This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Adaptive Cross-Modal Prototypes for Cross-Domain Visual-Language Retrieval

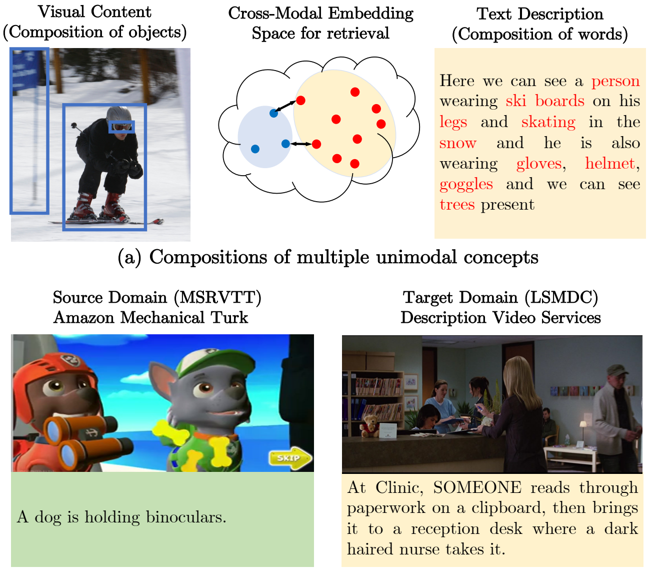
Yang Liu[1,3*] Qingchao Chen[2,4*] Samuel Albanie[3]
[1] Wangxuan Institute of Computer Technology, Peking University
[2] National Institute of Health Data Science, Peking University
[3] Visual Geometry Group, University of Oxford
[4] Department of Engineering Science, University of Oxford
qingchao.chen@eng.ox.ac.uk, {yangl,albanie}@robots.ox.ac.uk

Abstract

In this paper, we study the task of visual-text retrieval in the highly practical setting in which labelled visual data with paired text descriptions are available in one domain (the "source"), but only unlabelled visual data (without text descriptions) are available in the domain of interest (the "target"). We propose the ADAPTIVE CROSS-MODAL PROTOTYPES *framework which seeks to enable target domain retrieval by learning cross-modal visual-text representations while minimising both uni-modal and cross-modal distribution shift across the source and target domains. Our approach is built upon two key ideas: first, we encode the inductive bias that the learned cross-modal representations should be compositional with respect to concepts in each modality—this is achieved through clustering pretrained uni-modal features across each domain and designing a careful regularisation scheme to preserve the resulting structure. Second, we employ mutual information maximisation between cross-modal representations in the source and target domains during learning—this provides a mechanism that preserves commonalities between the domains while discarding signal in each that cannot be inferred from the other. We showcase our approach for the task of cross-domain visual-text retrieval, outperforming existing approaches for both images and videos.*

1. Introduction

Large-scale datasets of visual content paired with corresponding text descriptions have driven recent advances in cross-modal retrieval tasks such as image-text retrieval and video-text retrieval. In the last few years, approaches trained with such data have achieved a steady and significant improvement under retrieval tasks within the single-domain setting (in which training and inference take place on the same domain). However, in real-world applications, manual collection of paired visual content and text descriptions is a labour-intensive and time-consuming process, creating a significant barrier for the application of cross-modal retrieval methods to new domains.


Visual Content
(Composition of objects)
Cross-Modal Embedding
Space for retrieval
Text Description
(Composition of words)
Here we can see a person wearing ski boards on his legs and skating in the snow and he is also wearing gloves, helmet, goggles and we can see trees present
Source Domain (MSRVTT)
Amazon Mechanical Turk
Target Domain (LSMDC)
Description Video Services
SKIP
A dog is holding binoculars.
At Clinic, SOMEONE reads through paperwork on a clipboard, then brings it to a reception desk where a dark haired nurse takes it.

(a) Compositions of multiple unimodal concepts

(b) Cross domain shift

Figure 1: **Three challenges of cross-domain visual language retrieval.** *Top (a)*: Retrieval systems must be capable of generalising to novel compositions of multiple concepts, represented in both the visual and text domains; *Bottom (b)*: Retrieval systems must be robust to significant cross-domain shifts in both the visual and text distributions; *Top and bottom ((a) and (b))*: Retrieval systems must account for "reporting bias" (across both images and videos) in which only a subset of visual concepts are described in the corresponding text.

In this paper, we investigate the pragmatic question of how we can best learn knowledge on the "source" domain with paired data to generalize to other "target" domains

[*] Equally contributed first and corresponding authors.

without the prohibitive cost of additional data collection. Such a study sheds light on how well machines can understand visual and textual information in their generality, rather than learning and exploiting with domain-specific knowledge of the pairing.

The task of transferring a model that has been learned on a labelled source domain to an unlabelled target domain is known as Unsupervised Domain Adaptation (UDA). There has been a great deal of progress in this vein for uni-modal analysis, i.e., image classification [41], image segmentation [59], text sentiment classification [51], etc. However, relatively few works have attempted UDA for cross-modal tasks involving vision and free-form natural language descriptions—the topic we study in this paper.

To prosper in the UDA setting, a visual-text retrieval model must address three challenges (shown in Fig. 1):
(1) *Compositionality*. The model needs to encode complex semantic features with compositions of multiple visual entities (multiple words) as well as their relationships (as shown in Fig 1(a), which depicts an image from MS COCO [36] with its corresponding description provided by [53]). Effective retrieval on the target domain requires representations that enable novel combinations of visual entities and text which may not have been observed in the source domain.
(2) *Reporting bias*. Retrieval requires the model to solve a challenging set-to-set cross-modal matching problem (where multiple visual entities correspond to various words contained in free-form sentences), in which information across modality is only partial matched [28]. Examples of this effect can be seen in Fig 1(a) and Fig 1(b). Even for relatively dense descriptions such as the one associated to the image in Fig 1(a), the description is not exhaustive (in this case, the flag to the left of the skier is not described).
(3) *Visual and text domain shifts*. The retrieval model must be robust to domain shift in both visual content and written descriptions. For example, consider Fig 1(b), where we observe samples from strikingly different visual domains (cartoons and movies sourced from [70] and [55], resp.). In addition to "visual shifts", valid text queries can differ significantly in detail and manner: while both describe videos, the description on the left describes a single interaction while the description on the right conveys an ongoing set of interactions between people, objects and their environment.

To tackle these challenges, we propose the Adaptive Cross-Modal Prototypes (ACP) framework. The two key ideas underpinning this framework as follows. (1) To address the need for compositionality and achieve robustness to reporting bias, we propose to learn a cross-modal representation with carefully designed regularisation. Since data samples for text-video retrieval lack a natural discrete semantic class structure (unlike traditional UDA for classification, in which each visual input is mapped to one or more finite predefined categories), we first perform clustering on off-the-shelf uni-modal embeddings for visual content in the target domain and text in the source domain. We then attach prototypical networks to the cross-modal representation and task them with predicting, for each sample, the assignment probability of its uni-modal embedding to each of the cluster centres for samples within the same modality. The goal is to ensure that the relationships between categories discovered by the clustering are not lost in the cross-modal representation when it is trained with paired data on the source domain.This design is inspired in part by recent works highlighting the powerful generalisation capabilities of pretrained vision models to out of distribution data [40] and the ability of large-scale language models as few-shot learners [4], suggesting that knowledge of a vast array of concepts are likely already encoded among these features in a manner that enables their composition. (2) To minimise the influence of visual and text distribution shifts across domains, we employ mutual information maximization [29] between the predictions of the prototypical networks on the source and target domains. This aims to preserve commonalities between the domains while discarding signal in each that cannot be inferred from the other.

The contributions of this paper are as follows: (1) We propose a new framework, Adaptive Cross-Modal Prototypes, for cross-modal retrieval in the UDA setting by preserving semantic structure of compositional concepts from uni-modal data; (2) We demonstrate that maximising mutual information of the co-occurrence between source and target cross-modal prototype cluster assignment prediction is an effective mechanism to reduce domain shifts for both visual and text data; (3) Our method achieves improvements on three image-retrieval datasets and three video-retrieval datasets compared to a retrieval system trained only on the source domain, as well as alternative domain adaptation strategies, such as variants of maximum mean discrepancy [42], adversarial learning strategy [24] and transportation modelling [17].

2. Related Work

Visual-language cross-modal retrieval. A number of works on cross-modal *joint embedding* learning methods [33, 34, 35] have proven their effectiveness for image-text retrieval. There have also been extensions to video-text retrieval by employing temporal features [20, 47, 13, 50] in addition to other sensory modalities (such as sound [43, 44, 66, 21]). We employ the recent method of [37] as a testbed for our approach, motivated by its solid performance on several benchmarks. An important distinction with prior work such as that of [38] is that we consider the problem cross-domain visual-text task in which no paired video-text data is available on the target domain.
Domain adaptation in uni-modal applications: Several unsupervised domain adaptation techniques have been ex-

plored to align the cross-domain feature distribution, i.e., maximum mean discrepancy[40, 42], adversarial learning strategy[23, 41, 7, 10] and transportation modeling [12, 17]. More recently, advanced approaches are particularly designed for image classification [41, 42, 7, 48] and for image-image retrieval [52, 31, 32, 5, 22] and video classification [46]. Among them, [52] used natural language text to regularise an image-image retrieval system. These methods either rely on finite, predefined and discrete categorical labels or focus on uni-modal retrieval problem. Domain adaptation has also been an important research theme in the natural language processing community, leading to the development of approaches using auto-encoders [27, 9, 16], self-training [57, 56] and intelligent data selection [45, 65]. Differently from the works described above, our method is *cross-modal* in nature (linking visual content with natural language), which is more challenging due to the heterogeneous gap between different modalities.

Domain adaptation in cross-modal applications: Few works have considered deep UDA for cross-modal tasks. Among them, [6] investigated the problem of cross-dataset adaptation for visual question answering. [14] consider the problem of domain adaptation for image captioning (and hence a natural language output space). However, their approaches have to use both visual and text samples in the target domain (though without pairing information) and hence is *not* applicable to our setting where target domain language descriptions are not available. Recently, [11] introduce a UDA benchmark for video-text retrieval and study this task with a pseudo-labelling approach. We compare our approach with theirs in Sec. 4.

Robust domain adaptation with prototypes: Some works have revisited the classical technique of *self-labelling* [60, 71] for unsupervised domain adaptation with visual data via structured transduction losses [61] and self-training with multiple networks [58]. To avoid local optima associated with early mislabelling, techniques choose to improve robustness using pseudo-label prototypes [69, 49, 48, 7, 74] with moving averages. In particular, similarly to [48], we making use of clustering and mutual information maximisation. However, differently from each of these works (which classify images or pixels into a finite set of categories), we develop learned cross-modal prototypes that can preserve either visual or text data structures and leverage them to reduce cross-domain discrepancies. Our construction method of cross-modal prototypes is directly applicable to a natural language label space (Sec. 3.3).

3. Method

3.1. Problem Formulation

We consider the problem of learning a shared cross-modal embedding for visual content (such as an image or video) and natural language descriptions of the content in an Unsupervised Domain Adaptation (UDA) setting. Specifically, we assume access to a source domain $\mathcal{S} = \{v^s, \ell^s\}$ of paired visual samples and natural language fragments and a target domain $\mathcal{T} = \{v^t\}$ of unpaired visual samples. We further assume that the source and target domains are sampled from joint distributions $P(v^s, \ell^s)$ and $Q(v^t, \ell^t)$ respectively and that the i.i.d assumption does not hold across domains i.e. $P \neq Q$. Lastly, we assume access to a pretrained language model and generic pretrained visual descriptors (details are given in Sec. 4.1). Our objective is to learn a cross-modal embedding space such that distances within it respect the descriptions across both domain—v and ℓ should be embedded close together when ℓ describes v, and far apart otherwise.

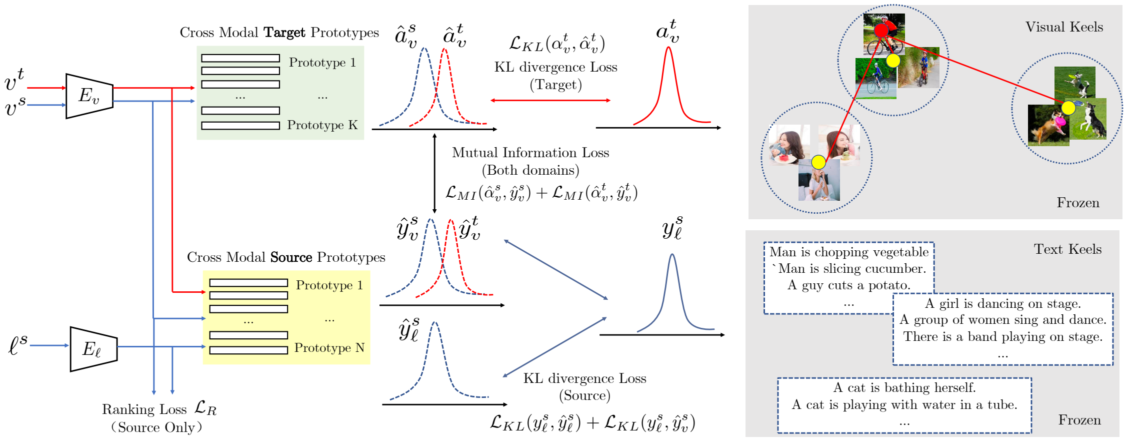
The overall framework of ADAPTIVE CROSS-MODAL PROTOTYPES (ACP) is illustrated in Fig 2, where blue and red arrows denote the flow of information from the source and target domains respectively. It is composed of six components, including the visual and text encoders E_v and E_ℓ, uni-modal visual and text keels K_v and K_ℓ, cross-modal source and target prototypical networks P_s and P_t. We discuss each of these components and their interactions in the following.

Visual and Text Encoders: Following the cross-modal approach popularised by [62], we use a visual encoder E_v and a text encoder E_ℓ to map each visual sample v and text description ℓ into a shared cross-modal embedding space, $E_v(v), E_\ell(\ell) \in \mathbb{R}^M$, where the visual and text embeddings are close to each other if and only if the text describes the visual input. We take advantage of paired data in the source domain to enforce a bidirectional ranking loss, $\mathcal{L}_R$ to align content and text descriptions as follows:

$$\mathcal{L}_R = \frac{1}{\mathcal{B}} \sum_{i=1, j\neq i}^{\mathcal{B}} [m + \xi^s_{i,j} - \xi^s_{i,i}]_+ + [m + \xi^s_{j,i} - \xi^s_{i,i}]_+ \quad (1)$$

where $\mathcal{B}$ is the size of each minibatch, m is a margin (set as a hyperparameter) and $\xi^s_{i,j} \triangleq \cos(E_v(v^s_i), E_\ell(\ell^s_j))$. Here $\cos(\cdot,\cdot)$ represents cosine similarity and $[\cdot]_+$ denotes the hinge function $\max(\cdot, 0)$.

Visual and Text Keels: In order to represent complex semantic features with compositions of multiple visual concepts (multiple words), we propose make use of readily available structural knowledge within each modality to construct visual and text keels. Specifically, we first chart the uni-modal data structure independently with *generic visual and text descriptors* – these are readily available "off-the-shelf" visual classification and sentence classification models that have been trained on labelled, large-scale uni-modal datasets available in the computer vision and natural language processing community. Inspired by the cluster assignment regularisation of [48] for open-set, uni-modal domain adaptation, we cluster the generic descriptors



v^t
v^s
E_v
Cross Modal **Target** Prototypes
Prototype 1
...
Prototype K
$\hat{a}_v^s$
$\hat{a}_v^t$
$\mathcal{L}_{KL}(\alpha_v^t, \hat{\alpha}_v^t)$
KL divergence Loss
(Target)
a_v^t
Visual Keels
Frozen
Mutual Information Loss
(Both domains)
$\mathcal{L}_{MI}(\hat{\alpha}_v^s, \hat{y}_v^s) + \mathcal{L}_{MI}(\hat{\alpha}_v^t, \hat{y}_v^t)$
Cross Modal **Source** Prototypes
Prototype 1
...
Prototype N
$\hat{y}_v^s$
$\hat{y}_v^t$
y_ℓ^s
$\hat{y}_\ell^s$
ℓ^s
E_ℓ
Ranking Loss $\mathcal{L}_R$
(Source Only)
KL divergence Loss
(Source)
$\mathcal{L}_{KL}(y_\ell^s, \hat{y}_\ell^s) + \mathcal{L}_{KL}(y_\ell^s, \hat{y}_v^s)$
Text Keels
Man is chopping vegetable
`Man is slicing cucumber.
A guy cuts a potato.
...
A girl is dancing on stage.
A group of women sing and dance.
There is a band playing on stage.
...
A cat is bathing herself.
A cat is playing with water in a tube.
...
Frozen

Figure 2: The components of the proposed ACP method (described in Sec. 3). Blue and red arrows denote the flow of information from the source and target domains respectively. 1) To represent semantic relationships between concepts in uni-modal data, we first construct visual and text keels and calculate the keel assignments a_v^t, y_ℓ^s. 2) To regularize the shared cross-modal feature learning, in addition to the conventional ranking loss $\mathcal{L}_R$, we propose to learn source and target prototypes constrained by preserving the relationships between uni-modal keels via $\mathcal{L}_{KL}$. 3) To reduce domain shift, we maximise mutual information between the source and target prototype assignments of the same data, no matter which domain the data comes from.

with Lloyds's algorithm [39] within each modality independently, to produce a set of centroids that we refer to as visual and text *keels* (the name reflects the intention that these centroids serve to stabilise the adaptation process). We then encode each sample by calculating its similarity distribution over all cluster centroids. This assignment is "static" in the sense that the pretrained descriptors are frozen and never fine-tuned, and the assignments therefore provide a domain-neutral (w.r.t source and target domain) signal to characterise uni-modal structural knowledge. Details about uni-modal keels are provided in sec 3.2.

Source and Target Prototypical Network: We next describe how to use uni-modal structural knowledge within each domain to effectively regularise the learning of the cross-modal embedding for retrieval. Specifically, we attach source and target prototypical network, P_s and P_t, (each contain a single linear projection) to the cross-modal embedding features, and task them with predicting a cluster assignment for each sample. We minimize the KL divergence loss $\mathcal{L}_{KL}$ to penalise the differences in this *prototype assignment* prediction given the cross-modal embedding with the *keel assignment* determined by the uni-modal keels. The goal of doing so is that the cross-modal embeddings should retain the local semantic relationships in the original uni-modal visual and textual space. Note that in our design, the target prototypical network is learned from the visual keels (constructed from target data), and the source prototypical network is driven by the text keels (constructed from source data). More intuition and details about the prototypical networks will be discussed in section 3.3.

Linking cross domain Prototypes: As the prototypical networks are driven by source and target samples respectively, the discrepancy of their assignments reflect domain shifts. The co-occurrence between these cluster assignments reveals the cross-domain underlying relationships. Specifically, for both the source and target samples, we regularise cross-modal feature learning by maximizing the mutual information (MI) [29, 48] between source and target prototype assignments (i.e. assignments obtained from the same sample should be predictable from one another, regardless of domain). This aims to help to help minimise domain shift in a cross-modal manner. More details about the mutual information maximisation will be discussed in section 3.4.

3.2. Uni-modal Compositional Keels

Text Keel Construction: Differently from a classification-based UDA setting, free-form text descriptions lack a clearly-defined, finite set of category labels. Thus we can not form the text keel by calculating the mean feature vector of instances within each category as did in [74, 7].

We propose to chart the uni-modal source text distribution with generic text descriptors–encoding source text descriptions ℓ^s with a "frozen" sentence-level language model pretrained on large corpus of free-form sentences (details are given in Sec. 4.1). The descriptors of the source text samples are then clustered into N clusters with Lloyd's algorithm [39]. Each cluster centroid is named as a text keel. We then encode each source text sample ℓ^s according to its relationship between the text keels $\{\mathfrak{L}_n\}_{n=1}^N$ by computing its probability of cluster assignment, $y^s = P_{\text{keel}}(\ell^s) \in R^N$.

Here, the n^{th} component of $P_{\text{keel}}(\ell^s)$ is defined as

$$P_{\text{keel}}(\ell^s)(n) = \frac{\exp(\cos(\ell^s, \mathfrak{L}_n))}{\sum_{n'} \exp(\cos(\ell^s, \mathfrak{L}_{n'}))}. \quad (2)$$

The role of the pretrained language model is twofold: (1) to encode sentences with similar semantics close together—in this way, each text cluster centroid represents how to describe a piece of visual content using compositions of description fragments; (2) to improve generalisation—large, pretrained language models exhibit remarkable few-shot learning capabilities [4], suggesting that representations encoded by pretrained models are sufficiently composable to generalise effectively.

Visual Keel Construction: Similar to text keel construction, we first chart the uni-modal visual data distribution with generic visual descriptors i.e. pretrained models from uni-modal perception tasks. After performing clustering algorithm to obtain *visual keels*, each target visual sample v^t is encoded via its relationship between the visual keels $\{\mathcal{V}_k\}_{k=1}^{K}$ by computing its probability of keel assignment, $a_v^t = P_{\text{keel}}(v^t) \in R^K$, where $P_{\text{keel}}(v^t)(k) = \exp(\cos(v^t, \mathcal{V}_k)) / \sum_{k'} \exp(\cos(v^t, \mathcal{V}_{k'}))$.

There are two main differences compared with the text keel construction: (1) To capture diverse composition of multiple visual concepts, we extract multiple generic visual descriptors rather than one, utilizing multiple perception models including those pretrained for object classification, action recognition, scene recognition (details are given in Sec. 4.1). Each visual keel then spans the following information: $\{$*what,how,where*$\}$ in the visual space, depicting compositions of multiple visual concepts. (2) To capture the distribution of visual concepts present in the target domain, the visual keel is constructed from the target samples rather than the source ones (used for the text construction)[1].

3.3. Source and Target Prototypical Network

We next describe how to regularise the shared cross-modal embedding space using the source text keels and target visual keels. To do so, we introduce the source and target prototypical networks P_s and P_t.

Source Prototypical Network: we attach a source prototypical network P_s (comprising a single linear projection, parameterized by $\mathcal{K} \in R^{N \times M}$) to the cross-modal embedding $E_v(v^s)$ and $E_\ell(\ell^s)$. The n^{th} row of the prototypical network weight matrix $\mathcal{K} \in R^{N \times M}$ represents the n^{th} source prototype $\mathcal{K}_n$, whose goal is to approximate centroids of the cross-modal embedding based on the guidance signal provided from uni-modal text keel. Mathematically, taking visual and text features $E_v(v^s)$ and $E_\ell(\ell^s)$ as inputs, the source prototypical network P_s aims to predict the N-dimensional probability vectors $\hat{y}_v^s = P_{\text{proto}}(v^s)$ and $\hat{y}_\ell^s = P_{\text{proto}}(\ell^s)$, representing their similarities with the source prototypes, where:

$$P_{\text{proto}}(v^s)(n) = \frac{\exp(cos(E_v(v^s), \mathcal{K}_n))}{\sum_{n'=1}^{N} \exp(cos(E_v(v^s), \mathcal{K}_{n'}))} \quad (3)$$

$$P_{\text{proto}}(\ell^s)(n) = \frac{\exp(cos(E_\ell(\ell^s), \mathcal{K}_n))}{\sum_{n'=1}^{N} \exp(cos(E_\ell(\ell^s), \mathcal{K}_{n'}))} \quad (4)$$

To integrate the structural knowledge of uni-modal data in the cross-modal embedding space, we minimise the source KL divergence loss $\mathcal{L}_s$ as shown in (5), penalizing the differences between "keel assignment" y_ℓ^s and the "prototype assignments" $\hat{y}_\ell^s$ and $\hat{y}_v^s$. Specifically, for each source text description ℓ^s, we use keel assignment y_ℓ^s obtained from uni-modal text keel as a "soft label" to guide the learning process of cross-modal text embedding $E_\ell(\ell^s)$. As the visual content v^s is paired with source text sample ℓ^s, we propose to propagate the acquired soft label from ℓ^s to v^s and use soft labelling consistency to regularise the learning process of cross-modal visual embedding $E_v(v^s)$.

$$\mathcal{L}_s = KL(y_\ell^s, \hat{y}_\ell^s) + KL(y_\ell^s, \hat{y}_v^s) \quad (5)$$

The key idea is that the learned source prototypes represent a robust "slow moving" (they evolve together with the cross-modal embedding space) characterisation of the visual-text relationships present in the source domain. Minimising the source KL divergence loss $\mathcal{L}_s$ regularises the cross-modal embeddings by requiring that they preserve local semantic relationships in the source text space.

Target Prototypical Network: We next describe how the uni-modal visual keels are used to regularise the cross-modal embedding space. We attach a target prototypical network P_t to the cross-modal embedding feature $E_v(v^t) \in R^M$ and task it with predicting the prototypical assignment of each target sample. The output of the target prototypical network, $P_{\text{proto}}(v^t)$, is defined as $\hat{a}_v^t = P_{\text{proto}}(v^t)$, where $P_{\text{proto}}(v^t)(k) = \exp(cos(E_v(v^t), \mathcal{W}_k)) / \sum_{k'} \exp(cos(E_v(v^t), \mathcal{W}_{k'}))$. Here $\mathcal{W}_k$ is the k^{th} target prototype (representing the k^{th} row of the target prototypical network parameter $\mathcal{W} \in R^{K \times M}$). Similar to the source prototypical network design, a target KL-divergence loss $\mathcal{L}_t$ between the prototype and keel assignment probabilities of target features $E_v(v^t)$ is defined as in (6), so that the target cross-modal prototypes are encouraged to retain structural knowledge of the uni-modal data in the target domain.

$$\mathcal{L}_t = KL(P_{\text{keel}}(v^t)||P_{\text{proto}}(v^t)) = KL(\alpha_v^t, \hat{\alpha}_v^t) \quad (6)$$

3.4. Maximising Mutual Information between Cross-Modal Prototypes

So far, we have introduced the losses that guide the learning of source and target cross-modal prototypes $\{\mathcal{K}_n\}_{n=1}^{N}$

[1] We use the source text to construct the text keel because in our setting, we do not have any target text in the training time.

and $\{\mathcal{W}_k\}_{k=1}^{K}$. For each source and target visual sample v^s and v^t, we can pass them through the source and target prototypical network P_s and P_t (calculating the similarity with prototypes $\{\mathcal{K}_n\}$ and $\{\mathcal{W}_k\}$) to obtain their cross-modal prototype assignments, denoted by $\hat{y}_v^s$, $\hat{y}_v^t$, $\hat{a}_v^s$, $\hat{a}_v^t$.

To reduce cross-domain shift, we propose to align the prototype assignments $\hat{y}_v$ and $\hat{a}_v$ of the same input data v regardless of which domain the sample originally comes from. As this operation is the same for v^s and v^t, we abbreviate the superscript in later descriptions in this section. Note that we do not use the constraint of perfectly aligning the source and target prototype assignments, because the source and target prototypes $\{\mathcal{K}_n\}$ and $\{\mathcal{W}_k\}$ may represent different (and even complementary) concepts. Instead, we regularise the feature learning process by maximizing the mutual information between source assignment $\hat{a}_v$ and target assignment $\hat{y}_v$ (i.e. not perfectly matched, but such that the two assignments can still be predicted from one another). This design is inspired by the observation that maximising mutual information preserves the common signal across the the domains, while discarding signal that occurs in one but not the other.

Although the MI between two random variables is hard to measure directly in high-dimension space, inspired by recent studies [1, 7] we adopt an objective function that implicitly maximizes the MI via an encoder discriminator architecture and an effective sampling strategy. In more detail, we draw positive and negative samples from the joint distribution $P(\hat{a}_v, \hat{y}_v)$ and the product of their marginal distributions $P(\hat{a}_v)P(\hat{y}_v)$ respectively. In our setting, positive samples are two assignments $(\hat{a}_{v_1}, \hat{y}_{v_1})$ that are predicted from the same input, while negative samples are $(\hat{a}_{v_1}, \hat{y}_{v_2})$ that predicted from different inputs. Given $\hat{a}_{v_1}$, the MI discriminator D_{MI} aims to distinguish whether $\hat{y}_{v_1}$ or $\hat{y}_{v_2}$ correspond to the same input. The MI discriminator first projects the prototype assignment $\hat{y}_v \in R^N$ to a vector $\hat{y}_v' \in R^K$ using a linear transformation $W \in \mathbb{R}^{K \times N}$, then calculates the similarity between $\hat{a}_v$ and $\hat{y}_v'$ via a dot product. The output of the MI discriminator is then given by: $D_{\text{MI}}(\hat{a}_v, \hat{y}_v) = \hat{a}_v^T W \hat{y}_v$. Various objective functions can be used to maximize $\text{MI}(\hat{a}_v, \hat{y}_v)$. In this paper, we follow the simple formulation used by [3, 29]. We adopt the standard binary cross-entropy (BCE) loss as shown in (7) where the output of D_{MI} is activated by a sigmoid function.

$$\begin{aligned}\mathcal{L}_{MI}(\hat{a}_v, \hat{y}_v) = \mathbb{E}_{X_P}[log(\sigma_{\text{sig}}(D_{\text{MI}}(\hat{a}_{v_1}, \hat{y}_{v_1}))] \\ + \mathbb{E}_{X_N}[log(1 - \sigma_{\text{sig}}(D_{\text{MI}}(\hat{a}_{v_1}, \hat{y}_{v_2})))],\end{aligned} \tag{7}$$

where X_P and X_N contains a set of positive and negative pairs from both source and target domains and $\sigma_{\text{sig}}(z) = \frac{1}{1+e^{-z}}$. As described at the start of this section, the final MI loss can be divided into two parts: $\mathcal{L}_{\text{MI}}(\hat{a}_v^s, \hat{y}_v^s)$ and $\mathcal{L}_{\text{MI}}(\hat{a}_v^t, \hat{y}_v^t)$.

3.5. Objective Functions

During training we minimize the sum of the above losses, with respect to the visual and text encoders E_v and E_ℓ, source and target prototypical networks P_s and P_t and mutual information discriminator D_{MI}. In more detail, the overall training objective of our proposed approach integrates the bidirectional ranking loss $\mathcal{L}_R$ in (1) on the source data, KL divergence loss $\mathcal{L}_s$ and $\mathcal{L}_t$ described in (5)(6) and the mutual information estimation loss L_{MI} in (7) on both source and target data:

$$\mathcal{L} = \sum_{v^s, \ell^s} (\mathcal{L}_R + \lambda_1 \mathcal{L}_s) + \lambda_2 \sum_{v^t} \mathcal{L}_t + \lambda_3 \sum_{v^s, v^t} \mathcal{L}_{\text{MI}} \tag{8}$$

where λ_1, λ_2 and λ_3 are set as hyperparmeters.

3.6. Analytical Motivation

The proposed ACP framework is motivated by the domain adaptation theory proposed by [2] which enables analysis of components, as shown by [7]. Let S, T denote the source and target domain respectively, and write $\mathcal{H}$ to denote the hypothesis class. We note that target domain *risk* $\epsilon_T(h)$, associated with hypothesis $h \in \mathcal{H}$ can be bounded by a sum comprising three components:

$$\epsilon_T(h) \le \epsilon_S(h) + \tfrac{1}{2} d_{H\Delta H}(S, T) + E$$

where $\epsilon_S(h)$ denotes source domain risk, $d_{H\Delta H}(S, T)$ measures the cross-domain discrepancy and $E = \epsilon_S(h^*, f_S) + \epsilon_T(h^*, f_T)$ represents the shared error of the ideal joint hypothesis h^*, where f_S and f_T are labelling functions for source and target respectively. In ACP, the $\epsilon_S(h)$ term is minimized by the source domain bidirectional loss $\mathcal{L}_R$. In addition, we employ MI maximisation to minimize the domain discrepancy $d_{H\Delta H}(S, T)$ [59]. However, only minimizing the first two terms is insufficient since E can grow when h cannot simultaneously reduce discrepancies with f_T and f_S [7, 67].

Due to the absence of the target domain text labels, we introduce more labelling functions to bound the shared error E, including f_S and f_T as the functions to calculate the uni-modal keel assignments, $f_{\hat{S}}$ and $f_{\hat{T}}$ to predict the cross-modal prototype assignments. More specifically, the shared error is bounded via triangle inequality as follows: $E \le \epsilon_S(h, f_S) + \epsilon_{\hat{T}}(h, f_{\hat{T}}) + 2\epsilon_{\hat{T}}(f_S, f_{\hat{T}}) + \epsilon_{\hat{T}}(f_T, f_{\hat{T}}) \le \epsilon_{\hat{S}}(h, f_{\hat{S}}) + \epsilon_{\hat{S}}(f_S, f_{\hat{S}}) + \epsilon_{\hat{T}}(h, f_{\hat{T}}) + 2\epsilon_{\hat{T}}(f_{\hat{S}}, f_{\hat{T}}) + \epsilon_{\hat{T}}(f_T, f_{\hat{T}}) + C$.
We note that minimizing ϵ_S via ranking loss $\mathcal{L}_R$ alone will overfit to source samples. We employ cross-modal prototypical assignments as extra label information and minimize the terms $\epsilon_{\hat{T}}(h, f_{\hat{T}})$ and $\epsilon_{\hat{S}}(h, f_{\hat{S}})$. To minimize the terms $\epsilon_{\hat{S}}(f_S, f_{\hat{S}})$ and $\epsilon_{\hat{T}}(f_T, f_{\hat{T}})$ in the above derivation, we reduce the source and target KL divergence loss $\mathcal{L}_s$ and $\mathcal{L}_t$, by aligning the cross-modal prototypical assignments with uni-modal keel assignments. To minimize

Table 1: Performance when adapting Open Narratives → COCO Narratives and COCO → COCO Narratives

Method	Open Narr→ Coco Narr						COCO → Coco Narr					
	t2v			v2t			t2v			v2t		
	R1	R10	MR	R1	R10	MR	R1	R10	MR	R1	R10	MR
SCAN [34]	17.4	52.6	9	18.2	54.8	7	22.3	72.9	5	24.2	74.9	4
VSRN [35]	19.6	54.7	7	20.8	59.6	7	25.1	75.4	**4**	26.3	79.0	**3**
Baseline [37]	19.6	56.4	7	20.5	58.9	7	24.5	75.8	**4**	26.0	78.2	**3**
CDAN [41]	20.6	**59.2**	6	20.0	60.4	7	22.2	73.3	5	20.2	75.2	5
CORAL [63]	19.4	58.3	7	20.2	60.4	7	25.4	74.6	**4**	26.8	80.2	**3**
DANN [24]	19.0	58.4	7	19.6	58.9	7	24.8	76.8	**4**	22.2	75.6	4
MMD [40]	17.3	50.8	9	18.1	56.7	7	22.6	72.0	5	19.9	73.8	5
OT [25]	20.3	57.1	8	20.2	60.4	7	25.0	75.6	**4**	26.4	79.6	**3**
ACP (ours)	**22.3**	57.9	**6**	**21.8**	**62.5**	**6**	**27.3**	**77.9**	4	**28.0**	**80.5**	**3**

the discrepancy between two cross-modal prototypical assignments $\epsilon_{\hat{T}}(f_{\hat{S}}, f_{\hat{T}})$, we use the $\mathcal{L}_{\mathrm{MI}}$ to encourage that the two domains' cross-modal prototypical assignments are predictable with each other.

4. Experiments

4.1. Implementation details

Datasets: For image-text retrieval, we use three datasets: COCO [15], COCO Narratives [53] and Open Narratives [53] to assess our approach. We also perform four transfer tasks among three video-text retrieval benchmarks from different domains: MSRVTT [70], MSVD [8], LSMDC [55]. More details of dataset statistics are reported in the supplementary material.

Evaluation Metrics: We use standard retrieval metrics (following [73, 43, 37, 44, 72]) to evaluate text-to-visual (t2v) retrieval and visual-to-text (v2t) retrieval. We measure rank-based performance by $R@K$ (where higher is better) and also report Median Rank (MR, lower is better).

Generic Visual Embeddings: We represent each target sample v^t as an concatenated feature of object (what), motion (how) and scene (where) features for clustering. These features are the outputs of three pretrained models trained for object classification (a ResNext-101 [68] pretrained on ImageNet [18]) and scene recognition (a Dense Net-161 [30] pretrained on Places365 [76]). For videos (but not images), an additional action classification model is also used (an R(2+1)D model [64] trained on IG-65m [26]), In this way, each learned visual keel can be deemed as a configuration of $\{what, how, where\}$ contained in the visual space.

Implementation Details: For fair comparison, we use the same architectural design of the encoders E_v and E_t as [37] and re-implemented it with the generic visual embeddings described above to serve as the baseline. The number of visual keels $K = 1024$ and text keels $N = 512$ (we pick $K > N$ to reflect the "semantic gap" between visual and written content [75]). The language model used to encode is the Sentence-Transformer [54]. More implementation details are reported in supplementary material.

4.2. Results

We choose methods from both cross-modal retrieval and UDA fields for comparisons. For adaptation modules, we compared with the UDA methods that are applicable to cross-modal retrieval task, including MMD [40], D-CORAL [63], DANN [24], Optimal Transport (OT) [25], CDAN [41][2] and CAPQ [11][3]. Implementation details of these comparison methods are in the supplementary material. For retrieval methods, we adopted the state of-the-art model in [37] as our baseline, and compare additionally to method SCAN [34],VSRN [35], MoEE [43] and MMT[21].

Image-Text Retrieval: The results of two transfer directions on three image retrieval benchmarks are shown in Table.1. Our approaches outperforms latest image-text retrieval approaches, i.e., SCAN [34],VSRN [35] and CE [37].It can be seen that conventional UDA approaches achieve similar performances to the baseline model, which suggest that they are inadequate to align domain shifts for retrieval task. As a more controlled study, for the same target set COCO Narratives, the methods always achieved better results using COCO than Open Narratives as the source domain. It indicates the importance of reducing the visual domain shift (images in COCO and COCO Narratives are the same). When adapting Open Narratives → COCO Narratives (a large visual domain shift), our proposed ACP outperforms others on all metrics, verifying its effectiveness. In the case of large language domain shifts (e.g. annotation shift, i.e,COCO → COCO Narratives), the superiority of ACP demonstrates that the features extracted from pretrained language model generalize well to novel concepts.

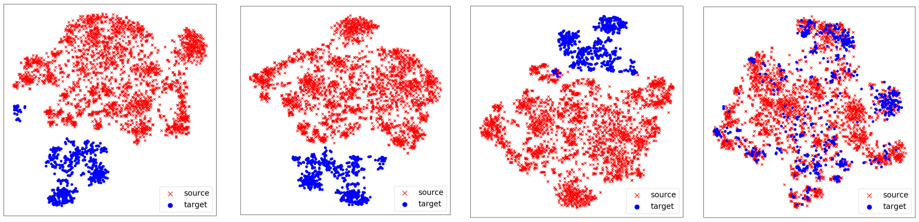
Video-Text Retrieval: Our method achieved strong results on 4 transfer directions on three video retrieval datasets in Tables. 2 and 3. More experiments on other transfer directions are given in supplementary material. The proposed ACP outperforms all methods on all transfer tasks. These results suggest that the proposed approach is able to learn the cross-modal features that are both transferable and discriminative for target domain retrieval.

4.3. Discussions and Analysis

Ablation Study: We conduct a detailed ablation study by examining the effectiveness of each proposed component in Table. 4. The CE model [37] serves as the *source-only* (Baseline) for the comparison. The three different variants of the proposed model all boost the retrieval metrics compared to the baseline. *Adding* either cross-modal source and target prototypes improve the performance over 1.2 on $R@1$. This verifies the effectiveness of preserving semantic

[2] As CDAN cannot be directly applied to our setup, we incorporate the proposed uni-modal static assignment as categorical labels and report corresponding results.

[3] We re-implement CAPQ and report the performance using the same features for a fair comparison.



(a) Baseline-SRC-ONLY (b) CORAL (c) DANN (d) ACP (Ours)

Figure 3: The t-SNE visualization of the shared embedding features using the proposed and the baseline methods.

Table 2: Performance when adapting MSRVTT → MSVD and MSVD → MSRVTT

Method	MSRVTT→ MSVD						MSVD→ MSRVTT					
	t2v			v2t			t2v			v2t		
	R1	R10	MR	R1	R10	MR	R1	R10	MR	R1	R10	MR
Baseline [37]	14.2	52.3	9	16.6	50.0	10	3.6	17.2	98	2.5	13.5	117
MMT [21]	14.7	53.9	9	17.9	50.8	9	3.8	17.4	98	2.5	12.9	119
MoEE [43]	14.0	53.4	10	16.7	48.7	10	3.1	17.1	102	2.3	12.4	123
CAPQ [11]	15.0	53.5	10	18.2	51.0	9	3.9	17.0	100	2.7	14.2	115
CDAN [41]	12.9	48.3	11	13.2	41.5	18	3.9	17.7	98	2.7	13.9	115
CORAL [63]	11.8	46.1	13	12.5	42.3	19	3.3	16.0	104	2.4	13.4	125
DANN [24]	12.1	47.1	12	11.8	36.9	23	3.8	17.4	100	2.5	13.1	120
MMD [40]	13.6	50.5	10	16.9	46.7	14	3.4	15.9	104	2.3	13.4	126
OT [25]	12.6	47.9	12	13.0	38.4	19	3.7	16.8	98	2.6	13.6	120
ACP(ours)	**16.6**	**55.2**	**8**	**22.1**	**52.5**	**8**	**4.4**	**17.9**	**97**	**3.1**	**15.3**	**111**

Table 3: Performance when adapting MSVD → LSMDC and LSMDC → MSVD

Method	MSVD→ LSMDC						LSMDC→ MSVD					
	t2v			v2t			t2v			v2t		
	R1	R10	MR	R1	R10	MR	R1	R10	MR	R1	R10	MR
Baseline [37]	1.6	11.2	160	2.8	8.7	194	8.3	36.8	21	9.4	34.3	31
MMT [21]	1.8	10.9	181	2.6	8.1	188	8.9	38.1	18	10.7	37.1	26
MoEE [43]	1.4	10.6	195	0.9	3.6	271	8.3	36.7	21	9.7	36.7	29
CAPQ [11]	2.2	11.5	163	2.7	10.5	158	10.2	39.9	18	12.5	38.2	24
CDAN[41]	2.0	8.9	185	1.2	7.3	288	8.0	37.5	20	10.8	37.1	24
CORAL [63]	1.3	10.6	172	1.7	9.1	205	9.5	39.9	18	11.9	35.5	27
DANN [24]	2.0	8.1	185	0.3	4.3	308	8.1	37.2	20	10.4	36.0	25
MMD [40]	1.2	5.7	251	1.0	5.0	284	9.7	40.7	17	12.1	36.1	27
OT [25]	1.7	10.6	181	1.9	9.1	217	8.3	36.7	21	9.7	36.7	29
ACP(ours)	**2.6**	**12.1**	**161**	**3.4**	**9.9**	**160**	**10.8**	**41.9**	**16**	**13.1**	**40.2**	**20**

Table 4: Ablation study of ACP model performances on MSRVTT → MSVD, using different losses.

Method	t2v				v2t			
	R1	R5	R10	MR	R1	R5	R10	MR
Baseline ($\mathcal{L}_R$)	14.2	38.0	52.3	9	16.6	37.5	50	10
Source Prototype($\mathcal{L}_R + \mathcal{L}_s$)	15.7	39.8	54.1	9	20.4	41.5	51.3	9
Target Prototype($\mathcal{L}_R + \mathcal{L}_t$)	15.4	40.3	53.6	9	19.8	42.0	51.8	9
All ($\mathcal{L}_R + \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_{MI}$)	16.6	42.4	55.2	8	22.1	42.9	52.5	8

structures from uni-modal keels when learning the cross-modal features. *Adding* the mutual information loss further boosts the performance and achieves the best performance, verifying the effectiveness of reducing cross-domain discrepancies by aligning the source and target prototype assignments. In Fig. 4, we provide the ACP performances using different pre-trained language models and the convergence analysis. We observed that the language model matters a lot in the retrieval task. Finally, our method converges to better solution and faster compared with others.

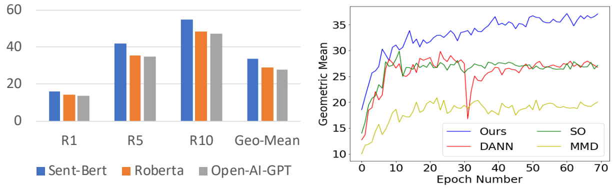


Figure 4: ACP performance using different language models and the convergence analysis.

A-distance: The A-distance [2] is calculated based on the following formula: $d_A = 2(1 - 2\theta)$, where θ is the generalization error of a kernel SVM trained on the binary problem of discriminating the source and target joint video-text features. The A-distance of the baseline method is 1.701, while the one of ACP is 1.592. The smaller A-distance suggests that the joint features of our ACP can close the cross-domain gap more effectively. In addition, the A-distance of video features learned by the Baseline method is 1.536, while ACP achieves 1.471. We also observed a smaller cross-domain distance of video features than video-text joint features, which demonstrates the difficulty of UDA with natural language label space.

Feature Visualization: In Fig.3, we visualise the joint video-text features on MSRVTT→MSVD, learned by Baseline, CORAL, DANN and our ACP method respectively using t-SNE [19]. The cross-domain joint video-text features learnt by our method in (d) are clearly clustered tighter than other models, which suggests the effectiveness of aligning the cross-domain video-text distributions, even if without the target domain text queries in advance.

5. Conclusion

In this work, we propose ADAPTIVE CROSS-MODAL PROTOTYPES (ACP), a framework for cross-domain visual-text retrieval. Quantitative and qualitative evaluation on both image and video retrieval demonstrate the strengths of our proposed approach over strong UDA baselines.

Acknowledgements. The authors acknowledge the support of the Beijing Advanced Discipline Construction Project BMU2019GJJXK001, PKU-Baidu Fund(2019BD017), WICT grant 6202000014/017 and EPSRC Programme Grant Visual AI EP/T028572/1.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation. *International Conference on Machine Learning*, 2018. 6

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 6, 8

[3] Philemon Brakel and Yoshua Bengio. Learning Independent Features with Adversarial Nets for Non-linear ICA. *arXiv preprint arXiv:1710.05050*, 2017. 6

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020. 2, 5

[5] Zhangjie Cao, Mingsheng Long, Chao Huang, and Jianmin Wang. Transfer adversarial hashing for hamming space retrieval. In *AAAI Conference on Artificial Intelligence*, 2018. 3

[6] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018. 3

[7] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 3, 4, 6

[8] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the ACL: Human Language Technologies*, pages 190–200, 2011. 7

[9] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *International Conference on Machine Learning*, 2012. 3

[10] Qingchao Chen and Yang Liu. Structure-aware feature fusion for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10567–10574, 2020. 3

[11] Qingchao Chen, Yang Liu, and Samuel Albanie. Mind-the-Gap! Unsupervised Domain Adaptation for Text-Video Retrieval. In *AAAI Conference on Artificial Intelligence*, 2021. 3, 7, 8

[12] Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, and Kevin Chetty. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 6, 2018. 3

[13] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 2

[14] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–530, 2017. 3

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 7

[16] Stéphane Clinchant, Gabriela Csurka, and Boris Chidlovskii. A domain adaptation regularization for denoising autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 26–31, 2016. 3

[17] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017. 2, 3

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 7

[19] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International conference on machine learning*, pages 647–655, 2014. 8

[20] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016. 2

[21] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. *European Conference on Computer Vision*, 2020. 2, 7, 8

[22] Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1869–1871, 2018. 3

[23] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *International Conference on Machine Learning*, 2014. 3

[24] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2, 7, 8

[25] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Sinkhorn-autodiff: Tractable wasserstein learning of generative models. *arXiv preprint arXiv:1706.00292*, 2017. 7, 8

[26] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019. 7

[27] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. *International Conference on Machine Learning*, 2011. 3

[28] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA, 2013. Association for Computing Machinery. 2

[29] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019. 2, 4, 6

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 7

[31] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015. 3

[32] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1654–1662, 2017. 3

[33] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[34] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2, 7

[35] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019. 2, 7

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[37] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *BMVC*, 2019. 2, 7, 8

[38] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2

[39] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4

[40] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *International Conference on Machine Learning*, 2015. 2, 3, 7, 8

[41] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1640–1650. Curran Associates, Inc., 2018. 2, 3, 7, 8

[42] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017. 2, 3

[43] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2, 7, 8

[44] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 2, 7

[45] Robert C Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224. Association for Computational Linguistics, 2010. 3

[46] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. 2020. 3

[47] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer, 2016. 2

[48] Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. Exploring category-agnostic clusters for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13867–13875, 2020. 3, 4

[49] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2239–2247, 2019. 3

[50] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *ICLR*, 2021. 2

[51] Viswa Mani Kiran Peddinti and Prakriti Chintalapoodi. Domain adaptation in sentiment analysis of twitter. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, pages 44–49, 2011. 2

[52] Jose Costa Pereira and Nuno Vasconcelos. Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems. *Computer Vision and Image Understanding*, 124:123–135, 2014. 3

[53] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 7

[54] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 7

[55] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 2, 7

[56] Guy Rotman and Roi Reichart. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713, 2019. 3

[57] Sebastian Ruder and Barbara Plank. Strong baselines for neural semi-supervised learning under domain shift. *ACL*, 2018. 3

[58] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2988–2997. JMLR. org, 2017. 3

[59] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2, 6

[60] H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. 3

[61] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2110–2118, 2016. 3

[62] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. 3

[63] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016. 7, 8

[64] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 7

[65] Marlies van der Wees, Arianna Bisazza, and Christof Monz. Dynamic data selection for neural machine translation. *ACL*, 2017. 3

[66] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 450–459, 2019. 2

[67] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 09–15 Jun 2019. 6

[68] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 7

[69] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432, 2018. 3

[70] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 2, 7

[71] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. 3

[72] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 7

[73] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017. 7

[74] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 433–443, 2019. 3, 4

[75] Rhong Zhao and William I Grosky. Bridging the semantic gap in image retrieval. In *Distributed multimedia databases: Techniques and applications*, pages 14–36. IGI Global, 2002. 7

[76] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 7