Information Sciences 609 (2022) 897–912

Contents lists available at ScienceDirect

Information Sciences

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

Toward a perceptive pretraining framework for Audio-Visual Video Parsing

Jianning Wu [a,1], Zhuqing Jiang [a,b,*,1], Qingchao Chen [c], Shiping Wen [d], Aidong Men [a], Haiying Wang [a]

[a] *School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China*
[b] *Beijing Key Laboratory of Network System and Network Culture, Beijing University of‘ Posts and Telecommunications, Beijing, China*
[c] *National Institute of Health Data Science, Peking University, Beijing, China*
[d] *Australian AI Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia*

ARTICLE INFO

Article history:
Received 10 November 2021
Received in revised form 18 July 2022
Accepted 24 July 2022
Available online 26 July 2022

Keywords:
Multi-modal
Max-margin
Domain-adaptation
Video
AVVP

ABSTRACT

Audio-Visual Video Parsing (AVVP) is a new multi-modal weakly supervised task which aims to detect and localize events leveraging the partial alignment of audio and visual streams and weak labels. We identified two significant challenges in the AVVP: Cross-mode semantic misalignment and Contextual audio-visual dataset bias. For challenge 1, the existing methods tend to leverage the temporal similarity of the features. However, it is inappropriate for our AVVP task because multi-modal features with the same label do not always have the same semantics. Thus, we propose an instance-adaptive multi-modal time series max-margin loss (MTSM) which uses the temporal information to align features adaptively. Furthermore, to restrict the inescapable noise introduced during the feature fusion, we reuse the expression of MTSM in the single-mode. For the second challenge, we argue that bias mitigation should seek help from model generalization. Thus, we propose collocating pre-trained models: either" traverse" or based on domain-adaptation. First, we prove a hypothesis and then propose a method based on the Alternating Direction Method of Multipliers(ADMM) to decouple the optimal pre-trained model collocation solution, which reduces the time consumption. Experiments show that our method outperforms the contrastive methods.

1. Introduction

Benchmarks of audio-visual applications [1–4] keep being updated with the help of multi-modal ideas [5–7] and pre-trained models [8]. These impressive progresses have been driven by large quantities of labelled data(e.g., YouTube-8 M [9]) in the context of supervised learning.

To mitigate the data eagerness, Audio-Visual Video Parsing (AVVP) [10] has been proposed to localize events leveraging the partial alignment of the audio and visual streams in a weakly supervised way.

Many AVVP tasks share one concern – multi-modal feature fusion [11,12]. We argue that two factors affect feature fusion, intrinsic and acquired. The intrinsic factors include the model, the dataset, and training prerequisites. The acquired ones are

[*] Corresponding author.
E-mail addresses: jianningwu@bupt.edu.cn (J. Wu), jiangzhuqing@bupt.edu.cn (Z. Jiang).
[1] The first two authors contribute equally to this work

https://doi.org/10.1016/j.ins.2022.07.144

mainly involved in the training stage, e.g. selection of loss functions and adjustment of hyper-parameters. From these perspectives, we identified two challenges in the AVVP.

1)*Cross-modal semantic misalignment.* Fig. 1 shows an example of the misalignment: the feature similarity of synchronous video and audio clips with the same label is weaker than that of asynchronous with different labels. The misalignment is accentuated in the weakly supervised tasks. Existing methods alleviate this problem by encouraging the similarity of synchronous features while suppressing the dissimilarity of asynchronous features. However, synchronously paired audio-visual semantics are unlikely to be precisely injective mapped (Definition 2) as they may be composed of multiple concepts. Thus, the designed alignment using many-to-many cross-modal mapping is prone to noise.
2)*Contextual and audio-visual Dataset Bias.* There are biases between the datasets of pre-trained models and the datasets of AVVP tasks. Thus, a dedicated selection of pre-trained models is essential to relieving the bias. That is, the audio-visual dataset bias raises the model generalization problems. Specifically, the key to cross-modality bias is a cooperative selection of the audio-visual pre-trained models. We name this cross-modal model searching problem the model collocation problem. How to search and select the optimal combination of cross-modal pre-trained models remains an open question. Because simple trial-and-error costs unaffored computation to find the optimal combination.

To tackle the mentioned challenges, we propose a perceptive pre-training framework. 1) *First,* we propose to learn a joint semantic embedding space for the AVVP misalignment. To be specific, instead of explicitly learning an injective mapping to completely ”align” cross-modal representations, we preserve the temporal semantic structure of each single-modality in the cross-modal latent space. 2) *Second,* We propose a new scheme to search for the optimal combination of pre-training models. More specifically, inspired by the idea of the Alternating Direction Method of Multipliers (ADMM), we decouple the whole optimization into two sub-problems – collocation selection and model training; and then optimize the two sub-problems alternately with time reduced and accuracy maintained.

The main contributions of this paper are:

1. We identify the problems of feature misalignment and the introduction of multiplicative noise and then propose Multimodal Time-series Max-margin losses (MTSM) to solve them.
2. We prove that the existing training methods ignore the importance of combining different pretrained models in the AVVP task.
3. Given the problems mentioned in 2, inspired by the domain adaption, we prove a hypothesis and then combine it with the ADMM strategy to shorten the time.

2. Related works

With the development of single-modal task research, the multi-modal task has attracted more and more attention due to its universality. Many sub-tasks have been derived.[13–20].

Audio-Visual Video Parsing (AVVP) [10] is a derivative task of weakly supervised temporal action localization. It uses the natural audio image flow of video to integrate multi-modal information to complement audio and image information to complete the temporal event location on the audio and image modes. AVVP [10] has a wide potential application in downstream video understanding tasks. (such as monitoring analysis, video summarization [21–23], and retrieval [24–26]). It is a newly introduced multi-modal task that detects and localizes events within the audio and visual streams. Tian [10] proposes this task and presents an LLP data set based on the non-strict symmetry of audio and video time domains, which is the only data set for this task so far. Tian [10] also proposes a hybrid attention network (HAN) framework to solve this problem. Compared with the single-mode temporal localization methods, HAN uses complementary information between both modes, which is better than these single-mode temporal localization methods. Wu [27], based on Tian [10], uses the method of contrast learning to regard different modal features of simultaneous order as positive sample pairs and non-simultaneous order features as negative sample pairs and proposes a contrast loss to enhance the constraint between different modal features. However, this loss based on the contrast learning idea enhances the constraint between the features of the simultaneous order. Furthermore, it treats the features of the non-simultaneous order as negative sample pairs, which is inappropriate for determining the positive and negative sample pairs. We also recognize this problem and design an adaptive construction method of positive and negative sample pairs based on max-margin.

Besides, in the field of image segmentation, Li [15] proposed proposes a novel Context-based Tandem Network (CTNet) by interactively exploring the spatial contextual information and the channel contextual information. This is also inspirable for video tasks because of the inter-frame redundancy of video. In the uni-model domain, Yu [18] devise a Hierarchical Deep Word Embedding (HDWE) model by integrating sparse constraints and an improved RELU operator to address click feature prediction from visual features. It is remarkable for multi-modal feature fusion of multi-modal tasks.

In addition, we notice that the existing multi-modal processing framework introduced some multiplication, which will introduce bad multiplicative noise to the output. For multi-modal tasks, this problem is more significant. Therefore, we also refer to the thought of max-margin, for the introduction of the multiplicative noise is restrained.

Meanwhile, we notice that, like most multi-modal problems, AVVP uses low-dimensional features extracted by the pre-training network as model input instead of high-dimensional and sparse source data. Thus, it will save the training time of

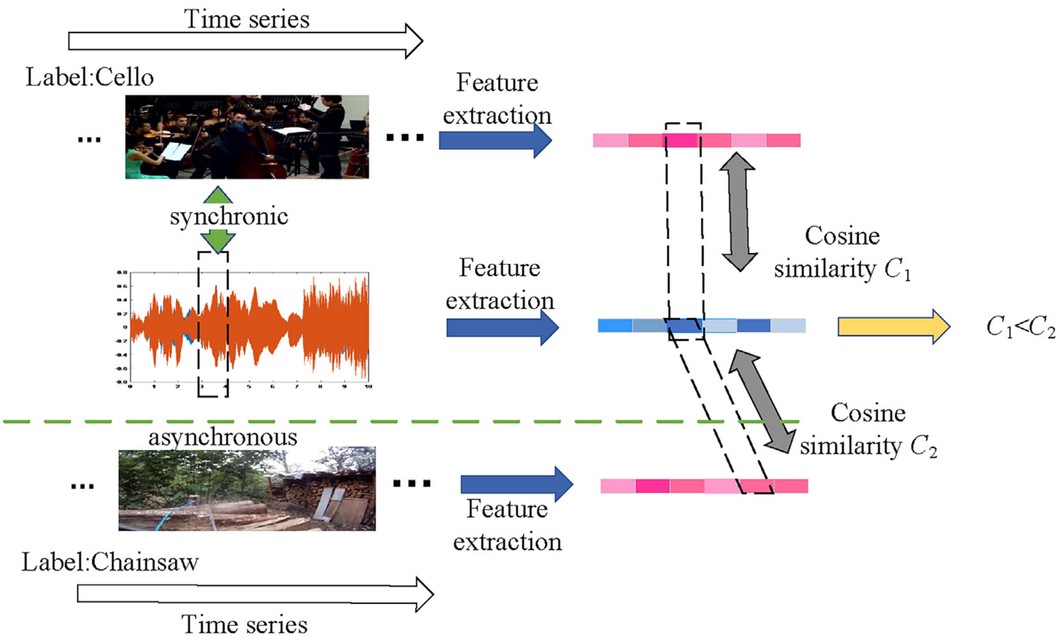


Fig. 1. We searched for cosine similarity for all audio-visual similarities in the LLP dataset. This graph shows the similarities between a typical audio clip and two frames. C_1 is the similarity between the audio in the figure and the synchronous video frame. They have the same label.C_2 is the similarity between the audio in the graph and the unrelated video frames in the data set, and they have different labels. The results show that $C_1 < C_2$ and C_2 is the maximum similarity corresponding to this audio clip.

the model. Up to now, there are a lot of choices for pre-training networks. However, the existing multi-modal methods are primarily for the design of the feature extraction networks in each mode but lack consideration for the collocation of feature networks of different modes.

We demonstrate the need for collocation by examining the differences in the performance of different pre-trained models on single-mode and multi-mode tasks. Moreover, inspired by the domain adaptation [28] we prove a hypothesis based on domain adaptive theory and experiments and propose a fast solution to the above problem based on this hypothesis. In this solution, we use the idea of ADMM.

3. Approach

In this section, we elaborate on our method. First, we summarize the whole section(Section 3.1). Secondly, we describe our proposed method. In order, it consists of two main parts: (i) a multi-modal time-series max-margin loss for better alignment(Section 3.2). (ii) a novel way to select the collocation of feature extraction networks(Section 3.3).

3.1. Overview

As shown in Fig. 2, our framework is composed of two single-modality feature extraction networks F_1, F_2, projection layers L_1, L_2, cross-modal fusion module T and the classifier C. Given the audio-visual pair (x_a, x_v), the features $F_1(x_a)$ and $F_2(x_v)$ are adjusted using L_1 and L_2 so that $T(L_1(F_1(x_v)))$ and $T(L_2(F_2(x_a)))$ are supplementary with each other via T. In the end, audio, visual, and audio-visual instances are classified using C. In addition, we use supervised BCE(Binary Cross Entropy) loss and self-supervised MTSM loss for the training. Since our goal is to deliver the label and period of each video event, We also adopt sparse labelling for the presence of video events and F1 scores to evaluate the performance. A detailed illustration of the two challenges is as follows.

3.1.1. Challeng 1: Cross-Modal Misalignment

.

Definition 1. Cross-modal semantic misalignment:

We define semantically injective (one to one) mappings in the AVVP task as f, g, if $d(f(x^i), g(y^i)) < d(f(x^i), g(y^j)), for\ i \neq j$, where f, g are audio and visual encoders, x, y are paired video and audio data, i and j represent corresponding labels.

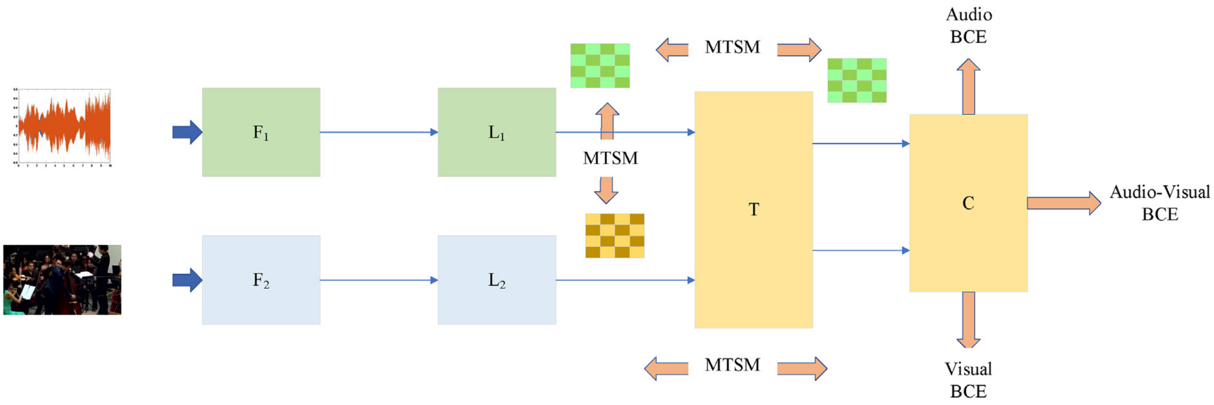


Fig. 2. The framework of our task.

Definition 2. Injective mapping functions $f < g$:

$d(f(x_i), g(y_i)) < d(f(x_i), g(y_j)), fori \neq j$, where f, g are audio and visual encoders, x, y are video and audio data, i and j represent two instances.

Definition 3. The target of AVVP:

AVVP aims to predict segment-level video instance tags by using event-level video tags. Taking the image mode as an example, the target of AVVP is to predict the label of each image frame, while an event consists of multiple image frames.

An initiative of the solution is to find a mapping of instances from individual data spaces(audio data space and visual data space) to a shared embedding space, so that related instances are mapped to nearby places in the space. However, the audio-visual instances are not injective in the space since they contain multiple concepts.

Given that temporal semantic structure in a single-modal space should be consistent with the one in the cross-modal feature space, we are able to implicitly regularizes the space by aligning the structures from different spaces. Specifically, a module called MTSM using an adaptive margin loss which requires that cross-modal feature similarities should follow some margins.

The adaptive margin is calculated by the single-modal similarities scaled by a variable $T_{ij} = \frac{1}{|i-j|^3}$, which guarantees that the cross-modal features similarities should be consistent with the single-modal at time-stamp (i and j).

Besides, feature noise is especially obvious for the weakly supervised task. Assuming that n_1, n_2 are the noise in the audio and visual features, and σ indicates the variance. The transformer module multiplies the audio-visual features and introduces multiplicative noise $\sigma(n_1 \ * n_2), \sigma(n_1 \ * n_2) > \sigma(n_1) + \sigma(n_2)$. Based on the MTSM, the pre-fusion features (which are not polluted by other modal noises) are utilized to constrain the post-fusion features, so as to suppress the cross-modal noise.

3.1.2. Challenge 2 pretraining models collocation

To our best knowledge, the optimal collocation of pre-trained single-modality models is out of the community scope. Nevertheless, experiments are designed to show that collocation plays a vital role in the AVVP. For example, PolyNet performs better than SENet154 for solely visual feature extraction, but when facing the AVVP, the collocation of PolyNet and VGGish is inferior to the collocation of SENet154 and VGGish (as shown in Tables 1 and 2). One method to tackle the challenge is evaluating different pre-trained model collocations, which is time-consuming. While inspired by the ADMM, we propose a divide-and-conquer scheme, which firstly decouples the whole optimization into two consecutive sub-problems (collocation selection and model training); and secondly, optimizes the sub-problems alternately. This scheme reduces the costs while retaining accuracy.

3.2. MTSM

As shown in 3.1, this subsection aims to solve the misalignment across the modes and the problems of various noises introduced by the multiplication operator(which is typical in Transformer).

To solve the above problems, we introduce the principle of max-margin in SVM (Support Vector Machines) In max-margin theory,as shown in Fig. 3the distances between positive sample scores y and negative sample scores y' are limited by the hyper-parameter m. Following this, the max-margin loss function is defined as Eq. 1

$$I(y, y') = \max(0, m - y + y') \tag{1}$$

Table 1
The possibility that the most similar segment-level pair that has the same labels.

	unimodal	multimodal
recall-top1	0.9714	0.2918

Table 2
The fine-tuning visual pretrained feature extraction networks test on the labeled images from LLP dataset.

	VGG19_bn	ResNet152	SENet154	PolyNet	OpenL3
accuracy	59.5	62.1	62.6	69.3	58.6

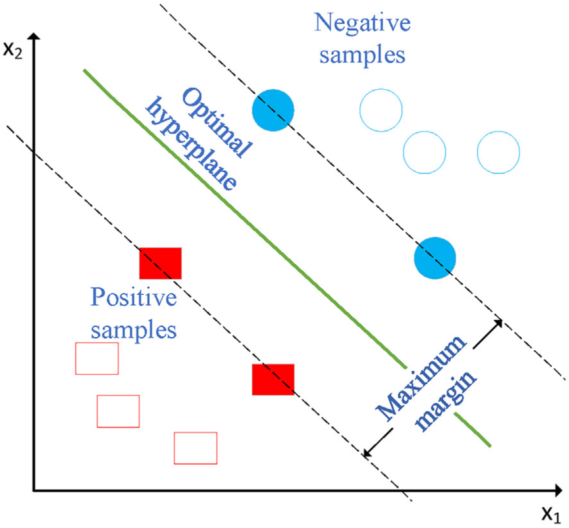


Fig. 3. An example of max-margin thinking on a two-dimensional plane. Max-margin makes the distance between positive and negative samples greater than m.

One of our core ideas is to re-align the similarity between the features of different modes before using the Transformer structure so that features with the same label (thus sharing the same semantics) have a greater similarity. This coincides with the idea of Max-margin.

For a sample pair z^1 in mode 1 and z^2 in mode 2, they form a positive pair if they have the same labels. But for our weakly-supervised task, we lack the segment-level labels, so we could not refer to each pair as negative or positive directly. Based on the phenomenon that the semantic(label) similarity of cross-modal features has a similar temporal variation trend as that of the single modal feature. We use the feature similarity map in a single mode to constrain the feature similarity across modes. Our loss function is defined as 2

$$\begin{aligned} l_1 &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\left|-\mathrm{sim}\left(z_i^1, z_j^2\right) + m_{i,j}\right| \\ m_{i,j} &= T_{i,j}\cdot \mathrm{sim}_{\mathrm{detach}}\left(z_i^1, z_j^1\right) \end{aligned} \tag{2}$$

where z_i^1 represents the ith features of $1th$ mode; $\mathrm{sim}_{\mathrm{detach}}$ means the detach of the similarity; $T_{i,j}$ is a time lag coefficient which increases as the time lag decreases.

In that way, we does not clearly indicate a sample pair as negative or positive. On the contrary,we defined m as a self-adaptive scores $T_{i,j}\cdot\mathrm{sim}_{\mathrm{detach}}(z_i^1, z_j^1)$ where $\mathrm{sim}_{\mathrm{detach}}$ represents a detach ofsim; when $m_{i,j} > \mathrm{sim}(z_i^1, z_j^2)$ we figure the pair as positive and vice versa. Experimentally, we set $T_{i,j} = \frac{1}{|i-j|^3}$.

We use prior information between features in the single-mode to align the feature similarity and label similarity corresponding to the multi-model. While narrowing the gap between feature similarity and label similarity, the information contained in the feature difference value is re-extracted.

Besides, to release the multiple noises introduced by the multiple operators in the Transformer, we use the original single-mode information to restrict the fused features. For the AVVP task, there are two modes($n = 2$). The loss function is defined as Eq. 3

$$l_2 = \sum_{n=1}^{2} \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} \left| -\mathrm{sim}\left(z_i^n, z_j^n\right) + \mathrm{sim}_{\mathrm{detach}}\left(f_i^n, f_j^n\right) \right| \tag{3}$$

where f is the corresponding fused feature after the Transformer.

Finally, our MTSM loss is defined as

$$l = l_1 + \alpha l_2 \tag{4}$$

where α is a weight coefficient and we set it as 1 experimentally.

3.3. Methods for the collocation of pre-trained feature extraction networks

3.3.1. Task objective function

Assuming that M pre-trained extraction networks are pre-trained for each of the N modalities' data, the following objective of the task in this section is to search for the optimal collocation of pre-trained networks for feature fusion:

$$\underset{j_1,\cdots j_N}{argmax}(\max_{\mathrm{E}}(\psi_1^{j_1}(x_1), \cdots, \psi_{\mathrm{i}}^{j_{\mathrm{i}}}(x_1), \cdots, \psi_{\mathrm{N}}^{j_{\mathrm{N}}}(x_1), \theta)) \tag{5}$$

Here ψ_i represents the ith modality and ψ^{j_i} represents the $j_i th$ feature network where, $j_1, \ldots, j_N \in [1, M]$; E is the evaluation metric of the task performances, such as the F1 scores in our task. θ denotes the model parameters.

3.3.2. The traversal method

We propose a solution by slightly changing the optimization objective as in Eq. 6, where we first calculate the performance corresponding to all collocations and then compare them to select the largest one. Therefore, the found collocation through the previous searching operation can be regarded as the alternative optimal collocation. The flow chart is shown in Fig. 4(a)

$$\underset{\phi}{\operatorname{argmax}}\, \mathrm{E}(\phi_i, \theta_{m_i}); \theta_{m_i} = \underset{\theta}{\operatorname{argmax}}\, \mathrm{E}(\phi_i, \theta) \tag{6}$$

where $\phi_i \in \{\psi_1^{j_1}(x_1), \psi_2^{j_2}(x_1), \cdots \psi_{\mathrm{i}}^{j_{\mathrm{i}}}(x_1), \cdots, \psi_{\mathrm{N}}^{j_{\mathrm{N}}}(x_1)\}$; it is one of M^N collocations of the pre-trained feature extraction network.

Assuming that it costs T_0 to train one of the pre-trained networks and T_1 in the inference, it costs $M^N \ast (T_0 + T_1)$ to traversal all cases. As the traversal method has to evaluate the task performance of all the collocations, the previously mentioned method, as shown in Eq. 6 is time-consuming.

3.3.3. The domain-adaptation-based Solution

Inspired by the idea of ADMM [29], the key to decoupling the subtask "find the optimal collocation" from the task "calculate the task performance of all collocations" is to find more effective metrics to compare collocations.

Therefore, we propose a conjecture: under the condition that the task model and the hyper-parameters are fixed, the collocation corresponding to better performance during training and testing on the test set is more likely to be our ideal collocation. We empirically prove it using experimental results in Section 4.2.1.

In this subsection, we explain and justify the conjecture implications using LDA(Linear Discriminant Analysis) as follows. Specifically, we applied the singular value decomposition (SVD) to compute respectively all singular values and eigenvectors of the training set feature matrix as F_s=$[f_1^s \ldots f_b^s]$and testing set feature matrix $F_t = [f_1^t \ldots f_b^t]$:

$$\begin{aligned} F_s &= U_s \Sigma_s V_s^\top \\ F_t &= U_t \Sigma_t V_t^\top \end{aligned} \tag{7}$$

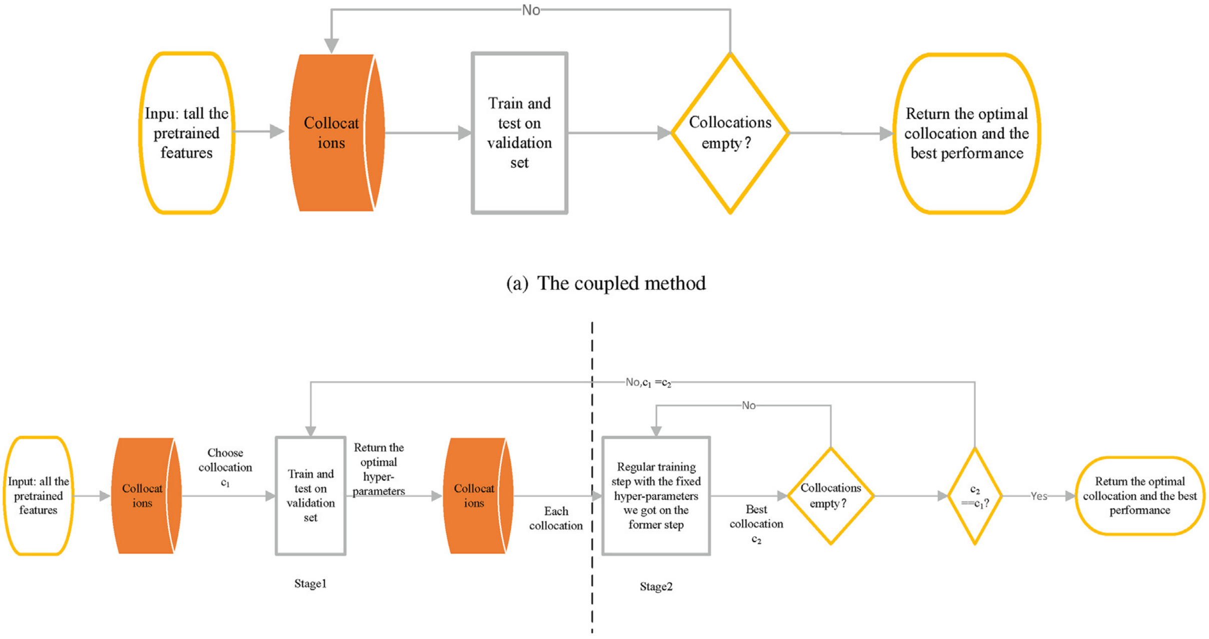
where b is the batch size.

Then we can calculate the corresponding angle [28] between two eigenvectors at the same singular value index, and its cosine value is as follows:

$$\cos(\psi_i) = \frac{\langle u_{s,i}, u_{t,i} \rangle}{\|u_{s,i}\| \|u_{s,i}\|} \tag{8}$$

where $u_{s,i}$ is the ith eigenvector in U_s with the i th largest singular value in the source feature matrix.

Existing study [28] showed that $\psi_1$represents the transferability of the feature representation. Based on the previous study [28], different features extracted from the same data set, $\cos(\psi_1)$ should be the similar but still different due to different collocation.


No
Inpu: tall the pretrained features
Collocat ions
Train and test on validation set
Collocations empty ?
Return the optimal collocation and the best performance
No,c1 =c2
No
Input: all the pretrained features
Collocat ions
Choose collocation c1
Train and test on validation set
Return the optimal hyper-parameters
Collocat ions
Each collocation
Regular training step with the fixed hyper-parameters we got on the former step
Best collocation c2
Collocations empty ?
c2 ==c1?
Yes
Return the optimal collocation and the best performance
Stage1
Stage2

(a) The coupled method

(b) The decoupled method

Fig. 4. Two methods we proposed to figure out the best collocation.

We have found a metric that can evaluate the performance of collocation and decouple it from the normal training process. Drawing on the idea of ADMM, we divide the optimization process of the task objective into two parts: finding the optimal collocation and calculating the performance corresponding to the collocation.

Fig. 4(b) shows our method vividly.

The above demonstration and subsequent experiments are conducted, conditioning that the test set is available during the training process. Therefore, we used the validation set as the alternative to the test set. We subsequently also verify the robustness of the above experimental hypothesis when the validation data distribution in the experiment varies against the final test data. (Fig. 6).

Therefore, we choose the validation set instead, and the entire progress is shown as Algorithm 1

Algorithm 1 Our domain-adaptation-based algorithm

```
1: Choose one collocation to start
2: Optimize the super-parameters corresponding to this collocation
3: Fix the super-parameters in 2 and train models with all the collocations under the fixed super-parameters on the
   validation set.
4: If there is another collocation which has a better performance than the initial one, change our selection to this one
   and return to 2
5: If our selection is the best one, return this collocation
```

4. Experiments

In Section 4.1, we introduce the experiments setups; In Section 4.2, we verify our approaches and explore the feasibility; In Section 4.3We perform the ablation experiments on our method; Our method is superior to the comparisons.4.4.

4.1. Implementation details

We consider the same experimental settings as Tian [10], with details as the following.

Dataset We evaluate our method on the LLP dataset, which is designed for AVVP. LLP contains 11,849 YouTube video clips spanning over 25 categories for a total of 32.9 h collected from AudioSet [4], therein including a wide range of video events (e.g., human speaking, singing, baby crying, dog barking, violin playing, car running, and vacuum cleaning, etc.) from diverse domains (e.g., human activities, animal activities, music performances, vehicle sounds, and domestic environments). Input video period is set as 10 s, with each video split into non-overlapping snippets of eight frames per second.

Pretrained feature extraction networks We use two famous audio feature extraction networks (VGGish [4] and OpenL3-audio [8]) and five popular RGB image feature extraction networks (Resnet152 [30], VGG19_bn [31], PolyNet [32], SENet154 [33] and OpenL3-image [8]).

Training setting Batch size and epochs are 16 and 40. The initial learning rate is 3E-4 and drops by 1/10 every ten epochs, and the models are optimized by Adam (Adaptive momentum). It is running on one NVIDIA 2080Ti.

Baselines We compare our method with the existing AVVP method [10] proposed by Tian. Since our contribution does not involve network design, our experiment is carried out on the network of baseline [10].

Evaluation Metrics As Tian [10] did, we evaluate them by parsing three modalities (individual audio, visual, and audio-visual events) from segment-level and event-level. We use F-scores [34] as metrics for both levels. The segment-level F-score evaluates snippet-wise event labelling performance. For event-level F-score, we extract events by concatenating consecutive positive snippets in the same categories (mIoU = 0.5 as the threshold). Besides, we also aggregate the F-scores of each level to evaluate the general parsing performance. Type@AV averages audio, visual, and audio-visual event F-scores and Event@AV computes a weighted F-score considering all audio and visual events of each sample rather than directly averaging the results as the Type@AV.

4.2. Experimental exploration

In this section, the misalignment between feature similarity and label similarity in different modes is verified, and we introduce MTSM loss to solve it. (4.2.1). Then, we verify the importance of the pre-trained model collocation and verify the reasonability of our method proposed in 3.3.3(4.2.2).

4.2.1. The misalignment between feature similarity and label similarity in different modes

To verify the existence of misalignment, we calculate the similarities within and between the extracted audio features and visual features on the LLP dataset(as shown in Table 1). Specifically, within one mode, if the two features have the maximum similarity, the probability is 97.14% that they belong to the same label. While in the multi-modes, it drops to 29.18 %.

Therefore, the misalignment between features similarity and label similarity is more severe in the multi-mode than in the single-mode. In this case, the direct use of the Transformer is not appropriate.

4.2.2. The collocation of pre-trained models

First, we verify the significance of the collocation of pre-trained models, and the details are as follows: We create an image dataset by collecting all the labelled frames from the LLP dataset and compare the classification accuracy of some fine-tuned image feature extraction networks. The results are shown in Table 2. It reflects the ability of the network individually to extract the modal data information. Moreover, we verify that the extracted features are visibly divisible(Table 4), hereby showing the strong extraction ability of the networks. Also, we can infer that the extracted features vary obviously according to the feature networks used.

In general, the collocation of pre-trained networks influences feature fusion(as shown in Table 3) from two aspects: one is the ability of an individual network to extract the single-modal information(Table 2), the other is the fusion ability of multi-modal features, namely mutual understanding ability among pre-trained models. By comparing Table 2 and Table 3, we found that some networks with high accuracy in the single-modal task were not always superior when dealing with the multi-modal task. Take PolyNet as an example. In Table 2, we compared the accuracy of visual pre-trained models in the visual mode, and PolyNet achieved the highest accuracy. While in Table 3, with a fixed audio pre-trained network, PolyNet

Table 3
The final performance in the AVVP task with different pre-trained feature extraction networks.Audio feature neworks are all the same: VGGish.

	VGG19_bn	ResNet152	SENet154	PolyNet	OpenL3
Type@AV	54.1	53.7	53.5	52.8	52.4

Table 4
The accuracy of distinguishing extracted features of the same original data among different feature extraction networks. We used pre-trained ResNet101 as the discriminator.

	ResNet	VGG19_bn	SENet
PolyNet	99.9	99.9	99.8
SENet	99.9	99.9	
VGG19_bn	99.8		

fell to the second-lowest. It also shows that two networks with excellent performance in single-mode may not keep winning when collocated, that is, $1 + 1\neg > 0.9 + 0.9$. Thus, we argue that feature fusion is not only related to the ability of the single-modal pre-trained models but also related to the mutual understanding of the multi-modal pre-trained models. Therefore, the collocation of pre-trained models, which takes both aspects into account, is of significance.

Secondly, we verify the reliability of our method: We compare the results of our proposed method with the traditional network collocation to verify the reliability of our method.

As shown in Fig. 5, VGGish and OpenL3 are selected as the audio feature extraction networks to combine with different image feature extraction networks. (audio feature extraction networks have a very narrow range of candidates than the image, and we choose the most widely used two.).

The experimental results verify that our method has consistent results with the traditional traversal method. Yet our methods can reduce the time complexity(we make the comparisons in a smaller dataset than the traditional method and we decouple "choosing an optimal collocation of pre-trained feature extraction networks" and "evaluating the optimal collocation").

Besides, the hypothesis in 3.3.3 is verified as follows: In Fig. 5, we have proved the rationality of the hypothesis in 3.3.3by comparing the performance of different collocations. Now we further verify the robustness.

In Fig. 7, we explore the robustness of the above hypothesis. Specifically, we orderly extracted 25, 20, 15, and 10 types of data from the LLP data set as new training sets and retained the original validation and test set to explore the trends among the performances of 6 collocations on the test set. The result is shown in Fig. 6, and there is what we find:

- With the reduction of the types in the training set, the similarity between the training set and the test set decreases. The performance trend among the six collocations is similar. This phenomenon verifies the robustness of our hypothesis.
- With the reduction of the types of retained data in the training set, the gap between the performance corresponding to different collocations gradually decreases. It suggests that the more similar the training set is to the test set, the easier it is to find the optimal collocation.
- As the difference between the training set and test set becomes larger, it calls for a higher extraction ability of the pre-trained models.

Finally, we tested the total time consumption of the two methods, including the collocations of five visual feature extractors and two audio extractors, and the results are shown in the Table 5. And as the number of collocations we consider increases, that percentage continues to grow. This reflects the effectiveness of our proposal regarding the decoupled method.

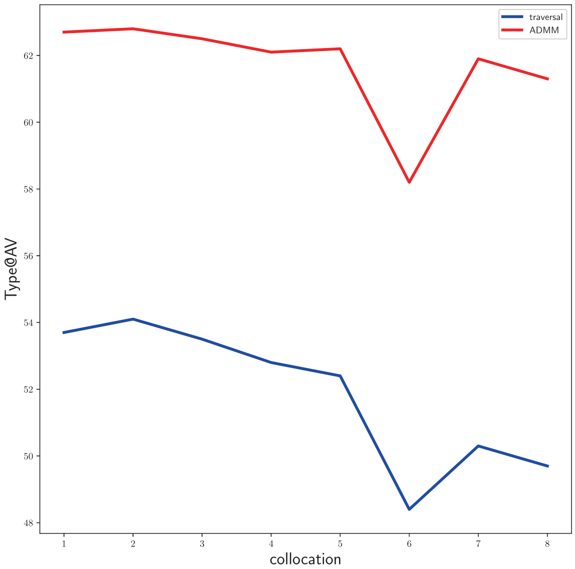


Fig. 5. The comparison between the traversal method and domain-adaptation-based method in evaluating the collocation of various feature extraction networks. The vertical axis is the evaluation index, and the horizontal axis is the collocations of each mode's pre-trained models.The horizontal axis shows the following eight different combinations with the training network (from right to left): 1.VGGish(A) and ResNet152(V); 2.VGGish(A) and VGG19_bn(V); 3. VGGish(A) and SENet154(V); 4.VGGish and PolyNet(A); 5.VGGish(A) and OpenL3(V); 6.OpenL3(A) and OpenL3(V); 7.OpenL3(A) and VGG19_bn(V); 8. OpenL3(A) and ResNet152(V). The blue line represents the traversal method, and the orange line represents the domain-adaptation-based method. **The trends of the two curves are similar.**

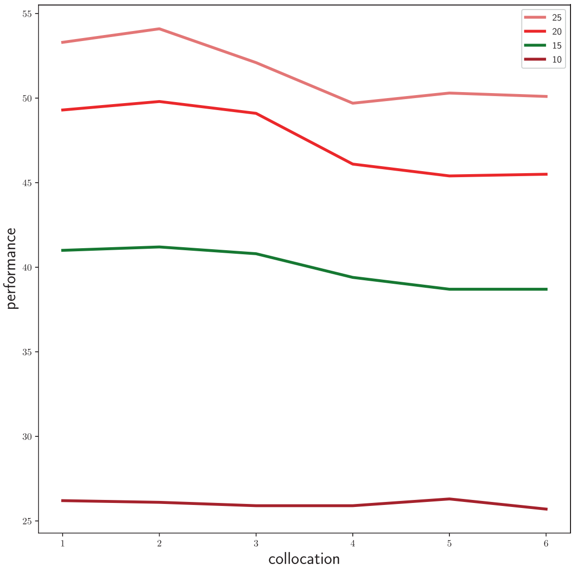


Fig. 6. Inquiry: if the distribution difference between the training set and the test set changes, whether the assumed robustness will change. The horizontal axis is the collocations of 6 different pre-trained models. The vertical axis is the performance of the test set. Different coloured lines correspond to different parts of the training set.

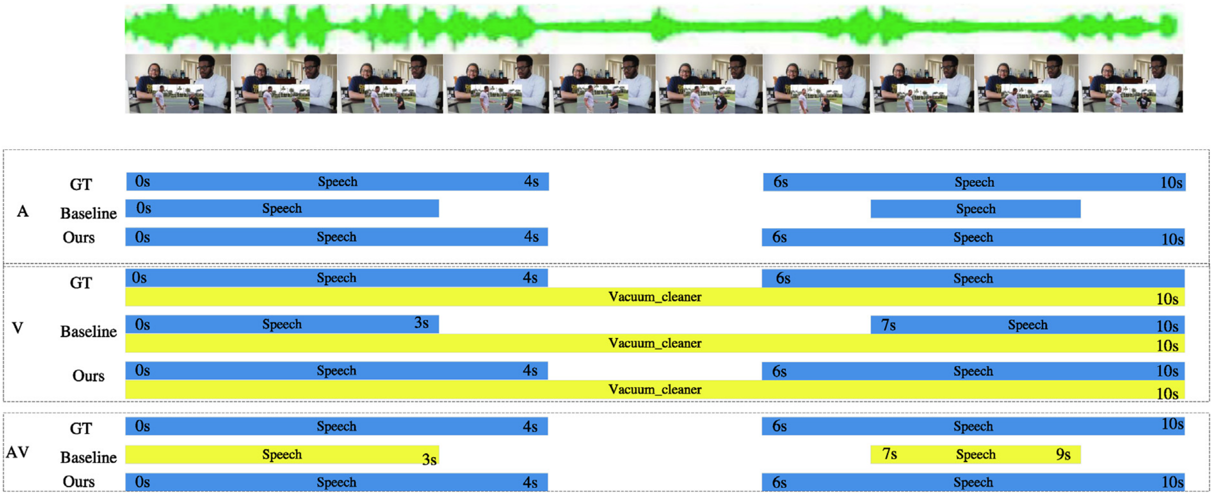


Fig. 7. A comparison of our approach and baseline on a typical example.

Table 5
The total time consumption of the two methods, including the coupled method and the decoupled method.

the coupled method	the decoupled method	the time ratio of the two methods
33887.77	11706.04	2.895

4.3. Ablation studies

We conduct ablation studies to show the effectiveness of individual parts of our method. Table 6 shows our model" Baseline + M" (the models with the MTSM loss) outperforms the baseline by 2.1% on the audio-visual event parsing. Specifically, for the visual event parsing, the model with MTSM loss improves the performance by 1.7% (from 48.7% to 50.8%) at the event level and 3.1% (from 51.5 % to 54.4 %) at the segment level. It validates that the idea of the max-margin improves the model for AVVP.

Table 6
Ablation studies of the proposed modules. Audio-visual video parsing accuracy (%) is reported on the LLP test dataset. “M” denotes the proposed MSTM loss for temporal localization. “A” is our domain-adaptation-based strategy.

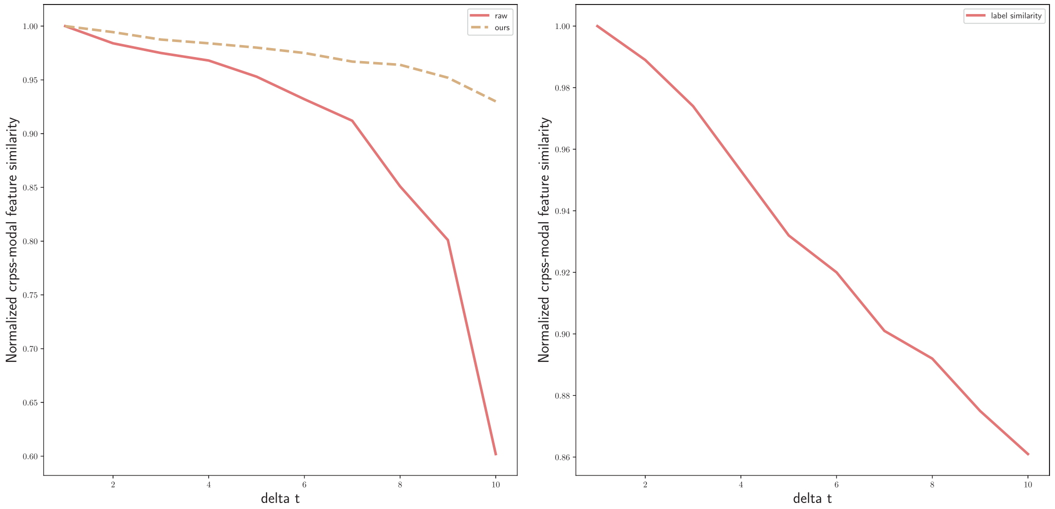
Event type	Methods	Segment-level	Event-level
Audio-visual	Baseline	50.2	42.7
	Baseline + M	50.5	44.4
	Baseline + A	49.0	42.9
	Baseline + M+A	**50.7**	**44.6**
Audio	Baseline	60.4	51.1
	Baseline + M	61.6	52.4
	Baseline + A	61.0	**52.7**
	Baseline + M+A	**62.1**	52.7
Visual	Baseline	51.5	48.7
	Baseline + M	54.1	50.4
	Baseline + A	52.3	48.5
	Baseline + M+A	**54.4**	**50.9**
Type@AV	Baseline	53.3	47.5
	Baseline + M	55.4	49.1
	Baseline + A	54.1	48.0
	Baseline + M+A	**55.7**	**49.4**
Event@AV	Baseline	55.0	47.8
	Baseline + M	56.4	48.8
	Baseline + A	55.6	49.0
	Baseline + M+A	**56.7**	**49.3**

Compared to the baseline, our model with the features extracted by the best collocation of pre-trained models (“Baseline + A”) shows outperformance on all of the metrics. By comparing the model “Baseline + M + A” and the model “Baseline + M”, we can find a better collocation of pre-trained models also prompt the performance by about 0.5 % on most evaluation metrics. Therefore, the collocation can naturally be combined with existing methods.

Additionally, we take an insight into the way the MTSM loss boosts the Baseline.

Fig. 7 is an example to show the merit of MTSM loss.

Furthermore, we demonstrate the tendency of multi-modal label similarity (Fig. 8(b)) and feature similarity(Fig. 8(a)) along time. Delta T is the time difference. Fig. 8(b) shows that as the time difference becomes larger, the probability for the segments to have the same label becomes smaller. This fast-decline tendency should apply when feature similarity is discussed. In Fig. 8(a), our method shows a more obvious decline than the baseline, thereby proving our method extends the variation of feature similarity as time goes by.



(a) Timeline variations in the similarity of features which is the input of the Transformer

(b) Variation law of similarity among features’ label on Time Axis (25 Categories in averaged)

Fig. 8. The changes in the similarity of the feature and the similarity of the label corresponding to the feature among different modes with the change of time. We take the similarity as one if both features’ labels are the same and vice versa.

Table 7
Ablation studies of two parts in the MSTM losses we propose. “M-1” denotes the cross-modal Max-margin loss for the alignments between different modes; “M-2” denotes the single-modal max-margin loss to restrict the multiple noises introduced by the HAN.

Event type	Methods	Segment-level	Event-level
Audio-visual	Baseline	50.2	42.7
	Baseline + M-1	49.7	43.8
	Baseline + M-2	49.9	43.1
	Baseline + M	**50.5**	**44.4**
Audio	Baseline	60.4	51.1
	Baseline + M-1	**61.7**	**52.8**
	Baseline + M-2	56.1	48.5
	Baseline + M	61.6	52.4
Visual	Baseline	51.5	48.7
	Baseline + M-1	**53.1**	**49.4**
	Baseline + M-2	53.1	47.7
	Baseline + M	**54.1**	**50.4**
Type@AV	Baseline	53.3	47.5
	Baseline + M-1	55.1	48.9
	Baseline + M-2	53.0	46.4
	Baseline + M	**55.4**	**49.1**
Event@AV	Baseline	55.0	47.8
	Baseline + M-1	56.2	**48.9**
	Baseline + M-2	51.5	46.4
	Baseline + M	**56.4**	48.8

Table 8
Audio-visual video parsing accuracy (%) of different methods on the LLP test dataset.Comparisons with the state-of-the-art methods of the audio-visual video parsing task on the LLP test dataset.

Event type	Methods	Segment-level	Event-level
Audio-visual	AVE [2]	50.0	41.7
	AVSDN [6]	47.2	40.4
	HAN(baseline) [10]	50.2	42.7
	MA-C [27]	49.7	43.8
	Ours	**50.7**	**44.6**
Audio	TALNet [35]	50.0	41.7
	AVE [2]	47.2	40.4
	AVSDN [6]	47.8	34.1
	HAN(baseline) [10]	60.4	51.1
	MA-C [27]	61.9	**52.8**
	Ours	**62.1**	52.7
Visual	STPN [36]	46.5	41.5
	CMCS [37]	48.1	45.1
	AVE [2]	37.1	34.7
	AVSDN [6]	52.0	46.3
	HAN(baseline)	51.5	48.7
	MA-C [27]	53.1	49.4
	Ours	**54.4**	**50.9**
Type@AV	AVE [2]	39.9	35.5
	AVSDN [6]	45.7	35.6
	HAN(baseline) [10]	53.3	47.5
	MA-C [27]	54.9	48.7
	Ours	**55.7**	**49.4**
Event@AV	AVE [2]	41.6	36.5
	AVSDN [6]	45.7	35.6
	HAN(Baseline) [10]	55.0	47.8
	MA-C [27]	56.2	49.0
	Ours	**56.7**	**49.3**

In addition, we also explore the influence of the values of the two hyper-parameters, loss proportionality factor β and the max-margin factor $T_{i,j}$, mentioned in the method on the effect of the method. The specific experimental results are shown in Tables 9 and 10. Besides, we separately analyzed the trend of the ”Segment-level Type@Avg”, which we mainly considered in

Table 9
How the metrics vary with the values of $T_{i,j}$.

	1	2	3	4	5
Audio Segment-level	61.1	62.0	**62.1**	60.3	61.4
Visual Segment-level	52.8	53.2	**54.4**	52.7	53.4
Audio-Visual Segment-level	49.6	50.2	**50.9**	49.6	49.9
Segment-level Type@Avg	54.5	55.1	**55.7**	54.2	54.9
Segment-level Event@Avg	55.7	56.3	**56.4**	55.1	55.9
Audio Event-level	51.8	**52.8**	52.7	51.0	52.5
Visual Event-level	47.1	48.0	**50.4**	48.6	49.1
Audio-Visual Event-level	42.2	43.7	**44.6**	43.1	43.3
Event-level Type@Avg	47.0	48.2	**49.4**	47.5	48.3
Event-level Event@Avg	46.8	48.3	**49.3**	47.8	48.5

Table 10
How the metrics vary with the exponential value of β

	0.1	0.3	0.5	0.8	1	1.2	1.4	2
Audio Segment-level	61.3	60.4	61.4	60.6	**62.1**	61.2	61.5	61.3
Visual Segment-level	54.3	**54.9**	53.6	54.7	54.4	53.3	53.3	53.6
A-V Segment-level	51.0	**51.7**	50.0	50.9	50.9	50.0	49.7	50.1
Segment-level Type@Avg	55.5	55.6	55.0	55.4	**55.7**	54.8	54.8	55.0
Segment-level Event@Avg	56.4	55.6	56.1	**55.6**	56.4	55.7	56.1	56.1
Audio Event-level	52.2	51.4	52.7	51.5	**52.7**	52.2	52.5	51.9
Visual Event-level	50.3	**51.0**	49.0	50.1	50.4	49.1	48.9	49.5
A-V Event-level	44.8	45.6	43.7	44.0	**44.6**	43.5	43.4	44.5
Event-level Type@Avg	49.1	49.4	48.5	48.5	**49.4**	48.3	48.3	48.6
Event-level Event@Avg	48.7	48.3	48.6	47.8	**49.3**	48.2	48.4	48.6

training with the change of parameters 9. As shown in the Fig. 9 and Tables 9 and 10, each indicator varies with the parameters, and most of them show oscillatory changes. Compared with the value of $T_{i,j}$, each metric has little change with the proportional coefficient β. The optimal exponential value of $T_{i,j}$ is affected by many factors, such as the video's frame extraction rate and the audio's sampling frequency. Therefore, personalized settings are required for different data sources.

Separately, we explore the effects of the two parts of MSTM loss. Results are shown in Table 7. "M-1" denotes the cross-modal Max-margin loss for the alignments between different modes; "M-2" denotes the single-modal max-margin loss to restrict the multiple noises introduced by the HAN. Thus, "M-1" contributes to our task, while we notice that "M-2" alone impair the performance of the model. The reason for the incompetence of "M-2" is the lack of pre-constraints on multi-modal features. That is, the inputs lack the fusion information from each mode.

4.4. Experimental comparison

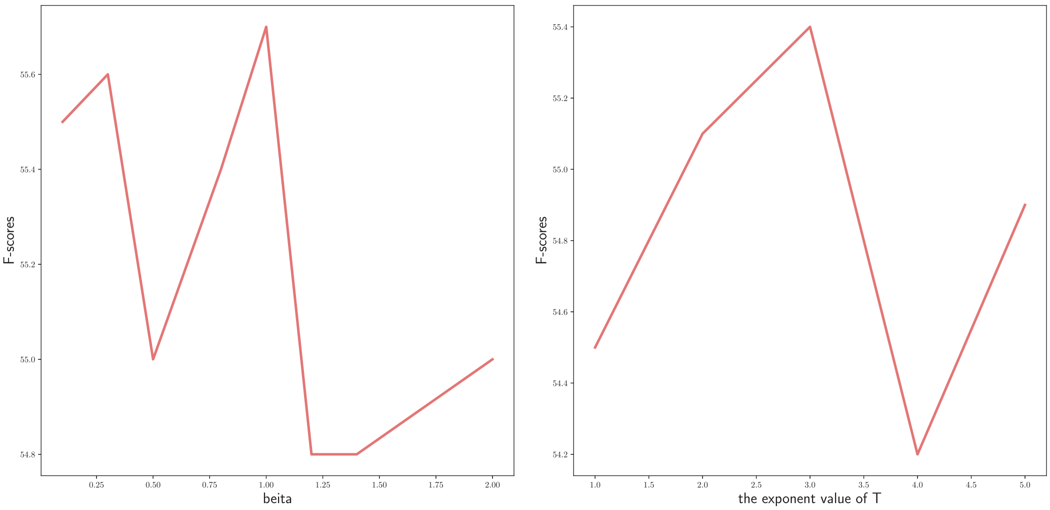
We compare our model with temporal action localization methods STPN [36] and CMCS [37], weakly-supervised sound detection method TALNet [35], and state-of-the-art audio-visual event parsing methods including AVE [2], AVSDN [6], HAN [10], and MA-C [27]. All the models are trained by the LLP training dataset for fair comparisons.

The results are shown in Table 8. We use the same metrics as the baseline, with more details in Section 4.1. Specifically, our MTSM outperforms the baseline method by 2.3 points(from 53.3 to 55.6) in the segment level and 1.9 points(from 47.5 to 49.4). Compared with the state-of-the-art method MA-C, We also have an advantage of around 0.5 points.

From my perspective, the value of our work is beyond the mentioned promotion. It lies in that we explore the specific manifestation of two problems common in multi-modal tasks.

4.5. Cross validation

Traditional cross-validation is hard to be realized for our weakly-supervised task. Because the training set only includes the Event-level labels, our training set and the validation set are not swappable.



(a) The curve of the metric ”Segment-level Type@Avg” as the value of β changes

(b) The curve of the metric ”Segment-level Type@Avg” as the exponent value of $T_{i,j}$ changes

Fig. 9. The influence of the values of the two hyper-parameters including loss proportionality factor β and the max-margin factor $T_{i,j}$.

Table 11
We divide the training set into five parts. Each time choose four parts to train the model. We don’t change the setting of validation and testing.

	1st	2nd	3rd	4th	5th
Audio Segment-level	60.4	59.3	60.3	59.8	60.5
Visual Segment-level	53.9	53.8	54.7	54.0	53.4
A-V Segment-level	50.4	50.7	51.0	49.7	49.7
Segment-level Type@Avg	54.9	54.6	55.3	54.5	54.5
Segment-level Event@Avg	55.8	54.9	55.7	55.3	55.4
Audio Event-level	50.6	50.5	51.7	50.3	51.5
Visual Event-level	49.7	49.4	50.6	49.3	49.4
A-V Event-level	43.5	44.4	44.4	42.9	44.0
Event-level Type@Avg	47.9	48.1	48.9	47.5	48.3
Event-level Event@Avg	47.6	46.9	48.1	47.0	47.8

Therefore, we divide the training set into five equal parts by referring to the method of fivefold cross-validation. We choose four parts for training, and we train the model five times. The generalization is observed as the training sets change. The results are shown in Table 11. It shows that most of the metrics stay almost unchanged with the change of training set.

5. Conclusion and future work

This work explores the factors influencing the AVVP task from the two aspects: innate(pre-training model) and acquired (training process). On this basis, we propose three challenges that the AVVP task faces:1. Cross-mode semantic misalignment. 2. Contextual and Audio-Visual Dataset Bias.

For the first challenge, we refer to the idea of max-margin learning commonly used in machine learning and use the inherent timing information of data to propose an adaptive multi-modal time-series max-margin loss. Given the second challenge, we proved the importance of the collocation of the pre-trained extraction networks to the multi-modal problem and gave a simple solution to this problem. Referring to the idea of ADMM, we decoupled the solution of the optimal collocation from the general training process and reduced the training time.

The background of the weakly supervised task in AVVP gives a higher requirement for feature fusion. However, challenges 1,2 we proposed are common in the multi-modal tasks, and thus our methods have a good reference value for other multimodal problems.

CRediT authorship contribution statement

Jianning Wu: Conceptualization, Methodology, Software, Writing-original-draft, Data-curation. **Zhuqing Jiang:** Writing-review-editing, Methodology, Visualization, Conceptualization, Supervision. **Qingchao Chen:** Writing-review-editing,

Methodology, Visualization, Conceptualization. **Shiping Wen:** Writing-review-editing. **Aidong Men:** Software, Validation. **Haiying Wang:** Writing-review-editing.

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.

Acknowledgements

Funding: This work was supported by the Beijing Advanced Discipline Construction Project [Grant No. BMU2019GJJXK001]; and the National Natural Science Foundation of China [Grant No. 62002026].

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, Louis-Philippe Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[2] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 247–263, 2018.
[3] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in the wild. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops, 2019.
[4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, R. Wade Lawrence, Channing Moore, Manoj Plakal, Marvin Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 776–780.
[5] Wu Yu, Linchao Zhu, Yan Yan, Yi Yang, Dual attention matching for audio-visual event localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6292–6300.
[6] Yan-Bo Lin, Yu-Jhe Li, Yu-Chiang Frank Wang, Dual-modality seq2seq network for audio-visual event localization, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2002–2006.
[7] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David Crandall, and Luc Van Gool. A survey on deep learning technique for video segmentation. arXiv preprint arXiv:2107.01153, 2021.
[8] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3852–3856.
[9] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675, 2016.
[10] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: weakly-supervised audio-visual video parsing. arXiv preprint arXiv:2007.10558, 2020.
[11] Huaping Liu, Amir Hussain, Shuliang Wang, Multi-modal fusion, Inf. Sci. 432 (2018) 462.
[12] Xinjian Gao, Mu. Tingting, John Y. Goulermas, Meng Wang, Attention driven multi-modal similarity learning, Inf. Sci. 432 (2018) 530–542.
[13] Zechao Li, Jinhui Tang, Tao Mei, Deep collaborative embedding for social image understanding, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2018) 2070–2083.
[14] Zechao Li, Jinhui Tang, Liyan Zhang, Jian Yang, Weakly-supervised semantic guided hashing for social image retrieval, Int. J. Comput. Vision 128 (8) (2020) 2265–2278.
[15] Zechao Li, Yanpeng Sun, Liyan Zhang, Jinhui Tang, Ctnet: Context-based tandem network for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[16] Chaoqun Hong, Yu. Jun, Jian Zhang, Xiongnan Jin, Kyong-Ho Lee, Multimodal face-pose estimation with multitask manifold deep learning, IEEE Trans. Ind. Inform. 15 (7) (2018) 3952–3961.
[17] Yu. Jun, Dacheng Tao, Meng Wang, Yong Rui, Learning to rank using user clicks and visual features for image retrieval, IEEE Trans. Cybern. 45 (4) (2014) 767–779.
[18] Yu Jun, Min Tan, Hongyuan Zhang, Dacheng Tao, Yong Rui, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
[19] Chaoqun Hong, Yu. Jun, Jian Wan, Dacheng Tao, Meng Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670.
[20] Chaoqun Hong, Yu. Jun, Dacheng Tao, Meng Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Industr. Electron. 62 (6) (2014) 3742–3751.
[21] Zhong Ji, Yaru Ma, Yanwei Pang, Xuelong Li, Query-aware sparse coding for web multi-video summarization, Inf. Sci. 478 (2019) 152–166.
[22] Chunlei Chai, Lu. Guoliang, Ruyun Wang, Chen Lyu, Lei Lyu, Peng Zhang, Hong Liu, Graph-based structural difference analysis for video summarization, Inf. Sci. 577 (2021) 483–509.
[23] Carolina L. Bez, João B.O. Souza Filho, Luiz G.L.B.M. de Vasconcelos, Thiago Frensch, Eduardo A.B. da Silva, Sergio L. Netto, Multimodal soccer highlight identification using a sparse subset of frames integrating long-term sliding windows, Inf. Sci. 578 (2021) 702–724.
[24] Qibing Qin, Lei Huang, Zhiqiang Wei, Jie Nie, Kezhen Xie, Jinkui Hou, Unsupervised deep quadruplet hashing with isometric quantization for image retrieval, Inf. Sci. 567 (2021) 116–130.
[25] K. Nalini Sujantha Bel, I. Shatheesh Sam, Black hole entropic fuzzy clustering-based image indexing and tversky index-feature matching for image retrieval in cloud computing environment, Inf. Sci. 560 (2021) 1–19.
[26] Xu. Wang, Hu. Peng, Liangli Zhen, Dezhong Peng, Drsl: Deep relational similarity learning for cross-modal retrieval, Inf. Sci. 546 (2021) 298–311.
[27] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1326–1335, 2021.
[28] Xinyang Chen, Sinan Wang, Mingsheng Long, Jianmin Wang, Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation, in: International conference on machine learning, PMLR, 2019, pp. 1081–1090.
[29] Roland Glowinski, Americo Marroco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique 9 (R2) (1975) 41–76.
[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[32] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, Da.hua. Lin, Polynet: A pursuit of structural diversity in very deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 718–726.

[33] Hu. Jie, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[34] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, Metrics for polyphonic sound event detection, Appl. Sci. 6 (6) (2016) 162.

[35] Yun Wang, Juncheng Li, Florian Metze, A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 31–35.

[36] Phuc Nguyen, Ting Liu, Gautam Prasad, Bohyung Han, Weakly supervised action localization by sparse temporal pooling network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6752–6761.

[37] Daochang Liu, Tingting Jiang, Yizhou Wang, Completeness modeling and context separation for weakly supervised temporal action localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1298–1307.