Bridging Weak and Strong Modalities via Adversarial Learning

Abstract

Recognizing faces across modalities is a task arising from real-world applications. Whereas there have been extensive studies on domain adaptation or cross-modality recognition, existing methods usually depend on an implicit but crucial assumption, namely, the samples in different modalities are comparably informative and they can be transformed from one modality to the other via a deterministic transform. Yet, this is not always the case in practice. Take the task of matching sketches to photos for instance, the sketch modality is often much weaker in the sense that a sketch is intuitively much less informative than a photo. We aim to tackle this problem under the imbalanced setting as described above. Specifically, we argue that the transformation from the weak modality to the strong modality is not one-to-one and thus not appropriate to be formulated as a deterministic function but can be described by a conditional distribution. Following this rationale, we propose a new framework for cross-modality recognition, where we resort to adversarial learning to derive a conditional distribution that bridges both modalities. Given a new query in the weak modality, this framework can sample multiple possible counterparts in the strong modality to approximate the conditional density, thus turning the recognition task into a density estimation problem. On large face databases, our approach outperforms various baselines in multiple metrics.

Introduction

Recent years have witnessed remarkable progress in face recognition thanks to the advances in deep learning techniques (Sun, Wang, and Tang 2014; Schroff, Kalenichenko, and Philbin 2015). However, this problem is still not completely solved. Practical application of face recognition still faces a number of challenges. For example, a face can be captured in various modalities, *e.g.* RGB images, infrared images, sketches, and 3D models. Hence, *cross-modality matching*, a challenging task where the query is presented in a modality different from that of the gallery instances, is often required in real-world practice. In this work, we aim to explore an effective way to tackle this problem.

Over the past decade, a number of studies have been done on this topic. A widely adopted strategy is to learn a *common representation* to bridge the gap between modalities (Ganin and Lempitsky 2015). This approach has been shown to be quite effective when the involved modalities are not too far apart in visual characteristics. Sharing a representation would become increasingly difficult when the modalities differ significantly. Domain adaptation methods, which consider each modality as a domain and attempt to learn an adaptation path from one to the other, also provide natural solutions to this problem. Yet, most methods developed in this category assume a one-to-one mapping between the two domains. Whereas such methods can work well in the situations where the modalities are balanced, namely the samples in both modalities are comparably informative, they would face difficulties when one modality is substantially weaker than the other, *e.g. sketches* vs. *normal photos*.

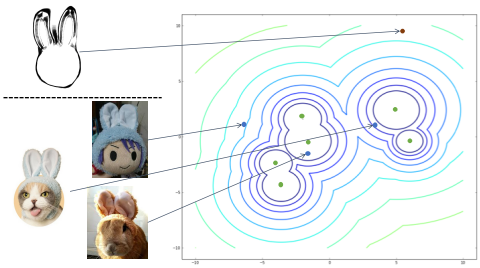


Figure 1: The illustration of bridging weak and strong modality. Left above is a sketch of a rabbit and left below are the corresponding RGB images. The blue points are the feature extracted from images, while the red point is the feature extracted directly from the sketch. It can be found that the sketch feature is far from the feature of other RGB images. Given the red point, our approach can sample a series of green points, so as to match the correct blue point in the strong modality.

In this work, we are motivated to tackle the cross-modality matching problem under an imbalanced setting – there exist a strong modality and a weak one, where the samples in the weak modality are significantly less informative than those in the strong modality. The key to this problem is still to bridge the gap between modalities, and more importantly, is how to build the bridge under the imbalanced setting as described above. In our search for an effective solution, we found that the *one-to-one assumption* that lies behind many domain adaptation methods may be flawed under this setting. As illustrated in Figure 2a, due to the lack of information, an instance in the weak modality

may have multiple distinct counterparts in the strong modality that are equally plausible. In this sense, the bridge from the weak domain to the strong one should be formulated as a *conditional distribution* instead of a *deterministic transform* like in previous literatures (Ganin and Lempitsky 2015; Tzeng et al. 2017).

Following this rationale, we propose a new framework for cross-modality matching. Given a query in the weak modality, the framework yields a conditional distribution over the strong modality, draws multiple samples therefrom to approximate the density, and performs the matching by computing the conditional density at each gallery instance. Also, inspired by the recent success of generative adversarial networks, we develop an algorithm to estimate the underlying model via adversarial learning. In this algorithm, the model is simultaneously learned to match across modalities and discriminate between classes.

In summary, the main contributions of this work lie in three aspects. First, we systematically study the cross-modality matching problem under an imbalanced setting that involves a weak modality and strong modality. Such a setting is commonly seen in practice but rarely explored in previous research. Second, we propose to bridge the weak and strong domains via a conditional distribution instead of a deterministic transform. This allows one to make multiple hypotheses when seeking an optimal match, thus improving the matching accuracy. Third, we propose a large sketch face dataset and evaluate the proposed method on several sketch face databases. Our approach outperforms previous approaches consistently in multiple performance metrics.

Related Work

Existing methods to reduce the cross-modality gap primarily focus on seeking a transform from one modality to another or a common space shared by different modalities. Besides cross-modality matching methods, the other related topic is cross-domain adaptation. In the view of formulation, these methods can be summarized into three categories below:

Traditional domain adaptation A conventional pipeline for traditional domain adaptation (Gopalan, Li, and Chellappa 2011)consists of two phases. In the first phase, a feature extractor is trained from the samples in the source modality. Then, with the feature extractor fixed, they learn a mapping from one domain to the other at the feature level. Following the remarkable success of Convolutional Neural Networks, various fine-tuning methods built on top of traditional techniques and CNNs have been explored. One representative way is to regularize a target classifier against the source classifier during training (Bergamo and Torresani 2010; Aytar and Zisserman 2011). Under the assumption that the label space of target data is a subspace of source labeled data, Tzeng *et al* (Tzeng et al. 2015) proposed to guide the training of the classifier in the target domain with source domain data. In recent work, Sun *et al* (Sun, Feng, and Saenko 2016) proposed to transform the source features to target space by aligning the second-order statistics between these two domains, in an unsupervised manner.

Common representation This kind of methods map samples from different modalities into a common latent space. Specifically, a domain-invariant representation is learned for all modalities so that the gap among modalities is reduced and representations can be measured directly. Chopra *et al* (Chopra, Balakrishnan, and Gopalan 2013) proposed to interleave samples from different domains to jointly learn the representations. This idea was followed by imposing additional loss to reduce some specific measure among different domains. Tzeng *et al* (Tzeng et al. 2014) obtained a domain-invariant representation by training with the Maximum Mean Discrepancy(MMD) which attempts to reduce the differences between two domains. Long *et al* (Long et al. 2015) further explored the idea by first embedding features in reproducing kernel Hilbert space and then applying MMD to minimize the distance between higher order statistic of two domains. Liong *et al* proposed DCML (Liong et al. 2017), which takes advantage of the pairs comprised of samples from both modalities and mapped them into a shared feature space. A recent work (Aytar et al. 2017) proposed a modality tuning method to learn aligned representations without such pair data across modalities.

Adversarial methods Another important idea is to bridge different modalities via adversarial learning. DANN (Ajakan et al. 2014; Ganin and Lempitsky 2015) is an early attempt along this line. With a gradient reversal layer, it intends to minimize the accuracy of domain discrimination while maximizing the classification performance among semantic categories. After the emergence of GAN (Goodfellow et al. 2014), some works have attempted to use GAN to generate target domain images conditioned on source domain images. Isola (Isola et al. 2017) applied conditional adversarial networks to many image-to-image translation tasks and proved that it is an effective approach for synthesizing image in another domain from its corresponding image. Zhu (Zhu et al. 2017) addressed the problem of modeling a distribution of possible output images in a conditional generative modeling setting. By training on corresponding pairs from different domains, they obtain realistic and diverse outputs at pixel-level. Recently, Tzeng *et al* (Tzeng et al. 2017) observed that modeling the image distribution is not strictly necessary to achieve domain adaptation, and proposed to optimize generated feature distribution until it is indistinguishable from the feature distribution of source domain. Nevertheless, it is devised for the purpose of learning domain-invariant feature, but not implicitly producing a distribution for matching.

Differences Our approach differs essentially from previous works in two aspects: (1) Our approach is devised to tackle the challenges arising in unbalanced cases, where the samples in one domain are significantly less informative and no corresponding pairs exist between two domains. This setting is commonly seen in real-world practice but has not been extensively explored in previous work. (2) In the view of technical formulation, given a sample in the source modality, our approach yields a conditional distribution at feature level over the target modality instead of a transformed sample.

(a) The leftmost and second left column are sketches and photos with the same identity. The other two columns contain photos with different identities but similar to the sketches. It illustrates that there may exist more than one identities in strong modality can be similar to the sketch in the weak modality.

(b) A glimpse of the training set. The left column is a sketch drawn based on the middle column. The right column is another photo of the same person. The corresponding photos used for drawing sketch will be excluded during training.

Cross-modality Matching

Problem Formulation

Cross-modality matching is a task commonly seen in real-world applications. More precisely, we consider two modalities $\mathcal{W}$ and $\mathcal{S}$ which share the same set of labels, denoted by $\mathcal{Y}$. In one modality $\mathcal{S}$, we have a gallery set of instances whose labels are known, denoted as $\mathcal{G} = \{(\mathbf{x}_1^g, y_1), \ldots, (\mathbf{x}_N^g, y_N)\}$, where N is the size of the gallery set and $y_i \in \mathcal{Y}$ is the label of the gallery instance $\mathbf{x}_i^g$. The query is in another modality $\mathcal{W}$, which is denoted by $\mathbf{x}^q$. Then the task can be defined as follows. Given a query $\mathbf{x}^q$, determine its label. Generally, this can be formulated into a matching problem as

$$\hat{y} = y_{\hat{i}}, \quad \text{with } \hat{i} = \operatorname*{argmax}_i s(\mathbf{x}^q, \mathbf{x}_i^g). \tag{1}$$

Here, we compute the *matching score* $s(\mathbf{x}^q, \mathbf{x}_i^g)$ between the query $\mathbf{x}^q$ and each gallery instance $\mathbf{x}_i^g$. Then we predict $\hat{y}$, the label of $\mathbf{x}^q$, to be the label of the gallery instance that yields the highest matching score. The key question that we intend to answer in this paper is *how to compute the matching score effectively*.

Bridge Weak and Strong Modalities

A natural way to solve the problem above is to learn a transform to map $\mathbf{x}^q$ to modality $\mathcal{S}$ and then measure the distance between the transformed query and the gallery instances in the same space. This is the approach adopted in many domain adaptation methods (Tzeng et al. 2014; Long et al. 2015; Liong et al. 2017; Aytar et al. 2017).

We aim to tackle the setting that two modalities are imbalanced. In this setting, $\mathcal{W}$ and $\mathcal{S}$ refers to weak modality and strong modality respectively, and the instances in $\mathcal{W}$ are less informative than those in $\mathcal{S}$. Such a setting is not uncommon in practice, *e.g.* sketches *vs.* photos, grayscale images *vs.* color images, and low-resolution images *vs.* high-resolution ones. Previous approaches may fail in this case due to the weak modality does not provide enough information to identify its counterparts in the strong modality. As shown in Figure 2a, for an instance in the weak modality $\mathcal{W}$, there can be more than one instances in the strong modality $\mathcal{S}$ that can match it. In other words, the transformation between $\mathcal{W}$ and $\mathcal{S}$ is *not one-to-one*, and therefore it is not appropriate to be formulated as a deterministic function.

Considering the uncertainties discussed above, it is more appropriate to formulate the relation between $\mathcal{W}$ and $\mathcal{S}$ as a conditional distribution. To be more specific, given a query $\mathbf{x}^q \in \mathcal{W}$, the corresponding counterparts in $\mathcal{S}$ constitute a conditional distribution over $\mathcal{S}$, denoted as $p_{\mathcal{S}|\mathcal{W}}(\cdot \mid \mathbf{x}^q)$. From this perspective, a natural criterion for matching between $\mathbf{x}^q \in \mathcal{W}$ and $\mathbf{x}^g \in \mathcal{S}$ is the probability of $\mathbf{x}^g$ *w.r.t.* the conditional distribution, *i.e.* $p_{\mathcal{S}|\mathcal{W}}(\mathbf{x}^g \mid \mathbf{x}^q)$.

Matching by Generation

Now, we face another challenging problem, that is, how to estimate the conditional distribution $p_{\mathcal{S}|\mathcal{W}}(\cdot \mid \mathbf{x}^q)$. For complicated samples like images, it is very difficult to formulate the distribution in a parametric form while preserving sufficient expressive power. Inspired by recent works in *Generative Adversarial Networks (GAN)* (Goodfellow et al. 2014; Isola et al. 2017; Zhu et al. 2017; Arjovsky, Chintala, and Bottou 2017), we explore an generative approach. The basic idea is to learn a conditional generator G with adversarial learning, and draw multiple samples therefrom to approximate the density $p_{\mathcal{S}|\mathcal{W}}$ using the Parzen-window method.

Moreover, this is a discriminative problem – our goal is not to generate plausible images. Hence, we can perform the generation at the feature level instead of pixel level. In this work, we choose to perform the cross-modality generation for top-level features, *i.e.* those derived from the last convolution layer of feature extractor. The benefits of this choice lie in two aspects: 1) High-level features usually have lower dimension. Therefore learning a generator for them is generally easier and less costly. 2) High-level features are closer to the semantic space. At this level, the gap between modalities is smaller and thus easier to bridge.

Let $F_{\mathcal{W}}$ denotes the CNN for extracting features for the weak modality $\mathcal{W}$, $F_{\mathcal{S}}$ denotes the CNN for the strong modality $\mathcal{S}$. With a learned conditional generator G, the matching can be done as follows. The gallery features are denoted as $\{\mathbf{f}_i^g\}_{i=1:N}$ with $\mathbf{f}_i^g = F_{\mathcal{S}}(\mathbf{x}_i^g)$. Given a query $\mathbf{x}^q \in \mathcal{W}$, we first compute its feature as $\mathbf{f}^q = F_{\mathcal{W}}(\mathbf{x}^q)$, and then sample multiple counterparts of $\mathbf{f}^q$ in the feature space of $\mathcal{S}$, which are denoted by $\{\mathbf{g}_j\}_{j=1:m}$ with $\mathbf{g}_j = G(\mathbf{f}^q, \mathbf{z}_j)$. Here, $\mathbf{z}_1, \ldots, \mathbf{z}_m$ are random vectors independently sampled from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

With this set of *generated counterparts*, we can approximate the conditional distribution $p_{\mathcal{S}|\mathcal{W}}(\cdot|\mathbf{f}^q)$ using the Parzen window method as

$$p_{\mathcal{S}|\mathcal{W}}(\mathbf{f}|\mathbf{f}^q) \simeq \frac{1}{m}\sum_{j=1}^{m}\phi(\mathbf{f}, \mathbf{g}_j), \tag{2}$$

where ϕ is a density kernel function given by

$$\phi(\mathbf{f}, \mathbf{g}_j) \propto \exp\left(-\frac{\|\mathbf{f} - \mathbf{g}_j\|^2}{2\sigma^2}\right). \tag{3}$$

Here, σ is a decision parameter that controls the width of each density kernel. Therefore the *matching score* $s(\mathbf{x}^q, \mathbf{x}_i^g)$ can be defined by $p_{\mathcal{S}|\mathcal{W}}(\mathbf{f}_i^g|\mathbf{f}^q)$, *i.e.* the value of this conditional probability.

Learning

The cross-modality matching model comprises two feature extractors $F_{\mathcal{W}}$ and $F_{\mathcal{S}}$ and a cross-modality generator G. To learn the conditional generator G via adversarial learning, we introduce a discriminator D to discriminate between real features and the fake ones. To enhance the discriminative power of the learned feature, we introduce a classifier C to provide supervisory signals. To diversify the generated feature, a latent code encoder E is devised to recover the input random vectors. In this section, we first present the objective functions for model learning, and then describe the end-to-end training process in our approach - Weak-Strong GAN(WSGAN).

Objectives

The objective functions of our architecture include three components: the adversarial loss, the classification loss and the noise regression loss.

Adversarial loss. To learn the conditional generator G, we adopt the Wasserstein loss (Arjovsky, Chintala, and Bottou 2017), which has been shown to be quite effective in generative learning. The objective is thus expressed as:

$$\mathcal{L}_A^d = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{W}}\Big[D\Big(G\big(F_{\mathcal{W}}(\mathbf{x}),\mathbf{z}\big)\Big)\Big] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{S}}\Big[D(F_{\mathcal{S}}(\mathbf{x}))\Big], \quad (4)$$

where D is the discriminator to tell between the real features in $\mathcal{S}$ and the generated ones. In adversarial learning, we minimize Eq.(4) *w.r.t.* D and maximize the first term in Eq.(4) *w.r.t.* G. Here, the first term is denoted by $\mathcal{L}_A^g$. This way encourages G to learn the feature distribution of the strong modality. On the other hand, the process of minimizing $\mathcal{L}_A^d$ improves D's discriminative power.

Classification loss. Although the adversarial loss pushes the marginal distribution of the generated features to be closer to that of the real features in $\mathcal{S}$, it does not necessarily align individual classes. Therefore, we explicitly introduce classification loss to enhance the discriminative power *between classes*, as follows.

$$\mathcal{L}_C^g = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathcal{W}}}\Big[l\big(G(F_{\mathcal{W}}(\mathbf{x}),\mathbf{z}),y;\mathbf{W}_C\big)\Big], \quad (5)$$

$$\mathcal{L}_C^d = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}_{\mathcal{S}}}\Big[l\big((F_{\mathcal{S}}(\mathbf{x}),y;\mathbf{W}_C\big)\Big]. \quad (6)$$

Here, we introduce a fully-connected layer to connect the features in $\mathcal{S}$, be it real or generated, to class labels. This layer in itself comes with a coefficient matrix $\mathbf{W}_C$. In the formulation above, l is the standard cross-entropy loss, $\mathcal{D}_{\mathcal{W}}$ and $\mathcal{D}_{\mathcal{S}}$ are respectively the labeled training sets in weak and strong modalities.

Noise Regression loss. As our condition contains strong identity information, G is easy to fall into a local minimum where G ends up as an identity mapping therefore ignoring the latent vector $\mathbf{z}$.

$$\mathcal{L}_E = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{W},\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big|\Big|E\Big(G\big(F_{\mathcal{W}}(\mathbf{x}),\mathbf{z}\big)\Big) - \mathbf{z}\Big|\Big|_2, \quad (7)$$

The latent code regression, introduced in (Chen et al. 2016), attempts to recover the latent vector from the output feature. With a latent code $\mathbf{z}$ randomly drawn from the multivariate Gaussian distribution, we use an encoder E to encourage the generator network to keep the information from $\mathbf{z}$.

Overall objective. In our algorithm, the training of generator and the training of discriminator are respectively done in two different phases, namely the *G-phase* and the *D-phase*. Incorporating adversarial loss, classification loss and noise regression loss above, we can derive the two joint objective functions, one for a phase, as:

$$\begin{aligned}\mathcal{J}_G(G,F_{\mathcal{W}}) = &-\mathcal{L}_A^g(G,F_{\mathcal{W}}) + \lambda\mathcal{L}_C^g(G,F_{\mathcal{W}},\mathbf{W}_C)\\ &+\mu\mathcal{L}_E(G,F_{\mathcal{W}},E),\end{aligned} \quad (8)$$

$$\begin{aligned}\mathcal{J}_D(D,E,F_{\mathcal{S}},\mathbf{W}_C) = &\mathcal{L}_A^d(D,F_{\mathcal{S}},G,F_{\mathcal{W}}) + \nu\mathcal{L}_C^d(F_{\mathcal{S}},\mathbf{W}_C)\\ &+\mu\mathcal{L}_E(G,F_{\mathcal{W}},E).\end{aligned} \quad (9)$$

Here λ, ν and μ are weights that control the relative importance of different losses. Both objective functions are minimized during training.

End-to-End Training

As shown in Fig. 3, the training procedure alternates between D-phase and G-phase respectively for learning the discriminator D and generator G. The other components are also updated along the way. The whole process can be optimized *end-to-end* via back-propagation, which allows us to reduce the distribution gap while simultaneously exploiting class labels to enhance the discriminative power.

The training in D-phase is driven by the objective $\mathcal{J}_D$ in Eq.(9). In this phase, to train the discriminator D, the feature extracted from training samples in the strong modality are taken as the real samples while those from the fixed G are taken as fake ones. $F_{\mathcal{S}}$ and D are optimized with gradients from both strong and weak modality samples. The classifier $\mathbf{W}_C$ and latent code encoder E are updated by the samples from the strong modality and weak modality respectively.

The training in G-phase is driven by the objective $\mathcal{J}_G$ in Eq.(8). During the training of G, only samples from the weak modality are involved. Conditioned on the feature extracted from these samples, G generates a series of *counterparts* in $\mathcal{S}$ by feeding G with random vectors. With the adversarial loss from the fixed D, the latent code regression loss from the fixed E and the classification loss from the fixed $\mathbf{W}_C$, G and $F_{\mathcal{W}}$ are optimized jointly. They are supervised to generate feature that are both similar to real distribution and discriminative in the strong modality. In this phase, the classifier $\mathbf{W}_C$ is fixed to avoid the undesirable influence from the generated features which are not always reliable especially in early stages.

Technical Details

To better optimize the architecture, different components in the framework are updated with different optimizers. The $F_{\mathcal{W}}$, $F_{\mathcal{S}}$, $\mathbf{W}_C$ and E are optimized by SGD, while the G, D are updated via RMSProp.as suggested in WGAN (Arjovsky, Chintala, and Bottou 2017). To balance the progress in both phases, D-phase is invoked 5 times more frequently than G-phase. Moreover, we clip the weights of discriminator to be in the range of $[-0.01, 0.01]$, which is consistent with WGAN. For balancing the influence of different losses, λ and ν is set to 0.1 and μ is set to 0.01 in our experiments.

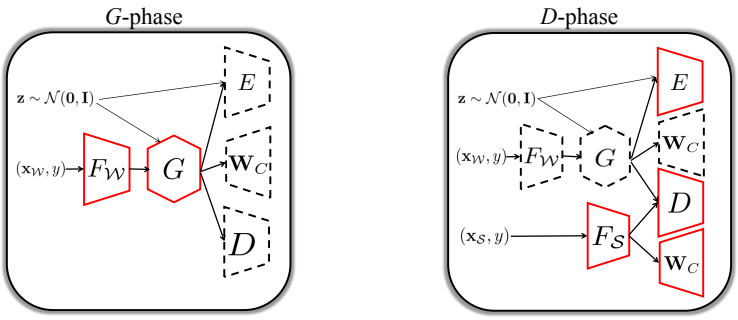


Figure 3: An overview of our proposed architecture. In D-phase, D learns the distribution difference from samples from both modalities. $\mathbf{W}_C$ is trained with informative samples from strong modality to provide reliable discriminative signals for samples from weak modality. In G-phase, the network are trained from adversarial and classification signals jointly. The component that are active for updating is marked by red solid box, while the dotted box indicates that the parameters of the component is fixed in this phase.

Experiments

We test our method on large-scale face databases for two tasks, namely face retrieval and verification. We not only compare our approach with various methods, but also study some crucial design choices, via a series of ablation studies.

Experimental Settings

Datasets To evaluate our method, we constructed a data set with two modalities, namely *images* and *sketches*. They are respectively treated as the *strong* and *weak* modalities in our formulation. The dataset is constructed on two public face databases, *MS-Celeb-1M* and *LFW*. *MS-Celeb-1M* (Guo et al. 2016) is a large-scale face dataset, which contains $100K$ identities, each with about 100 facial images. This is used as the source of our training set. *LFW* (Huang et al. 2007) is the *de facto* standard testing set for face verification, which contains $13,233$ facial images from $5,749$ identities. This is used as the source of our testing set. We chose $10K$ and $1K$ identities from these two datasets respectively and ensure each identity has more than 2 images. We then selected one image from each identity and asked painters to draw a corresponding sketch for each image, obtaining $10K$ sketches of *MS-Celeb-1M* and $1K$ sketches of *LFW*.

The training and testing sets were organized as follows to facilitate the research in cross-domain matching. The training set was selected from *MS-Celeb-1M* and the sketches derived therefrom. Particularly, we selected $48K$ classes, containing the drawn $10K$ classes, as the training samples in the strong modality. The drawn sketches are used as the training samples in the weak modality. These two modalities differ significantly in the number of images ($2.4M$ *vs.* $10K$) and the average number of images per identity (67 *vs.* 1). Note that the original images used as the source of the sketches were excluded from the strong modality, so as to avoid corresponding pairs of images and sketches that are directly aligned. The testing set, derived from *LFW*, contains a gallery set of $12,233$ images in the strong modality and a query set of $1,000$ images in the weak modality. We divided *LFW* into three parts: (1) For identities with only one image, we kept their original images. (2) For identities which have drawn sketches, we took the corresponding sketches as the query samples in the weak modality. (3) All the remaining images, except those serving as the sources of sketches, were also incorporated into the gallery set. Note that the parts (2) and (3) share the same $1,000$ identities.

Besides the large dataset collected by ourself, we also conduct experiments on existing datasets, namely PRIP-HDC, CUFS and CUFSF. PRIP-HDC (Klum et al. 2014) contains 47 pairs of hand-drawn composites-mugshot pairs, which are drawn based on the verbal description by the eyewitness. CUFS (Wang and Tang 2009) has 311 pairs in total, including 188 faces from the Chinese University of Hong Kong(CUHK) student database and 123 faces from the AR database (Martinez 1998). CUFSF (Zhang, Wang, and Tang 2011) comprises $1,194$ persons from the FERET database (Phillips et al. 2000). Except a face photo from FERET, each person has a sketch with shape exaggeration drawn by an artist when viewing the photo.

Evaluation. We consider two tasks in our evaluation:
(1) Face retrieval. In face retrieval, for each query in weak modality, we select top k identities from the gallery. The performance is measured by *recall@K*, *i.e.* the fraction of predictions where the true identity occurs in the top-k list.
(2) Face verification. For cross-modality face verification, we are asked to determine whether a query in the weak modality has the same identity as another sample in the strong modality. We form the set of testing pairs, where each pair comprises one sample from the query set and the other from the gallery set. The testing model will give a matching score for each pair. We use a metric widely adopted in practice, namely the *true positive rate* under the condition that the *false positive rate* is fixed to be 0.01 or 0.001.

Network architecture details. Our feature extraction network is based on ResNet-50, with input size reduced to 112x112 and a 256-dimension feature added before the classifier. Since the input of both modalities is facial images, the feature extraction network is shared between two modalities. The generator and discriminator used in our architecture are simply a fully-connected network with two hidden layers, each followed by BN (Ioffe and Szegedy 2015) and LReLU (He et al. 2015). In all settings, we set the mini-batch size to 512 and the learning rate to 0.005.

Method Comparison

We compare the proposed method with a series of baselines. These methods are briefly described below.

(1) Pretraining. This scheme trains the CNN-based feature extractor, using *only* the samples in the strong modality. The trained feature extractor will be used to extract features for the samples in both modalities. It serves as a reference to demonstrate the gap between the two modalities.

(2) Jointly training (Chopra, Balakrishnan, and Gopalan 2013), a classical method used for learning common representation. It trains the model from scratch on a mixed dataset that comprises all training samples from both modalities.

(3) Fine-tuning, another popular method to learn shared representation. It first pretrains on the strong modality, and then fine-tunes on either the weak modality or both. We consider three different schemes for fine-tuning: *(a) Fixed Fine-tuning*, which fine-tunes on the weak modality, updating only the last two layers with others fixed, *(b) End-to-end Fine-tuning*, which fine-tunes on the weak modality in an end-to-end manner,*(c) Joint Fine-tuning*, which fine-tunes on *both* modalities in an end-to-end manner.

(4) DANN (Ganin and Lempitsky 2015), an adversarial learning method to learn a shared feature representation via a gradient reversal layer. Following the original unsupervised design, *DANN(U)*, only domain label is used for samples in the weak domain. This framework can be easily adapted to incorporate class labels for both domains. We refer to the adapted version as *DANN(S)*.

(5) SDT(Tzeng et al. 2015), simultaneously optimizes domain invariance feature via domain confusion loss, while transferring information from source domain to targeted domain in the form of a cross entropy soft label loss.

(6) JAN(Long et al. 2017), reduces the distribution distance between features from two modalities by aligning the joint distribution of activations from multiple layers. The original method, denoted as *JAN(U)*, is devised for unsupervised adaptation. With class labels on both modalities, we adapt the method to supervised scenario and refer to it as *JAN(S)*.

(7) Cross-modal (Aytar et al. 2017), a simple yet effective way to capture modality-specific information and obtain a shared representation. The method consists of two stages. It firstly initializes all modalities network with the model pretrained on strong modality. With the shared high-level layers fixed, it trains on all modalities by classification for a given number of iterations. In the second phase, it trains in an end-to-end manner with the statistical regularization on the shared layers. We refer to the result of two stages as *Cross-modal(A)* and *Cross-modal(A+B)* respectively.

Results As described in previous section, we train all models on *MS-Celeb-1M* and test them on different benchmarks, namely, a testing protocol on *LFW* designed by us and three existing sketch face datasets with different kinds of sketches. Tab. 1 and Tab. 2 show some quantity results for different methods. Here, the *performance* is measured by different cross-modality matching metrics.

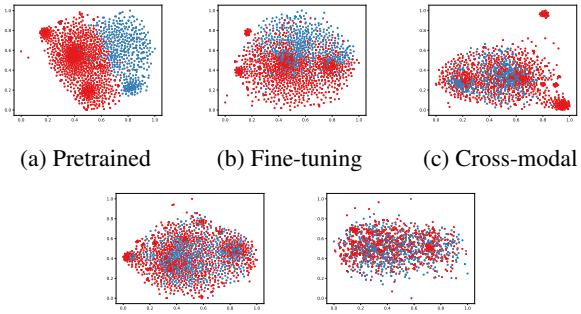
The results show: (1) For *Pretraining*, lacking information about the other modality harms the cross-modality matching. (2) For *Jointly Training* and different schemes of *Fine-tuning*, utilizing classification signals greatly boost the performance. Notice that *Fixed Fine-tuning* may be limited by the expressive power of the model since it fails to capture the modality-specific information from low-level layers of the network. While *End-to-End Fine-tuning* is likely to suffer from overfitting problem due to the limited amount of labeled data. (3) *DANN(U)* only learns the marginal distribution of weak modality and does not perform very well. On the other hand, *DANN(S)* works much better than the unsupervised version, which indicates that class information from the weak modality may be a crucial factor for cross-modality matching. (4) *SDT* simultaneously optimize modality invariance and aligns classes between modalities. It reduces the modalities gap to some extent, but the performance may be limited by ignoring the unbalanced informativeness between modalities. (5) Compared with *DANN(U)*, although *JAN(U)* learns better marginal distribution, lacking of class labels also results in inferior cross-matching performance. Once exploiting class labels, as *JAN(S)* demonstrates, it enhances discriminative power by a large margin. (6) *Cross-modal* employs two networks to learn specific low-level representations while sharing high-level representations. The advantage over other baselines indicates that applying specific network to capture low-level information may be important for learning aligned cross-modality representation. (7) *WSGAN* (Our method) can benefit from generating a set of feature from the learned generator. For each

Method	Recall@K		TPR@FPR	
	5	50	1e-3	1e-2
Pretraining	0.382	0.677	0.387	0.671
Jointly Training	0.538	0.796	0.538	0.791
Fixed Fine-tuning	0.51	0.777	0.486	0.761
End-to-End Fine-tuning	0.483	0.738	0.448	0.721
Jointly Fine-tuning	0.43	0.718	0.433	0.718
DANN(U)	0.413	0.695	0.389	0.675
DANN(S)	0.502	0.771	0.513	0.768
SDT	0.549	0.786	0.531	0.78
JAN(U)	0.486	0.757	0.457	0.738
JAN(S)	0.545	0.802	0.536	0.792
Cross-modal(A)	0.548	0.797	0.531	0.789
Cross-modal(A+B)	0.545	0.801	0.534	0.793
WSGAN†	**0.567**	0.816	0.542	**0.803**
WSGAN	0.561	**0.821**	**0.545**	**0.803**

Table 1: Comparison of cross-modality matching among different methods. WSGAN† denotes our approach without applying noise regression during training.

Method	PRIP-HDC (47 Pairs)		CUFS (311 Pairs)		CUFSF (1194 Pairs)	
	1	5	1	5	1	5
Pretraining	0.15	0.34	0.63	0.83	0.393	0.585
SDT	0.19	0.404	0.79	0.92	0.493	0.706
JAN(S)	0.19	0.447	0.75	0.94	0.499	0.708
Cross-modal	0.17	**0.458**	0.73	0.92	0.478	0.689
WSGAN	**0.21**	0.443	**0.87**	**0.98**	**0.502**	**0.721**

Table 2: Recall@K on different face sketch databases.



(a) Pretrained (b) Fine-tuning (c) Cross-modal

(d) WSGAN(m=1) (e) WSGAN(m=10)

Figure 4: The figure shows t-SNE visualizations of feature from two modalities. The blue points refer to the feature of sketches and the red points refer to the feature of photos. Although retrieval performance of (c) and (d) is close, the feature distributions are different. (c) covers a large number of real points but still miss some groups of samples. As for (d), the points of both modalities distribute evenly in the feature space. By drawing more counterparts, (e) illustrates more similar distribution to the real one.

Method	Recall@K		TPR@FPR	
	5	50	1e-3	1e-2
$\mathcal{L}_C$	0.552	0.781	0.534	0.785
$\mathcal{L}_A + \mathcal{L}_C (m = 1)$	0.563	0.816	0.541	0.802
$\mathcal{L}_A + \mathcal{L}_C (m = 10)$	0.567	0.816	0.542	0.803
$\mathcal{L}_A + \mathcal{L}_C + \mathcal{L}_E (m = 1)$	0.554	0.815	0.538	0.796
$\mathcal{L}_A + \mathcal{L}_C + \mathcal{L}_E (m = 10)$	0.561	0.821	0.545	0.803

Table 3: Influence of objectives on cross-modality matching.

query image, we generate 10 strong modality feature and report the average result of 5 times testing to avoid the influence of input random vectors. The consistent performance margin on different benchmarks indicates the advantage of our approach. Fig. 4 qualitatively shows that the generated feature of WSGAN has close distribution to the real one.

Ablation Studies

In this section, we study how different factors influence the performance of our approach.

Influence of different objectives Table 3 shows the performance influenced by different losses. (1) Although the adversarial loss itself can not enhance the discriminative power, it brings a consistent gain about 2% when jointly trained with classification loss. It indicates that the adversarial loss provides useful signal for shaping the overall distribution when learning discriminative feature. (2) However, the model only trained by adversarial loss and classification loss explore a limited generated space. As shown in Table 3, even the number of generated feature is 10 times larger, the performance improves slightly. Adding the noise regression loss forces the generator to cover more possible latent space around an identity, and thus benefiting the final result from the increasing number of generated feature.

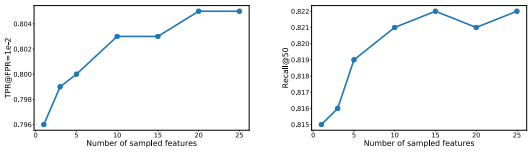


Figure 5: Performance vs. The number of sampled feature m

	intra-sim	inter-sim
Pretraining	-	0.159±0.11
$\mathcal{L}_A + \mathcal{L}_C$	0.972±0.009	0.037±8e-5
$\mathcal{L}_A + \mathcal{L}_C + \mathcal{L}_E$	0.755±0.054	0.048±0.002
Real	0.696±0.068	0.054±0.002

Table 4: Variance of generated feature on LFW.

Variance of generated feature Two metrics are used to analyze the variance of the generated feature, namely *intra-sim* and *inter-sim*. The *intra-sim* computes the average cosine similarity of generated feature from the same identity, while *inter-sim* calculates the average cosine similarity of generated feature between identities. Note that each identity has only one sketch, deterministic methods like *Pretraining* cannot compute the *intra-sim*. The variance of feature from gallery photos is used as a reference value of real distribution. As shown in Table 4, when training without noise regression, our method closes the gap of *inter-sim* between pretrained model and real distribution The model trained with all objectives increases the intra-variance of output feature, approximating the real distribution on both metrics.

The number of sampled feature m With a learned generator G, given a sample in the weak modality, multiple different counterparts can be sampled in the strong modality. Here, we study how m, the number of sampled feature, affects the performance. As shown in Fig. 5, the performance improves as m increases. This clearly suggests the importance of viewing the bridge as a conditional distribution instead of a deterministic transform. As m increases beyond 10, the performance gradually saturates.

Conclusions

We presented a study on cross-modality matching under the condition that the samples in one modality are less informative than those in the other. This is a setting commonly arising from practical applications, but has rarely been explored in previous works. To tackle the problem, we proposed a new formulation, which considers the bridge between the weak and the strong modalities as a conditional distribution instead of a deterministic transform. We learn a generator for producing counterparts in the strong modality for each given query in the weak modality via adversarial learning, thus converting the matching problem into density computation. On large-scale face datasets, our approach outperformed a series of state-of-the-art methods.

References

Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; and Marchand, M. 2014. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Aytar, Y., and Zisserman, A. 2011. Tabula rasa: Model transfer for object category detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2252–2259. IEEE.

Aytar, Y.; Castrejon, L.; Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2017. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*.

Bergamo, A., and Torresani, L. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in neural information processing systems*, 181–189.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2172–2180.

Chopra, S.; Balakrishnan, S.; and Gopalan, R. 2013. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189.

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

Gopalan, R.; Li, R.; and Chellappa, R. 2011. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 999–1006. IEEE.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 87–102. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint*.

Klum, S. J.; Han, H.; Klare, B. F.; and Jain, A. K. 2014. The facesketchid system: Matching facial composites to mugshots. *IEEE Transactions on Information Forensics and Security* 9(12):2248–2263.

Liong, V. E.; Lu, J.; Tan, Y.-P.; and Zhou, J. 2017. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia* 19(6):1234–1244.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 97–105.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*.

Martinez, A. M. 1998. The ar face database. *CVC Technical Report24*.

Phillips, P. J.; Moon, H.; Rizvi, S. A.; and Rauss, P. J. 2000. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence* 22(10):1090–1104.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *AAAI*, volume 6, 8.

Sun, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1891–1898.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Tzeng, E.; Hoffman, J.; Darrell, T.; and Saenko, K. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4068–4076.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. *arXiv preprint arXiv:1702.05464*.

Wang, X., and Tang, X. 2009. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(11):1955–1967.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 513–520. IEEE.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 465–476.