Cross-Modal Knowledge Transfer Without Task-Relevant Source Data

Sk Miraj Ahmed[1], Suhas Lohit[2], Kuan-Chuan Peng[2], Michael J. Jones[2], and Amit K. Roy-Chowdhury[1]

[1] University of California, Riverside, CA 92507, USA
{sahme047@,amitrc@ece.}ucr.edu
[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
https://www.merl.com {slohit,kpeng,mjones}@merl.com

Abstract. Cost-effective depth and infrared sensors as alternatives to usual RGB sensors are now a reality, and have some advantages over RGB in domains like autonomous navigation and remote sensing. As such, building computer vision and deep learning systems for depth and infrared data are crucial. However, large labeled datasets for these modalities are still lacking. In such cases, transferring knowledge from a neural network trained on a well-labeled large dataset in the source modality (RGB) to a neural network that works on a target modality (depth, infrared, etc.) is of great value. For reasons like memory and privacy, it may not be possible to access the source data, and knowledge transfer needs to work with only the source models. We describe an effective solution, SOCKET: SOurce-free Cross-modal KnowledgE Transfer for this challenging task of transferring knowledge from one source modality to a different target modality without access to task-relevant source data. The framework reduces the modality gap using paired task-irrelevant data, as well as by matching the mean and variance of the target features with the batch-norm statistics that are present in the source models. We show through extensive experiments that our method significantly outperforms existing source-free methods for classification tasks which do not account for the modality gap.

1 Introduction

Depth sensors like Kinect and RealSense, LIDAR for measuring point clouds directly, or high resolution infra-red sensors such as from FLIR, allow for expanding the range of applications of computer vision compared to using only visible wavelengths. Sensing depth directly can provide an approximate three-dimensional picture of the scene and thus improve the performance of applications like autonomous navigation, while sensing in the infra-red wavelengths can allow for easier pedestrian detection or better object detection in adverse atmospheric conditions like rain, fog, and smoke. These are just a few examples.

Building computer vision applications using the now-straightforward supervised deep learning approach for modalities like depth and infrared needs large

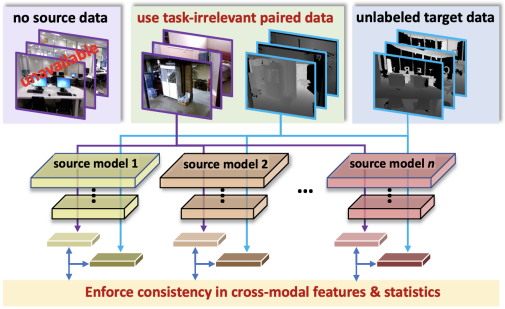

Fig. 1. SOCKET: We describe the problem of single/multi-source cross-modality knowledge transfer using no data used to train the source models. To effectively perform knowledge transfer, we minimize the modality gap by enforcing consistency of cross modal features on **task-irrelevant** paired data in feature space, and by **matching the distributions** of the unlabeled task-relevant features and the source features

amounts of diverse labeled data. However, such large and diverse datasets do not exist for these modalities and the cost of building such datasets can be prohibitively high. In such cases, researchers have developed methods like knowledge distillation to transfer the knowledge from a model trained on a modality like RGB, where large amounts of labeled data are available, to the modality of interest like depth [1].

In contrast to prior work, we tackle a novel and challenging problem in the context of cross-modal knowledge transfer. We assume that we have access only to (a) the source models trained for the task of interest (TOI), and (b) unlabeled data in the target modality where we need to construct a model for the same TOI. The key aspect is that we assume we have **no access to any data in the source modality** for TOI. Such a problem setup is important in cases where memory and privacy considerations do not allow for sharing the training data from the source modality; only the trained models can be shared [2–5].

We develop **SOCKET**: **SO**urce-free **C**ross-modal **K**nowledg**E** **T**ransfer as an effective solution to this problem for bridging the gap between the source and target modalities. To this end, we show that employing an external dataset of source-target modality pairs, which are not relevant to TOI – which we call Task-Irrelevant (TI) data – can help in learning an effective target model by bringing the features of the two modalities closer. In addition to using TI data, we encourage matching the statistics of the features of the unlabeled target data – which are Task-Relevant (TR) by definition – with the statistics of the source data which are available to us from the normalization layers that are present in the trained source model.

We provide important empirical evidence showing that the modality-shift from a source modality like RGB to a target modality like depth can be much more challenging than a domain shift from one RGB dataset to another. This shows that the proposed framework is necessary to help minimize the modality

gap, so as to make the knowledge transfer more effective. Based on the above ideas, we show that we can improve on existing state-of-the-art methods which were devised only for cross-domain setting in the same modality. We summarize our main contributions below:

1. We formulate a novel problem for knowledge transfer from a model trained for a source modality to a different target modality without any access to task-relevant source data and when the target data is unlabeled.
2. In order to bridge the gap between modalities, we propose a novel framework, SOCKET, for cross-modal knowledge transfer without access to source data (a) using an external task-irrelevant paired dataset, and (b) by matching the moments obtained from the normalization layers in the source models with the moments computed on the unlabeled target data.
3. Extensive experiments on multiple datasets – both for knowledge transfer from RGB to depth, and from RGB to IR, and both for single-source and multi-source cases – show that SOCKET is useful in reducing the modality gap in the feature space and produces significantly better performance (improvement of as high as 12% for some cases) over the existing source-free domain adaptation baselines which do not account for the modality difference between the source and target modalities.
4. We also show empirically that, for the datasets of interest, the problem of knowledge transfer between modalities like RGB and depth is harder than domain shifts in the same modality such as sensor changes and viewpoint shifts, considered previously in literature.

2 Related work

Cross-modal distillation methods. Cross-modal knowledge distillation (CMKD) methods aim to learn representations for a modality which does not have a large amount of labeled data from a large labeled dataset of another modality [1]. These methods have been used for a variety of practical computer vision and learning tasks [6–9]. Most of these works assume access to task-relevant paired data across modalities [1, 8, 10, 11]. A recent line of work relaxed this assumption in the context of domain generalization, where one does not have access to the Task-Relevant paired data on the target domain but has access to them for the source domain [12]. There also exist some works regarding domain translation across modalities for better classification of indoor scenes [13–15]. However these methods consider UDA across domains, where the target domain has unlabeled RGB-D pairs instead of a single modality. All of the above works either utilize the Task-Relevant paired data for cross modal knowledge transfer [1], or consider cross modal paired data as a domain [12, 13]. There are also works in zero-shot domain adaptation that utilize external task-irrelevant paired data [16] but need access to the source data. Our work takes steps to allow for different source and target modalities, and can perform effective knowledge transfer without access to the TR paired data between source and target.

Table 1. We compare the proposed work SOCKET with existing problem settings in literature for knowledge transfer across different domains and modalities. The competitive settings described in this table are: (1) UDA (Unsupervised Domain Adaptation), DT (Domain Translation) [13–15, 17–20] [$\mathcal{C}_1$], (2) MSDA (Multi-source domain adaptation) [21] [$\mathcal{C}_2$], (3) SFDA (Source free single source DA) [3, 22–26] [$\mathcal{C}_3$], (4) MSFDA (Source free multi-source DA) [4] [$\mathcal{C}_4$], (5) CMKD (Cross modal knowledge distillation) [1, 6–8] [$\mathcal{C}_5$], and (6) ZDDA (Zero shot DA) [16] [$\mathcal{C}_6$], respectively. We group citations into [$\mathcal{C}_1$] to [$\mathcal{C}_6$] based on problem settings. Only SOCKET allows cross-modal knowledge transfer from multiple sources without any access to relevant source training data for an unlabeled target dataset of a different modality

Property \ Problem setting	UDA+DT [$\mathcal{C}_1$]	MSDA [$\mathcal{C}_2$]	SFDA [$\mathcal{C}_3$]	MSFDA [$\mathcal{C}_4$]	CMKD [$\mathcal{C}_5$]	ZDDA [$\mathcal{C}_6$]	SOCKET
Multiple sources	✗	✓	✗	✗	✗	✗	✓
No source data	✗	✗	✓	✓	✗	✗	✓
Unlabeled target data	✓	✓	✗	✓	✗	✗	✓
Different target modality	✗	✗	✗	✗	✓	✓	✓
Usage of Task-Irrelevant Data	✗	✗	✗	✗	✗	✓	✓

Unsupervised domain adaptation methods without source data. Most UDA methods that have been used for a wide variety of tasks [17–20] need access to the source data while adapting to a new target domain [21, 27]. To combat the storage or privacy issue regarding the source data, a new line of work named Hypothesis Transfer Learning (HTL) [2, 5] has emerged recently, where one has access only to the trained source model instead of the source data [3, 4]. Here, people have explored adapting target domain data, which has limited labels [2] or no labels at all [3] in the presence of both single source [3, 22, 23] or multiple source models [4]. [3, 26] adapts a single source model to an unlabeled target domain via information maximization and an iterative self-supervised pseudo-label based cross entropy loss. [22] ensured that the adapted source model performs well, both on source and target domains, while [23] proposed a source free domain adaptation (SFDA) method by encouraging label consistency among local target features. [24] proposed to add an extra classifier for refinement of the source decision boundary, while [25] proposed a more robust adaptation method which works well in the presence of noisy pseudo-labels. The authors in [4] proposed fusion of multiple source models with appropriate weights so as to minimize the effect of negative transfer, which we refer to as multiple source free domain adaptation (MSFDA) in Table 1. Both these methods do not work well in a regime where the unlabeled target set is from a different modality than the source. We solve this problem by modality gap reduction via feature matching of the task-irrelevant external data, as well as data statistics matching between the source and target modalities.

Table 1 summarizes the related work and compares them with SOCKET.

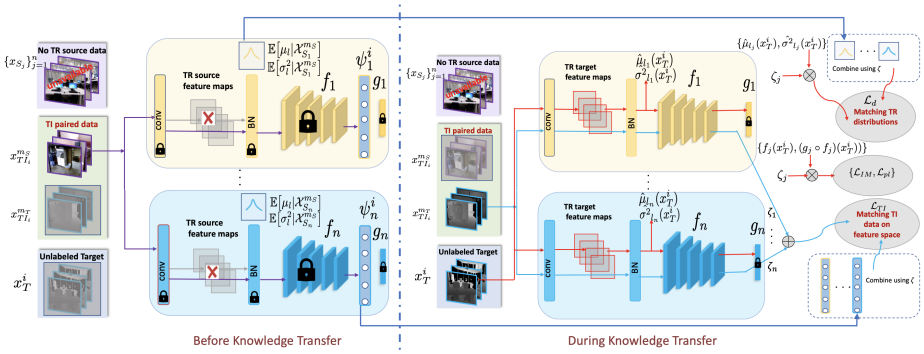

Fig. 2. SOCKET description: Our framework can be split into two parts: (i) Before Knowledge Transfer (left): We freeze the source models and pass the task-irrelevant (TI) source data through the source feature encoders to extract the TI source features. As task-relevant (TR) source feature maps are not available, we extract the stored moments of its distribution from the BN layers. (ii) During Knowledge Transfer (right): We freeze only the classification layers and feed the TI and unlabeled TR target data through the models to get batch-wise TI target features and the TR target moments, respectively, which we match with pre-extracted source features and moments to jointly train all the feature encoders along with the mixing weights, ζ_k's. The final target model is the optimal linear combination of the updated source models

3 Problem setup and notation

We address the problem of source-data free cross-modality knowledge transfer by devising specialized loss functions that help reduce the gap between source and target modality features. We focus on the task of classification where both the source and target data belong to the same N classes. Let us consider that we have n source models of the same modality (*e.g.*, RGB). We denote the trained source classifiers as $\{\mathcal{F}_{S_k}^{m_S}\}_{k=1}^{n}$, where S_k denotes the k-th source model and m_S represents the modality on which the source models were trained. The source models are denoted as $\mathcal{F}_{S_k}^{m_S}$ which are trained models that map images from the source modality distribution $\mathcal{X}_{S_k}^{m_S}$ to probability distribution over the classes. $\{x_{S_k}^i, y_{S_k}^i\}_{i=1}^{n_k} \sim \mathcal{X}_{S_k}^{m_S}$ are the data on which the k-th source model was trained, n_k being the number of training data points corresponding to the k-th source. In our problem setting, at the time of knowledge transfer to the target modality, the source data are unavailable for all the sources.

We also have access to an unlabeled dataset in the target modality $\{x_T^i\}_{i=1}^{n_T} \sim \mathcal{X}_T^{m_T}$, where n_T is the number of target samples. Note that the target modality, m_T, is different from the source modality. Traditional source free UDA methods try to mitigate domain shift by adapting the source models to unlabeled target data that belong to the same modality [3, 4]. As we will show, applying these methods directly to the cross-modal setting results in poor performance. Hence, we propose to solve this problem using two novel losses as regularization terms which minimize the modality gap between source and target modalities. Our

goal is to learn a target classifier $\mathcal{F}_T^{m_T}$, that adapts well on a target distribution obtained from a different sensor modality (*e.g.*, depth or NIR).

To train $\mathcal{F}_T^{m_T}$, we employ (a) methods that enable learning feature embeddings for the target modality that closely match with the source modality embeddings, which we group under modality-specific losses, since it bridges the gap between two different modalities; (b) modality-agnostic loss terms which operate only on the unlabeled target data and do not take into account shift in modality.

We split each of the source models into two blocks – *feature encoder* and *classifier*. For the k-th source model, we denote these blocks as f_k and g_k, respectively. The function $f_k : \mathbb{R}^{H\times W} \rightarrow \mathbb{R}^{\eta}$ maps the input image to an η dimensional feature vector and $g_k : \mathbb{R}^{\eta} \rightarrow \mathbb{R}^{N}$ maps those features to the probability distributions over the N classes, the maximum of which is treated as the classifier prediction. We can thus write $\mathcal{F}_{S_k}^{m_S} = g_k \circ f_k$, where "$\circ$" is function composition. Since the classifier layer g_k contains the information about unseen k-th source domain distribution, following the protocol of [4], we freeze all the g_k's and train the target specific feature encoders by optimizing over all f_k's.

4 Cross-Modal Feature Alignment

Traditional source free UDA methods [3, 4] use domain specific but modality-agnostic losses which do not help in reducing the feature distance between the source and target modalities. In order to train the target model, $\mathcal{F}_T^{m_T}$, with reduced modality-gap, we propose SOCKET, which uses *task-irrelevant feature matching* and *task-relevant distribution matching* which are described next.

4.1 Task-irrelevant feature matching

Capturing the mapping between two modalities effectively requires lots of paired data from both modalities [28]. For our task of interest, we do not have task relevant (TR) data on the source side. As a result, it is not possible to match the target modality with the source modality by using the data from task relevant classes directly. Hence, we propose to use **Task-Irrelevant (TI) paired data** from both modalities to reduce modality gap. TI data contain only classes that are completely **disjoint** from the TR classes and can be from any external dataset. For modalities like RGB-depth and RGB-IR, we can access a large amount of paired TI data that contain classes with no privacy concerns, which are available in public datasets or can be collected using multi-modal sensors. Moreover there are many real world applications where pairwise TI data can be collected and used beyond RGB-D or RGB-IR, such as autonomous driving, adpatation of LiDAR data, medical applications [29]. We denote paired TI data as $\{x_{TI_i}^{m_S}, x_{TI_i}^{m_T}\}_{i=1}^{n_{TI}}$, where $x_{TI_i}^{m_S}$ is the i-th TI data point from source modality and $x_{TI_i}^{m_T}$ is its paired counterpart from the target modality, n_{TI} the total number of pairs. We compute our proposed loss $\mathcal{L}_{TI}$ using TI data as follows:

Step 1: We feed source modality images of the TI dataset through each of the source models to pre-compute features that are good representations of modality

m_S. We denote the i-th TI source feature extracted from source j as ψ_j^i:

$$\psi_j^i = f_j(x_{TI_i}^{m_S}). \tag{1}$$

Step 2: During the knowledge transfer phase, we feed the target modality images of the TI dataset which are encouraged to match the corresponding pre-extracted source modality features. We do so by minimizing $\mathcal{L}_{TI}$ defined below with respect to the parameters in the feature encoders for the target modality:

$$\mathcal{L}_{TI} = \sum_{i=1}^{n_{TI}} \sum_{j=1}^{n} \left\| \zeta_j (\psi_j^i - f_j(x_{TI_i}^{m_T})) \right\|^2. \tag{2}$$

4.2 Task-relevant distribution matching

In the task-irrelevant feature matching, we match the TI features of two modalities in the feature space. Even if this captures some class independent cross modal mapping between source and target modalities, it has no information about the *TR-class conditional cross modal mapping*. By this term we refer to the cross modal relationship between source and target, given the relevant classes. Assuming that the marginal distribution of the source features across the batches can be modeled as Gaussian, such feature statistics can be fully characterized by its mean and variance. We propose to match the feature statistics across the source and target, to reduce the modality gap further.

It might seem as though some amount of source data would be required to estimate the batch-wise mean and and variance of its feature map, but the running average statistics stored in the conventional BatchNorm (BN) layers are good enough to serve our purpose. The BN layers normalize the feature maps during the course of training to mitigate the covariate shifts [30, 31]. As a result it is able to capture the channel-wise feature statistics cumulatively over all the batches, which gives rise to a rough estimate of the expected mean and variance of the batch-wise feature map, at the end of training. Let us consider that the BN layer corresponding to the l-th convolution layer ($\mathcal{B}_l$) has r_l nodes and there exist b number of such layers per source model. Then we refer to the expected batch-wise mean and variance of the l-th convolution layer of the k-th source model as $\mathbb{E}\big[\mu_l | \mathcal{X}_{S_k}^{m_S}\big] \in \mathbb{R}^{r_l}$ and $\mathbb{E}\big[\sigma_l^2 | \mathcal{X}_{S_k}^{m_S}\big] \in \mathbb{R}^{r_l}$.

Prior to the start of the knowledge transfer phase, we pre-extract the information about the source feature statistics from all of the pre-trained source models. During the knowledge transfer phase, for each iteration we calculate the batch-wise mean and variance of the feature map of target data from all the source models, linearly combine them according to the weights ζ_i and minimize the distance of this weighted combination with the weighted combination of the pre-computed source feature statistics. We calculate this loss $\mathcal{L}_d$ given by

$$\mathcal{L}_d = \sum_{l=1}^{b} \left(\left\| \sum_{j=1}^{n} \zeta_j \mathbb{E}\big[\mu_l | \mathcal{X}_{S_j}^{m_S}\big] - \sum_{j=1}^{n} \zeta_j \hat{\mu}_{l_j} \right\| + \left\| \sum_{j=1}^{n} \zeta_j \mathbb{E}\big[\sigma_l^2 | \mathcal{X}_{S_j}^{m_S}\big] - \sum_{j=1}^{n} \zeta_j \hat{\sigma^2}_{l_j} \right\| \right), \tag{3}$$

where $\mathbb{E}\big[\mu_l|\mathcal{X}_{S_j}^{m_S}\big]$ and $\mathbb{E}\big[\sigma_l^2|\mathcal{X}_{S_j}^{m_S}\big]$ are the running mean and variance of the batchnorm layer corresponding to the l-th convolution layer of source j, which we refer as $\mathcal{B}_l^j$, and $\hat{\mu}_{l_j} = \frac{1}{n_T}\sum_{k=1}^{n_T}\mathcal{B}_l^j(x_T^k)$ and $\hat{\sigma^2}_{l_j} = \frac{1}{n_T}\sum_{k=1}^{n_T}(\mathcal{B}_l^j(x_T^k) - \hat{\mu}_{l_j})^2$ denote the mean and variance of the target output from the same batchnorm layer. The losses $\mathcal{L}_{TI}$ and $\mathcal{L}_d$ minimize the modality gap between source and target. We name the combination of these two losses as *Modality Specific Loss* $\mathcal{L}_{ms} = \lambda_{TI}\mathcal{L}_{TI} + \lambda_d\mathcal{L}_d$, where λ_{TI} and λ_d are regularization hyper-parameters.

4.3 Overall optimization

The two proposed methods above help to reduce the modality gap between source and target without accessing task-relevant source data. In addition to them, we employ the unlabeled target data directly for knowledge transfer. Specifically, we perform *information maximization* along with minimization of a self-supervised *pseudo-label loss*, which have shown promising results in source-free UDA [3, 4] where the source and target modalities are the same.

Information Maximization (IM): IM is essentially the task of performing maximization of the mutual information between distribution of the target data and its labels predicted by the source models. This mutual information is a combination of a conditional and a marginal entropy of the target label distribution.

Motivated by [4], we calculate the *conditional entropy* $\mathcal{L}_{ent}$ and the marginal entropy termed as *diversity* $\mathcal{L}_{div}$ as follows:

$$\mathcal{L}_{ent} = -\frac{1}{n_T}\Big[\sum_{i=1}^{n_T}(\mathcal{F}_T^{m_T}(x_T^i))\log(\mathcal{F}_T^{m_T}(x_T^i))\Big], \mathcal{L}_{div} = -\sum_{j=1}^{N}\bar{p}_j\log\bar{p}_j, \quad (4)$$

where $\mathcal{F}_T^{m_T}(x_T^i) = \sum_{k=1}^{n}\zeta_k\mathcal{F}_{S_k}^{m_S}(x_T^i)$, ζ_k is the weight assigned to the k-th source such that $\zeta_k \geq 0$, $\sum_{k=1}^{n}\zeta_k = 1$ and $\bar{p} = \frac{1}{n_T}\sum_{i=1}^{n_T}\Big[\mathcal{F}_T^{m_T}(x_T^i)\Big] \in \mathbb{R}^N$ is the empirical label distribution. The *mutual information* is calculated as $\mathcal{L}_{IM} = \mathcal{L}_{div} - \mathcal{L}_{ent}$. Maximization of $\mathcal{L}_{IM}$ (or minimization of $-\mathcal{L}_{IM}$) ensures the target labels, as predicted by the sources, more confident and diverse in nature.

Pseudo-label loss: Maximizing $\mathcal{L}_{IM}$ helps to obtain labels that are more confident in prediction and globally diverse. However, that does not prevent mislabeling (*i.e.*, assigning wrong labels to the inputs), which leads to *confirmation bias* [32]. To alleviate this problem, we adopt a self supervised pseudo-label based cross entropy loss, inspired by [3, 4] (see the supplement for the exact details about computing the self-supervised pseudo-labels.) After calculating pseudo-labels we compute the *pseudo-label cross entropy* loss $\mathcal{L}_{pl}$ as follows:

$$\mathcal{L}_{pl} = -\frac{1}{n_T}\sum_{i=1}^{n_T}\sum_{k=1}^{K}\mathbf{1}\{\hat{y}_T^i = k\}\log\big[\mathcal{F}_T^{m_T}(x_T^i)\big]_k, \quad (5)$$

where $\hat{y}_T^i$ is the pseudo-label for the i-th target data point and $\mathbf{1}\{.\}$ is an indicator function that gives value 1 when the argument is true. Our final loss is the

combination of the above two losses. We call this combination *modality agnostic loss* $\mathcal{L}_{ma}$, which is expressed as $\mathcal{L}_{ma} = -\mathcal{L}_{IM} + \lambda_{pl}\mathcal{L}_{pl}$.

We calculate the overall objective function as the sum of *modality agnostic* and *modality specific* losses and optimize Eq. (6) using Algorithm 1.

$$\underset{\{f_j\}_{j=1}^{n}, \zeta}{\text{minimize}} \quad \mathcal{L}_{ma} + \mathcal{L}_{ms} \quad \text{s.t.} \sum_{k=1}^{n} \zeta_k = 1, \zeta_k \geq 0 \tag{6}$$

Algorithm 1: Algorithm to Solve Eq. (6)

Input: n source models trained on modality m_S $\{\mathcal{F}_{S_k}^{m_S}\}_{k=1}^{n} = \{g_k \circ f_k\}_{k=1}^{n}$, unlabeled target data $\{x_T^i\}_{i=1}^{n_T}$ from modality m_T, TI cross modal pairs $\{x_{TI_i}^{m_S}, x_{TI_i}^{m_T}\}_{i=1}^{n_{TI}}$, mixing weights $\{\zeta_k\}_{k=1}^{n}$, max number of epochs E, regularization parameters λ_{TI}, λ_d, number of batches B

Output: Optimal adapted feature enocoders $\{f_k^\star\}_{k=1}^{n}$, mixing weights $\{\zeta_k^*\}_{k=1}^{n}$

Initialization: Freeze final classification layers $\{g_k\}_{k=1}^{n}$, set $\zeta_k = \frac{1}{n}$ for all k

Calculate $\{\psi_j^i\}_{j=1}^{n}$ $\forall i \in [1, 2 \ldots, n_{TI}]$ using Eq. (1)

Retrieve $\mathbb{E}\big[\mu_l | \mathcal{X}_{S_j}\big]$ and $\mathbb{E}\big[\sigma_l^2 | \mathcal{X}_{S_j}\big]$ for all j and l as in Section 4.2

Knowledge Transfer Phase:

for *epoch* $= 1$ ***to*** E **do**

  for *iteration* $= 1$ ***to*** B **do**

    Sample a mini batch of target data and feed it through each of the source models

    Calculate loss terms in Eq. (2), (3), (4), and (5)

    Compute overall objective from Eq. (6)

    Update parameters in $\{f_j\}_{j=1}^{n}$ and $\{\zeta_k\}_{k=1}^{n}$ by optimizing (6)

    Make ζ non-negative by setting $\zeta_k := 1/(1 + e^{-\zeta_k})$

    Normalize ζ by setting $\zeta_k := \zeta_k / \sum_{i=1}^{n} \zeta_i$

  end

end

Final target model $\mathcal{F}_T^{m_T} = \sum_{k=1}^{n} \zeta_k^\star (g_k \circ f_k^\star)$

5 Experiments

We first describe the datasets, baselines and experimental details we employ. Next, we show results of single and multi-source cross modal transfer which show the efficacy of our method. In Section 5.3 we demonstrate experimentally why source free cross modal is a much harder problem compared to cross domain knowledge transfer. We conclude this section by performing analysis on different hyperparameters.

5.1 Datasets, baselines and experimental details

Datasets: To show the efficacy of our method we extensively test on publicly available cross-modal datasets. We show results on two RGB-D (RGB and

Depth) datasets – SUN RGB-D [33] and DIML RGB+D [34], and the RGB-NIR Scene (RGB and Near Infrared) dataset [35]. We summarize the statistics of the datasets in Table 2. In the supplement, we provide examples from each dataset and the list of classes which we use as TI and TR data in our experiments.

1. SUN RGB-D: A scene understanding benchmark dataset which contains 10335 RGB-D image pairs of indoor scenes. The dataset has images acquired from four different sensors named *Kinect version1 (kv1)*, *Kinect version2 (kv2)*, *Intel RealSense* and *Asus Xtion*. We treat these four sensors as four different domains. Out of total 45 classes, 17 common classes are treated as TR classes and the remaining 28 classes as TI classes. To train four source models, one for each domain, we use the RGB images from the TR classes, specific to that particular domain. We treat the TR depth images from each of the domains as the target modality data.
2. DIML RGB+D: This dataset consists of more than 200 indoor/outdoor scenes. We use the smaller sample dataset instead of the full dataset, which has 1500/500 RGB-D pairs for training/testing distributed among 18 scene classes. We split the training pairs into RGB and depth, and treat those two as source and target, respectively. The synchronized RGB-D frames are captured using Kinect v2 and Zed stereo camera [36–38].
3. RGB-NIR Scene: This dataset consists of 477 images from 9 scene categories captured in RGB and Near-infrared (NIR). The images were captured using separate exposures from modified SLR cameras, using visible and NIR [35]. We perform single source knowledge transfer from RGB to NIR and vice versa for this dataset. For all the datasets, TR/TI split is done according to Table 2.

Baseline Methods: The problem statement we focus on in this paper is new and has not been considered in literature before. As such, there is no direct baseline for our method. However, the closest related works are source free cross domain knowledge transfer methods that operates under both single and multi-source cases [3, 4, 22–26]. SHOT [3] and DECISION [4] are the best-known works on single source and multi-source SFDA and we compare against only these two methods. Unlike SOCKET, neither of these baselines employ strategies to overcome modality differences and use only the modality-agnostic loss $\mathcal{L}_{ma}$ for training the target models. Using scene classification as the task of interest, we will show that SOCKET outperforms these baselines for cross-modal knowledge transfer with no access to task-relevant source data. We provide details about the network architecture in the supplement. We note that there a few more recent works [22–26] which have shown small improvements over SHOT, and are orthogonal to the ideas in this paper. Incorporating these improvements for SOCKET as well can be interesting and consider this future work.

Performing knowledge transfer: Recall that we initialize the target models with the source weights and the classifier layers are frozen. The weights in the feature encoders and source mixing weight parameters (ζ_k's) in the case of multi-source are the optimization parameters. The values of various parameters like the learning rate are given in the supplement.

Table 2. Datasets statistics

	SUN-RGBD [33]	RGB-NIR Scene [35]	DIML [34]
Number of domains	4	1	1
Domain names	kv1,kv2,Realsense,Xtion	N/A	N/A
# of TR images for source training	1264,1234,238,2512	204	527
# of TR unlabeled images	1264,1234,238,2512	204	527
Number of TI paired images	1709	153	1088
Number of TR & TI classes	17 & 28	6 & 3	6 & 12
Modalities	RGB-D	RGB-NIR	RGB-D

Table 3. Results on the SUN RGB-D dataset [33] for the task of single-source cross-modal knowledge transfer from RGB to depth modality without access to task relevant source data. The rows represent RGB domains on which the source models are trained. The columns represent the knowledge transfer results on the depth domains for three methods – *Unadapted* shows results with unadapted source, SHOT[3] and SOCKET.

Source RGB \ Target depth	Kinect v1			Kinect v2			Realsense			Xtion		
	Unadapted	SHOT	SOCKET	Unadapted	SHOT	SOCKET	Unadapted	SHOT	SOCKET	Unadapted	SHOT	SOCKET
Kinect v1	14.8	16.7	**25.3**	14.6	20.3	**23.6**	9.0	11.9	**13.4**	7.1	15.3	**18.1**
Kinect v2	4.0	12.8	**13.6**	17.0	29.4	**35.2**	10.8	19.3	**22.8**	**10.6**	7.0	8.3
Realsense	2.0	7.9	**20.3**	7.1	18.4	**23.5**	14.7	27.4	**30.0**	5.1	9.5	**11.8**
Xtion	0.7	9.5	**14.2**	6.0	20.2	**24.2**	9.0	21.8	**23.5**	8.1	13.2	**22.2**
Average	5.4	11.7	**18.4**	11.2	22.1	**26.6**	10.9	20.1	**22.4**	7.7	11.3	**15.1**

λ_{pl} is set as 0.3 for all the experiments following [4]. For the regularization parameters λ_{TI} and λ_d of *modality specific* losses, we set them to be equal. We empirically choose those parameters in such a way so as to balance it with the *modality agnostic* losses such that no loss component overpowers the other by a large margin. Empirically we found that a range of $(0.1, 0.5)$ works best. All of the values in this range outperform the baselines and we report the best accuracies amongst those. For images from the modalities other than RGB, which are depth and NIR, we repeat the single-channel images into three-channel images, to be able to feed it through the feature encoders which are initialized from the source models trained on RGB images. We use a batch size of 32 for all of our experiments. We run our method 3 times for all experiments with 3 random seeds in PyTorch [39] and report the average accuracies over those.

5.2 Main results

Results on the SUN RGB-D dataset [33]: Our method is general enough to deal with any number of sources and we demonstrate both single and multi-source knowledge transfer. In Table 3, we show single source RGB to depth results for all of the four domains. Treating the unlabeled depth data of each domain as target, we adapt these using source models trained on RGB data from each of the four domains. It is easily evident from Table 3, that for the target domains Kinect V1, Kinect V2, Realsense and Xtion, SOCKET consistently outperforms the baseline by a good margin of 6.7%, 4.5%, 2.3%, and 3.8%, respectively, thus proving the efficacy of SOCKET in a source-free cross modal

Table 4. Results on the SUN RGB-D dataset [33] for the task of multiple cross-modal knowledge transfer from RGB to depth modality without access to task relevant source data. The rows show the six combinations of two trained source models on RGB data from four different domains. The columns represent the knowledge transfer results on the domain specific depth data for *DECISION*[4], the current SOTA for multiple source adaptation without source data, and SOCKET

Source RGB \ Target depth	Kinect v1		Kinect v2		Realsense		Xtion	
	DECISION	SOCKET	DECISION	SOCKET	DECISION	SOCKET	DECISION	SOCKET
Kinect v1 + Kinect v2	17.9	**19.5**	34.2	**36.6**	18.8	**19.8**	14.6	**18.0**
Kinect v1 + Realsense	12.6	**18.0**	23.3	**26.8**	24.3	**24.7**	10.9	**12.2**
Kinect v1 + Xtion	11.7	**23.9**	29.6	**35.7**	20.3	**21.1**	16.7	**20.0**
Kinect v2 + Realsense	7.4	**11.7**	22.7	**33.1**	28.4	**29.4**	6.9	**9.1**
Kinect v2 + Xtion	14.8	**16.2**	27.0	**31.0**	**25.4**	25.0	11.6	**18.3**
Realsense + Xtion	8.3	**10.7**	23.1	**25.2**	30.1	**31.5**	9.5	**10.8**
Average	12.1	**16.6**	26.7	**31.4**	24.6	**25.3**	11.7	**14.7**

Table 5. Classification accuracy (%) on DIML dataset with different TI data

TI data	Unadapted N/A	SHOT [3] N/A	SOCKET DIML RGB+D	SOCKET SUN RGB-D
RGB→Depth	26.9	41.4	**46.1**	**53.2**

setting. In some of the cases SOCKET outperforms the baseline by a very large margin, as high as 12.4% (Realsense-RGB to Kinect V1-depth). We show two-source RGB to depth adaptation results in Table 4. For four domains we get six two-source combinations, each of which is used for adaptation to depth data from all four domains. We see that in this case also, on average SOCKET outperforms the baseline for all four target domains by good margins. SOCKET shows good improvement for some individual cases like (Kinect v1 + Xtion)-RGB to Kinect v1 depth – improvement of 12.2% – and (Kinect v2 + Realsense)-RGB to Kinect v2 depth –improvement of 10.4%.

Results on the DIML RGB+D dataset [34]: We performed a single source adaptation experiment (Table 5) by restructuring the dataset according to Table 2. In Table 5, we use the TI data from both the DIML RGB+D as well as SUN RGB-D datasets in two separate columns, where the TI data of SUN RGB+D is the same that have been used for experiments related to the SUN RGB-D dataset. By doing so, we show that SOCKET can perform well even with TI data from a completely different dataset, and find that SOCKET has a gain of 4.7% and 11.8% over baseline for these two TI data settings, respectively.

Results on the RGB-NIR scene dataset [35]: We now show that SOCKET also outperforms baslines when the modalitiies are RGB and NIR using the RGB-NIR dataset. We follow the splits described in Table 2. We do experiments on both RGB to NIR and vice versa. The results are given in Table 6. For RGB to NIR transfer, SOCKET shows 3.5% improvement, while for NIR to RGB transfer, it shows 0.5% improvement over the competing method.

Table 6. Results on RGB-NIR dataset [35] for the task of single-source cross-modal knowledge transfer from RGB to NIR and vice versa without task-relevant source data

Setting \ Method	Unadapted	SHOT [3]	SOCKET
RGB → NIR	84.8	86.7	**90.2**
NIR → RGB	65.2	92.2	**92.7**

Table 7. Cross modal vs cross domain knowledge transfer for SUN RGB-D dataset scene classification using SHOT[3]: (1) The first columm shows the accuracies for RGB to depth transfer within the same domain. (2) The second column is generated by transferring knowledge from one RGB domain to other three RGB domains taking the average of the accuracies

Source	Cross-Modal	Cross-Domain
Kinect v1	16.7	24.5
Kinect v2	29.4	39.6
Realsense	27.4	29.7
Xtion	13.2	43.1
Average	21.7	34.2

5.3 Cross Modal vs Cross Domain

In order to show the importance of the novel problem we consider, we compare the single-source knowledge transfer results on the SUN RGB-D dataset for modality change vs domain shift in Table 7. We use SHOT [3] which is a source-free UDA method for this experiment. All the domain-specific source models are trained on RGB images. For domain shift, the targets are all the RGB images of the remaining 3 domains and we report the average over them. Domain shift involves changes in sensor configuration, viewpoints, etc. For modality change, the target data are depth images from the same domain. The scenes are the same as in the RGB source, except they are captured using the depth sensor. The table clearly shows that the accuracy drops by a large margin of 12.5% when we transfer knowledge across modalities instead of domains of the same modality. This shows that a cross-modal knowledge transfer is not the same as DA and a framework like SOCKET is necessary to reduce the modality gap.

5.4 Ablation and sensitivity analysis

Contribution of loss components: In Table 8, the first row has the result with just the *modality agnostic* loss $\mathcal{L}_{ma}$, whereas second and third row shows the individual effect of our proposed *modality specific* losses along with the $\mathcal{L}_{ma}$. For all cases, SOCKET outperforms the baseline and using both losses in conjunction with $\mathcal{L}_{ma}$ yields best results.

Effect of number of TI images: We randomly chose six classes from SUN RGB-D dataset as TI data. Table 9 clearly shows that increasing per class

Table 8. Ablation of contribution of our proposed novel loss components. The first accuracy column (a) corresponds to single source adaptation from RGB to depth on *kv2* domain, whereas the second column (b) shows the multi-source adaptation result from *kv1+xtion* to *kv1* domain of SUN RGB-D dataset. We show the accuracy gain over using $\mathcal{L}_{ma}$ only inside the parentheses

$\mathcal{L}_{ma}$	$\mathcal{L}_d$	$\mathcal{L}_{TI}$	(a) accuracy (%)	(b) accuracy (%)
✓			30.0	11.7
✓	✓		31.6 (↑1.6)	18.3 (↑6.6)
✓		✓	34.9 (↑4.9)	22.6 (↑10.9)
✓	✓	✓	**36.3** (↑6.3)	**23.9** (↑12.2)

Table 9. Left: Effect of number of TI data. We perform knowledge transfer from Kinect v1 RGB to unlabeled depth data. We use six random TI classes and vary the number of TI images per class from 0 to 60 in steps of 20. **Right: Effect of regularization hyper-parameters**. We perform Kinect v1 and Kinect v2 RGB to Kinect v1 depth transfer with varying $(\lambda_{TI}, \lambda_d)$ and tabulate the accuracy of SOCKET

Images per class	60	40	20	0
Accuracy (%)	25.0	22.5	20.3	16.7

$(\lambda_{TI}, \lambda_d)$	0.00	0.05	0.10	0.50	1.00
Kinect v1	16.1	15.0	16.6	23.4	21.0
Kinect v2	29.3	34.2	35.0	36.7	16.3

samples of TI data results in improving the scene-classification accuracy for RGB to depth transfer on the SUN RGB-D dataset. In short, for a fixed number of TI classes, the more TI images per class, the better SOCKET performs.

Effect of regularization parameters: In Table 9, we observe the effect of test accuracy vs the regularization hyper-parameters for our novel losses proposed as a part of SOCKET. We keep λ_{TI} and λ_d equal to each other for values between 0 to 1. Using the value of 0 is the same as using SHOT. From the table, we see that as the value of the parameter increases the accuracy also increases up to a certain point, and then it starts decreasing.

6 Conclusion

We identify the novel and challenging problem of cross-modality knowledge transfer with no access to the task-relevant data from the source sensor modality, and only unlabeled data in the target. We propose our framework, SOCKET, which includes devising loss functions that help bridge the gap between the two modalities in the feature space. Our results for both RGB-to-depth and RGB-to-NIR experiments show that SOCKET outperforms the baselines which cannot effectively handle modality shift.

Acknowledgements. SMA, SL, KCP and MJ were supported by Mitsubishi Electric Research Laboratories. SMA and ARC were partially supported by ONR grant N00014-19-1-2264 and the NSF grants CCF-2008020 and IIS-1724341.

Bibliography

[1] Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2827–2836

[2] Ahmed, S.M., Lejbolle, A.R., Panda, R., Roy-Chowdhury, A.K.: Camera on-boarding for person re-identification using hypothesis transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 12144–12153

[3] Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning, PMLR (2020) 6028–6039

[4] Ahmed, S.M., Raychaudhuri, D.S., Paul, S., Oymak, S., Roy-Chowdhury, A.K.: Unsupervised multi-source domain adaptation without access to source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 10103–10112

[5] Perrot, M., Habrard, A.: A theoretical analysis of metric hypothesis transfer learning. In: International Conference on Machine Learning, PMLR (2015) 1708–1717

[6] Thoker, F.M., Gall, J.: Cross-modal knowledge distillation for action recognition. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 6–10

[7] Dai, R., Das, S., Bremond, F.: Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 13053–13064

[8] Garcia, N.C., Bargal, S.A., Ablavsky, V., Morerio, P., Murino, V., Sclaroff, S.: Dmcl: Distillation multiple choice learning for multimodal action recognition. arXiv preprint arXiv:1912.10982 (2019)

[9] Wang, J., Tang, Z., Li, X., Yu, M., Fang, Q., Liu, L.: Cross-modal knowledge distillation method for automatic cued speech recognition. arXiv preprint arXiv:2106.13686 (2021)

[10] Sayed, N., Brattoli, B., Ommer, B.: Cross and learn: Cross-modal self-supervision. In: German Conference on Pattern Recognition, Springer (2018) 228–243

[11] Hoffman, J., Gupta, S., Leong, J., Guadarrama, S., Darrell, T.: Cross-modal adaptation for rgb-d detection. In: 2016 IEEE international conference on robotics and automation (ICRA), IEEE (2016) 5032–5039

[12] Zhao, L., Peng, X., Chen, Y., Kapadia, M., Metaxas, D.N.: Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6528–6537

[13] Ferreri, A., Bucci, S., Tommasi, T.: Translate to adapt: Rgb-d scene recognition across domains. arXiv preprint arXiv:2103.14672 (2021)

[14] Du, D., Wang, L., Wang, H., Zhao, K., Wu, G.: Translate-to-recognize networks for rgb-d scene recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 11836–11845

[15] Ayub, A., Wagner, A.R.: Centroid based concept learning for rgb-d indoor scene classification. arXiv preprint arXiv:1911.00155 (2019)

[16] Peng, K.C., Wu, Z., Ernst, J.: Zero-shot deep domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 764–781

[17] Hsu, H.K., Yao, C.H., Tsai, Y.H., Hung, W.C., Tseng, H.Y., Singh, M., Yang, M.H.: Progressive domain adaptation for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2020) 749–757

[18] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 7167–7176

[19] Paul, S., Tsai, Y.H., Schulter, S., Roy-Chowdhury, A.K., Chandraker, M.: Domain adaptive semantic segmentation using weak labels. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer (2020) 571–587

[20] Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning, PMLR (2018) 1989–1998

[21] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 1406–1415

[22] Yang, S., Wang, Y., van de Weijer, J., Herranz, L., Jui, S.: Generalized source-free domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 8978–8987

[23] Yang, S., Wang, Y., van de Weijer, J., Herranz, L., Jui, S.: Exploiting the intrinsic neighborhood structure for source-free domain adaptation. arXiv preprint arXiv:2110.04202 (2021)

[24] Yang, S., Wang, Y., van de Weijer, J., Herranz, L., Jui, S.: Casting a bait for offline and online source-free domain adaptation. arXiv preprint arXiv:2010.12427 (2020)

[25] Agarwal, P., Paudel, D.P., Zaech, J.N., Van Gool, L.: Unsupervised robust domain adaptation without source data. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2022) 2009–2018

[26] Liang, J., Hu, D., Wang, Y., He, R., Feng, J.: Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

[27] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1) (2016) 2096–2030

[28] Bridle, J.S., Heading, A.J., MacKay, D.J.: Unsupervised classifiers, mutual information and'phantom targets'. (1992)
[29] Kutbi, M., Peng, K.C., Wu, Z.: Zero-shot deep domain adaptation with common representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
[30] Ioffe, S., Normalization, C.S.B.: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167
[31] Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8715–8724
[32] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. (2017) 1195–1204
[33] Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 567–576
[34] Cho, J., Min, D., Kim, Y., Sohn, K.: Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset. Expert Systems with Applications **178** (2021) 114877
[35] Brown, M., Süsstrunk, S.: Multi-spectral sift for scene category recognition. In: CVPR 2011, IEEE (2011) 177–184
[36] Kim, Y., Ham, B., Oh, C., Sohn, K.: Structure selective depth superresolution for rgb-d cameras. IEEE Transactions on Image Processing **25**(11) (2016) 5227–5238
[37] Kim, S., Min, D., Ham, B., Kim, S., Sohn, K.: Deep stereo confidence prediction for depth estimation. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017) 992–996
[38] Kim, Y., Jung, H., Min, D., Sohn, K.: Deep monocular depth estimation via integration of global and local predictions. IEEE transactions on Image Processing **27**(8) (2018) 4131–4144
[39] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 8026–8037