Modality-Invariant Image Classification Based on Modality Uniqueness and Dictionary Learning

Seungryong Kim, *Student Member, IEEE,* Rui Cai, *Member, IEEE,* Kihong Park, *Student Member, IEEE,* Sunok Kim, *Student Member, IEEE,* and Kwanghoon Sohn, *Senior Member, IEEE*

***Abstract*—We present a unified framework for image classification of image sets taken under varying modality conditions. Our approach is motivated by a key observation that the image feature distribution is simultaneously influenced by the semantic-class and the modality category label, which limits the performance of conventional methods for this task. With this insight, we introduce modality uniqueness as a discriminative weight that divides each modality cluster from all other clusters. By leveraging the modality uniqueness, our framework is formulated as unsupervised modality clustering and classifier learning based on modality-invariant similarity kernel. Specifically, in the assignment step, training images are first assigned to the most similar cluster in terms of modality. In the update step, based on the current cluster hypothesis, the modality uniqueness and the sparse dictionary are updated. These two steps are formulated in an iterative manner. Based on the final clusters, a modality-invariant marginalized kernel is then computed, where the similarities between the reconstructed features of each modality are aggregated across all clusters. Our framework enables the reliable inference of semantic-class category for an image, even across large photometric variations. Experimental results show that our method outperforms conventional methods on various benchmarks, e.g., landmark identification under severely varying weather conditions, domain-adapting image classification, and RGB-NIR image classification.**

***Index Terms*—Image classification, unsupervised modality clustering, modality uniqueness, dictionary learning.**

I. Introduction

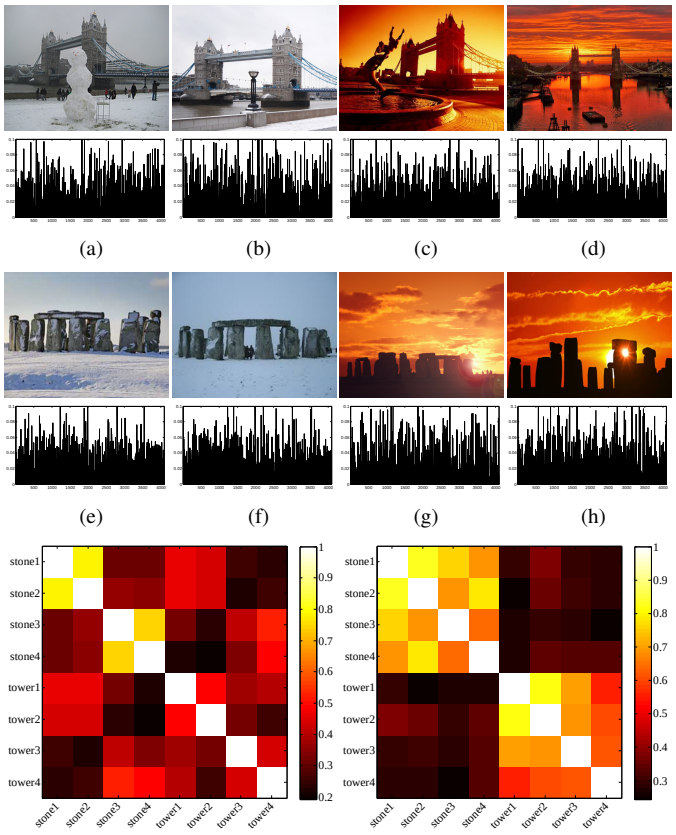
IMAGE classification for analyzing or classifying an image into semantically meaningful categories has been one of the most popular research topics in many computer vision and computational photography society [1], [2].

Conventionally, the bag-of-words (BoW) [4] approach and its variants based on local features, such as the scale invariant feature transform (SIFT) [5], have been mainstream methods for image classification [6]–[9]. Recently, with the availablity of large-scale training databases, *e.g.*, LabelME [10] and ImageNet [11], deep convolutional neural networks (CNNs) [1] methods have provided substantially improved performance for that task [2], [12]–[14]. More recently, many approaches have tried to use CNN activations as off-the-shelf features, followed by a classifier learning [2], [3], [15].

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2013R1A2A2A01068338).

S. Kim, K. Park, S. Kim, and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: {srkim89,khpark7727,kso428,khsohn}@yonsei.ac.kr).

R. Cai is with the Microsoft Research, Beijing, P.R. China, 100080 (e-mail: ruicai@microsoft.com).



(i) Conventional similarity kernel (j) Proposed similarity kernel

Fig. 1. Examples of similarity kernel for images under challengingly varying modality conditions. (a)-(d) Tower bridge images taken under snow condition in (a), (b) and sunset condition in (c), (d) with their corresponding deep CNN activation features (4096-d) [3]. (e)-(h) Stonehenge images taken under snow condition in (e), (f) and sunset condition in (g), (h). In conventional similarity kernel (*e.g.*, inner product) as in (i), the similarity between images derived from same modality is higher than that between images derived from a similar category, which limits the performance of conventional methods. Unlike this conventional kernel, the proposed kernel is robust to photometric variations.

To semantically classify an image, conventional methods have been based on a common assumption that images from the same semantic category in the training set have similar feature distributions, while images from different categories have different feature distributions [2], [4]. Based on this assumption, they tried to train classifiers (or decision boundaries) on the feature space that divide the training images into semantic categories. However, in real circumstances, an image might be taken under various modality[1] conditions, such

[1]In this paper, a modality is considered as terminology to denote a common image characteristic, which is similar to domain [16] or style [17].

as different scene radiance across lighting conditions, times, weathers, and seasons, or different camera specifications and settings [17], [18]. In these cases, the color statistics of images could be different across modality conditions, and further the feature distributions even from same semantic category cannot be coincident; rather, the feature distribution of images can be influenced by the modality condition corresponding to where the images were taken. Therefore, conventional approaches for image classification might provide unsatisfactory performance for a training image set built under severely varying conditions, since their feature-consistency assumption across semantic-class categories is no longer valid. Moreover, when training and testing image sets are established under different modality conditions, the accuracy of the image classification could also be degraded severely.

In these circumstances, robust local descriptor-based methods [5], [19] or robust global descriptor-based methods [1], [4] cannot be a fundamental solution for that task, since the modality variation explicitly influences the local and global feature distributions. With a similar problem setting as these cases, domain adaptation methods have been tried to solve domain-variation problems between training and testing image sets, or source and target domain image sets [16], [20]. In fact, domain adaptation methods consider explicitly divided image sets as source and target domain. However, in practice, a training image set can be thought as a collection of subsets with multiple unknown domains; thus, approaches for domain adaptation cannot be applied to general domain variation problems. Furthermore, these problems can be related to cross-domain matching [21], [22], which aims to find visually similar images across various domains. However, they only consider spatially and structually similar images.

To solve these ill-posed problems, our approach starts from a key observation that the image feature distribution is simultaneously influenced by semantic class and modality category, as shown in Fig. 1, which induces the inherent limitation of conventional image classification methods. Inspired by this insight, from the similar perspective of a decision boundary that divides training images in terms of semantic class labels, we discover that a decision boundary also exists, which divides images of one modality category from images of other modality categories.

Based on this observation, we propose a modality-invariant image classification framework that has two key ingredients, namely unsupervised modality clustering and modality-invariant similarity kernel based classifier learning. We initially divide the training images into modality clusters using a spectral clustering scheme. For unsupervised clustering, in the assignment step, each training image is assigned to the most similar cluster. In the update step, the modality uniqueness and dictionary are computed based on the current cluster hypothesis. These steps are iteratively formulated. Based on the final clusters and modality uniqueness/dictionary for each cluster, a marginalized similarity kernel is finally computed by aggregating the similarities between the reconstructed features of each cluster. Our framework enables us to robustly infer semantic category labels for images taken under severe modality variations. We compare our framework with conventional methods on novel landmark identification [23] under varying weather conditions, domain adaptation [24], and RGB and near-infrared (NIR) image classification [25].

The remainder of this paper is organized as follows. Section 2 introduces related work for the proposed method. Section 3 describes the proposed modality-invariant image classification framework. Experimental results and discussions are given in Section 4. Finally, the conclusion and suggestions for future work are given in Section 5.

A. Contributions

The contributions of our approach can be summarized as follows. First, to the best of our knowledge, our approach is the first attempt to solve the image classification problem under severe modality variations, which conventional methods cannot address. Second, we introduce modality uniqueness to encode a distinctive property of each modality, which is defined as a decision boundary to divide each modality-specific cluster from all other clusters. By leveraging this, we propose unsupervised modality clustering. Third, we define a novel similarity kernel to provide modality invariance for training a classifier. Fourth, we built a novel landmark-identification benchmark taken in severely varying weather conditions. Finally, we provide an intensive comparative study with state-of-the-art methods using various datasets.

II. Related Work

This section describes related works, including global feature descriptor, domain adaptation and generalization, cross-domain matching, and sparse dictionary learning.

A. Global Feature Descriptor

Conventionally, GIST [26] has been one of the primary global features for image classification; however, its performance is limited on small-scale databases. Subsequently, as a pioneering work, bag-of-words (BoW) [4] based methods have been the mainstream for that task. To alleviate the lack of spatial information in BoW, spatial pyramid matching (SPM) [27] was proposed, which leverages a multi-level grid technique. Furthermore, to solve the problem of a hard voting process in BoW, sparse coding (SC) [6], locality linear coding (LLC) [7], vectors of locally aggregated descriptor (VLAD) [8], Fisher vectors (FV) [9], and fast local-area-independent representations (FLAIR) [28] have been proposed. However, these methods have inherently limited performance, since they are derived from local descriptors, such as SIFT [5], which have limited discriminative power. Recently, image-classification performance has been impressively improved by leveraging the availability of large-scale training datasets [10], [11] and deep convolutional neural networks (CNNs) [1], [2], [12]–[14]. More recently, many approaches have been tried to use CNN activations as off-the-shelf features, followed by a classifier learning [2], [3], [15]. However, even CNN-based approaches cannot address severe modality variations in the training and testing image sets, as will be shown in our experiments.

B. Domain Adaptation

To deal with the domain variations between training and testing image sets, the domain adaptation has been widely studied in many research areas, *e.g.*, language processing [29], [30], machine learning [31], [32], and computer vision [33]. As a primary work, based on information-theoretic metric learning (ITML) [34], Saenko *et al.* [24] proposed a method that adapted object models acquired in a particular domain to a new domain by learning a transformation. The asymmetric regularized cross-domain transformation (ARC-t) [20] was also proposed. These methods have tried to solve problem settings where the source and target domains are explicitly determined, and the semantic-class labels are also known.

As a more challenging problem setting, unsupervised domain adaptation deals with the problem where only the source domain is labeled, whereas the target domain is not. Grassmann manifold-based methods were proposed to solve these problems, *e.g.*, the sampling geodesic flow (SGF) [16] and geodesic flow kernel (GFK) [35]. Fernando *et al.* [36] performed unsupervised domain adaptation based on subspace alignment. Zhu *et al.* [37] proposed a semi-supervised approach based on dictionary learning. Unlike these methods, our approach tries to solve the problem where no domain information is known for the training or testing images. Recently, an adaptive descriptor design (ADD) [17] was proposed to solve a similar problem by leveraging the kernel descriptor and multiple kernel learning (MKL) [38]; however, it also provided unsatisfactory results since it was defined with gradient-based kernel descriptors, whose limitation were studied in [39].

C. Domain Generalization

One of the works most related to our approach is domain generalization. In [40], Hoffman *et al.* proposed a constrained clustering method to discover the latent domains. In [41], Gong *et al.* partitioned the training samples from one domain into multiple domains by simultaneously maximizing the distinctiveness and learnability. Unlike these methods, our approach divides training images into modality-specific clusters and trains modality-invariant classifiers simultaneously.

On the other hand, several approaches have been proposed to classify an image according to its style. Using images from aesthetic visual analysis (AVA) datasets [42], several previous approaches have developed a system to classify images into classic painting styles [43], [44]. However, they considered only a handful of styles that are visually very distinct. Mensink [45] provided a larger artwork dataset, but did not consider the style classification. Furthermore, several approaches have been considered for weather classification [46]–[48], but they have focused on specific weather cues, which provide limited performance for general modality conditions.

D. Cross-Domain Matching

Many studies have been devoted to matching images across specific domains, including photos across different lighting conditions [49], sketches to photographs [50], [51], paintings to photographs [52], and computer graphic (CG) images to photographs [53]. However, these domain-specific solutions are not applicable to cases with multiple potential domains. For a more generalized solution, some methods focused on designing robust local descriptors, *e.g.*, self-similarity [19], across different visual domains. Furthermore, for that task, data-driven approaches [21], [22] spatially boost local descriptors by leveraging a linear classifier. Nevertheless, these approaches cannot be applied to general modality variation problems.

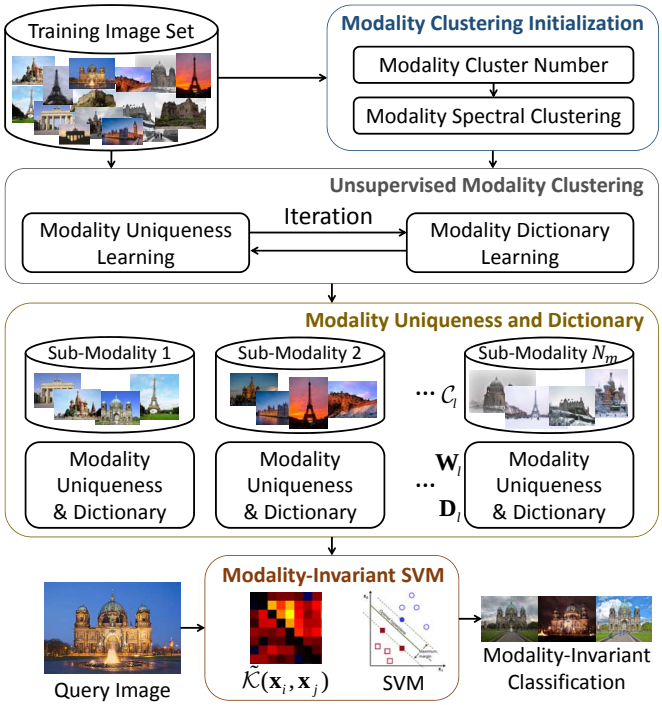


Fig. 2. Framework of the modality-invariant image classification. It consists of clustering initialization, unsupervised clustering, and modality-invariant kernel computation. Based on modality uniqueness and dictionary from each cluster, it divides the training image set into modality-specific clusters, and estimates modality-invariant similarity kernel to finally learn a classifier.

E. Sparse Dictionary Learning

Dictionary learning based on sparse coding has been proven to be very effective in image reconstruction [54]–[56], image de-noising [57], [58], image de-blurring [59], image inpainting [60]–[62], super resolution [59], [62], and image retrieval [63]. Among the many existing dictionary learning methods, the k-means singular-value decomposition (K-SVD) method [64] is one of the most widely used methods. For image classification in [63], sparse coding was applied to derive a compact yet discriminative image representation from multiple feature types for large-scale image retrieval. In [65], a clustering method using the sparse modeling and dictionary learning setting was introduced, where a set of dictionaraies was optimized to reconstruct signals in a sparse coding manner. In our approach, a sparse dictionary is used to encode a distinctive characteristic for each modality, and we further propose a novel fidelity term in dictionary learning to encode modality-specific information.

III. Modality-Invariant Image Classification

A. Problem Formulation and Overview

Let us consider global feature descriptors in a matrix form $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_{\mathcal{N}_t}] \in \mathbb{R}^{d \times \mathcal{N}_t}$ computed as $\mathbf{x}_k = \Psi(\mathbf{I}_k) \in \mathbb{R}^d$

for an image $\mathbf{I}_k$ from training image set $\mathbf{S}_t = \{(\mathbf{I}_k, y_k)\}$. Here, $\Psi(\cdot)$ is the global feature description operator. $k \in \Phi_t = \{1, ..., \mathcal{N}_t\}$ and $y_k \in \Phi_c = \{1, ..., \mathcal{N}_c\}$, where the number of training images and class categories are denoted as $\mathcal{N}_t$ and $\mathcal{N}_c$, respectively. Our approach aims to infer class category label y_i for given query image $\mathbf{I}_i$ in the testing image set by leveraging classifier $\mathcal{H}(\mathbf{S}_t)$ from training image set $\mathbf{S}_t$.

Unlike conventional image classification approaches, which provide limited performance for training and testing image sets under severe modality variations, our approach starts from the observation that training image set $\mathbf{S}_t$ is derived under multiple hidden modalities m_l for $l \in \Phi_m = \{1, ..., \mathcal{N}_m\}$, where $\mathcal{N}_m$ is the number of modality categories. By considering modality categories m_l, our image classification model can be formulated in a probabilistic aspect as follows:

$$\begin{aligned} y_i &= \arg\min_{c \in \Phi_c} \mathbf{P}(c|\mathbf{x}_i) \\ &= \arg\min_{c \in \Phi_c} \sum_{l \in \Phi_m} \mathbf{P}(m_l|\mathbf{x}_i)\mathbf{P}(c|\mathbf{x}_i, m_l), \end{aligned} \tag{1}$$

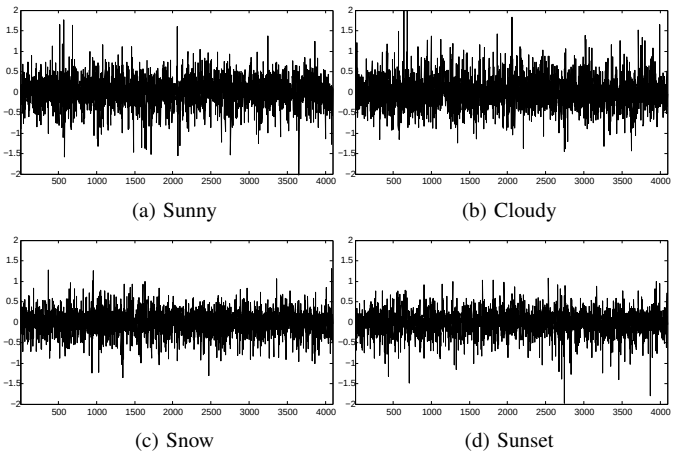
where $\mathbf{P}(c|\mathbf{x}_i)$ is the probability of class category c under $\mathbf{x}_i$, $\mathbf{P}(m_l|\mathbf{x}_i)$ is the probability of modality category m_l under $\mathbf{x}_i$, and $\mathbf{P}(c|\mathbf{x}_i, m_l)$ is the probability of class category c under m_l and $\mathbf{x}_i$. In discriminative learning, *e.g.,* support vector matchines (SVM), it is important to design a kernel function, where a high similarity is encoded for features from inner-class categories and a low similarity is encoded for features from inter-class categories [66], [67].

In a similar manner, based on the Bayesian theorem [66], we can reformulate such that $\mathbf{P}(c|\mathbf{x}, m) = \mathbf{P}(\mathbf{x}, m|c)\mathbf{P}(c)$ in (1) for each modality m with the probability $\mathbf{P}(m|\mathbf{x})$. To estimate a reliable y_i in (1), when $\mathbf{P}(c)$ is considered to be same for all c, features $\mathbf{x}_i$ and $\mathbf{x}_j$ from the same class category c and modality category m should have similar probabilities, $\mathbf{P}(\mathbf{x}_i, m|c)$ and $\mathbf{P}(\mathbf{x}_j, m|c)$. That is, under the same m and c, $\mathbf{x}_i$ and $\mathbf{x}_j$ should have a high similarity. Based on this, our approach leverages a kernel-embedding scheme where the kernel function $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ for $\mathbf{x}_i$ and $\mathbf{x}_j$ is formulated as

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l \in \Phi_m} \sum_{l' \in \Phi_m} \mathbf{P}(m_l|\mathbf{x}_i)\mathbf{P}(m_{l'}|\mathbf{x}_j)\mathcal{K}_{m_l, m_{l'}}(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

where $\mathbf{P}(m_{l'}|\mathbf{x}_j)$ is the probability of modality category $m_{l'}$ for $\mathbf{x}_j$. $\mathcal{K}_{m_l, m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$ is a joint-modality kernel function, which encodes the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ under modality categories m_l and $m_{l'}$, respectively. Using this aggregated kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, the effects of the modality variations in the training and testing image set can be reduced, and the further discriminative powers of a classifier can be enhanced. However, it is not easy to define $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ because the modality clusters from the training image set, and even the number of modalities are unknown. Furthermore, the modality probability $\mathbf{P}(m_l|\mathbf{x}_i)$ and the joint kernel function $\mathcal{K}_{m_l, m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$ are also hard to define.

To accomplish this task, we first introduce the modality uniqueness concept as a discriminative weight that divides each modality cluster from all other modality clusters, which will be discussed in Sec. III-B. We then estimate a sparse dictionary for each modality, where its fidelity term is weighted with the modality uniqueness. This modality uniqueness and dictionary are iteratively computed until the training images are divided optimally, which will be described in Sec. III-C. We finally define the similarity kernel to train a modality-invariant image classifier, which will be described in Sec. III-D. The number of modalities and initial cluster hypothesis are estimated in an initialization, which will be described in Sec. III-E. Our framework is summarized in Fig. 2.



(a) Sunny (b) Cloudy (c) Snow (d) Sunset

Fig. 3. Examples of the modality uniqueness $\mathbf{w}_l$ with deep CNNs activation features [1] for different modality (*e.g.*, weather) clusters. Different modality uniqueness encodes different modality characteristic.

B. Modality Uniqueness

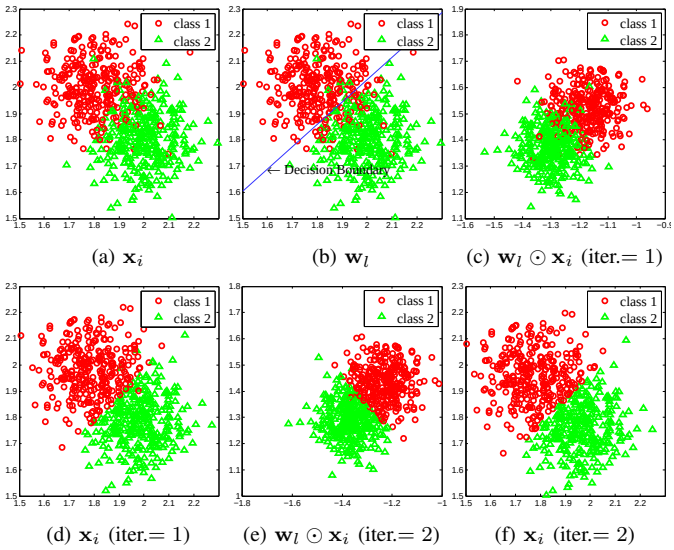
Encoding a distinctive visual property for each modality is not an easy task since it requires a detailed model. In this section, we instead encode this property of each modality in a data-driven manner. Our main observation is that each modality has a distinctive weight in a feature domain that distinguishes between the features of one modality cluster and all other modality clusters. We define this weight vector as the *modality uniqueness*. It is derived from similar intuitions with an exemplar-SVM [68] and data-driven uniqueness [69], [70]. Compared to these methods, the modality uniqueness is defined to encode modality-specific characteristics in a feature domain.

Specifically, it is defined as a discriminative weight vector to divide the features of cluster $\mathcal{C}_l$ for modality m_l from the features of sub-set $\mathbf{S}_t/\mathcal{C}_l$ that exclude images of $\mathcal{C}_l$ from $\mathbf{S}_t$. The modality uniqueness $\mathbf{w}_l \in \mathbb{R}^d$ for modality m_l is defined by exploiting discriminative learning, *i.e.,* SVM, with a linear decision boundary $\rho(\mathbf{w}_l, \mathbf{x}_i)$ defined such that

$$\rho(\mathbf{w}_l, \mathbf{x}_i) = \mathbf{w}_l^T \mathbf{x}_i, \tag{3}$$

where weight $\mathbf{w}_l$ indicates the contribution of a feature descriptor $\mathbf{x}_i$ for each component. Learning the modality uniqueness $\mathbf{w}_l$ amounts to minimizing objective function $\mathbf{E}(\mathbf{w}_l)$, composed of a fidelity function $\mathcal{L}(\mathbf{w}_l)$ and a regularization function $\mathcal{R}(\mathbf{w}_l)$, such that

$$\begin{aligned} \mathbf{E}(\mathbf{w}_l) &= \eta \mathcal{L}(\mathbf{w}_l) + \mathcal{R}(\mathbf{w}_l) \\ &= \eta \sum\nolimits_{i \in \Phi_t} \mathbf{h}(b_i^l \cdot \rho(\mathbf{w}_l, \mathbf{x}_i)) + \mathcal{R}(\mathbf{w}_l), \end{aligned} \tag{4}$$



(a) $\mathbf{x}_i$ (b) $\mathbf{w}_l$ (c) $\mathbf{w}_l \odot \mathbf{x}_i$ (iter.= 1)

(d) $\mathbf{x}_i$ (iter.= 1) (e) $\mathbf{w}_l \odot \mathbf{x}_i$ (iter.= 2) (f) $\mathbf{x}_i$ (iter.= 2)

Fig. 4. Toy examples of unsupervised modality clustering. For global feature descriptor $\mathbf{x}_i$ with initial 2 modality cluster class in (a), first of all, we compute modality uniqueness $\mathbf{w}_l$ in (b). With the modality uniqueness, the feature can be transformed in a modality-distinctive space $\mathbf{w}_l \odot \mathbf{x}_i$. With the modality uniqueness and dictionary, the modality cluster is divided more distinctively in (d) compared to (a). After evolving iterations, as shown in (e), (f), the modality cluster can be more distinctively divided.

where η represents a regularization parameter, and the hinge loss function $\mathbf{h}(\cdot) = \mathbf{max}(1 - \cdot)$. b_i^l represents 1 for images $\mathbf{I}_i$ from $\mathcal{C}_l$, and 0 otherwise.

In a matrix/vector form, (4) can be redefined as

$$\mathbf{E}(\mathbf{W}_l) = \eta \mathbf{tr}\{\mathbf{h}(\mathbf{B}_l(\mathbf{X}^{1/2})^T \mathbf{W}_l \mathbf{X}^{1/2})\} + \mathcal{R}(\mathbf{W}_l), \quad (5)$$

where $\mathbf{W}_l$ is the modality uniqueness in a matrix form, defined as $\mathbf{W}_l = \mathbf{diag}[w^l]_{d\times d}$. $\mathbf{B}_l = \mathbf{diag}[b_i^l]_{\mathcal{N}_t \times \mathcal{N}_t}$.

In an SVM [67], the regularization function is generally defined as l_2 norm, *i.e.*, $\mathcal{R}(\mathbf{W}_l) = \|\mathbf{W}_l\|_F^2$. In contrast, to estimate a modality-specific weight $\mathbf{W}_l$, our regularization function is derived from a fidelity term in sparse dictionary learning, which will be described in Sec. III-C. It is worth noting that compared to conventional domain- or style-specific cues [46]–[48], the modality uniqueness can be generally applied to any feature descriptors.

Fig. 3 represents the examples of the modality uniqueness $\mathbf{w}_l$. Each component of $\mathbf{w}_l$ represents the importance of the corresponding feature components towards encoding a distinctive property for each modality. As shown in Fig. 5, in a feature space of $\mathbf{x}_i$ weighted by modality uniqueness $\mathbf{w}_l$, *i.e.*, $\mathbf{w}_l \odot \mathbf{x}_i$, where $\odot$ indicates a component-wise product, the division of clusters can be more distinctive compared in the feature space of $\mathbf{x}_i$ itself. To leverage this property, the modality uniqueness will be incorporated in an energy function of the unsupervised modality-clustering scheme in Sec. III-C and, further, be used to define a modality-invariant kernel for classification in Sec. III-D.

C. Unsupervised Modality Clustering

Our approach assumes that the training image set consists of multiple hidden modalities, but we cannot have any prior information on the modality categories. In this section, we propose an unsupervised modality-clustering scheme, where the training image set $\mathbf{S}_t$ is divided into the modality cluster $\mathcal{C}_l$. We employ an iterative scheme for this clustering. Specifically, based on each modality cluster $\mathcal{C}_l$, we estimate the modality uniqueness $\mathbf{w}_l$, defined in Sec. III-B, and the sparse dictionary $\mathbf{D}_l$. Furthermore, to reduce the discrepancy of dictionary $\mathbf{D}_l$, we also propose an incoherent dictionary $\widetilde{\mathbf{D}}$ similar to [65]. All these parameters are iteratively updated.

For modality cluster $\mathcal{C}_l$, our energy function $\mathbf{E}(\mathbf{w}, \mathbf{D}, \widetilde{\mathbf{D}}, \alpha)$ is formulated to estimate modality uniqueness $\mathbf{w}$, dictionary $\mathbf{D}$, and incoherent dictionary $\widetilde{\mathbf{D}}$ simultaneously such that

$$\begin{aligned}\mathbf{E}(\mathbf{w}, \mathbf{D}, \widetilde{\mathbf{D}}, \alpha) = \sum_{l\in\Phi_m} \sum_{i\in\mathbf{C}_l} \left(\|\mathbf{w}_l \odot (\mathbf{x}_i - \mathbf{D}_l\alpha_i)\|_2^2 + \lambda|\alpha_i|_1\right) \\ + \tau\|\widetilde{\mathbf{D}} - \mathbf{D}_l\|_F^2 + \eta\mathcal{L}(\mathbf{w}_l) \\ \text{s.t.} \quad \forall l, u, \quad \|\mathbf{d}_{l,u}\|_2 < 1,\end{aligned} \quad (6)$$

where λ, τ, and η are parameters. $\|\cdot\|_2$, $|\cdot|_1$, and $\|\cdot\|_F$ denote l_2 norm, l_1 norm, and frobenius norm, respectively. $\mathbf{D}_l = [\mathbf{d}_{l,1}, ..., \mathbf{d}_{l,v}] \in \mathbb{R}^{d\times v}$, where v is a dimension of sparse coefficient. $\alpha_i \in \mathbb{R}^v$ is the sparse coefficient. $\mathcal{L}(\mathbf{w}_l)$ is the loss function for modality uniqueness $\mathbf{w}_l$ defined in (4).

For the sake of simplification, the energy function in (6) can be derived as a matrix/vector form denoted as $\mathbf{E}(\mathbf{W}, \mathbf{D}, \widetilde{\mathbf{D}}, \mathbf{\Lambda})$ with same constraints such that

$$\begin{aligned}\sum_{l\in\Phi_m} \|\mathbf{W}_l(\mathbf{X}_l - \mathbf{D}_l\mathbf{\Lambda}_l)\|_F^2 + \lambda|\mathbf{\Lambda}_l|_1 \\ + \tau\|\widetilde{\mathbf{D}} - \mathbf{D}_l\|_F^2 + \eta\mathbf{tr}\{\mathbf{h}(\mathbf{B}_l(\mathbf{X}^{1/2})^T\mathbf{W}_l\mathbf{X}^{1/2})\},\end{aligned} \quad (7)$$

where $\mathbf{X}_l = [\mathbf{x}_1, ..., \mathbf{x}_{\mathcal{N}_{\mathcal{C}_l}}] \in \mathbb{R}^{d\times\mathcal{N}_{\mathcal{C}_l}}$ is a sub-global feature descriptor for $\mathbf{I}_i$ for cluster $\mathcal{C}_l$, where $\mathcal{N}_{\mathcal{C}_l}$ is the number of samples in $\mathcal{C}_l$. $\mathbf{\Lambda}_l = [\alpha_1, ..., \alpha_{\mathcal{N}_{\mathcal{C}_l}}] \in \mathbb{R}^{v\times\mathcal{N}_{\mathcal{C}_l}}$ is the sparse coefficient matrix, and $|\mathbf{\Lambda}_l|_1 = \sum_{i\in\mathbf{C}_l} |\alpha_i|_1$.

Our energy function $\mathbf{E}(\mathbf{W}, \mathbf{D}, \widetilde{\mathbf{D}}, \mathbf{\Lambda})$ is formulated to have the following three desirable properties. First, the fidelity term for dictionary learning (*i.e.*, $\|\mathbf{W}_l \cdot (\mathbf{X}_l - \mathbf{D}_l\mathbf{\Lambda}_l)\|_F^2$) is weighted by modality uniqueness $\mathbf{W}_l$, which enables estimating the modality-specific dictionary $\mathbf{D}_l$ by considering important components from $\mathbf{W}_l$. Second, the fidelity term is further considered as a regularization function $\mathcal{R}(\mathbf{W}_l)$ for training modality uniqueness $\mathbf{W}_l$, which enables the estimation of the modality-specific weight $\mathbf{W}_l$ by considering the reconstruction error of each component between $\mathbf{X}_l$ and $\mathbf{D}_l\mathbf{\Lambda}_l$. Third, by introducing $\widetilde{\mathbf{D}}$, rather than independently estimating $\mathbf{D}_l$ and $\mathbf{\Lambda}_l$ for each $\mathbf{C}_l$, all $\mathbf{D}_l$ and $\mathbf{\Lambda}_l$ are learned simultaneously. This enables building a robust sparse dictionary that encodes not only the distinctive characteristics of each cluster, but also common ones across all clusters.

$\mathbf{E}(\mathbf{W}, \mathbf{D}, \widetilde{\mathbf{D}}, \mathbf{\Lambda})$ cannot be solved directly due to its non-convex property. Instead, we minimize this energy function such that modality uniqueness $\mathbf{W}_l$, dictionary $\mathbf{D}_l$, and incoherent dictionary $\widetilde{\mathbf{D}}$ are iteratively solved for each cluster $\mathbf{C}_l$. Following this, the cluster hypothesis $\mathbf{C}_l$ is estimated. For unsupervised modality clustering, our framework employs an iterative optimization using a Lloyd′s-type algorithm including assignment and update steps.

1) Assignment Step: In each step, we iteratively divide the training images into modality clusters. When modality uniqueness $\mathbf{W}_l$ and dictionary $\mathbf{D}_l$ are estimated and fixed in the previous iteration, all images in the training image set are divided into the cluster C_l such that the best cluster is determined by minimizing the reconstruction error

$$\mathcal{P}(\mathbf{x}_i; \mathbf{w}_l, \mathbf{D}_l) = \|\mathbf{w}_l \odot (\mathbf{x}_i - \mathbf{D}_l\alpha_i)\|_2^2 + \lambda|\alpha_i|_1, \quad (8)$$

where α_i is a sparse coefficient for $\mathbf{x}_i$ decomposed using $\mathbf{w}_l$ and $\mathbf{D}_l$. In experiments, we used SPArse Modeling Software (SPAMS) toolbox [71].

With reconstruction error $\mathcal{P}(\mathbf{x}_i; \mathbf{w}_l, \mathbf{D}_l)$, each modality-cluster hypothesis C_l is determined, such that

$$C_l = \{i|\mathcal{P}(\mathbf{x}_i; \mathbf{w}_l, \mathbf{D}_l) \leq \mathcal{P}(\mathbf{x}_i; \mathbf{w}_{l'}, \mathbf{D}_{l'}), \forall i, l' \in \Phi_m\}. \quad (9)$$

Modality cluster hypothesis C_l is then used to train modality uniqueness $\mathbf{w}_l$ and modality dictionary $\mathbf{D}_l$ in the update step, as described in Sec. III-C2.

2) Update Step: In the update step, the modality uniqueness and dictionary are computed fixing the cluster assignments $\mathbf{C}_l$ found in the assignment step as in Sec. III-C1. In the following, we will summarize how to estimate $\mathbf{W}_l$, $\mathbf{D}_l$, and $\mathbf{\Lambda}_l$ from $\mathbf{E}(\mathbf{W}, \mathbf{D}, \widetilde{\mathbf{D}}, \mathbf{\Lambda})$ in each iteration step.

Computing $\mathbf{W}_l$: As described in Sec. III-B, the modality uniqueness $\mathbf{W}_l$ for cluster $\mathbf{C}_l$ is computed as a discriminative weight as a decision boundary for dividing features of $\mathbf{C}_l$ from those of $\mathbf{S}_t/\mathbf{C}_l$. The energy function in (7) in terms of modality uniqueness $\mathbf{W}$ with fixing $\mathbf{D}$, $\widetilde{\mathbf{D}}$, and $\mathbf{\Lambda}$ can be derived as

$$\mathbf{E}(\mathbf{W}) = \sum\nolimits_{l \in \Phi_m} \mathbf{E}(\mathbf{W}_l). \quad (10)$$

This energy function $\mathbf{E}(\mathbf{W})$ can be solved independently for each $\mathbf{E}(\mathbf{W}_l)$. Unlike the conventional l_2 norm regularization $\|\mathbf{W}_l\|_F^2$ in SVM [67], the energy function $\mathbf{E}(\mathbf{W}_l)$ for modality uniqueness is defined as

$$\begin{aligned}\mathbf{E}(\mathbf{W}_l) = \eta\mathbf{tr}\{\mathbf{h}(\mathbf{B}_l(\mathbf{X}^{1/2})^T\mathbf{W}_l\mathbf{X}^{1/2})\} \\ + \|\mathbf{W}_l(\mathbf{X}_l - \mathbf{D}_l\mathbf{\Lambda}_l)\|_F^2.\end{aligned} \quad (11)$$

This novel SVM energy function enables us to encode more modality-specific properties on $\mathbf{W}_l$. Furthermore, it can still be easily solved using existing SVM solvers.

Specifically, with $\mathbf{V}_l = \mathbf{X}_l - \mathbf{D}_l\mathbf{\Lambda}_l$, it can be simplified as

$$\eta\mathbf{tr}\{\mathbf{h}(\mathbf{B}_l(\mathbf{X}^{1/2})^T\mathbf{W}_l\mathbf{X}^{1/2})\} + \|\mathbf{W}_l\mathbf{V}_l\|_F^2. \quad (12)$$

With a feature $\widetilde{\mathbf{X}} = \mathbf{V}_l^+\mathbf{X}$, where $\mathbf{V}_l^+$ means the Moore-Penrose pseudo inverse of $\mathbf{V}_l$, we learn $\widetilde{\mathbf{W}}_l = \mathbf{W}_l\mathbf{V}_l$ by minimizing the following energy function

$$\mathbf{E}(\widetilde{\mathbf{W}}_l) = \eta\mathbf{tr}\{\mathbf{h}(\mathbf{B}_l(\widetilde{\mathbf{X}}^{1/2})^T\widetilde{\mathbf{W}}_l\widetilde{\mathbf{X}}^{1/2})\} + \|\widetilde{\mathbf{W}}\|_F^2. \quad (13)$$

Using learned weight $\widetilde{\mathbf{W}}_l$, the final weight can be estimated such that $\mathbf{W}_l = \widetilde{\mathbf{W}}_l\mathbf{V}_l^+$. In experiments, we used LIB-SVM [72] to minimize the objective function $\mathbf{E}(\widetilde{\mathbf{W}}_l)$.

Computing $\mathbf{D}_l$ and $\widetilde{\mathbf{D}}$: For updating modality dictionary $\mathbf{D}$ and $\widetilde{\mathbf{D}}$ with fixing $\mathbf{W}$ and $\mathbf{\Lambda}$, our energy function in (7) in terms of dictionary $\mathbf{D}$ is derived as

$$\mathbf{E}(\mathbf{D}) = \sum\nolimits_{l \in \Phi_m} \mathbf{E}(\mathbf{D}_l). \quad (14)$$

This energy function $\mathbf{E}(\mathbf{D})$ can be solved independently for each $\mathbf{E}(\mathbf{D}_l)$ with fixed $\widetilde{\mathbf{D}}$ such that

$$\begin{aligned}\mathbf{E}(\mathbf{D}_l) =& \|\mathbf{W}_l(\mathbf{X}_l - \mathbf{D}_l\mathbf{\Lambda}_l)\|_F^2 + \tau\|\widetilde{\mathbf{D}} - \mathbf{D}_l\|_F^2 \\ & \text{s.t.} \quad \forall l, u, \quad \|\mathbf{d}_{l,u}\|_2 < 1,\end{aligned} \quad (15)$$

which is a quadratically constrained quadratic program (QCQP) with respect to $\mathbf{D}_l$, and the solutions can be found using the Lagrange dual technique [73]. In experiments, we used the CVX convex optimization toolbox [74]. This energy function is formulated with incoherent dictionary $\widetilde{\mathbf{D}}$; thus, it should be solved iteratively with a fixed $\widetilde{\mathbf{D}}$.

The incoherent dictionary $\widetilde{\mathbf{D}}$ can be simply computed using the following energy function

$$\mathbf{E}(\widetilde{\mathbf{D}}) = \sum\nolimits_{l \in \Phi_m} \tau\|\widetilde{\mathbf{D}} - \mathbf{D}_l\|_F^2. \quad (16)$$

By minimizing the energy function, the incoherent modality dictionary can be just computed as the average of $\mathbf{D}_l$ without any optimization scheme such that $\widetilde{\mathbf{D}} = \sum_l \mathbf{D}_l/\mathcal{N}_m$.

Computing $\mathbf{\Lambda}_l$: For updating sparse coefficient matrix $\mathbf{\Lambda}$ with fixing $\mathbf{W}$, $\mathbf{D}$, and $\widetilde{\mathbf{D}}$, our energy function in (7) in terms of a sparse coefficient $\mathbf{\Lambda}$ is derived as

$$\mathbf{E}(\mathbf{\Lambda}) = \sum\nolimits_{l \in \Phi_m} \mathbf{E}(\mathbf{\Lambda}_l). \quad (17)$$

This energy function $\mathbf{E}(\mathbf{\Lambda})$ can be solved independently for each $\mathbf{E}(\mathbf{\Lambda}_l)$ such that

$$\mathbf{E}(\mathbf{\Lambda}_l) = \|\mathbf{W}_l(\mathbf{X}_l - \mathbf{D}_l\mathbf{\Lambda}_l)\|_F^2 + \lambda|\mathbf{\Lambda}_l|_1. \quad (18)$$

It can be solved using a sparse coding solver. In experiments, we used the SPAMS toolbox [71].

After updating all $\mathbf{W}_l$, $\mathbf{D}_l$, $\widetilde{\mathbf{D}}$, and $\mathbf{\Lambda}_l$, our framework iteratively infers the modality cluster index $\mathcal{C}_l$ for the training image set as the assignment step. The assignment and update steps can be computed iteratively until they converge.

D. Modality-Invariant Marginalized Kernel

Many discriminative learning methods, such as SVM [67], have focused on computing a similarity kernel, *e.g.*, it can be simply defined as the inner production $\mathcal{K}^{\text{linear}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$. As a more complex case, the Gaussian kernel can be used such that $\mathcal{K}^{\text{Gaussian}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{exp}(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma)$. However, as described in Sec. III-A, leveraging these existing kernels directly is not well suited for training images under challenging modality variations.

In the following, we make use of the modality uniqueness and modality dictionary to design a new marginalized kernel as $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ in (2). It is formulated with the assumption that training images cannot belong to one specific modality cluster; rather, they might be derived from multiple modalities. Therefore, our kernel is computed by aggregating the similarities of each cluster with the corresponding probability.

First of all, based on modality uniqueness $\mathbf{w}_l$ and modality dictionary $\mathbf{D}_l$, the probability $\mathbf{P}(m_l|\mathbf{x}_i)$ in (2) can be defined with the reconstruction error $\mathcal{P}(\mathbf{x}_i, \mathbf{w}_l, \mathbf{D}_l)$ in (9) such that

$$\mathbf{P}(m_l|\mathbf{x}_i) = 1 - \frac{\mathcal{P}(\mathbf{x}_i; \mathbf{w}_l, \mathbf{D}_l)}{\sum_{l' \in \Phi_m} \mathcal{P}(\mathbf{x}_i; \mathbf{w}_{l'}, \mathbf{D}_{l'})}. \quad (19)$$

Here, it means that if $\mathbf{x}_i$ is derived from modality condition m_l, $\mathcal{P}(\mathbf{x}_i; \mathbf{w}_l, \mathbf{D}_l)$ would be small, then $\mathbf{P}(m_l|\mathbf{x}_i)$ is closed to 1. Otherwise, $\mathbf{P}(m_l|\mathbf{x}_i)$ is closed to 0.

Secondly, to derive joint modality-kernel $\mathcal{K}_{m_l,m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$ in (2), we also leverage modality uniqueness $\mathbf{w}_l$ and modality dictionary $\mathbf{D}_l$. With the latent modality m_l for $\mathbf{x}_i$ and $m_{l'}$ for $\mathbf{x}_j$, the reconstructed feature with the corresponding dictionary is $\mathbf{D}_l\alpha_i$ and $\mathbf{D}_{l'}\alpha_j$, respectively. Their weighted version using modality uniqueness can be denoted such that $\Omega(\mathbf{x}_i; m_l) \simeq \mathbf{w}_l \odot \mathbf{D}_l\alpha_i$ and $\Omega(\mathbf{x}_j; m_{l'}) \simeq \mathbf{w}_{l'} \odot \mathbf{D}_{l'}\alpha_j$, where α_i and α_j are sparse coefficients of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. As described in Sec. III-B, in the weighted feature space, modality-specific properties can be more boosted. Using the $\mathcal{K}^{\text{linear}}(\mathbf{x}_i, \mathbf{x}_j)$, our kernel function is computed as the inner product similarity of weighted features, $\Omega(\mathbf{x}_i; m_l)$ and $\Omega(\mathbf{x}_j; m_{l'})$, for each modality. Specifically, the joint-modality kernel $\mathcal{K}_{m_l,m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$ in (2) is defined such that

$$\begin{aligned}\mathcal{K}_{m_l,m_{l'}}(\mathbf{x}_i, \mathbf{x}_j) &= \mathbf{S}_{m_l,m_{l'}}\mathcal{K}^{\text{linear}}(\Omega(\mathbf{x}_i; m_l), \Omega(\mathbf{x}_j; m_{l'})) \\ &= \mathbf{S}_{m_l,m_{l'}}(\mathbf{w}_l \odot \mathbf{D}_l\alpha_i)^T(\mathbf{w}_{l'} \odot \mathbf{D}_{l'}\alpha_j),\end{aligned} \tag{20}$$

where $\mathbf{S}_{m_l,m_{l'}}$ is the similarity between two modalities, m_l and $m_{l'}$, with $0 \leq \mathbf{S}_{m_l,m_{l'}} \leq 1$. It should be noted that our joint modality kernel also can be formulated by using more complex kernel such as $\mathcal{K}^{\text{Gaussian}}(\mathbf{x}_i, \mathbf{x}_j)$, which might improve the performance while requiring more complexity. But, considering the trade-off between efficiency and accuracy, our kernel is defined using linear function, enough to provide satisfactory performance.

Now, based on the modality probability $\mathbf{P}(m_l|\mathbf{x}_i)$ and the joint modality kernel $\mathcal{K}_{m_l,m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$, the $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ in (2) can be defined. However, to formulate a general kernel function in discriminative learning, the kernel needs to be positive semi-definite (PSD) [67]. Since PSD kernels are closed under addition and multiplication, the $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ is PSD as long as the joint kernel $\mathcal{K}_{m_l,m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$ is PSD itself. In our case, a simple way to satisfy the PSD constraint consists of setting $\mathbf{S}_{m_l,m'_l} = 1$ when $m_l = m_{l'}$ and 0 otherwise similar to [67]. Then, $\mathcal{K}_{m_l,m_{l'}}(\mathbf{x}_i, \mathbf{x}_j)$ can be simplified to only consider the cases where $m_l = m_{l'}$ such that

$$\mathcal{K}_{m_l}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{w}_l \odot \mathbf{D}_l\alpha_i)^T(\mathbf{w}_l \odot \mathbf{D}_l\alpha_j), \tag{21}$$

which leads to the following final kernel $\widetilde{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j)$ such that

$$\widetilde{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j) = \sum\nolimits_{l \in \Phi_m} \mathbf{P}(m_l|\mathbf{x}_i)\mathbf{P}(m_l|\mathbf{x}_j)\mathcal{K}_{m_l}(\mathbf{x}_i, \mathbf{x}_j). \tag{22}$$

Note that $\widetilde{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j)$ based classifier learning can be considered as multiple kernel learning (MKL) [38] whose robustness has been proven. Compared to a simple linear combination [38], it is formulated as a probabilistic combination for each modality, which provides robustness for modality variations.

Using $\widetilde{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j)$, we finally train the modality-invariant SVM classifier $\mathcal{H}(\mathbf{S}_t)$ from training image set $\mathbf{S}_t$. Algorithm 1 summarizes our modality-invariant image categorization.

E. Unsupervised Modality Clustering Initialization

The iterative scheme in unsupervised modality clustering requires an initial cluster hypothesis. However, conventional clustering methods, such as k-means or spectral clustering [75], cannot estimate modality-specific clusters well, since global feature descriptors might be simultaneously influenced by two components: category-driven and modality-driven similarities. In this section, to divide the training image set into fully modality-specific image sub-sets, we propose a modality spectral clustering scheme.

First, for a training image set, we construct a similarity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the vertex set $\mathcal{V} = \{v_1, ..., v_{\mathcal{N}_t}\}$ on the feature descriptor $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_{\mathcal{N}_t}]$ and $\mathcal{E}$ is a set of links. An undirected edge exists if two vertices $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{V}$ are adjacent. In our approach, the k-nearest neighbor (k-nn) system is used to construct $\mathcal{E}$. To build a modality-specific image similarity graph $\mathcal{G}$, not a category-specific one, each query image finds the k-nn on different class training image sets. It is derived from the same insight that feature distribution can be determined by class categories and modality categories. By excluding images found by compulsively using the similarity of class categories, we can find the candidate images in terms of modality. Specifically, for feature $\mathbf{x}_i$ from corresponding class category y_i, the k-nn are found in other-class label image set $\mathbf{S}_t/\mathbf{S}_t^{y_i}$, where $\mathbf{S}_t^{y_i}$ is the training image set with the same class category as y_i. This simple scheme enables us to reduce the effect of category-specific similarities, and to focus on the effect of modality-specific similarities.

For the set of k-nn links, using the distance function as $\mathbf{d}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$, we compute a weighted adjacency function $\omega_{i,j} \in \mathcal{W}$, where $\mathcal{W}$ is a set of $\mathcal{N}_t \times \mathcal{N}_t$, such that

$$\omega_{i,j} = \mathbf{exp}\left(-\mathbf{d}(\mathbf{x}_i, \mathbf{x}_j)/\sigma\right), \quad \mathbf{x}_i, \mathbf{x}_j \in \mathcal{V}, \tag{23}$$

where σ is a range bandwidth. Then, the affinity and corresponding degree matrices of the graph $\mathcal{G}$ can be described as $\mathcal{W} = [\omega_{i,j}]_{i,j=1,...,\mathcal{N}_t}$ and $\mathcal{D} = \mathbf{diag}[d_i, ..., d_{\mathcal{N}_t}]$ where $d_i = \sum_j \omega_{i,j}$. With the graph $\mathcal{G}$ and its corresponding $\mathcal{W}$ and $\mathcal{D}$, we determine the number of latent modalities in the training image set, and divide the training image set into latent modality as following.

1) The Number of Modality Clusters: Based on the affinity matrix $\mathcal{W}$ from the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the number of clusters $\mathcal{N}_m$ can be estimated. Since each element of the affinity matrix $\mathcal{W}$ defines a distance in the feature space, the affinity matrix will have a block diagonal structure when there are definite groupings or clusters within the data sample [76]. It is worth noting that for affinity function $\mathcal{W}$ defined in (23), the following approximation holds, due to the convolution theorem for Gaussian [76] such that

$$\frac{1}{\mathcal{N}_t^2}\sum_{i \in \Phi_t}\sum_{j \in \Phi_t}\omega_{i,j} = \frac{1}{\mathcal{N}_t^2}\mathbb{1}_{\mathcal{N}_t}^T\mathcal{W}\mathbb{1}_{\mathcal{N}_t}, \tag{24}$$

where $\mathbb{1}_{\mathcal{N}_t}$ is the $\mathcal{N}_t \times 1$ dimensional vector with elements of value 1. An eigenvalue decomposition on the affinity matrix gives $\mathcal{W} = \mathbf{U}_\mathcal{W}\mathbf{\Sigma}_\mathcal{W}\mathbf{U}_\mathcal{W}^T$, where the columns of the matrix $\mathbf{U}_\mathcal{W}$ are the individual eigenvectors $u_i^\mathcal{W}$ of affinity and the diagonal matrix $\mathbf{\Sigma}_\mathcal{W}$ contains associated eigenvalues denoted as $\lambda_i^\mathcal{W}$. Then, we can rewrite $\mathbb{1}_{\mathcal{N}_t}^T\mathcal{W}\mathbb{1}_{\mathcal{N}_t}$ in (24) such that

$$\mathbb{1}_{\mathcal{N}_t}^T\left(\sum_{i \in \Phi_t}\lambda_i u_i u_i^T\right)\mathbb{1}_{\mathcal{N}_t} = \sum_{i \in \Phi_t}\lambda_i^\mathcal{W}\left(\mathbb{1}_{\mathcal{N}_t}^T u_i^\mathcal{W}\right)^2. \tag{25}$$

Algorithm 1: Modality-Invariant Image Categorization (MIIC)

Input: training image set $\mathbf{S}_t$.

Output: modality uniqueness $\mathbf{w}_l$, dictionary $\mathbf{D}_l$, incoherent dictionary $\widetilde{\mathbf{D}}$, cluster $\mathcal{C}_l$, SVM classifier $\mathcal{H}(\mathbf{S}_t)$.

Parameters and Notation

$\mathcal{N}_m$: the number of clusters.

$\mathbf{\Lambda}_l$: sparse coefficient matrix for cluster $\mathcal{C}_l$.

/ Initialization */*

Encode an image $\mathbf{I}_i$ from $\mathbf{S}_t$ into features $\mathbf{x}_i = \Psi(\mathbf{I}_i)$.

Construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from modality-specific k-nn images.

Determine the number of clusters $\mathcal{N}_m$.

Initialize cluster $\mathcal{C}_l$ from $\mathbf{S}_t$ using a modality spectral clustering.

Initialize modality uniqueness $\mathbf{w}_l$, dictionary $\mathbf{D}_l$ and sparse coefficient $\mathbf{\Lambda}_l$ for each cluster $\mathcal{C}_l$, and incoherent dictionary $\widetilde{\mathbf{D}}$.

while not converged **do**

  / Assignment Step */*

  Assign training images from $\mathbf{S}_t$ into cluster $\mathcal{C}_i$ using (9).

  / Update Step */*

  Estimate modality uniqueness $\mathbf{w}_l$ by optimizing (11).

  Estimate dictionary $\mathbf{D}_l$ by optimizing (15).

  Estimate incoherent dictionary $\widetilde{\mathbf{D}}$ by optimizing (16).

  Estimate sparse coefficient $\mathbf{\Lambda}_l$ by optimizing (18).

end while

Construct modality-invariant kernel $\widetilde{\mathcal{K}}(\mathbf{x}_i, \mathbf{x}_j)$ in (22).

Construct modality-invariant SVM classifier $\mathcal{H}(\mathbf{S}_t)$.

This indicates that if there are $\mathcal{N}_m$ distinct clusters within the data samples, there are $\mathcal{N}_m$ dominant terms $\lambda_i^{\mathcal{W}} \left(\mathbb{1}_{\mathcal{N}_t}^T u_i^{\mathcal{W}}\right)^2$ in the summation. In other words, the number of clusters $\mathcal{N}_m$ can be obtained by analyzing the dominant terms of $\lambda_i^{\mathcal{W}} \left(\mathbb{1}_{\mathcal{N}_t}^T u_i^{\mathcal{W}}\right)^2$. By plotting $\mathbf{log}(\lambda_i^{\mathcal{W}} \left(\mathbb{1}_{\mathcal{N}_t}^T u_i^{\mathcal{W}}\right)^2)$ with respect to i, a curve with an apparent elbow can be obtained. By finding the elbow of this plot, we can obtain the optimal number of clusters $\mathcal{N}_m$ prior to the classification.

2) Modality Spectral Clustering: Based on the number of clusters $\mathcal{N}_m$ as in Sec. III-E1, we initially divide the training images $\mathbf{S}_t$ into the initial modality cluster $\mathcal{C}_l$. We use a spectral clustering scheme [77] for the initial clustering. From graph $\mathcal{G}$, the un-normalized graph Laplacian matrix is computed as

$$\mathcal{L} = \mathcal{D} - \mathcal{W}. \tag{26}$$

An eigenvalue decomposition on the Laplacian matrix gives $\mathcal{L} = \mathbf{U}_{\mathcal{L}} \mathbf{\Lambda}_{\mathcal{L}} \mathbf{U}_{\mathcal{L}}^T$, wherethe columns of matrix $\mathbf{U}_{\mathcal{L}}$ are the eigenvectors $u_i^{\mathcal{L}}$ and the diagonal matrix $\mathbf{\Lambda}_{\mathcal{L}}$ contains the associated eigenvalues denoted as $\lambda_i^{\mathcal{L}}$. By using the first r eigenvectors $u_1^{\mathcal{L}}, ..., u_r^{\mathcal{L}}$, the features $\mathbf{x}_i$ for $i = 1, ..., \mathcal{N}_t$ are clustered using the k-means clustering into clusters $\mathcal{C}_1, ..., \mathcal{C}_{\mathcal{N}_m}$.

IV. Experimental Results

A. Experimental Environments

In the following experiments, our modality-invariant image classification framework was implemented with the same parameter settings for all datasets: $\{\lambda, \tau, \eta\} = \{0.5, 0.01, 0.1\}$. We employed cross-validation in order to estimate the parameters, and all parameters were fixed during the experiments. We implemented our approach in C++ on an Intel Core i7-3770 3.40 GHz CPU. As described in the above sections, any other global feature descriptors can be incorporated into our modality-invariant image classification (MIIC) approach. To evaluate our framework, we used various global feature descriptors, including BoW [4], GIST [26], ScSPM [6], LLC [7], VLAD [8], and CNN [1]. To evaluate the performance of the unsupervised clustering, our framework was compared to latent domain-clustering methods [40], [41], [79]. Furthermore, we examined the performance gains of the modality clustering in our approach, including the modality uniqueness, dictinoary, and incoherent dictionary. To evaluate our kernel embedding, we additionally examined the performance contributions of the kernel functions, including multi-modality projection and probability-based projection.

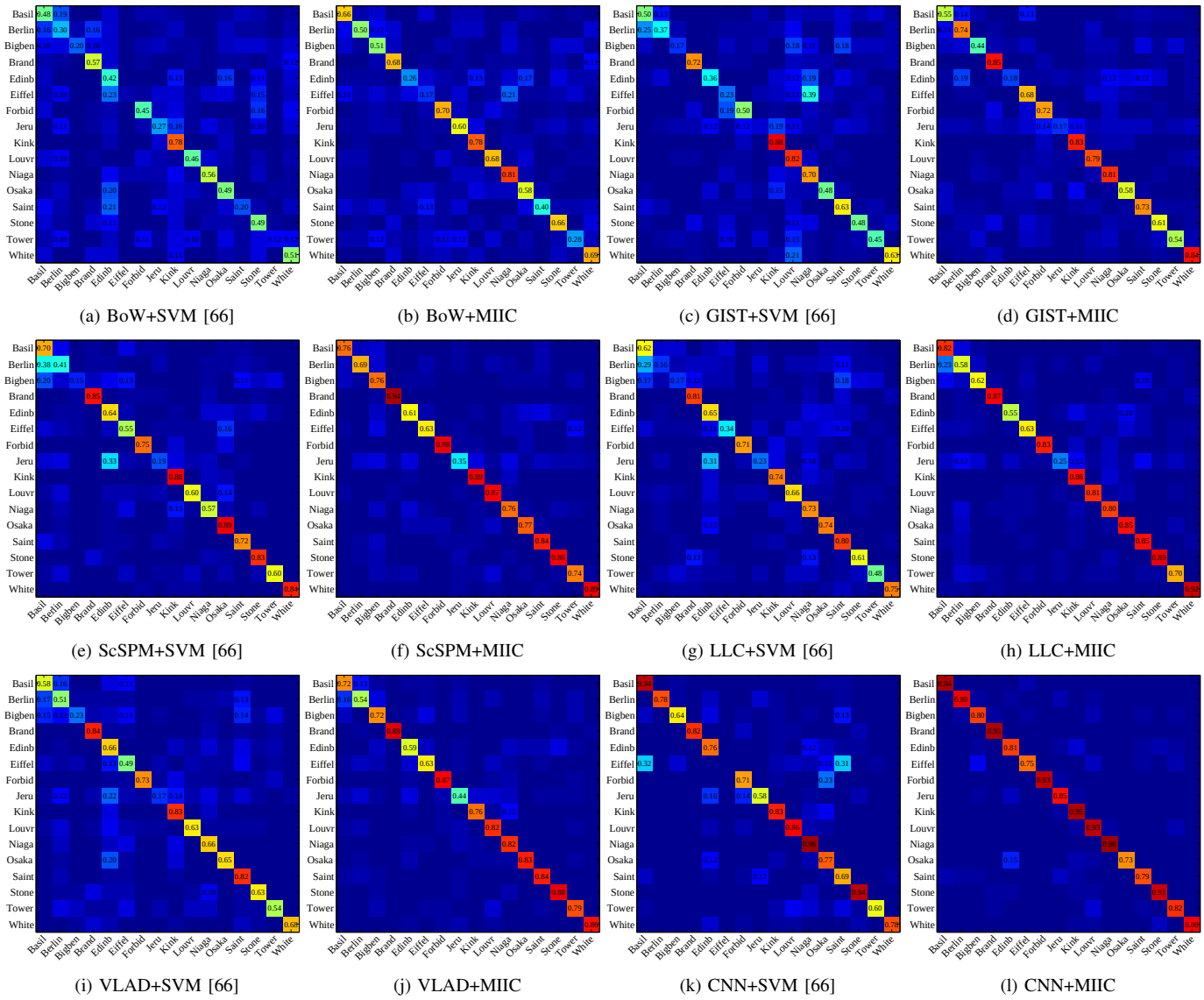
(a) Sunny (b) Cloudy

(a) Snowy (b) Sunset

Fig. 5. Examples of landmark identification images [23] under different image modality conditions, such as sunny, cloudy, snowy, and sunset.

TABLE I
Comparison of Quantitative Evaluation on Landmark Identification Under Challenging Weather Variations.

classifier	descriptor	dimension	mAP (%)
SVM w/linear kernel [66]	BoW [4]	1024d	36.15±0.7%
	GIST [26]	512d	48.14±1.4%
	ScSPM [6]	5120d	46.18±1.1%
	LLC [7]	5120d	50.12±1.0%
	VLAD [8]	8192d	52.72±2.1%
	CNN [1]	4096d	64.11±2.7%
SVM w/ Gaussian kernel [66]	BoW [4]	1024d	40.35±1.7%
	GIST [26]	512d	52.24±2.5%
	ScSPM [6]	5120d	51.51±2.1%
	LLC [7]	5120d	55.41±1.2%
	VLAD [8]	8192d	60.51±2.4%
	CNN [1]	4096d	70.51±3.1%
MIIC w/MKL [38]	BoW [4]	1024d	51.11±0.4%
	GIST [26]	512d	54.14±1.0%
	ScSPM [6]	5120d	57.11±1.1%
	LLC [7]	5120d	59.19±0.7%
	VLAD [8]	8192d	60.11±0.9%
	CNN [1]	4096d	73.16±1.8%
MIIC w/GMKL [78]	BoW [4]	1024d	50.82±1.3%
	GIST [26]	512d	56.10±1.7%
	ScSPM [6]	5120d	60.11±3.0%
	LLC [7]	5120d	65.11±1.1%
	VLAD [8]	8192d	68.11±2.3%
	CNN [1]	4096d	76.11±1.8%
MIIC	BoW [4]	1024d	**56.15±1.6%**
	GIST [26]	512d	**60.11±2.1%**
	ScSPM [6]	1024d	**68.58±1.9%**
	LLC [7]	1024d	**70.51±1.0%**
	VLAD [8]	1024d	**72.41±2.2%**
	CNN [1]	4096d	**87.51±2.3%**



(a) BoW+SVM [66] (b) BoW+MIIC (c) GIST+SVM [66] (d) GIST+MIIC

(e) ScSPM+SVM [66] (f) ScSPM+MIIC (g) LLC+SVM [66] (h) LLC+MIIC

(i) VLAD+SVM [66] (j) VLAD+MIIC (k) CNN+SVM [66] (l) CNN+MIIC

Fig. 6. Confusion matrix evaluations for landmark identification [23] under challenging weather conditions. For a variety of global feature descriptors such as BoW [4], GIST [26], ScSPM [6], LLC [7], VLAD [8], and CNN [1], our MIIC highly improves the classification performance compared to SVM.

For a qualitative evaluation of classification performance, the mean accuracy probability (mAP) was computed while varying the ratio of the training and testing image sets for each evaluation benchmark. Our method was evaluated for image classification on various benchmarks: landmark identification under challenging weather varitions, object recognition under varying domain configurations [24], and RGB-NIR image classification [25]. The novel landmark-identification benchmark can be found on our project page [23].

1) Global Feature Descriptor: To evaluate our framework, we utilized six global feature descriptors:

- **BoW**: Bag-of-words [4] represents an image as an orderless histogram of local features. In experiments, we computed a histogram for 1024-dimension quantized visual words for densely sampled SIFT features [5].
- **GIST**: The GIST descriptor [26] constructs a low-dimension description of the scene structure based on various orientation and scale filter responses. In experiments, we computed Gabor filters with three scales and eight initial parameters from Torralba [26], which provide the dimension of 512.
- **ScSPM**: Sparse-coding spatial-pyramid matching [6] extends SPM [27] by generalizing the vector quantization to sparse coding followed by multi-scale spatial max pooling. Similar to the BoW features, we used published code. The feature dimension is 5120.
- **LLC**: Locality linear coding [7] extends ScSPM by leveraging a simple but effective coding scheme called locality-constrained linear coding in the vector quantization coding in traditional SPM, whose feature dimension is 5120.
- **VLAD**: Vector of locally aggregated descriptors [8] divides an image into features using principal component analysis (PCA). In experiments, we used user-provided code, and the feature dimension is 8192.
- **CNN**: Deep convolutional neural networks [1] represents an image using convolutional responses in a deep architecture. In experiments, we utilized ImageNet based

TABLE II
COMPARISON OF QUANTITATIVE EVALUATION FOR DOMAIN ADAPTATION AS UNSUPERVISED SETTINGS.

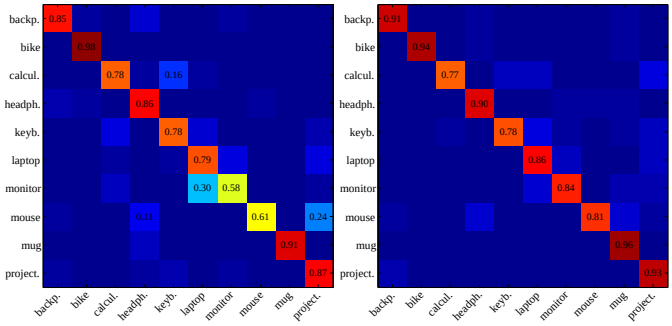
Domain		Unsupervised						
source	target	ARC-t [20]	K-SVD [64]	SGF [16]	GFK [35]	Sub. Int. [80]	ADD [17]	MIIC
Caltech	Amazon	18.4±1.2%	20.5±0.8%	36.8±0.5%	40.4±0.7%	45.4±0.3%	50.1±0.4%	**56.2±0.8%**
Caltech	DSLR	17.5±0.2%	19.8±1.0%	32.6±0.7%	41.1±1.3%	42.3±0.4%	41.5±0.3%	**45.1±0.2%**
Amazon	Caltech	23.5±0.2%	20.2±0.9%	35.3±0.5%	37.9±0.4%	40.4±0.5%	38.2±0.1%	**47.1±0.1%**
Amazon	Webcam	20.1±0.7%	16.9±1.0%	31.0±0.7%	35.7±0.9%	37.9±0.9%	32.5±0.7%	**39.4±0.4%**
Webcam	Caltech	21.2±1.0%	13.2±0.6%	21.7±0.4%	29.3±0.4%	36.3±0.3%	30.1±0.4%	**47.2±1.0%**
Webcam	Amazon	16.2±0.4%	14.2±0.7%	27.5±0.5%	35.5±0.7%	38.3±0.3%	30.6±0.1%	**47.9±0.2%**
DSLR	Amazon	22.2±0.1%	14.3±0.3%	32.0±0.4%	36.1±0.4%	39.1±0.5%	38.2±0.2%	**49.6±0.6%**
DSLR	Webcam	51.8±0.9%	46.8±0.8%	66.0±0.7%	79.1±0.7%	86.2±1.0%	66.1±0.2%	**88.2±0.4%**

TABLE III
COMPARISON OF QUANTITATIVE EVALUATION FOR DOMAIN ADAPTATION AS SEMI-SUPERVISED SETTINGS

Domain		Semi-supervised						
source	target	ARC-t [20]	K-SVD [64]	SGF [16]	GFK [35]	Sub. Int. [80]	ADD [17]	MIIC
Caltech	Amazon	28.7±1.0%	31.2±1.0%	40.2±0.7%	46.1±0.6%	50.0±0.5%	47.2±0.3%	**64.2±0.2%**
Caltech	DSLR	29.2±0.8%	34.6±1.0%	36.6±0.8%	55.0±0.9%	57.1±0.4%	50.1±0.6%	**65.1±0.2%**
Amazon	Caltech	29.2±1.1%	25.2±0.7%	37.7±0.5%	39.6±0.4%	41.5±0.8%	40.1±0.8%	**49.5±1.0%**
Amazon	Webcam	30.5±0.2%	42.7±0.6%	37.9±0.7%	56.9±1.0%	57.8±0.5%	50.7±0.1%	**64.2±0.3%**
Webcam	Caltech	20.2±0.5%	23.4±0.4%	29.2±0.7%	32.8±0.7%	40.6±0.4%	40.2±0.8%	**54.3±0.2%**
Webcam	Amazon	30.5±1.0%	32.9±0.7%	38.2±0.6%	46.2±0.7%	51.5±0.6%	49.2±0.2%	**63.5±0.8%**
DSLR	Amazon	29.7±0.2%	31.2±0.2%	39.2±0.7%	46.2±0.6%	50.3±0.2%	47.1±0.7%	**64.9±0.2%**
DSLR	Webcam	50.1±0.4%	49.9±1.4%	69.5±0.5%	80.2±0.4%	**82.8±1.0%**	69.1±0.1%	79.2±0.4%

(a) Amazon (b) DSLR (c) Webcam (d) Caltech

Fig. 7. Examples of domain adapting images under varying domains such as Amazon, DSLR, Webcam, and Caltech [24].



(a) CNN+SVM [66] (b) CNN+MIIC

Fig. 8. Confusion matrix evaluations for domain-adapting object recognition.

learning parameters by using MatConvNet [81]. For a feature descriptor, we used the final activations from the fully-connected layer, thus the dimension is 4096.

B. Landmark Identification Across Weather Variations

1) Landmark Identification Benchmark: To evaluate the image classification performance under modality variations, we constructed a novel landmark-identification benchmark taken under varying weather conditions. The database consists of 17 landmark images taken under several weather conditions, *e.g.*, sunny, cloudy, snowy, and sunset. These images were found on *Flickr* [82], *Google* [83], and *Bing* [84]. Each landmark dataset consists of 150 images, for a total 2550 images. Examples of landmark images under weather variations are shown in Fig. 5. The landmark-identification benchmark can be accessed on our project page [23].

2) Evaluation: We evaluated our MIIC by incorporating various global feature descriptors, *e.g.*, BoW [4], GIST [26], ScSPM [6], LLC [7], VLAD [8], and CNN [1]. Since the main performance gain of our framework comes from a novel similarity kernel function, our method was evaluated compared to SVM with linear (SVM w/linear kernel) and Gaussian kernel functions (SVM w/Gaussian kernel) [66]. Furthermore, to evaluate the modality uniqueness weights and modality probability in the proposed kernel, we formulated our MIIC as a varying aggregation scheme, such that the linear kernels between reconstructed features for each modality were aggregated with MKL (MIIC w/MKL) [38] and general MKL (MIIC w/GMKL) [78]. Table I shows a comparison of the qualitative evaluation, and Fig. 6 shows the confusion-matrix evaluation for landmark identification.

As expected, compared to conventional methods [4], [6]–[8], [26], CNN-based methods showed the best classification performance [1]. However, the performance of global descriptors combined with SVM using a linear kernel was limited. Even though the SVM with a Gaussian kernel improved the

TABLE IV
COMPARISON OF QUANTITATIVE EVALUATION ON DOMAIN-ADAPTING OBJECT RECOGNITION.

classifier	descriptor	dimension	mAP (%)
SVM w/linear kernel [66]	BoW [4]	1024d	37.21±1.2%
	GIST [26]	512d	54.97±3.5%
	ScSPM [6]	5120d	54.11±1.1%
	LLC [7]	5120d	57.11±1.0%
	VLAD [8]	8192d	59.11±3.4%
	CNN [1]	4096d	69.11±1.1%
MIIC w/MKL [38]	BoW [4]	1024d	39.11±0.2%
	GIST [26]	512d	56.36±1.5%
	ScSPM [6]	5120d	56.81±2.2%
	LLC [7]	5120d	59.72±2.0%
	VLAD [8]	8192d	61.73±1.4%
	CNN [1]	4096d	72.21±2.1%
MIIC w/GMKL [78]	BoW [4]	1024d	42.11±2.8%
	GIST [26]	512d	60.17±0.5%
	ScSPM [6]	5120d	62.83±1.2%
	LLC [7]	5120d	60.21±2.0%
	VLAD [8]	8192d	62.83±1.4%
	CNN [1]	4096d	76.21±0.7%
MIIC	BoW [4]	1024d	**44.25±1.0%**
	GIST [26]	512d	**62.21±3.1%**
	ScSPM [6]	5120d	**67.18±1.2%**
	LLC [7]	5120d	**68.21±2.2%**
	VLAD [8]	8192d	**70.11±1.8%**
	CNN [1]	4096d	**84.11±1.3%**

(a) RGB (b) NIR

Fig. 9. Examples of RGB-NIR images for image classification.

classification performance to some extent, it also showed unsatisfactory performance. Compared to these methods, our MIIC method dramatically improved the classification performance when combined with any descriptor.

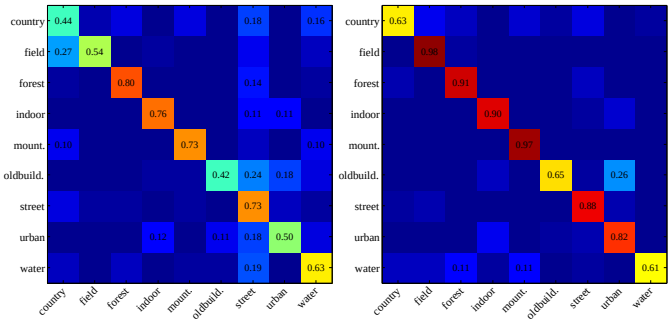
C. Domain-Adapting Object Recognition

1) Benchmark: To evaluate our framework compared with other methods for domain adaptation, we used the domain-adaptation dataset introduced by [24], where images from the same object categories are from different sources (called domains). The first three datasets were collected by [24], which include images from *amazon.com* (Amazon), taken by *digital single-lens reflex* (DSLR) and *webcam* (Webcam). The fourth dataset is Caltech-256 (Caltech). Each dataset constitutes one domain. Following the experimental settings in [16], we evaluated our approach for object recognition on four datasets, with a total of 2533 images from 10 categories.

2) Evaluation: For a fair evaluation with existing methods, we first evaluated the performance of domain-adapting image classification using the BoW feature [4]. The proposed MIIC was compared with ARC-t [20], K-SVD [64], SGF [16], GFK [35], subspace interpolation (Sub. Int.) [80], and adaptive descriptor design (ADD) [17]. Following [16], we used a

TABLE V
COMPARISON OF QUANTITATIVE EVALUATION ON RGB-NIR IMAGE CLASSIFICATION.

classifier	descriptor	dimension	mAP (%)
SVM linear kernel [66]	BoW [4]	1024d	35.65±0.7%
	GIST [26]	512d	50.74±1.3%
	ScSPM [6]	5120d	49.51±2.1%
	LLC [7]	5120d	50.11±0.2%
	VLAD [8]	8192d	62.17±0.4%
	CNN [1]	4096d	68.51±2.1%
MIIC	BoW [4]	1024d	**56.15±1.6%**
	GIST [26]	512d	**60.11±2.1%**
	ScSPM [6]	5120d	**68.58±1.9%**
	LLC [7]	5120d	**70.51±1.0%**
	VLAD [8]	8192d	**72.41±2.2%**
	CNN [1]	4096d	**87.51±2.3%**



(a) CNN+SVM [66] (b) CNN+MIIC

Fig. 10. Confusion matrix evaluations for RGB-NIR image classification.

speeded-up robust features (SURF) to extract interest points, and built BoW feature whose dimension is 800. We report the performance for eight different pairs of source and target combinations. In unsupervised settings, we randomly selected 8 labeled images per category for Webcam/DSLR/Caltech and 20 for Amazon as the source domain. To completely evaluate existing methods, we also carried out experiments in a semi-supervised setting where we additionally sampled 3 labeled images per category from the target domain. We ran 20 different trials corresponding to different selections of labeled data from the source and target domains. The average recognition rate and standard deviation are reported in Table II and Table III for unsupervised and semi-supervised setting, respectively. It is worth noting that our MIIC method does not use any domain labels; rather, it estimates the optimal modality clusters based on the training set. Compared to existing domain-adaptation methods, *e.g.*, ARC-t [20], K-SVD [64], SGF [16], GFK [35], Sub. Int. [80], and ADD [17], our MIIC method provided satisfactory image classification performance for both unsupervised and semi-supervised settings.

Furthermore, similar to the experimental settings in Sec. IV-B1, we measured the mAP while varying the feature descriptors. We evaluated our MIIC by incorporating various feature descriptors, *e.g.*, BoW [4], GIST [26], ScSPM [6], LLC [7], VLAD [8], and CNN [1]. As expected, compared to these global descriptors combined with a linear-kernel SVM, which shows limited classification performance, our MIIC method dramatically improved the classification performance when combined with any global descriptor.

TABLE VI
EVALUATION OF COMPONENT CONTRIBUTION ON UNSUPERVISED MODALITY CLUSTERING IN MIIC.

methods	Amazon	Caltech	DSLR	Webcam
Kulis et al. [79]	35.17%	45.29%	31.32%	30.72%
Gong et al. [41]	50.12%	52.78%	53.94%	50.12%
Hoffman et al. [40]	60.11%	56.72%	59.14%	60.70%
MIIC w/dic.	79.21%	59.42%	79.24%	69.12%
MIIC w/inco. dic.	80.12%	**68.11**%	71.24%	74.19%
MIIC	**82.17**%	63.97%	**92.43**%	**77.12**%

TABLE VII
EVALUATION OF COMPONENT CONTRIBUTION ON MARGINALIZED KERNELIZATION IN MIIC.

descriptors	SVM w/lin. [66]	MIIC w/mult.-lin.	MIIC w/prob.-lin.	MIIC
BoW [4]	36.15±0.7%	50.38±1.0%	52.70±1.6%	**56.15**±**1.6**%
GIST [26]	48.14±1.4%	53.71±1.1%	57.79±0.1%	**60.11**±**2.1**%
ScSPM [6]	46.18±1.1%	54.11±0.9%	62.72±1.9%	**68.58**±**1.9**%
LLC [7]	50.12±1.0%	57.88±2.1%	66.71±1.2%	**70.51**±**1.0**%
VLAD [8]	52.72±2.1%	60.72±1.2%	67.91±1.7%	**72.41**±**2.2**%
CNN [1]	64.11±2.7%	77.11±1.3%	83.72±1.7%	**87.51**±**2.3**%

D. RGB-NIR Image Classification

1) Benchmark: To evaluate our image classification method for multi-spectral images, we adopted the RGB-NIR database [25], as shown in Fig. 9. It consists of 477 images with 9 categories as follows: country, field, forest, indoor, mountain, old building, street, urban, and water. The images were processed using automatic white balancing for the RGB components, and the NIR components were equally weighted with standard gain control and gamma correction [25]. We performed image classification on this dataset of 477 images, randomly selecting 99 images for testing (11 per category) and using the remaining images for training. We repeated all our experiments using 10 trials with a randomly selected training/test ratio, following the experimental settings of [25].

2) Evaluation: Similar to the above experiments, we evaluated MIIC by incorporating various global descriptors, *e.g.*, BoW [4], GIST [26], ScSPM [6], LLC [7], VLAD [8], and CNN [1]. We compared the MIIC with SVM using a linear kernel [66]. Fig. 10 shows the confusion-matrix evaluations for the RGB-NIR image classification, and Table V shows the comparison of quantitative evaluation for the RGB-NIR image classification. As expected, the global features combined with SVM [66] provided limited performance, since the similarity from semantic categories across RGB-NIR images might be lower than the similarity within RGB (or NIR) images. Global features combined with MIIC outperformed hose combined with SVM [66].

E. Component Contribution Analysis

1) Component Contribution on Modality Clustering: In this section, we analyzed the performance gain of our unsupervised modality clustering. Our clustering method consisted of three key ingredients: modality uniqueness, dictionary, and incoherent dictionary. In this context, we analyzed the accuracy gain of our framework on the domain adaptation benchmark in Table VI, for which we know the ground truth domain labels, *e.g.*, Amazon, Caltech-256, DSLR, and Webcam. We evaluated our modality-clustering method performance compared to latent domain-clustering methods [40], [41], [79]. The performance gain of our clustering was evaluated with 1) only the dictionary (MIIC w/dic.), 2) dictionary with incoherent dictinary (MIIC w/inco. dic.), and 3) dictionary and incoherent dictinary with modality uniquess (MIIC), which is our final method. For fair evaluation of our framework, we assigned the ground truth domain labels in initialization, split the training image set into specific domains using our method, and then estimated the clustering performance on testing image sets with ground truth domain labels. As baselines [40], [41], [79], we used implementation settings similar to [40]. In these settings, our clustering framework explicitly outperformsed the conventional domain-clustering methods [40], [41], [79]. By leveraging the modality uniqueness and dictionary, our clustering framework provided reliable performance.

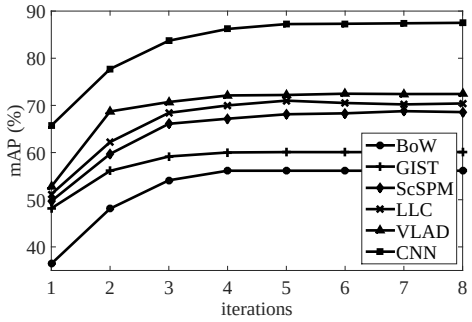


Fig. 11. Convergence analysis of our MIIC framework. MIIC was converged into globally optimal accuracy after some iterations.

2) Component Contribution on Marginalized Kernelization: In this section, to evaluate the performance gain of our modality-invariant marginalized kernelization, we compared SVM using a linear kernel [66], multiple linear kernels on modality clusters (MIIC w/mult.-lin.), probabilistic multiple linear kernels on modality clusters (MIIC w/prob.-lin.), and our kernel function (MIIC) for various global features in Table VII. We analyzed the performance using the landmark-identification benchmark [23]. For all global features, the multiple linear kernel scheme outperformed SVM with a linear kernel, which was also shown in MKL [38]. Using the probability for the modality cluster, the clustering performance was highly improved. By futher measuring the kernel function with reconstruced features on each modality cluster, our MIIC showed satisfactory performances.

3) Convergence Analysis: To evaluate the convergence of our MIIC framework as an iterative scheme, we measured the mAP on the landmark-identification benchmark [23], evolving the number of iterations. It should be noted that one iteration means that all update steps in Sec. III-C are processed. For each modality, if the optimal modality uniqueness and dictionary can be estimated, our energy function in (7) for unsupervised modality clustering cannot vary any more, which provides optimal clustering results. Fig. 11 shows the convergence analysis of our modality-clustering scheme, while varying the global feature descriptors. Our MIIC framework converged to a global minimum after some iterations.

V. Conclusion

The modality-invariant image classification (MIIC) framework was proposed for classifying images taken under varying modality conditions. Based on the observation that semantic and modality category labels simultaneously influence the image feature distribution, the modality uniqueness concept was introduced to encode each distinctive property for each modality. By leveragining this, unsupervised modality clustering and modality-invariant similarity kernel-based classifier learning were represented. The optimal cluster hypothesis and their correspoding modality uniqueness and dictionary were determined iteratively, and the modality-invariant marginalized kernel was computed based on the final clusters. Our MIIC method was validated on an extensive set of experiments. In future work, MIIC can potentially benefit from a large-scale image classification.

References

[1] K. Alex, S. Ilya, and E. H. Geoffrey, "Imagenet classification with deep convolutional neural networks," *In Proc. of NIPS*, 2012.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional neworks for visual recognition," *TPAMI*, vol. 37, no. 9, pp. 1904–1916, 2015.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *In: ICML*, 2014.

[4] G. Csurka, C. Dance, W. J. Fan, L. X., and C. Bray, "Visual categorization with bags of keypoints," *In Proc. on ECCV*, 2004.

[5] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *In Proc. of CVPR*, 2009.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *In Proc. of CVPR*, 2010.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," *In Proc. of CVPR*, 2010.

[9] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," *In Proc. of CVPR*, 2007.

[10] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *IJCV*, 2008.

[11] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *In Proc. on CVPR*, 2009.

[12] D. Z. Matthew and R. Fergus, "Visualizing and understanding convolutional networks," *In Proc. on ECCV*, 2014.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, and R. Fergus, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv:1312.6229*, 2013.

[14] A. V. K. Chatfield, K. Simonyan, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *ArXiv:1405.3531*, 2014.

[15] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," *In Proc. of ECCV*, 2014.

[16] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," *In Proc. of ICCV*, 2011.

[17] Z. Guo and Z. J. Wang, "An adaptive descriptor design for object recognition in the wild," *In Proc. of ICCV*, 2013.

[18] M. Grossberg and S. Nayar, "Modeling the space of camera response functions," *TPAMI*, vol. 26, no. 10, 2004.

[19] E. Schechtman and M. Irani, "Matching local self-similarities across images and videos," *In Proc. of CVPR*, 2007.

[20] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," *In Proc. of CVPR*, 2011.

[21] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *TOG*, vol. 30, no. 6, p. 154, 2011.

[22] G. Sun, S. Wang, X. Liu, Q. Huang, Y. Chen, and E. We, "Accurate and efficient corss-domain visual matching leveraging multiple feature representations," *The Visual Computer*, vol. 29, no. 6-8, pp. 565–575, 2013.

[23] http://diml.yonsei.ac.kr/∼srkim/MIIC/.

[24] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," *In Proc. of ECCV*, 2010.

[25] M. Brown and S. Susstrunk, "Multispectral sift for scene category recognition," *In Proc. of CVPR*, 2011.

[26] A. Oliba and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[27] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *In Proc. of CVPR*, 2006.

[28] K. E. A. van de Sande, C. G. M. Snoek, and A. W. M. Smeulders, "Fisher and vlad with flair," *In Proc. of CVPR*, 2014.

[29] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," *In Proc. of EMNLP*, 2006.

[30] H. Daumé III, "Frustratingly easy domain adaptation," *In Proc. of ACL*, 2007.

[31] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," *In Proc. of NIPS*, 2006.

[32] S. Ben-David, J. Blitzer, K. Crammer, and K. Pereira, "Analysis of representations for domain adaptation," *In Proc. of NIPS*, 2007.

[33] S. A. Winder and M. Brown, "Learning local image descriptors," *In Proc. of CVPR*, 2007.

[34] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. PAMI*, vol. 39, no. 12, pp. 2143–2157, 2009.

[35] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," *In Proc. of CVPR*, 2012.

[36] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," *In Proc. of ICCV*, 2013.

[37] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *IJCV*, vol. 109, no. 1–2, pp. 42–59, 2004.

[38] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, no. 2, pp. 2211–2268.

[39] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," *In Proc. of CVPR*, 2015.

[40] J. Hoffman, B. Kulis, T. Darrel, and K. Saenko, "Discovering latent domains for multisource domain adaptation," *In Proc. of ECCV*, 2012.

[41] B. Gong, K. Grauman, and F. Sha, "Reshaping visual datasets for domain adaptation," *In Proc. of NIPS*, 2013.

[42] N. Murray, D. Barcelona, L. Marchesotti, and F. Perronnin, "Ava: A lage-scale database for aesthetic visual analysis," *In Proc. of CVPR*, 2012.

[43] D. Keren, "Painter identification using local features and naive bayes," *In Proc. of ICPR*, 2002.

[44] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, "Impressionisom, expressionism, surrealism: Automated recognition of painters and schools of art," *ACM TAP*, vol. 7, no. 2, 2002.

[45] T. Mensink and J. V. Gemert, "The rijksmuseum challenge: Museum-centered visual recognition," *In Proc. of ICML*, 2014.

[46] X. Yan, Y. Luo, and X. Zheng, "Weather recognition based on images captured by vision system in vehicle," *In Proc. of CVPR*, 2009.

[47] J. F. Lalonde, A. Efros, and S. Narasimhan, "Estimating the natural illumination conditions from a single outdoor image," *IJCV*, 2012.

[48] C. Le, D. Lin, J. Jia, and C. K. Tang, "Two-class weather classification," *In Proc. of CVPR*, 2014.

[49] H. Y. Chong, S. J. Gortler, and T. Zickler, "A perception-based color space for illumination-invariant image processing," *ToG*, 2008.

[50] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large scale sketch-based image search," *In Proc. of CVPR*, 2011.

[51] M. Eltz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *TVCG*, vol. 17, no. 11, pp. 1624–1636.

[52] B. C. Russell, J. Sivic, J. Ponce, and H. Dessales, "Automatic alignment of paintings and photographs depicting a 3d scene," *3dRR*, 2011.

[53] G. Mark, "Mercer kernel-based clustering in feature space," *IEEE Transactions on Neural Networks*, vol. 31, no. 3, pp. 780–1285.

[54] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. IP*, vol. 17, no. 1, pp. 53–69, 2008.

[55] W. Dong, L. Zhang, and G. Shi, "Centralized sparse representation for image restoration," *In Proc. of ICCV*, 2011.

[56] G. Peyre, "Sparse modeling of textures," *Journal of Mathematical Imaging and Vision*, vol. 34, no. 1, pp. 17–31, 2009.
[57] M. Elad, B. Matalon, and M. Zibulevsky, "Image denoising with shrinkage and redundant representations," *In Proc. of CVPR*, 2006.
[58] M. Elad and A. M, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. IP*, vol. 15, no. 12, pp. 3736–3745, 2006.
[59] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. IP*, vol. 20, no. 7, pp. 1838–1857, 2011.
[60] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising - part 1: theory," *IEEE Trans. IP*, vol. 15, no. 3, pp. 539–554, 2006.
[61] ——, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising - part 2: adaptive algorithms," *IEEE Trans. IP*, vol. 15, no. 3, pp. 555–571, 2006.
[62] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," *In Proc. of CVPR*, 2008.
[63] T. Ge, Q. Ke, and J. Sun, "Sparse-coded features for image retrieval," *In Proc. of BMVC*, 2013.
[64] M. Aharon, M. Elad, and A. Bruckstein, "The k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. SP*, vol. 54, no. 11, pp. 4311–4322, 2006.
[65] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," *In Proc. of CVPR*, 2010.
[66] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167.
[67] B. Fernando, E. Fromont, D. Muselet, and M. Sebban, "Discriminative feature fusion for image classification," *In Proc. of CVPR*, 2012.
[68] T. Malisiewicz, A. Gupta, and A. Efros, "Ensemble of exemplar-svms for object detection and beyond," *In Proc. of ICCV*, 2011.
[69] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *ToG*, 2011.
[70] G. Sun, S. Wang, X. Liu, Q. Huang, Y. Chen, and E. Wu, "Accurate and efficient cross-domain visual matching leveraging multiple feature representations," *Vis. Comput.*, 2013.
[71] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictioary learning for sparse coding," *In Proc. of ICML*, 2009.
[72] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Trans. IST*, vol. 2, no. 3, pp. 1–27, 2011.
[73] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *In Proc. of NIPS*, 2006.
[74] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[75] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
[76] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comput.*, 1974.
[77] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *In Proc. of NIPS*, 2002.
[78] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," *In: ICML*, 2009.
[79] B. Kulis and M. I. Jordan, "Revisiting k-means: New algorithms via bayesian nonparametrics," *In Proc. of ICML*, 2012.
[80] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," *In Proc. of CVPR*, 2013.
[81] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," *CoRR*, vol. abs/1412.4564, 2014.
[82] https://www.flickr.com/.
[83] https://images.google.com/.
[84] https://www.bing.com/.