Neurocomputing 440 (2021) 207–219

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-level Alignment Network for Domain Adaptive Cross-modal Retrieval

Jianfeng Dong [a,d,1], Zhongzi Long [b,1], Xiaofeng Mao [c], Changting Lin [a,e,*], Yuan He [c], Shouling Ji [b,d]

[a] Zhejiang Gongshang University, China
[b] Zhejiang University, China
[c] Alibaba Group, China
[d] Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, China
[e] The State Key Laboratory of Information Security (Institute of Information Engineering, Chinese Academy of Sciences), China

ARTICLE INFO

Article history:
Received 28 November 2020
Revised 8 January 2021
Accepted 30 January 2021
Available online 17 February 2021
Communicated by Zidong Wang

Keywords:
Cross-modal retrieval
Domain adaptation
Cross-dataset training
Adversarial learning

ABSTRACT

Cross-modal retrieval is an important but challenging research task in the multimedia community. Most existing works of this task are supervised, which typically train models on a large number of aligned image-text/video-text pairs, making an assumption that training and testing data are drawn from the same distribution. If this assumption does not hold, traditional cross-modal retrieval methods may experience a performance drop at the evaluation. In this paper, we introduce a new task named as *domain adaptive cross-modal retrieval*, where training (source) data and testing (target) data are from different domains. The task is challenging, as there are not only the semantic gap and modality gap between visual and textual items, but also domain gap between source and target domains. Therefore, we propose a *Multi-level Alignment Network* (MAN) that has two mapping modules to project visual and textual modalities in a common space respectively, and three alignments are used to learn more discriminative features in the space. A semantic alignment is used to reduce the semantic gap, a cross-modality alignment and a cross-domain alignment are employed to alleviate the modality gap and domain gap. Extensive experiments in the context of domain-adaptive image-text retrieval and video-text retrieval demonstrate that our proposed model, MAN, consistently outperforms multiple baselines, showing a superior generalization ability for target data. Moreover, MAN establishes a new state-of-the-art for the large-scale text-to-video retrieval on TRECVID 2017, 2018 Ad-hoc Video Search benchmark.

1. Introduction

As data from different modalities, such as text, images, and videos, are growing at an unprecedented rate, cross-modal retrieval [1–6] has attracted increasing attention in the multimedia community [7–9]. The existing efforts in this research direction mainly focus on the tasks of image-text retrieval [10–12] and video-text retrieval [13–15]. Take the image-text retrieval as an example, the goal of this task is to retrieve images which are semantically related to a given text query, or retrieve text describing the content of a given image query.

Recently, with the development of deep learning technologies, great progress has been achieved for cross-modal retrieval [10–18]. The key of cross-modal retrieval is to learn a common space where the similarity between different modalities can be directly computed. Based on the recurrent neural network and convolutional neural network, Dong et al. [13] first propose a multi-level video encoding and multi-level text encoding to obtain strong video and sentence representations, then project videos and text into a common space by two linear transformations. In [17], Chen et al. utilize a graph convolutional network to model the connection between words, and generate hierarchical textual embeddings via attention-based graph reasoning. Instead of using one specific sentence encoder, Li et al. [18] propose to utilize diverse sentence encoders and learn multiple encoder-specific common spaces to measure cross-modal similarity. More recently, Gabeur et al. [16] introduce a multi-modal Transformer with four stacked Transformer layers [19] to jointly encode the different modalities, such as motion, audio and appearance, in videos, which allows each of them to attend to the others. Besides the deep learning technologies, this success is also attributed to the availability of

* Corresponding author at: 18 Xuezheng St, Jianggan District, Hangzhou, Zhejiang, China.
E-mail address: linchangting@gmail.com (C. Lin).
[1] Jianfeng Dong and Zhongzi Long contribute equally to this work.

https://doi.org/10.1016/j.neucom.2021.01.114
0925-2312/

large-scale human-annotated cross-modal datasets such as the MS-COCO [20] and MSR-VTT [21] datasets. For instance, MSR-VTT totally has 10,000 videos, and each video is annotated with 20 crowd-sourced natural language sentences that briefly describe the main objects and their relations, scenes and activities in the video. To annotate such a kind of video-text dataset, one has to watch the videos, listen to the audios, and carefully utilize natural language sentences to describe the content of the watched videos. Therefore, collecting large-scale annotated cross-modal datasets is time-consuming and laborious.

In real application scenarios, if we would like to build a cross-modal retrieval for a specific new domain (target domain), a straightforward way is to collect a great number of labeled training data which have the same data distribution as the unlabeled data of the target domain, thus train a model on the collected data. However, as mentioned before that collecting a large-scale annotated cross-modal dataset is time-consuming and laborious, this solution is to some extent suboptimal. Instead of collecting a new dataset, another solution is to utilize off-the-shelf labeled cross-modal data, but such data usually show different data distribution with the target domain. As shown in Fig. 1, two fashion images from different datasets show a clear difference. In this context, most existing cross-modal retrieval models will likely experience a significant performance drop (as shown in Fig. 2), as they make an assumption that training and testing data are drawn from the same distribution. Based on this assumption, cross-modal retrieval models [13,24,25] typically mainly focus on the differences between data of different modalities. For example, in order to learn the common space, Dong et al. [24] employ a mean squared error based loss to reduce the distance of relevant cross-modal data pairs in the common space. In [25], Faghri et al. use a triplet ranking loss to make the distance between the relevant cross-modal data pair larger than that between irrelevant ones. As these methods do not consider the domain gap between the source and target data, it hurts their generalization ability to the new target domain. Therefore, how to use off-the-shelf labeled datasets to build a cross-modal retrieval model for a new target domain is still an open question. To promote this direction, we introduce a task called *domain adaptive cross-modal retrieval*, where given several labeled source datasets which have different data distribution with the unlabeled target dataset, it asks to build a cross-modal retrieval model which adapts to the target domain.

Focusing on visual and textual modalities, in this paper we propose a *Multi-level Alignment Network* (MAN) for the domain adaptive cross-modal retrieval task. MAN first maps visual and text modalities into a common space by a visual encoding and a text encoding, respectively, and three alignments are employed to alleviate the above gaps in the mapped common space: a semantic alignment for reducing the semantic gap, a cross-modality alignment for alleviating the modality gap, and a cross-domain alignment is employed to reduce the domain gap. Specifically, for the semantic alignment, we use a triplet ranking loss [25] to make the semantically relevant data near and semantically irrelevant data far away in the common space. For the domain gap, we utilize a number of discriminators to distinguish features from different domains while the mapping encoders confuse them by adversarial learning, which pushes the distribution of source data to well align with target data. For the modality gap, adversarial learning is sim-

Fig. 1. Sample images associated with the corresponding textual description from (a) Fashion200k and (b) DeepFashion Datasets, showing a clear difference across the datasets.

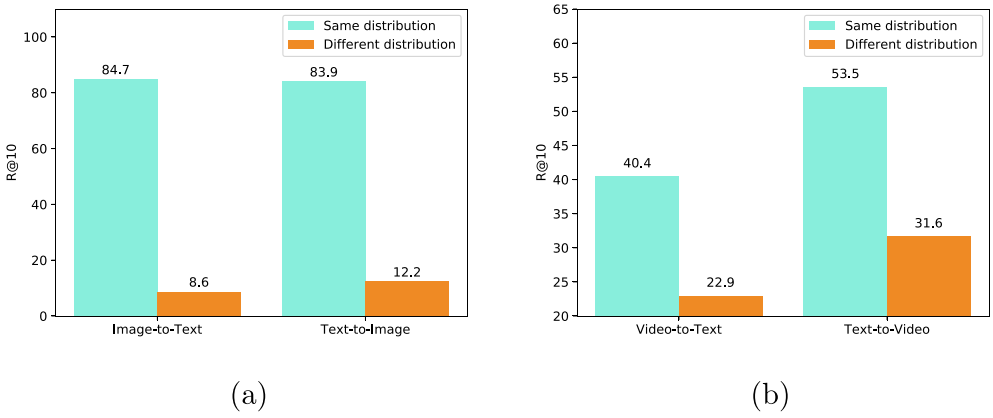


Fig. 2. The performance significantly decrease when the training data and test data have different distribution in the context of (a) image-text retrieval tested on the Fashion200K dataset [22] and (b) video-text retrieval tested on VATEX [23].

ilarly employed to align the distribution of data of different modalities.

To sum up, this paper makes the following contributions:

- We introduce a *domain adaptive cross-modal retrieval* task, which aims to utilize off-the-shelf labeled datasets to build a cross-modal retrieval model that adapts to the target domain. The task is a cross-domain and cross-modal related task, so it is very challenging but valuable for practical applications.
- We propose a *Multi-level Alignment Network* (MAN), which learns visual-semantic embeddings cross domains and modalities by three alignment modules. Our model is orthogonal to visual and textual encoder, allowing us to flexibly embrace state-of-the-art visual and textual encoders.
- Extensive experiments in the context of domain adaptive video-text and image-text retrievals verify the effectiveness of our proposed MAN. Moreover, MAN establishes a new state-of-the-art for the large-scale text-to-video retrieval on TRECVID 2017, 2018 Ad-hoc Video Search benchmark.

The rest of the paper is organized as follows. Section 2 reviews the methods related to cross-modal retrieval and domain adaptation. In Section 3, we first formally define the problem of domain adaptive cross-modal retrieval, followed by the model structure and model training description of our proposed Multi-level Alignment Network. The experimental results and analysis in the context of video-text retrieval and image-text retrieval are provided in Section 4. Finally, Section 5 concludes our work and gives our future work.

2. Related work

2.1. Cross-modal retrieval

The existing efforts in cross-modal retrieval mainly focus on the tasks of image-text retrieval [10–12,26] and video-text retrieval [13–15,27,28,79]. Despite the different forms of these two retrieval paradigms, they essentially share a similar methodology. The key of these retrieval paradigms is to compute the cross-modal relevance between two modalities. The common solution [25,14,12] is to project the data of different modalities into a common space, thus measure cross-modal relevance in the common space by a standard distance metric, *e.g.,* cosine distance.

In the context of image-text retrieval, Frome et al. [29] first propose a visual-semantic embedding model to map images and text into a common space, which utilizes pre-trained visual and textual models. In a follow-up work, Kiros et al. [30] extend the model by encoding images with a convolutional neural network (CNN) and encoding text with a Long Short-Term Memory (LSTM), which can be trained in an end-to-end manner. Using the similar model structure with [30], Faghri et al. [25] improve the training strategy by hard negatives mining, and Zhang et al. [31] propose a cross-modal projection matching loss and a cross-modal projection classification loss to learning more discriminative features. Considering videos are more complex than images, the majority of video-text retrieval methods focus on video representation [13,14,32,33]. For instance, Dong et al. [13] propose a dual encoding network to encode videos and text into powerful dense representations of their own. Based on [13], Wu et al. [34] propose a unified dual-task learning framework to increase the interpretability of the model. Antoine et al. [14] introduce gated embedding units to encoding videos and texts and train the model with bidirectional max-margin ranking loss. Recently, Liu et al. [32] propose a collaborative gating to fuse multiple different features, such as visual, motion, audio features, to obtain a strong video representation.

The above cross-modal retrieval methods are based on an assumption that training and testing data are drawn from the same distribution. However, in real application scenarios, it is not easy to hold this assumption. The approaches mentioned above will likely perform badly when the assumption does not hold. Different from the above methods, our proposed model allows training and testing data can be drawn from the different distributions. It is worth noting that the most similar work to ours is [35] which also allows training and testing data can be drawn from the different distributions. But this work [35] needs the labeled image-text pairs from both source and target domains for training. By contrast, our proposed method does not need the labeled training pairs of the target domain while only utilizes the existing labeled source datasets, which further reduces the cost of collecting training data.

2.2. Domain adaptation

Domain adaptation methods aim at addressing the domain shift problems [36,37]. According to whether the labels are available of target data, domain adaptation methods are typically categorized into two groups: unsupervised domain adaptation [38,39] and supervised domain adaptation [40,41]. For unsupervised domain adaptation, labeled data in the target domain are unavailable during training. By contrast, supervised domain adaptation has an access to the labeled data of the target domain for training. Although these two groups of methods have different settings, they both try to reduce the gap between the source domain and the target domain. Most domain adaptation methods seek to reduce the distribution discrepancy between source and target features, by minimizing several distribution discrepancy measures such as maximum mean discrepancy [42–44], correlation alignment function [45,46] or optimal transport [47]. For instance, Sun et al. align two domains by reducing their mean and covariance difference of feature distribution [45]. Additionally, some works [48,40,49] are proposed to learn domain-invariant representations between the source domain and target domain by adversarial learning. They adopt a similar idea with Generative Adversarial Networks (GAN) [50,51] by introducing domain discriminators into their architectures. The domain discriminators are optimized to distinguish different domains, while the feature extractors are optimized in the opposite direction. Through adversarial training, it becomes difficult for the domain discriminators to distinguish different domains, resulting in that the domain gap is reduced.

Common domain adaptation methods typically train models on a source dataset, while recently we notice an increasing use of multiple source datasets as the training data [52–55]. Such methods are known as the multi-source domain adaption method. For instance, Xu et al. [53] propose a deep cocktail network which solves the multi-source domain adaptation problem by a k-way domain discriminator and category classifier. Zhao et al. [54] try to align the distribution between the sources and targets through adversarial learning. Most recently, in [55], Peng et al. transfer knowledge learned from multiple labeled source domains to an unlabeled target domain by dynamically aligning moments of their feature distributions.

In this paper, our introduced domain adaptive cross-modal retrieval task is an unsupervised domain adaptation problem, where no labeled video/image-text pairs of the target domain are available. Besides, as the above methods are designed for classification tasks, they cannot be utilized directly for domain adaptive cross-modal retrieval task. Therefore, targeting domain adaptive cross-modal retrieval, this paper proposes MAN which aligns the feature distribution among domains and modalities simultaneously.

3. Material and method

In this section, we first formally define the problem of domain adaptive cross-modal retrieval, followed by the model structure and model training description of our proposed MAN. For the ease of reference, main abbreviations used in this work is listed in Table 1.

3.1. Domain adaptive cross-modal retrieval

In the task of domain adaptive cross-modal retrieval, we focus on visual and textual data, specifically for images and text, or videos and text. In this task, we are provided with k labeled source datasets $\left\{(\mathscr{V}_j^s, \mathscr{T}_j^s) = \{(v_{ji}^s, t_{ji}^s)\}_{i=1}^{m_j}\right\}_{j=1}^{k}$, where m_j indicates the number of semantically relevant images-sentence/video-sentence pairs (v_{ji}^s, t_{ji}^s) of the j-th source dataset. Additionally, we are also given an unlabeled target dataset $\left\{\mathscr{V}^t = \{v_i^t\}_{i=1}^n, \mathscr{T}_i^t = \{t_i^t\}_{j=1}^n)\right\}$ with a collection of images/videos $\mathscr{V}^t$ and a collection of textual sentences $\mathscr{T}^t$. Note that the target images and sentences are unpaired, which means the relevant annotation are not available for the target dataset. Based on the above source datasets and the target dataset, domain adaptive cross-modal retrieval asks to learn a cross-modal retrieval model which can search relevant images/video by textual query or search relevant sentence by image/video query in the context of the target domain. Fig. 3 illustrates a toy example to show difference between traditional cross-modal retrieval and domain adaptive cross-modal retrieval.

Table 1
Main abbreviations used in this paper.

Abbreviation	Description
MAN	Multi-level Alignment Network
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
FC	Fully Connected layer
BoW	Bag-of-Words
MLP	Multilayer Perceptron
ReLU	Rectified Linear Unit
mAP	Mean Average Precision
infAP	Inferred Average Precision
Med r	Median rank
SumR	Sum of the recalls
AVS	Ad-hoc Video Search

3.2. Network architecture

Fig. 4 illustrates the framework of our proposed multi-level alignment network for domain adaptive cross-modal retrieval, which consists of three components: a visual encoder to extract features of visual items, a textual encoder to extract features of sentences, and a common space learning module to align the cross-modality and cross-domain representations in the learned common space.

For image-text retrieval, given an image, we adopt a CNN model pre-trained on ImageNet and utilize its output of the last pooling layer as the image encoding feature; given a sentence, we employ a bidirectional LSTM and further use a max pooling layer to aggregate the hidden states of all time steps, and the output is regarded as the sentence encoding feature. As images and sentences are of different modalities, their encoding features are not directly comparable. Hence, a fully connected layer is further employed over the encoding features to project them into a common space where the image-text similarity can be directly computed by a standard distance metric, *e.g.,* cosine distance. For ease of reference, we merge feature extraction and projection process into mapping module $\phi(\cdot), \psi(\cdot)$ for visual items and text, respectively. Note that for both target and source data, we share the mapping modules to map them into the common space. Such a design is expected to transfer the knowledge learned in the source domain to the target domain. For video-text retrieval, we use a multi-level video encoding and a multi-level sentence encoding derived from [13] to encode and project videos and text into a common space, respectively.

In this paper, we focus on the common space learning module, and describe our proposed module in the following sections. It is worth noting that our proposed method is orthogonal to the visual and textual encoder, allowing us to flexibly embrace state-of-the-art visual and textual encoders.

3.3. Multi-level alignment

For the domain adaptive cross-modal retrieval, besides the well-known semantic gap and modality gap between visual and text items in cross-modal retrieval, there is a domain gap between different datasets, which makes the problem more challenging. In this paper, we propose a multi-level alignment to increase the generality of mapped visual and textual features, which contains a semantic alignment to reduce the semantic gap and a cross-modality alignment to alleviate the modality gap, and a cross-domain alignment to reduce the domain gap.

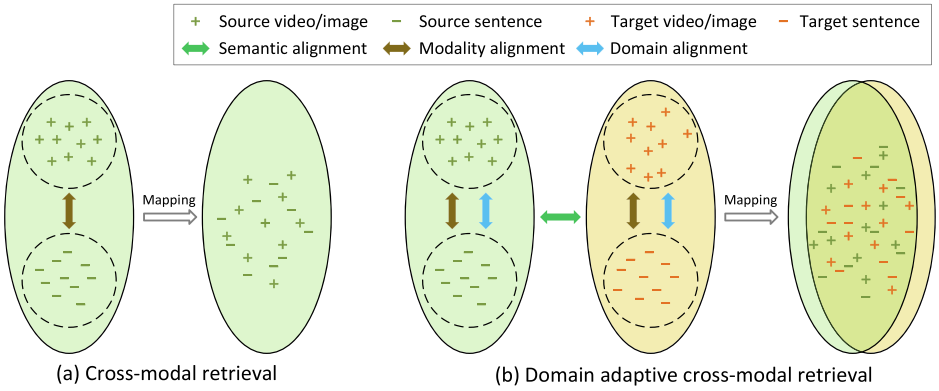


Fig. 3. A toy example showing the difference between traditional cross-modal retrieval and domain adaptive cross-modal retrieval. (a) Traditional cross-modal retrieval aims to map data of different modalities into a common space where semantically relevant data are near and irrelevant ones are far away. But it does not consider the case of data from different domains. (b) The goal of domain adaptive cross-modal retrieval is to learn a common space not only for the data of different modalities, but also for the data of different domains. To achieve it, we propose MAN which reduces the semantic gap, the modality gap and the domain gap by three alignments.

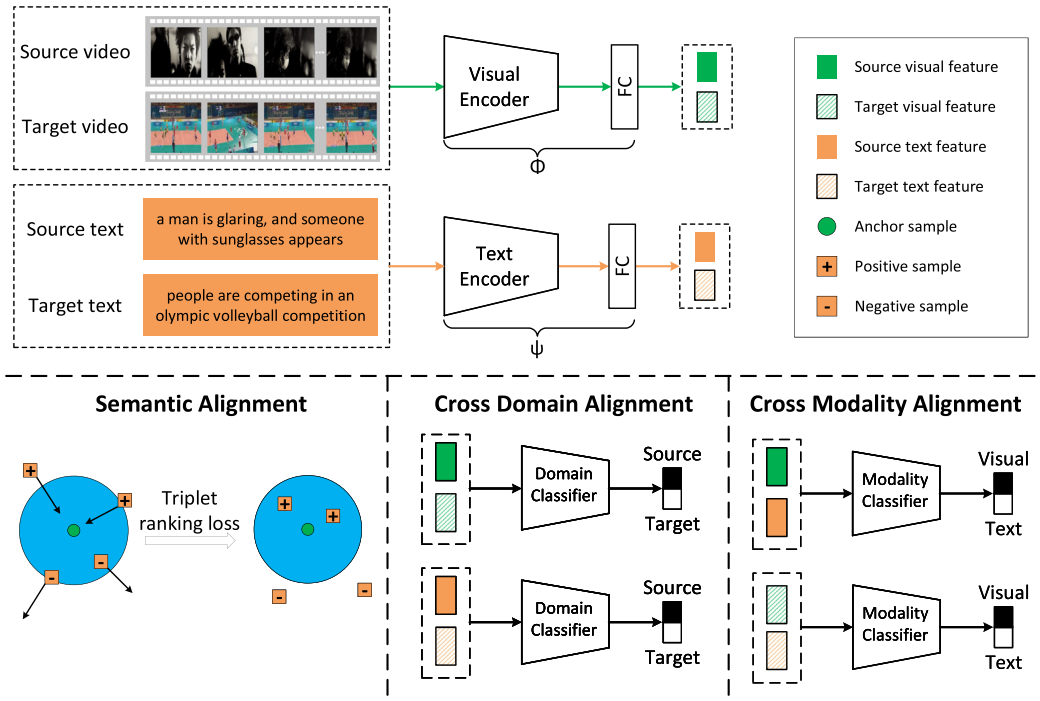


Fig. 4. Illustration of our proposed *Multi-level Alignment Network* (MAN) for domain adaptive cross-modal retrieval. MAN first projects visual and text modalities in a common space by a visual mapping and a text mapping respectively, and three alignments are employed: a semantic alignment to reduce the semantic gap, a cross-modality alignment to alleviate the modality gap, and a cross-domain alignment is employed to reduce the domain gap.

Semantic Alignment. The basic requirement of mapped features in the common space is semantically discriminative, which makes the retrieval model able to find the semantically relevant items to the given query. To this end, the popular way is to use a triplet ranking loss which aims to enforce semantically relevant (positive) visual-text pairs close while semantically irrelevant (negative) ones far away in the common space. Following the good practice of using a triplet ranking loss with the hardest negative example mining[56,25,13], we employ this loss over k source datasets. Note that as the target dataset is unlabeled, we can not employ the triplet ranking loss over the target dataset. Specifically, given k source datasets, the triplet ranking loss is defined as:

$$\begin{aligned}\mathscr{L}_{tri} &= \sum_{j=1}^{k}\sum_{i=1}^{m_j}[\max(0,\alpha - s(\phi(v_{ji}^s),\psi(t_{ji}^s)) + s(\phi(v_{ji}^{s^-}),\psi(t_{ji}^s))) \\ &+ \max(0,\alpha - s(\phi(v_{ji}^s),\psi(t_{ji}^s)) + s(\phi(v_{ji}^s),\psi(t_{ji}^{s^-})))],\end{aligned} \tag{1}$$

where α is the margin constant, (v_{ji}^s, t_{ji}^s) are i-th relevant visual-text pairs from j-th source dataset, $\phi(v_{ji}^s)$ and $\psi(t_{ji}^s))$ are their corresponding mapped feature vectors in the common space and $s(\cdot,\cdot)$ indicates the similarity metric function. We implement $s(\cdot,\cdot)$ with the cosine similarity as it normalizes feature vectors and is found to be better than the dot product or Euclidean distance according to our preliminary experiments. Moreover, $v_{ji}^{s^-}$ and $t_{ji}^{s^-}$ respectively denote the negative visual item and textual item for the positive pair (v_{ji}^s, t_{ji}^s). The two negatives are not randomly sampled, we choose the most similar yet negative items in the current mini-batch. Instead of applying a computationally expensive scheme that samples negative samples in the whole instance space, we perform sampling in the mini-batch for high efficiency.

Cross Domain Alignment. As the above semantic alignment only considers the source datasets while ignores the target dataset, the feature learned with only the semantic alignment may not work well for the target domain. To alleviate it, we additionally introduce a cross domain alignment which aims to make the learned feature domain-invariant. The goal of cross domain alignment is learning a representation indistinguishable between the source domains and the target domain, thus make the model learned on the labeled source domain work well for the target domain.

Inspired by the training idea of GAN [50] which has been successfully used for a number of feature alignment works [48,57,11], we also use it to align the features between the source and target domains. A GAN model is typically comprised of a generator G and a discriminator D, which is usually trained with a two-player adversarial game with G and D where G tries to fool D while D tries to make accurate predictions. In our model, we regard two mapping modules, $\phi(\cdot)$ and $\psi(\cdot)$, described in Section 3.2 as the generator which generates feature representations of source and target data. Besides, we additionally introduce domain discriminators that predict whether the input feature is drawn from the source or the target domain. Although we have k source datasets, we perform the same process for every source dataset. Specifically, for the j-th source dataset, we equip it with a domain discriminator f_j^v for the visual modality and a domain discriminator f_j^t for the textual modality. The discriminators are implemented with binary classifiers that take a mapped feature vector in the common space and outputs a scalar indicating the probability of the input from the source dataset. To train the discriminators, we use the cross-entropy loss. For the j-th source dataset, the loss is the summation of cross-entropy losses respectively for the visual and textual modalities:

$$\begin{aligned}\mathscr{L}_d^j &= -\sum_{i=1}^{m_j}\log f_j^v(v_{ji}^s) - \sum_{i=1}^{n}\log(1 - f_j^v(v_{ji}^t)) - \sum_{i=1}^{m_j}\log f_j^t(t_{ji}^s) \\ &- \sum_{i=1}^{n}\log(1 - f_j^t(t_{ji}^t)),\end{aligned} \tag{2}$$

Finally, given k source datasets, the whole loss of the cross domain alignment is given by:

$$\mathscr{L}_d = \sum_{j=1}^{k} \mathscr{L}_d^j. \tag{3}$$

We wish the mapped feature from different domains are aligned as much as possible, so that they cannot be distinguished by domain classifiers. In other words, two mapping modules, $\phi(\cdot)$ and $\psi(\cdot)$, try to maximize the loss of Eq. 3. By contrast, the domain discriminators try to minimize the loss. Since the optimization goals of mapping modules and discriminators are opposite, the training process runs as a minimax game of the two concurrent sub-processes:

$$\theta_D^* = \underset{\theta_D}{argmin}\ \mathscr{L}_d\ , \quad \theta_G^* = \underset{\theta_G}{argmax}\ \mathscr{L}_d, \tag{4}$$

where θ_D and θ_G denote all the trainable parameters of the domain classifiers and the mapping modules, respectively. In practice, maximizing $\mathscr{L}_d$ directly is hard, so we insert a gradient reversal layer [38] before the discriminators to reverse the gradient of θ_D. Therefore, the minimax optimization can be performed simultaneously by only minimizing $\mathscr{L}_d$.

Note that we have also tried a $(k+1)$-class classifier (each class corresponds to a dataset) as the discriminator that predicts which dataset the input feature comes from. But we found that their performance slightly worse than multiple binary classifiers in our preliminary experiments. We attribute it to that the multi-class discriminator is expected not only to discriminate the target and source datasets, but also to discriminate different source datasets, which may increase the optimization difficulty of the learning objective.

Cross Modality Alignment. Features of different modalities usually have inconsistent distributions and representations. Previous methods [12,25,24] typically project data from different modalities into a common space by modeling the consistency of the corresponding video and text pairs, while do not take care of the distribution consistency between different modalities. Moreover, although the triplet ranking loss to some extent reduces the cross modality gap, it is only employed for the labeled source domain. Therefore, the cross modality gap of the target domain still exists. Therefore, similar to the cross domain alignment, we additionally introduce a cross modality alignment to learn modality-invariant features. Specifically, we introduce two modality discriminators, one is for the source domains g^s and the other is for the target domain g^t. The cross-entropy loss is also employed to train the modality discriminator. Formally, given k source datasets and a target dataset, the loss of the cross modality alignment is defined as:

$$\begin{aligned}\mathscr{L}_m = &-\sum_{j=1}^{k}\sum_{i=1}^{m_j}\left[\log g^s(v_{ji}^s) + \log(1 - g^s(t_{ji}^s))\right] \\ &- \sum_{i=1}^{n}\left[\log g^t(v_i^t) + \log(1 - g^t(t_i^t))\right],\end{aligned} \tag{5}$$

The optimization process is the same as Eq. 4, and a gradient reversal layer is also employed.

3.4. Joint Training and Inference

The overall loss of our proposed model is the sum of the triplet ranking loss ($\mathscr{L}_{tri}$), the cross-domain adversarial loss ($\mathscr{L}_d$), and the cross-modality adversarial loss ($\mathscr{L}_m$), that is:

$$\mathscr{L} = \mathscr{L}_{tri} + \gamma\mathscr{L}_d + \epsilon\mathscr{L}_m, \tag{6}$$

where γ and ϵ are the trade-off coefficients. Our proposed MAN can be trained in an end-to-end manner by minimizing the overall loss of $\mathscr{L}$. Note that gradient reversal layers are used in the adversarial losses.

After the model being trained, we can use two mapping modules, $\phi(\cdot)$ and $\psi(\cdot)$, with the trained parameters to conduct cross-modal retrieval for the target domain. Specifically, given a visual item v^t and a sentence t^t from the target domain, we use cosine similarity over their corresponding mapped feature to measure their similarity, that is:

$$s(v^t, t^t) = \frac{\phi(v^t)\cdot\psi(t^t)}{\|\phi(v^t)\|\cdot\|\psi(t^t)\|}. \tag{7}$$

For text-to-image retrieval, given a sentence query, we sort all the candidate images in descending order in terms of their cosine similarity with the given sentence query. For image-to-text retrieval, given an image query, we sort all the candidate sentences in a similar way. For text-to-video retrieval and video-to-text retrieval, similar process are performed based on the two learned mapping modules for videos and text.

4. Experiments

In order to verify the effectiveness of our proposed MAN for domain-adaptive cross-model retrieval, we evaluate it in the context of video-text retrieval and image-text retrieval.

4.1. Experiments on Domain Adaptive Video-text Retrieval

4.1.1. Experimental Setup

Datasets. As there are no existing domain adaptive video-text retrieval benchmarks, we build a retrieval benchmark based on three common video-text retrieval datasets: VATEX, TGIF, and MSR-VTT. VATEX and TGIF are used as the source datasets, and MSR-VTT is regarded as the target dataset. An overview of the three datasets used in the domain adaptive video-text retrieval experiments is given in Table 2.

VATEX [23]. VATEX is a new large-scale multilingual video description dataset. It contains over 41,250 videos and 825,000 captions in both English and Chinese (10 English and 10 Chinese captions for each video). We have successfully downloaded 5,464 videos, and we utilize all these videos with the corresponding English captions as our source dataset.

TGIF [58]. The Tumblr GIF (TGIF) dataset contains 100K animated GIFs collected from Tumblr and 120 K sentences describing the visual content of the animated GIFs. It is used as another source dataset.

MSR-VTT [21]. The MSR-VTT dataset, originally developed for video captioning, consists of 10K web video clips and 200K natural sentences describing the visual content of the clips. The average number of sentences per clip is 20. We use the official data partition, *i.e.,* 6,513 clips for training, 497 clips for validation, and the remaining 2,990 clips for testing.

Here, we consider data of different datasets being from different domains, as a specific dataset is typically collected in a distinct manner. For the above three datasets, the sources of videos are different: videos in VATEX and MSR-VTT are collected from YouTube while videos in TGIF are derived from Tumblr. It usually causes variances in video frame rate, compression rate or visual quality, etc. Besides, the length of videos in VATEX is much longer than that of TGIF and MSR-VTT.

Evaluation Metrics. Following previous works [15,13] for video-text retrieval, we report rank-based performance metrics, namely $R@K$ ($K = 1, 5, 10$) and Median rank (Med r). $R@K$ is the percentage of test queries for which at least one relevant item is found among the top-K retrieved results. Med r is the median rank of the first relevant item in the search results. Higher $R@K$, mAP,

Table 2
Three datasets used in the domain adaptive video-text retrieval experiments.

Dataset	Domain	#Videos	#Sentences
VATEX	Source	5,464	54,640
TGIF	Source	100 K	120 K
MSR-VTT	Target	10 K	200 K

and lower Med r mean better performance. For overall comparison, we report the sum of all recalls (SumR).

Implementations. Before proceeding to the experiments, we first detail our implementations. Our implementation is modified based on [13]. For text preprocessing, we first convert all words to the lowercase and then replace words that occurring less than five times in the training set with a special token. Each word in the sentence is then initialized to a dense vector using a word2vec [59] model provided by [24], which trained word2vec on English tags of 30 million Flickr images. For videos, we sample video frames uniformly with a pre-specified interval of 0.5 s, and extract its feature by pre-trained ResNext-101 provided by [56]. For ease of reference, we name the feature as ResNext-101. For the text and video encoding, we adopt multi-level text encoding and video encoding [13] as its state-of-the-art performance for video-text related tasks. The dimensionality of the common space is set to 2,048. For all the discriminators, we utilize a three-layer MLP with a structure of 2048–2048-2, and a ReLU is used for hidden layer activation and a Softmax is used for the output layer.

During training, we use Adam [60] with an initial learning rate of 0.0001. The learning rate decays every epoch by a multiplier of 0.99. We set the max epochs as 50 and mini-batch size as 128, and the margin α of triplet ranking loss as 0.2. For trade-off coefficients, we set $\gamma = \epsilon = 0.01$. Following previous work [24], we take an adjustment schedule that once the validation loss does not decrease in three consecutive epochs, we divide the learning rate by 2. Early stop occurs if the validation performance does not improve in ten consecutive epochs.

4.1.2. Performance comparison

As there are no existing works targeting the domain adaptive video-text retrieval, we compare our proposed method with the general cross-modal retrieval methods. We select the following five methods, considering their source codes publicly available:

- VSE++ [25]: A state-of-the-art text-image retrieval model, which is commonly used as the strong baseline model for text-video retrieval. We replace its image-side branch with a mean pooling on frame-level features followed by a FC layer.
- MEE [61]: It learns multiple common spaces for similarity measurement, and the weighted sum of similarities in the multiple spaces is regarded as the final video-text similarity.
- Howto100m [14]: It projects videos and text into a common space by a gated embedding module respectively, and a triplet ranking loss is employed to train the model.
- W2VV++ [56]: It jointly utilizes BoW, word2vec, and GRU to extract multi-granularity features of text, and also uses a triplet ranking loss for model training.
- DualEncoding [13]: It utilizes a multi-level video encoding and multi-level text encoding to encode video and sentence, and a FC layer is respectively further employed to map two modalities into a common space. A triplet ranking loss is also used to train the model.

For a direct comparison, we have re-trained the above five methods with their open-sourced codes, and using the same ResNext-101 feature. As the above methods ignore the domain gap, we simply combine all the source datasets we used as the training data.

Learning with multiple sources. Table 3 summarizes the performance of different models learning with multiple sources. Our proposed MAN consistently outperforms five cross-modal retrieval models. Note that the compared five models ignore the domain gap between the source domain and target domain, while our model jointly considers the semantic gap, domain gap, and modality gap by the multi-level alignment. The result verifies the effectiveness of our multi-level alignment for domain adaptive video-text retrieval. Additionally, DualEncoding can be regarded as a degraded version of our model without the cross-domain alignment and cross-modality alignment. The better performance of our model than DualEncoding shows the importance of these two alignments.

Fig. 5 shows some text-to-Video retrieval examples obtained by our proposed MAN and the baseline method DualEncoding, where the baseline method only uses a triplet ranking loss for training and ignores the domain gap. Our proposed MAN gives better performance. For example, for the query Q1 of "a person made a paper plane", the results returned by our model are all about *making a paper plane*. But, some videos returned by the baseline are about "a real plane". The results further show the effectiveness of MAN for domain adaptive video-text retrieval.

Learning with a single source. Although our proposed MAN is designed for learning with multiple sources, it is also able to learn with a single source dataset. Table 4 shows the performance comparison with the other five methods using TGIF or VATEX as the source dataset. Our MAN again performs the best. The result also verifies the effectiveness of MAN for domain adaptive video-text retrieval with a single source dataset.

4.1.3. Ablation studies

In order to investigate the contribution of each component in our proposed method, we perform ablation studies in the context of video-text retrieval. Table 5 summarizes the results. The performance of the degraded variants is clearly worse than our full MAN, showing the importance of each component. To be specific, removing any particular loss degrades the performance, and the variants without the triplet ranking loss $\mathscr{L}_{tri}$ performs the worst. As $\mathscr{L}_{tri}$ is employed for semantic alignment, the result shows that the semantic alignment is essential for domain adaptive video-text retrieval. Additionally, we also observe that the variant without both $\mathscr{L}_m$ and $\mathscr{L}_d$ losses performs worse than that removing one of them in terms of SumR. The results not only show the importance of these two losses, but also verify their complementarity for the domain adaptive video-text retrieval task.

Additionally, we report another degraded variant MAN_s which has the same model structure with MAN, while only roughly creates an aggregated source dataset with the two source datasets. Its worse performance reveals the necessity of distinguishing multiple source datasets for video-text retrieval learned from multiple sources.

4.2. Experiments on Ad-hoc Video Search

4.2.1. Experimental setup

In this experiment, we compare our MAN with state-of-the-art methods in TRECVID Ad-hoc Video Search (AVS) task that is the closest task to our evaluation setting. Given an ad hoc query, the task aims to return a list of 1,000 shots from the test collection ranked according to their relevance with the given query. As TRECVID does not specify training data, in which we need to train a generalized model using other existing datasets and transfer learned knowledge from the existing datasets to the AVS task.

Table 3
Performance comparison on domain adaptive video-text retrieval with multiple source datasets. *Source*: TGIF + VATEX, Target: MSR-VTT. Larger R@{1,5,10} and smaller Med r indicate better performance. Our proposed method MAN performs the best.

	Text-to-Video Retrieval				Video-to-Text Retrieval				SumR
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
VSE++ [25]	7.9	20.0	27.6	54.0	16.3	33.2	41.6	18.0	146.6
Howto100m [14]	8.3	20.4	28.1	51.0	15.5	32.9	42.4	17.0	147.6
MEE [61]	8.8	22.2	30.6	41.0	16.3	34.5	44.3	15.0	156.7
W2VV++ [56]	9.3	23.4	31.2	37.0	16.0	34.8	44.9	15.0	159.6
DualEncoding [13]	9.3	23.1	31.5	38.0	16.5	35.2	45.4	14.0	161.0
MAN (this work)	**10.1**	**23.8**	**32.8**	**35.0**	**18.0**	**36.9**	**46.6**	**13.0**	**168.2**

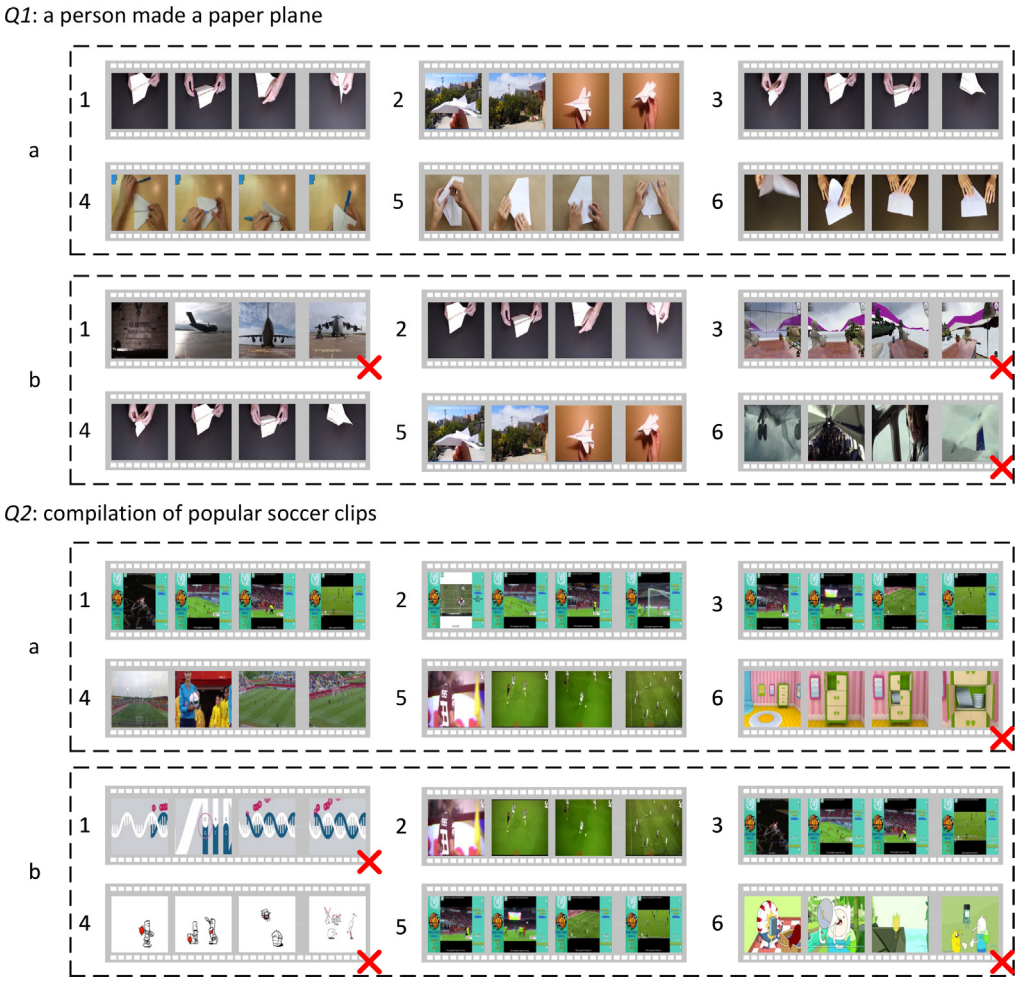


Fig. 5. Text-to-Video retrieval examples obtained by our proposed MAN and the baseline method DualEncoding. These two models are trained on TGIF + VATEX and tested on MSR-VTT. Red × indicates videos are not semantically with the input query. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Datasets. IACC.3 is the official test set for the TRECVID AVS task 2016–2018 [62]. The set contains 4,593 Internet archive videos (600 h) with Creative Commons licenses in MPEG-4/H.264 format. Video duration ranges from 6.5 min to 9.5 min, with a mean duration of approximately 7.8 min. Automated shot boundary detection has been performed by the task organizers, resulting in 335,944 video clips in total. Per year TRECVID specifies 30 distinct queries of varying complexity. Following the previous work [13], we utilize the joint collection of MSR-VTT and TGIF as the training data. Therefore, MSR-VTT and TGIF are regarded as the source datasets, and IACC.3 is the target dataset. For video feature, we utilize 2048-dim ResNext-101 provided by [56].

Evaluation Metrics. We report inferred Average Precision (infAP), the official performance metric used by the TRECVID AVS task. The overall performance is measured by averaging infAP scores over the queries.

Baselines. For method comparison, we include the top 3 entries of each year, *i.e.*, [63–65] for 2016, [66–68] for 2017 and [69–71] for 2018. Besides we include publications that report performance on the tasks, *i.e.*, [72,73]. Among them, [69] fuses three W2VV++ variants with different settings. [70] uses two attention networks, besides the classical concept-based representation. [71] is based with VSE++. Notice that visual features and training data used by these methods vary, meaning the conclusions drawn from this

Table 4
Performance comparison on domain adaptive video-text retrieval with a single source dataset. Target: MSR-VTT. Our proposed method MAN consistently performs better.

	Text-to-Video Retrieval				Video-to-Text Retrieval				SumR
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
Source: TGIF									
MEE [61]	5.9	16.3	23.2	78.0	10.3	23.0	32.0	37.0	110.7
VSE++ [25]	6.6	17.0	23.8	76.0	12.3	28.0	36.6	25.0	124.2
Howto100m [14]	6.9	17.5	24.3	72.0	12.5	27.6	36.6	24.0	125.3
W2VV++ [56]	8.6	21.4	29.7	43.0	14.9	31.5	42.1	17.0	148.2
DualEncoding [13]	8.5	21.5	29.7	42.0	15.1	32.4	42.3	17.0	149.5
MAN (this work)	**9.1**	**22.5**	**30.6**	**39.0**	**17.1**	**34.8**	**44.2**	**15.0**	**158.3**
Source: VATEX									
VSE++ [25]	4.2	12.4	18.3	115.0	7.4	19.0	26.0	50.0	87.3
Howto100m [14]	4.9	14.1	20.5	94.0	7.6	20.5	29.6	40.0	97.0
MEE [61]	5.0	14.9	21.6	77.0	7.5	20.2	28.1	46.0	97.3
W2VV++ [56]	5.4	15.2	22.3	75.0	9.7	23.7	31.4	36.0	107.7
DualEncoding [13]	5.6	15.9	22.9	74.0	9.7	23.6	31.6	36.0	109.3
MAN (this work)	**5.9**	**16.8**	**23.5**	**65.0**	**10.4**	**24.7**	**32.3**	**31.0**	**113.6**

Table 5
Ablation studies of our proposed MAN on domain adaptive video-text retrieval with multiple source datasets. *Source*: TGIF + VATEX, Target: MSR-VTT. $\mathscr{L}_{tri}$ is the triplet ranking loss, $\mathscr{L}_d$ is the cross-domain adversarial loss, and $\mathscr{L}_m$ denotes the cross-modality adversarial loss. Our full model performs better than degraded ones.

	Text-to-Video Retrieval				Video-to-Text Retrieval				SumR
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
MAN (Full model)	**10.1**	**23.8**	**32.8**	**35.0**	**18.0**	**36.9**	**46.6**	**13.0**	**168.2**
MAN w/o $\mathscr{L}_{tri}$	0.0	0.2	0.4	1433.0	0.0	0.2	0.4	2690.0	1.2
MAN w/o $\mathscr{L}_m$	9.7	23.6	32.1	36.0	17.3	36.1	46.4	13.0	165.2
MAN w/o $\mathscr{L}_d$	9.7	23.7	32.2	36.0	17.9	36.7	46.6	13.0	166.8
MAN w/o $\mathscr{L}_m$ and $\mathscr{L}_d$	9.3	23.1	31.5	38.0	16.5	35.2	45.4	14.0	161.0
MAN_s	9.3	23.3	32.0	37.0	17.1	35.9	45.5	14.0	163.1

comparison are at a system level. So we also compare VSE++ [25], W2VV [24], W2VV++ [56] and Dual encoding [13] using the same training data and the same ResNeXt-101 feature.

4.2.2. Experimental results

Table 6 summarizes the performance of different methods on the TRECVID 2016, 2017, and 2018 AVS tasks (also include the top 3 entries of each year). Our proposed method MAN surpasses the state-of-the-art in terms of the overall infAP. Particularly, MAN achieves the infAP of 0.248 at 2017 AVS task, which exceeds recent methods with a large margin. It is worth noting that VSE++ [25], W2VV++ [56] and Dual encoding [13] use the same training data and video feature, but none of them consider the domain gap between the source and target domains. The higher overall performance of our proposed model shows its superior generalization ability for AVS task.

4.3. Experiments on domain adaptive image-text retrieval

4.3.1. Experimental setup

Datasets. For domain adaptive image-text retrieval, we build an evaluation benchmark based on CUHK-PEDES, Fashion200K, and DeepFashion. CUHK-PEDES and Fashion200K are used as the source datasets, and DeepFashion is regarded as the target dataset. An overview of the three datasets used in the domain adaptive image-text retrieval experiments is given in Table 7.

CUHK-PEDES [74]. The CUHK-PEDES dataset contains 40,206 pedestrian images, each of which is described by two textual descriptions. There are 34,054 images and corresponding 68,108 sentence descriptions in the training set. The validation set and test set consist of 3,078 and 3,074 images, respectively. Since we select the dataset as source training data, only the image-text pairs in the training set are used.

Fashion200K [22]. The Fashion200K dataset contains over 200,000 clothing images, and each image is described by one textual description. The dataset is split into 172,049 images for training, 12,164 images for validation, and 25,331 images for testing. Similarly, we only use the image-text pairs in the training set as the source data.

DeepFashion [75]. The DeepFashion dataset is a large-scale fashion-related benchmark for various tasks, such as fashion attribute prediction, landmark detection. Here, we use the Fashion Synthesis benchmark which was originally developed for the fashion image generation task, while has also been used for image-text retrieval recently [76]. The dataset totally has 78,979 fashion images ranging from well-posed shop images to unconstrained consumer photos. Each image is annotated with one sentence caption which describes the visual content of the clothes in the images. These images are officially divided into 70,000 for training and 8,979 for evaluation. As there is no validation set available, we randomly sample 1,000 images from the test set as the validation set, and another 1,000 images as our final test set.

Evaluation Metrics. We use the $R@K$ ($K = 1, 5, 10$) as the performance metrics. For overall comparison, we also report the sum of all recalls (SumR).

Implementations. For text encoding, we first convert all words to the lowercase and then replace words that occurring less than five times in the training set with a special token. We set the hidden-state size of each LSTM of bidirectional LSTM model to be 512. For image encoding, we resize images to 224×448, and horizontally flip them with a possibility of 0.5. We use MobileNet [77] to extract deep features of images. The dimensionality of the common space is set to 512. For all discriminators in the cross-domain and cross-modality alignments, we use the same structure in video-text retrieval. During training, we use Adam [60] with an initial learning rate of 0.0002. The learning rate decays every epoch by a multiplier of 0.95. We set the max epochs as 300, mini-

Table 6
Performance comparison with the state-of-the-art methods on the TRECVID 2016/ 2017/ 2018 AVS tasks.

	TRECVID edition			
	2016	*2017*	*2018*	**overall**
Top-3 TRECVID finalists:				
Rank 1	0.054 [63]	0.206 [66]	0.121 [69]	–
Rank 2	0.051 [64]	0.159 [67]	0.087 [70]	–
Rank 3	0.040 [65]	0.120 [68]	0.082 [71]	–
Literature methods:				
W2VV [24]	0.050	0.071	0.022	0.048
Markatopoulou et al. [73]	0.064	–	–	–
VideoStory [72]	0.087	0.150	–	–
VSE++ [25]	0.123	0.154	0.074	0.117
W2VV++ [56]	0.137	0.168	0.088	0.131
DualEncoding [13]	**0.159**	0.208	0.116	0.161
MAN (this work)	0.145	**0.248**	**0.126**	**0.173**

Table 7
Three datasets used in the domain adaptive image-text retrieval experiments.

Dataset	Domain	#Images	#Sentences
CUHK-PEDES	Source	34,054	68,108
Fashion200k	Source	209,544	209,544
DeepFashion	Target	78,979	78,979

batch size as 16, and margin in triplet ranking loss as 0.2. For trade-off coefficients, we empirically set $\gamma = 0.001, \epsilon = 0.001$. Early stop occurs if the validation performance does not improve in ten consecutive epochs.

Baselines. Besides the baselines VSE++ and W2VV++ compared in the video-text retrieval, we also compare two common image-text retrieval models, *i.e.,* CMPC + CMPM [31], VSA-AE-MMD [78]. CMPC + CMPM has two same mapping modules to project images and text into a common space, but does not consider the domain gap between the source and target domains. VSA-AE-MMD employs a maximum mean discrepancy to align the source and target domains, but its method only works with one source domain. For CMPC + CMPM, we use the open-sourced code and re-train with the same datasets. Since there is no public code of VSA-AE-MMD, we re-implement it by ourselves and use the same image and text encoding.

4.3.2. Performance comparison

Table 8 summarizes the performance on the target dataset Fashion200K, where models are trained on two source datasets of Fashion200K and CUHK-PEDES. Our proposed MAN achieves the best overall performance SumR of 71.1. Among all the compared models, CMPC + CMPM performs the worst, whose two mapping structures for visual and textual modalities are the same as our MAN. The result again verifies the effectiveness of our multi-level alignment for domain adaptive cross-modal retrieval. For text-to-image retrieval, W2VV++ is better than our model. W2VV ++ uses a strong multi-level text encoding strategy including BoW, word2vec, and bidirectional GRU, while our MAN only utilizes a bidirectional LSTM. We attribute the higher performance of W2VV++ to its strong text encoding. Since this paper mainly focuses on how to align the cross-domain and cross-modality representations, we leave the exploration of strong visual and textual encodings for future study. Table 9 shows the performance comparison of domain adaptive image-text retrieval with a single

Table 8
Domain adaptive image-text retrieval learning with multiple sources. *Source*: Fashion200K + CUHK-PEDES, Target: DeepFashion.

	Text-to-Image Retrieval			Image-to-Text Retrieval			SumR
	R@1	R@5	R@10	R@1	R@5	R@10	
CMPC + CMPM [31]	0.3	1.8	4.0	2.8	10.9	18.0	37.8
VSE++ [25]	2.8	9.6	15.4	3.8	14.3	21.1	67.0
W2VV++ [56]	**3.1**	**11.7**	**18.5**	2.6	12.5	20.3	68.7
MAN (this work)	2.9	10.1	16.8	**4.2**	**15.1**	**22.0**	**71.1**

Table 9
Domain adaptive image-text retrieval with a single source. Target: DeepFashion.

	Text-to-Image Retrieval			Image-to-Text Retrieval			SumR
	R@1	R@5	R@10	R@1	R@5	R@10	
Source: Fashion200K							
VSE++ [25]	1.3	5.5	8.6	2.5	7.2	12.2	37.6
VSA-AE-MMD [78]	1.8	4.8	9.1	1.2	6.5	9.3	32.7
CMPC + CMPM [31]	1.2	4.0	7.4	0.8	4.9	9.4	27.7
W2VV++ [56]	1.4	3.7	7.0	1.9	5.0	9.7	28.7
MAN (this work)	**2.0**	**8.3**	**12.7**	**2.5**	**8.9**	**14.2**	**48.6**
Source: CUHK-PEDES							
VSE++ [25]	2.2	10.4	16.6	2.6	10.6	18.5	60.9
VSA-AE-MMD [78]	**3.1**	11.2	15.4	**3.6**	11.3	17.2	61.8
CMPC + CMPM [31]	2.9	**12.5**	**19.3**	2.5	9.4	16.3	62.9
W2VV++ [56]	1.8	8.2	13.9	2.6	10.7	18.3	55.5
MAN (this work)	2.7	11.3	18.2	3.3	**11.9**	**20.4**	**67.8**

source dataset. Again, MAN gives the best SumR score. The results further show the effectiveness of our model for domain adaptive image-text retrieval.

5. Conclusion

In this paper, we introduce a new task *domain adaptive cross-modal retrieval*, where the training data and the test data have different distributions. Due to the semantic gap, the domain gap and the modality gap exist in the task, this task is very challenging. Targeting this task, we propose a Multi-level Alignment Network, which learns well-aligned visual-semantic embeddings cross domains and modalities by three alignment modules. Our model is orthogonal to the visual and textual encoders, allowing us to flexibly embrace state-of-the-art visual and textual encoder structures. Extensive experiments in the context of video-text retrieval and image-text retrieval verify the effectiveness of our proposed MAN. In the future, we will explore the visual and textual encoder structures that are suitable for domain adaptive cross-modal retrieval. While in this paper we demonstrate the domain adaptation idea in the specific case of cross-modal retrieval, they can in principle generalize to other retrieval based tasks, such as video-to-video retrieval.

CRediT authorship contribution statement

Jianfeng Dong: Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Project administration. **Zhongzi Long:** Data curation, Software, Validation, Visualization, Writing - original draft. **Xiaofeng Mao:** Methodology, Data curation, Software, Formal analysis, Writing - original draft, Resources. **Changting Lin:** Writing - original draft, Writing - review & editing, Resources. **Yuan He:** Writing - review & editing, Resources. **Shouling Ji:** Writing - review & editing, Resources, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

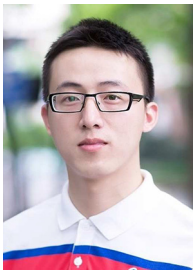
Acknowledgments

This work was partly supported by the National Key Research and Development Program of China under No. 2018YFB0804102 and No. 2020YFB2103802, NSFC under No. 61902347, No. 61772466, U1936215, and U1836202, the Zhejiang Provincial Natural Science Foundation under No. LQ19F020002,No. LR19F020003 and No. LQ21F020010, the Public Welfare Technology Research Project of Zhejiang Province under No. LY21F020010, the Science and Technology Program of Zhejiang Province under No. 2021C01120, Alibaba-Zhejiang University Joint Institute of Frontier Technologies, and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

References

[1] K. Wang, Q. Yin, W. Wang, S. Wu, L. Wang, A comprehensive survey on cross-modal retrieval, arXiv preprint arXiv:1607.06215 (2016)..
[2] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges, TCFVT 28 (9) (2017) 2372–2385.
[3] X. Xu, K. Lin, L. Gao, H. Lu, H.T. Shen, X. Li, Learning cross-modal common representations by private-shared subspaces separation, IEEE Trans. Cybern. (2020).
[4] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, H.T. Shen, Universal weighting metric learning for cross-modal matching, CVPR (2020) 13005–13014.
[5] E. Yang, C. Deng, C. Li, W. Liu, J. Li, D. Tao, Shared predictive cross-modal deep quantization, IEEE Trans. Neural Networks Learn. Syst. 29 (11) (2018) 5292–5303.
[6] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, X. Gao, Pairwise relationship guided deep hashing for cross-modal retrieval, AAAI (2017) 1618–1625.
[7] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3d room layout estimation from a single rgb image, IEEE Trans. Multimedia 22 (11) (2020) 3014–3024.
[8] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[9] C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, Y. Zhang, Depth image denoising using nuclear norm and learning graph model, ACM Trans. Multimedia Computing, Commun., Appl. 16 (4) (2020) 1–17.
[10] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, CVPR (2016) 5005–5013.
[11] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H. T. Shen, Adversarial cross-modal retrieval, in: ACM Multimedia, 2017, pp. 154–162..
[12] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, J. Shao, Camp: Cross-modal adaptive message passing for text-image retrieval, in: ICCV, 2019, pp. 5764–5773..
[13] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, X. Wang, Dual encoding for zero-example video retrieval, CVPR (2019) 9346–9355.
[14] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, in: ICCV, 2019, pp. 2630–2640..
[15] N.C. Mithun, J. Li, F. Metze, A.K. Roy-Chowdhury, Learning joint embedding with multimodal cues for cross-modal video-text retrieval, ICMR (2018) 19–27.
[16] V. Gabeur, C. Sun, K. Alahari, C. Schmid, Multi-modal transformer for video retrieval, in: ECCV, Vol. 5, 2020..
[17] S. Chen, Y. Zhao, Q. Jin, Q. Wu, Fine-grained video-text retrieval with hierarchical graph reasoning, CVPR (2020) 10638–10647.
[18] X. Li, F. Zhou, C. Xu, J. Ji, G. Yang, Sea: Sentence encoder assembly for video retrieval by textual queries, IEEE Trans. Multimedia (2020).
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, NIPS (2017) 5998–6008.
[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, ECCV (2014) 740–755.
[21] J. Xu, T. Mei, T. Yao, Y. Rui, MSR-VTT: A large video description dataset for bridging video and language, in: CVPR, 2016, pp. 5288–5296..
[22] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, L. S. Davis, Automatic spatially-aware fashion concept discovery, in: ICCV, 2017, pp. 1463–1471..
[23] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, W. Y. Wang, Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, in: ICCV, 2019, pp. 4581–4591..
[24] J. Dong, X. Li, C.G. Snoek, Predicting visual features from text for image and video caption retrieval, IEEE Trans Multimedia 20 (12) (2018) 3377–3388.
[25] F. Faghri, D.J. Fleet, J.R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives, BMVC (2018).
[26] J. Dong, X. Li, D. Xu, Cross-media similarity evaluation for web image retrieval in the wild, IEEE Trans. Multimedia 20 (9) (2018) 2371–2384.
[27] W. Wang, J. Gao, X. Yang, C. Xu, Learning coarse-to-fine graph neural networks for video-text retrieval, IEEE Trans. Multimedia (2020).
[28] Z. Feng, Z. Zeng, C. Guo, Z. Li, Exploiting visual semantic reasoning for video-text retrieval, arXiv preprint arXiv:2006.08889 (2020)..
[29] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., Devise: A deep visual-semantic embedding model, NeurIPS (2013) 2121–2129.
[30] R. Kiros, R. Salakhutdinov, R. S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, arXiv preprint arXiv:1411.2539 (2014)..
[31] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, ECCV (2018) 686–701.
[32] Y. Liu, S. Albanie, A. Nagrani, A. Zisserman, Use what you have: Video retrieval using representations from collaborative experts, arXiv preprint arXiv:1907.13487 (2019)..
[33] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, T.-S. Chua, Tree-augmented cross-modal encoding for complex-query video retrieval, SIGIR (2020) 1339–1348.
[34] J. Wu, C.-W. Ngo, Interpretable embedding for ad-hoc video search, in: ACM Multimedia, 2020, pp. 3357–3366..
[35] X. Huang, Y. Peng, Deep cross-media knowledge transfer, CVPR (2018) 8837–8846.
[36] G. Wilson, D.J. Cook, A survey of unsupervised deep domain adaptation, ACM Trans. Intell. Syst. Technol. (TIST) 11 (5) (2020) 1–46.
[37] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.
[38] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: ICML, 2015, pp. 1180–1189..
[39] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, CVPR (2018) 3723–3732.
[40] S. Motiian, M. Piccirilli, D. A. Adjeroh, G. Doretto, Unified deep supervised domain adaptation and generalization, in: ICCV, 2017, pp. 5715–5725..
[41] L. Hedegaard, O. A. Sheikh-Omar, A. Iosifidis, Supervised domain adaptation using graph embedding, arXiv preprint arXiv:2003.04063 (2020)..
[42] M. Long, Y. Cao, J. Wang, M. I. Jordan, Learning transferable features with deep adaptation networks, arXiv preprint arXiv:1502.02791 (2015)..
[43] M. Long, H. Zhu, J. Wang, M.I. Jordan, Unsupervised domain adaptation with residual transfer networks, NeurIPS (2016) 136–144.

[44] M. Long, H. Zhu, J. Wang, M. I. Jordan, Deep transfer learning with joint adaptation networks, in: ICML, 2017, pp. 2208–2217..
[45] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, arXiv preprint arXiv:1511.05547 (2015)..
[46] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, ECCV (2016) 443–450.
[47] B. Bhushan Damodaran, B. Kellenberger, R. Flamary, D. Tuia, N. Courty, Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation, in: ECCV, 2018, pp. 447–463..
[48] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, J. Zheng, Temporal attentive alignment for large-scale video domain adaptation, in: ICCV, 2019, pp. 6321–6330..
[49] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, The Journal of Machine Learning Research 17 (1) (2016) 2096, 2030.
[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, NeurIPS (2014) 2672–2680.
[51] C. Wang, C. Xu, X. Yao, D. Tao, Evolutionary generative adversarial networks, IEEE Trans. Evol. Comput. 23 (6) (2019) 921–934.
[52] S. Sun, H. Shi, Y. Wu, A survey of multi-source domain adaptation, Information Fusion 24 (2015) 84–92.
[53] R. Xu, Z. Chen, W. Zuo, J. Yan, L. Lin, Deep cocktail network: Multi-source unsupervised domain adaptation with category shift, CVPR (2018) 3964–3973.
[54] H. Zhao, S. Zhang, G. Wu, J.M. Moura, J.P. Costeira, G.J. Gordon, Adversarial multiple source domain adaptation, NeurIPS (2018) 8559–8570.
[55] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, Moment matching for multi-source domain adaptation, in: ICCV, 2019, pp. 1406–1415..
[56] X. Li, C. Xu, G. Yang, Z. Chen, J. Dong, W2VV++: fully deep learning for ad-hoc video search, in: ACM Multimedia, 2019, pp. 1786–1794..
[57] N. Sarafianos, X. Xu, I.A. Kakadiaris, Adversarial representation learning for text-to-image matching, ICCV (2019) 5814–5824.
[58] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, J. Luo, Tgif: A new dataset and benchmark on animated gif description, in: CVPR, 2016, pp. 4641–4650..
[59] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013)..
[60] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014)..
[61] A. Miech, I. Laptev, J. Sivic, Learning a Text-Video Embedding from Incomplete and Heterogeneous Data, arXiv preprint arXiv:1804.02516 (2018)..
[62] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. Smeaton, Y. Graham, et al., Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search, in: TRECVID Workshop, 2018..
[63] D.-D. Le, S. Phan, V.-T. Nguyen, B. Renoust, T. A. Nguyen, V.-N. Hoang, T. D. Ngo, M.-T. Tran, Y. Watanabe, M. Klinkigt, et al., Nii-hitachi-uit at trecvid 2016, in: TRECVID Workshop, 2016..
[64] M. Foteini, M. Anastasia, G. Damianos, M. Theodoros, K. Vagia, I. Anastasia, S. Symeonidis, Iti-certh participation in trecvid 2016, in: TRECVID Workshop, 2016..
[65] J. Liang, J. Chen, P. Huang, X. Li, L. Jiang, Z. Lan, P. Pan, H. Fan, Q. Jin, J. Sun, et al., Informedia@ trecvid 2016, in: TRECVID Workshop, 2016..
[66] C. G. M. Snoek, J. Dong, X. Li, X. Wang, Q. Wei, W. Lan, E. Gavves, N. Hussein, D. C. Koelma, A. W. M. Smeulders, University of amsterdam and renmin university at trecvid 2016: Searching video, detecting events and describing video, in: TRECVID Workshop, 2016..
[67] K. Ueki, K. Hirakawa, K. Kikuchi, T. Ogawa, T. Kobayashi, Waseda_meisei at trecvid 2017: Ad-hoc video search., in: TRECVID Workshop, 2017..
[68] P. A. Nguyen, Q. Li, Z.-Q. Cheng, Y.-J. Lu, H. Zhang, X. Wu, C.-W. Ngo, Vireo@ trecvid 2017: Video-to-text, ad-hoc video search, and video hyperlinking., in: TRECVID Workshop, 2017..
[69] X. Li, J. Dong, C. Xu, J. Cao, X. Wang, G. Yang, Renmin university of china and zhejiang gongshang university at trecvid 2018: Deep cross-modal embeddings for videotext retrieval, in: TRECVID Workshop, 2018..
[70] P.-Y. Huang, J. Liang, V. Vaibhav, X. Chang, A. Hauptmann, Informedia@ trecvid 2018: Ad-hoc video search with discrete and continuous representations, in: TRECVID Workshop, 2018..
[71] M. Bastan, X. Shi, J. Gu, Z. Heng, C. Zhuo, D. Sng, A. Kot, Ntu rose lab at trecvid 2018: Ad-hoc video search and video to text, in: TRECVID Workshop, 2018..
[72] A. Habibian, T. Mensink, C.G. Snoek, Video2vec embeddings recognize events when examples are scarce, IEEE Trans. Pattern Anal. Mach. Intell. 39 (10) (2016) 2089–2103.
[73] F. Markatopoulou, D. Galanopoulos, V. Mezaris, I. Patras, Query and keyframe representations for ad-hoc video search, ICMR (2017) 407–411.
[74] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, CVPR (2017) 1970–1979.
[75] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: CVPR, 2016, pp. 1096–1104..
[76] S. Zhu, R. Urtasun, S. Fidler, D. Lin, C. Change Loy, Be your own prada: Fashion synthesis with structural coherence, in: ICCV, 2017, pp. 1680–1688..
[77] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017)..
[78] A. Carraggi, M. Cornia, L. Baraldi, R. Cucchiara, Visual-semantic alignment across domains using a semi-supervised approach, ECCV (2018) 1–16.
[79] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, M. Wang, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–16, https://doi.org/10.1109/TPAMI.2021.3059295.



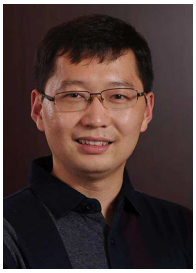
Jianfeng Dong received the B.E. degree in software engineering from Zhejiang University of Technology in 2009, and the Ph.D. degree in computer science from Zhejiang University in 2018, all in Hangzhou, China. He is currently a Research Professor at the College of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China. His research interests include multimedia understanding, retrieval and recommendation. He was awarded the ACM Multimedia Grand Challenge Award in 2016. Has won a number of international competitions including the TRECVID 2016, 2017, 2018 Video-to-Text (VTT) Matching and Ranking task, the MSR Bing Image Retrieval Challenge at ACM Multimedia 2015, and so on.

Zhongzi Long received his B.S. degree in Computer Science from Zhejiang University, Hangzhou, China in 2020. He is currently a research assistant in NESA Lab, Zhejiang University, researching on cross-modal retrieval.

Xiaofeng Mao received the B.S. degree in Internet of things Engineering from Northeastern University in 2016. and the master?s degree in Computer Science and Technology from Harbin Engineering University in 2019. He is currently a Algorithm Engineer at Alibaba Group, Hangzhou, China. His research interests include adversarial machine learning, multimedia understanding, image recognition.

Changting Lin received the Ph.D?s degree in computer science from the Zhejiang University in 2018. His is currentlyis currently an assistant professor in the School of Computer and information Engineering at Zhejiang Gongshang University, China. His research interests include IoT, AI, Blockchain and SDN.

Yuan He received his B.S. degree and Ph.D. degree from Tsinghua University, P.R. China. Currently, he is a Senior Staff Engineer in the Security Department of Alibaba Group, and working on artificial intelligence based content moderation and intellectual property protection systems. His research interests include computer vision, machine learning and AI security.



Shouling Ji is a ZJU 100-Young Professor in the College of Computer Science and Technology at Zhejiang University and a Research Faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology. He received a Ph.D. in Electrical and Computer Engineering from Georgia Institute of Technology and a Ph.D. in Computer Science from Georgia State University. His current research interests include AI and Security, Data-driven Security and Data Analytics. He is a member of IEEE and ACM and was the Membership Chair of the IEEE Student Branch at Georgia State (2012–2013).