solid line denotes the average performance over ten trials and the shaded region denotes the standard error deviation. We observe that the performance is quite stable when the embedding space dimension changes. This result suggests that if the learned embedding space is discriminative for the source domain, then our method can successfully match the target domain distribution to the source distribution in the embedding. We conclude that for computational efficiency, it is better to select the embedding dimension to be as small as possible. For this reason, we set the dimension of the embedding to be 8 for the other our experiments in this paper as we conclude from Figure 4 that increasing the dimension beyond 8 is not helpful.

Finally, we also performed an experiment to serve as an ablation study for our framework. Our previous experiments demonstrate that the first three terms in Eq. (3) are all important for successful knowledge transfer. We explained that the fourth term is important for class-conditional alignment. We solved Eq. (3) without considering the fourth term to study its effect. We have presented the Umap visualization of the datasets in the embedding space in Figure 5. We observe that as expected the embedding is discriminative for EO dataset and predicted labels are close to the real data labels as the classes are separable. However, despite following a similar marginal distribution in the embedding space, the formed SAR clusters are not class-specific. We can see that in each cluster, we have data points from both classes and as a result the SAR classification rate is poor. This result demonstrates that all the terms in Eq. (3) are important for the success of our algorithm.

6. Conclusions

In this paper, we developed a method to classify SAR images using deep neural network when only few-labeled samples are accessible. Our idea is tackle the problem of label scarcity through transferring knowledge from EO domain, where sufficient labeled data is accessible. We map the data samples from both domains to a shared embedding space, such that the data distributions are matched, We demonstrated that our approach is effective and competitive. Future work extends to online domain adaptation scenarios where the goal is to use sequential training, rather joint training which can improve learning speed and the need to store EO data. Choosing the proper source domain is another area which requires more investigation.

References

[1] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. 3

[2] M. Carriere, M. Cuturi, and S. Oudot. Sliced wasserstein kernel for persistence diagrams. *arXiv preprint arXiv:1706.03358*, 2017. 3

[3] S. Chen, H. Wang, F. Xu, and Y. Jin. Target classification using the deep convolutional networks for sar images. *IEEE Trans. on Geo. and Remote Sens.*, 54(8):4806–4817, 2016. 2

[4] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE TPAMI*, 39(9):1853–1865, 2017. 3, 5

[5] B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. *arXiv preprint arXiv:1803.10081*, 2018. 5

[6] H. Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009. 3

[7] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006. 3

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009. 1

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3, 5

[10] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. 2009. 2, 3

[11] R. Hammell. Ships in satellite imagery, 2017. data retrieved from Kaggle, https://www.kaggle.com/rhammell/ships-in-satellite-imagery. 6

[12] Z. Huang, Z. Pan, and B. Lei. Transfer learning with deep convolutional neural network for sar target classification with limited labeled data. *Remote Sensing*, 9(9):907, 2017. 2

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015. 2

[15] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. 5

[16] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced-wasserstein auto-encoders. *International Conference on Learning Representation (ICLR)*, 2019. 3

[17] S. Kolouri, G. K. Rohde, and H. Hoffman. Sliced Wasserstein distance for learning Gaussian mixture models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–, 2018. 2, 5

[18] V. Koo, Y. Chan, G. Vetharatnam, M. Y. Chua, C. Lim, C. Lim, C. Thum, T. Lim, Z. bin Ahmad, K. Mahmood, et al. A new unmanned aerial vehicle synthetic aperture radar for environmental monitoring. *Progress In Electromagnetics Research*, 122:245–268, 2012. 1

[19] H. Lang, S. Wu, and Y. Xu. Ship classification in sar images improved by ais knowledge transfer. *IEEE Geoscience and Remote Sensing Letters*, 15(3):439–443, 2018. 2

[20] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1417, 2014. 5

[21] H. Maitre. *Processing of Synthetic Aperture Radar (SAR) Images*. Wiley, 2010. 1

[22] D. Malmgren-Hansen, A. Kusk, J. Dall, A. Nielsen, R. Engholm, and H. Skriver. Improving sar automatic target recognition models with transfer learning from simulated data. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1484–1488, 2017. 1, 2

[23] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 7

[24] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 6670–6680, 2017. 2

[25] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 2

[26] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011. 2, 3

[27] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011. 5

[28] A. Redko, I.and Habrard and M. Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer, 2017. 2, 5

[29] M. Rostami, D. Huber, and T.-C. Lu. A crowdsourcing triage algorithm for geopolitical event forecasting. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 377–381. ACM, 2018. 1

[30] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 3

[31] C. Schwegmann, W. Kleynhans, B. Salmon, L. Mdakane, and R. Meyer. Very deep learning for ship discrimination in synthetic aperture radar imagery. In *IEEE International Geo. and Remote Sensing Symposium*, pages 104–107, 2016. 1, 6

[32] R. Shang, J. Wang, L. Jiao, R. Stolkin, B. Hou, and Y. Li. Sar targets classification based on deep memory convolution neural networks and transfer parameters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2834–2846, 2018. 2

[33] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 2, 3, 5

[34] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang. Sar target detection based on ssd with data augmentation and transfer learning. *IEEE Geoscience and Remote Sensing Letters*, 2018. 2

[35] J. Zhang, D., W. Heng, K. Ren, and J. Song. Transfer learning with convolutional neural networks for sar ship recognition. In *IOP Conference Series: Materials Science and Engineering*, volume 322, page 072001. IOP Publishing, 2018. 2