Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding

Xiang Fang1* Daizong Liu2* Wanlong Fang3 Pan Zhou1† Yu Cheng4 Keke Tang5 Kai Zou6
1Huabei Key Laboratory of Distributed System Security, Huabei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology 2Peking University 3Henan University 4The Chinese University of Hong Kong 5Guangzhou University 6Protogolabs Inc.
xfang9508@gmail.com dzliu@stu.pku.edu.cn wanlongfang@gmail.com
panzhou@hust.edu.cn chengyu@cse.cuhk.edu.hk tangbohutbh@gmail.com kz@protogolabs.com

Abstract

This paper addresses the task of temporal sentence grounding (TSG). Although many respectable works have made decent achievements in this important topic, they severely rely on massive expensive video-query paired annotations, which require a tremendous amount of human effort to collect in real-world applications. To this end, in this paper, we target a more practical but challenging TSG setting: unsupervised temporal sentence grounding, where both paired video-query and segment boundary annotations are unavailable during the network training. Considering that some other cross-modal tasks provide many easily available yet cheap labels, we tend to collect and transfer their simple cross-modal alignment knowledge into our complex scenarios: 1) We first explore the entity-aware object-guided appearance knowledge from the paired Image-Noun task, and adapt them into each independent video frame; 2) Then, we extract the event-aware action representation from the paired Video-Verb task, and further refine the action representation into more practical but complicated real-world cases by a newly proposed copy-paste approach; 3) By modulating and transferring both appearance and action knowledge into our challenging unsupervised task, our model can directly utilize this general knowledge to correlate videos and queries, and accurately retrieve the relevant segment without training. Extensive experiments on two challenging datasets (ActivityNet Captions and Charades-STA) show our effectiveness, outperforming existing unsupervised methods and even competitively beating supervised works.

1 Introduction

As a popular yet challenging natural language processing task, temporal sentence grounding (TSG) ...

(Fang et al., 2022, 2023a; Liu et al., 2023c) has drawn increasing attention in recent years. TSG aims to locate a temporal video segment with an activity that semantically corresponds to a given sentence query. As shown in Figure 1(a), only a short video segment semantically matches the query, while most of the video contents are query-irrelevant. Clearly, TSG tries to break through the barrier between computer vision and natural language processing techniques for more challenging cross-modal grounding (Li et al., 2023b,a, 2022; Wang and Shi, 2023; Wang et al., 2021a, 2020c,a,b).

Most previous TSG works (Liu et al., 2023a; Fang et al., 2023b; Zhang et al., 2021; Fang et al., 2020, 2021b,a; Fang and Hu, 2020; Liu et al., 2021b, 2020b, 2022a, 2021a, 2022d, 2021c) are under a fully-supervised setting, where each frame is manually annotated as query-relevant or query-
irrelevant. Despite the decent progress, these data-hungry methods severely rely on numerous frame-query annotations, which are significantly labor-intensive and time-consuming to collect. To alleviate annotation reliance to a certain extent, some recent works (Tan et al., 2021; Chen et al., 2019; Mithun et al., 2019; Ma et al., 2020; Lin et al., 2020; Song et al., 2020; Zhang et al., 2020d) explore a weakly-supervised setting to only leverage the coarse-grained video-query annotations instead of the fine-grained frame-query annotations. Unfortunately, this weakly-supervised setting still requires expensive video-query annotations and its performance validates unsatisfactory.

To this end, in this paper, we try to tackle a more practical but challenging setting for the TSG task, i.e., unsupervised TSG, which excludes both coarse-grained video-query annotations and fine-grained frame-query annotations. Therefore, how to guide the unsupervised model to extract the vision-language correlations becomes a crucial problem. Luckily, unlike the expensive annotated TSG task requiring long sentences and complicated videos, some cross-modal tasks (e.g., image retrieval and video retrieval) always provide a massive number of cheap and fully-annotated datasets, i.e., matched image-noun and paired video-verb. Hence, we pose a brand-new idea: can we transfer the annotation knowledge from other cheap cross-modal annotations to the complicated and challenging unsupervised TSG task? Our main motivation is that we can first learn simple image-noun and video-verb correlations to extract the visual appearance and action knowledge, respectively. Then, by generalizing such cross-modal knowledge and transferring it into unsupervised TSG, we can not only well correlate the queries with corresponding videos without large-scale video-query pairs, but also match them with image-aware frames without frame-query pairs. Therefore, the current problem becomes how to separately collect the appearance and action knowledge and adaptively transfer the knowledge into our unsupervised TSG task.

To tackle these issues, we propose a novel Cross-Modal Knowledge Transferring (CMKT) network, which fine-tunes and transfers the extracted appearance and action knowledge from other cross-modal tasks to search the best-matching video activity related to given queries in the unsupervised TSG setting. As shown in Figure 1(b), our CMKT contains three modules: (i) In the appearance knowledge module, we try to learn the object-guided matched region-noun information from the Image-Noun task to model the appearance information. (ii) In the action knowledge module, single-action and multi-action branches are designed to simulate complex video scenes. In the single-action branch, we extract the pure action information based on a set of labeled single-action videos. In the multi-action branch, since real-world videos often contain multiple actions, we introduce a clip-level copy-paste strategy to construct the action-hybrid videos to adapt to more complex scenes. (iii) Finally, we refine both the appearance and action knowledge, and aggregate and transfer them into our complicated unsupervised TSG for inference without further training.

Our main contributions are summarized as follows:

• To the best of our knowledge, we make the first attempt to transfer the cross-modal knowledge from other tasks into the TSG task. Without any TSG annotation, we can directly generalize the collected pre-trained appearance and action information for precise grounding, which eliminates complex training.

• We carefully design the appearance and action knowledge collection modules to learn simple knowledge from other cheap tasks. To handle the complicated video with multiple action contexts, we introduce a copy-paste idea to synthesize various multi-action clips, which improves the generalization ability.

• Comprehensive experiments on two challenging datasets (ActivityNet Captions and Charades-STA) show the effectiveness of our proposed CMKT.

2 Related Works

Most existing TSG methods are under the fully-supervised setting (Tang et al., 2021; Liu et al., 2023b; Zhang et al., 2021; Cao et al., 2020; Lei et al., 2020; Liu et al., 2020a, 2022b; Liu and Hu, 2022; Liu et al., 2022e), where all video-query pairs and precise segment boundaries are manually annotated. Based on the top-down approach, some methods (Anne Hendricks et al., 2017; Chen et al., 2018; Zhang et al., 2020b) first pre-define multiple segment proposals and then align these proposals with the query for cross-modal semantic matching based on the similarity. Finally, the
3 Our Method

3.1 Overview

Problem definition. Given a set of untrimmed videos \( V = \{V_h\}_{h=1}^{NV} \) and sentence queries \( Q = \{Q_g\}_{g=1}^{NQ} \), we denote \( V_h \) and \( Q_g \) as the \( h \)-th video and the \( g \)-th query, \( NV \) and \( NQ \) as the number of videos and queries, respectively. For each query, unsupervised TSG has no annotation information about its corresponding related video and detailed activity location. Obviously, our unsupervised TSG is more challenging than previous supervised TSG, since we drop all annotated information between \( V \) and \( Q \) including their correspondence and annotated segment boundaries.

Overall pipeline. To tackle the challenging unsupervised setting, we propose a novel Cross-Modal Knowledge Transfer (CMKT) network shown in Figure 2, which first designs two task-agnostic knowledge collection modules to learn both appearance and action knowledge from other labeled yet cheap cross-modal tasks, then refine and transfer their knowledge into our complicated TSG task.

For the appearance knowledge module, we first collect a set of conceptual nouns and their corresponding image regions from public Image-Noun datasets (e.g., Visual Genome (VG) (Krishna et al., 2017)), where a noun corresponds to multiple relevant regions. We denote these region-level images as \( V^a = \{v_i^a\}_{i=1}^{N^a} \) and their corresponding words (object nouns) as \( W^a = \{w_j^a\}_{j=1}^{N^a} \), where \( N^a \) and \( N^Q \) are the number of regions and nouns, respectively. Then, we explore the region-level
appearance knowledge by the region-word pair information \((w^v_j, \bar{v}^a_j)\), where \(\bar{v}^a_j\) is the consensus of all the regions corresponding to \(w^v_j\). For noun \(w^v_j\), we compute \(\bar{v}^a_j\) by averaging the representations of all relevant regions. With pairwise nouns and their consensus region representations, we can obtain general cross-modal appearance knowledge.

In the action knowledge module, we are given a massive number of conceptual verbs and their corresponding action-pure clips from popular Video-Verb datasets (e.g., Kinetics (Carreira and Zisserman, 2017)), where a verb corresponds to many labeled clips. These clips and corresponding verbs are denoted as \(V^m = \{v^m_{i,j}\}_{j=1}^{N^m}\) and \(W^v = \{w^v_{j}\}_{j=1}^{N^v}\) respectively, and \(N^m\) and \(N^v\) are the number of clips and verbs. We tend to learn the clip-level action knowledge \((w^v_j, \bar{v}^a_j)\), where \(\bar{v}^a_j\) is the consensus of all the regions corresponding to \(w^v_j\). Besides obtaining the basic action feature via averaging over all relevant single-action clips, we synthesize the multi-action clips by the copy-paste approach (Dvornik et al., 2018; Xu et al., 2021; Rong et al., 2021) for simulating more complicated scenes. By refining both single- and multi-action representations, we obtain the final action knowledge. Finally, a knowledge transfer module is designed to transfer appearance and action knowledge into our unsupervised TSG.

### 3.2 Appearance Knowledge Collection

Since the cheap Image-Word datasets provide sufficient known image-noun pairs, we can leverage them to extract the simple cross-modal appearance knowledge. However, our desired appearance knowledge has two key properties: 1) Region-aware: Rather than extract visual backgrounds, appearance knowledge is required to accurately align the region-level foreground objects and corresponding nouns. It is more fine-grained than other methods that roughly link the whole image to words. 2) One-to-one matching: Instead of aligning a noun to multiple related images in a one-to-many manner, we focus on cross-modal one-to-one matching to alleviate the appearance variations of objects by aligning the noun and its corresponding consensus region. Therefore, we collect all the nouns and their related regions on the VG dataset, which contains diverse contents that are useful to improve knowledge generalization.

**Noun-aware textual encoder.** Given \(N^n\) nouns as the label set of \(N^v\) region-level images, we utilize Glove (Pennington et al., 2014) to embed each noun into a dense vector \(\{w^n_j\}_{j=1}^{N^n}\), where \(w^n_j \in \mathbb{R}^d\) is the \(j\)-th noun feature and \(d\) is the feature dimension.

**Appearance-aware visual encoder.** Given \(N^a\) region-level images, we use Faster-RCNN (Ren et al., 2015) to extract their features \(\{a^n_j\}_{j=1}^{N^n}\), where \(a^n_j \in \mathbb{R}^d\) is the \(j\)-th appearance feature.

**Appearance knowledge acquisition.** For the \(j\)-th noun \(w^v_j\), we can obtain its prototypical region representation \(\bar{v}^a_j\) by averaging all relevant region features \(\{v^a_{j,k}\}_{k=1}^{K^n_j}\):

\[
\bar{v}^a_j = AVP(v^a_{j,1}, \ldots, v^a_{j,K^n_j}),
\]

where \(\bar{v}^a_j\) is the consensus appearance feature and \(AVP(\cdot)\) denotes the average pooling, \(K^n_j\) is the total related region number. To ensure that \(\bar{v}^a_j\) shares the same semantics with its corresponding noun-aware features, we learn two parametric transformation matrices \(B^a\) and \(B^n\) via a regularization loss (Lin et al., 2017) \(L_{app}\) to enforce consensus region feature \(v^a_j\) as close as possible to corresponding noun feature \(w^n_j\):

\[
L_{app} = ||B^a\bar{V}^a - B^nW^n||_F^2,
\]

where \(||\cdot||_F\) is Frobenius norm, \(\bar{V}^a = \{\bar{v}^a_j\}_{j=1}^{N^n} \in \mathbb{R}^{d \times N^n}\) and \(W^n = \{w^n_j\}_{j=1}^{N^n} \in \mathbb{R}^{d \times N^n}\).

### 3.3 Action Knowledge Collection

Similarly, we can utilize cheap Video-Verb datasets to extract cross-modal action knowledge by learning the video-verb pairs. To meet the requirement of diverse actions in the TSG task, we collect the verbs and their paired trimmed videos from a highly diverse video dataset Kinetics (Carreira and Zisserman, 2017). In the TSG task, it is difficult to obtain the specific feature of every single action from multi-action videos. Thus, we simulate various action videos to better generalize the mixed action features. To collect different types of action knowledge, we design two branches: single- and multi-action branches. In the single-action branch, since all the clips share the same-class action information, we aim to obtain the pure average action feature of a given verb. For the multi-action branch, we refine the consensus action feature by merging multiple action clips into a hybrid one with a copy-paste strategy.

**Verb-aware textual encoder.** Given \(N^v\) verbs, we use the Glove network (Pennington et al., 2014) to embed each verb into a dense vector \(W^v = \{w^v_j\}_{j=1}^{N^v} \in \mathbb{R}^{d \times N^v}\).
Action-aware visual encoder. Given $N^m$ labeled trimmed videos, we first extract their clip-wise features by a pre-trained 3D network (Tran et al., 2015), and then employ a multi-head self-attention (Vaswani et al., 2017) module to capture the long-range dependencies among video frames. We denote the extracted video features as $V^m = \{v^m_i\}_{i=1}^{N^m} \in \mathbb{R}^{d \times N^m}$.

Action knowledge acquisition. In the single-action branch, we are given a set of verbs $\{w^v_j\}_{j=1}^{N^v}$ and each verb $w^v_j$ corresponds to multiple action-pure videos $\{v^m_{j,p}\}_{p=1}^{N^m}$, where $N^m$ is the number of action videos corresponding to $w^v_j$. For verb $w^v_j$, we directly utilize the average pooling to obtain its average action feature: $(v^m_{j,p})' = AVP(v^m_{j,p}), \ldots, v^m_{j,N^m})$.

Since real-world videos often contain multiple actions, in the multi-action branch, we introduce a copy-paste strategy to augment the multi-action videos. As shown in Figure 2, given any two trimmed action videos, we randomly copy some clips from a video, and paste them onto the other video by replacing the same number of clips. For action-pure video $v^m_{j,p}$ with $N^m_{j,p}$ clips (corresponding to verb $w^v_j$) and another action-pure video $v^m_{x,z}$ (irrelevant to verb $w^v_j$), we randomly replace $P_{j,p}$ clips of $v^m_{j,p}$ with $P_{j,p}$ clips from $v^m_{x,z}$. The augmented multi-action video is denoted as $(v^m_{j,p})^*$. For convenience, we design a metric to denote the copy-paste amount: paste rate $\mu = P_{j,p}/N^m_{j,p} \in (0, 1)$, i.e., the percent of replaced clips in the original video $v^m_{j,p}$. After the copy-paste process, since the augmented video $(v^m_{j,p})^*$ still contains verb-related action information (related to verb $w^v_j$), we obtain the corresponding average action feature by the average pooling: $(v^m_{j,p})'' = AVP((v^m_{j,p})^*), \ldots, v^m_{j,N^m})$. $(v^m_{j,p})''$ denotes the generated action feature of $v^m_{j,p}$ after the copy-paste operation. By combining single-action $(v^m_{j,p})'$ and multi-action $(v^m_{j,p})''$, the final consensus action feature is obtained by:

$$\tilde{v}^m_j = (1 + \mu)(v^m_{j,p})' + (1 - \mu)(v^m_{j,p})'', \quad (3)$$

where $\mu$ is used to adaptively control the balance between the action-pure information and hybrid action information.

To generalize the above action knowledge, we design an inter-action loss $\mathcal{L}_{inter}$ and an intra-action loss $\mathcal{L}_{intra}$ for constraint. For the inter-action loss, in different augmented videos, we take the matched action-verb pairs as positive samples and the unmatched action-verb pairs as negative samples, and use the weighted binary cross-entropy loss to supervise the verb-relevance $\{r^{uv}_{j} \}_{j=1}^{N^v}$.

$$\mathcal{L}_{inter} = \sum_{j=1}^{N^v} (-y_j \log(r^{uv}_j) - (1 - y_j) \log(1 - r^{uv}_j)), \quad (4)$$

where $y_j$ is the label of the $j$-th sample that equals 1 for the matched action-verb pairs and 0 for the unmatched action-verb pairs; $r^{uv}_{j}$ is obtained by:

$$r^{uv}_{j} = \frac{v^m_j(w^v_j)^T}{||v^m_j||_2 ||w^v_j||_2}.$$ 

For the intra-action loss, in each multi-action video, we take verb-related video clips as positive samples and define verb-irrelevant video clips as negative samples. We adopt a hinge loss for supervision:

$$\mathcal{L}_{intra} = \sum_{j=1}^{N^v} \max(0, \beta - r^{uv}_j + \hat{r}^{uv}_j), \quad (5)$$

Eq. (5) tries to align the same-class action features and push the diff-class action features away.

3.4 Knowledge Transfer to Unsupervised TSG

In previous sections, we can obtain consensus appearance knowledge and consensus action knowledge, which are important for comprehending the target moments in our TSG task. However, we cannot directly utilize the collected knowledge in our TSG task due to (i) the domain gap between Image-Noun/Video-Verb task and the TSG task, and (ii) more complicated scenes in the TSG task. Thus, we aim to fine-tune collected knowledge to fit the TSG domain by knowledge transfer. In the following, we will illustrate how we transfer the collected knowledge into the TSG task.

TSG query encoder. Given a set of queries $Q = \{Q_g\}_{g=1}^{N^q}$, to keep the query features same as the knowledge-based one, we utilize the Glove network to embed each word into a dense vector. For $Q_g$ with $N^q_g$ words, its word-level features are denoted as $Q_g = \{f_{g,j}\}_{j=1}^{N^q_g} \in \mathbb{R}^{d \times N^q_g}$. We utilize the NLP tool spaCy (Honnibal and Montani, 2017) to parse nouns from the given query, then remove duplicate nouns. The textual features of the reserved nouns are denoted as $f_{g,j}^{n} \in \mathbb{R}^{d}$, the $j$-th noun feature and $N^q_g$ is the reserved noun number. Similarly, we parse verbs and obtain the reserved verb feature set $\{f_{g,j}^{v}\}_{j=1}^{N^q_g}$, where $f_{g,j}^{v} \in \mathbb{R}^{d}$ is the $j$-th reserved verb feature and $N^q_g$ is the reserved verb number.

TSG video encoder. Given $N^v$ videos $V = \{V_h\}_{h=1}^{N^v}$, we denote $V_h$ as the $h$-th video and $T_h$ as its frame number. For the appearance feature,
we utilize Faster-RCNN (Ren et al., 2015) built on a ResNet50 (He et al., 2016) backbone to detect $K$ objects on each video to obtain a set of appearance features $V^n_i = \{f^n_{h,i,k}\}_{k=1}^{T_h,i,k}$. For the action feature, we first divide video $V_h$ into $T_h/8$ clips. Then, we utilize a pre-trained C3D network (Tran et al., 2015) and a multi-head self-attention to extract the clip-aware features $V^m_i = \{f^m_{h,i,k}\}_{k=1}^{T_h,i,k}$. **Fitting the TSG domain via knowledge transfer.** To transfer the pre-trained knowledge into the TSG task, we fine-tune the collected appearance and action knowledge based on the principle of cross-modal cycle consistency. Given the noun features $\{f^n_{g,j}\}_{j=1}^{N_g}$ and verbs features $\{f^v_{g,j}\}_{j=1}^{N_g}$, we search each $f^n_{g,j}$ in the appearance knowledge and each $f^v_{g,j}$ in the action knowledge with the cosine similarity, then re-represent them with the collected consensus appearance features $\tilde{V}^n_j$ and action features $\tilde{V}^v_j$, respectively. For convenience, we denote $A_g = \{\tilde{V}^n_j\}_{j=1}^{N_g} \in \mathbb{R}^{d \times N_g}$ and $M_g = \{\tilde{V}^v_j\}_{j=1}^{N_g} \in \mathbb{R}^{d \times N_g}$. As for the visual appearance and action features, we also enhance them via the collected knowledge, respectively.

To fine-tune consensus features ($A_g$ and $M_g$), we design a cross-modal cycle-consistent loss to make collected knowledge more applicable to the TSG task. For appearance knowledge, we first compute the cosine similarity matrix $S_g^n = \{(v^{n}_{j}v^m_{j})\}_{j=1}^{N_g}$ between noun-guided consensus appearance features and video appearance features. Then, we combine all appearance features by treating similarities as weights to obtain reconstructed nouns features:

$$\begin{align*}
\tilde{Q}^n_g &= W^n_g C^n \sigma_r(S^n_g)^T, \\
S^n_g &= (W^n_g A_g)^T W^n_g C^n, \\
\end{align*}$$

where $C^n = \{V^n_{h}\}_{h=1}^N$, $W^n_g$ is a learnable transformation matrix, $\sigma_r(\cdot)$ is the softmax operation along the column dimension, and $\tilde{Q}^n_g$ is the reconstructed noun representation. Similarly, we combine all action features to obtain the reconstructed verbs features:

$$\begin{align*}
\tilde{Q}^v_g &= W^v_g C^v \sigma_r(S^v_g)^T, \\
S^v_g &= (W^v_g M_g)^T W^v_g C^v, \\
\end{align*}$$

where $C^m = \{V^m_{h}\}_{h=1}^N$, $W^v_g$ is a learnable transformation matrix, $\tilde{Q}^v_g$ is the reconstructed verb features, and $S^v_g \in \mathbb{R}^{d \times N_g}$ is the similarity matrix between transformed verb and action features.

Our cycle-consistent loss aims to (i) align the semantics of original nouns and their reconstructed ones, (ii) pull verbs and their reconstructed ones together. Thus, we construct two indicator matrices $\mathbf{Y}^a_g \in \mathbb{R}^{N_g \times N_g}$ and $\mathbf{Y}^m_g \in \mathbb{R}^{N_g \times N_g}$ to evaluate if each noun/verb and its reconstructed one are the same or not as:

$$
\begin{align*}
\mathbf{Y}^a_g &= \sigma((\tilde{Q}^n_g)^T (W^n_g A_g))^T, \\
\mathbf{Y}^m_g &= \sigma((\tilde{Q}^v_g)^T (W^v_g M_g))^T. \\
\end{align*}
$$

With $\mathbf{Y}^a_g$ and $\mathbf{Y}^m_g$, we optimize $W^n_g$ and $W^v_g$ by minimizing the cross-entropy loss:

$$
\mathcal{L}_{clc} = -\sum_{g=1}^{N_g} \sum_{j=1}^{N_g} y^{a,g,j} \log(\tilde{y}^{a,g,j}) + \sum_{j=1}^{N_g} y^{m,g,j} \log(\tilde{y}^{m,g,j}),
$$

$y^{a,g,j}$ and $y^{m,g,j}$ are one-hot ground-truth vectors, where the $z$-th and $x$-th value are one, and rest values are zeros. By $\mathcal{L}_{clc}$, we use $W^n_g$ and $W^v_g$ to transfer all consensus appearance and action knowledge into TSG.

### 3.5 Knowledge-based Grounding

By pre-training the appearance knowledge by $\mathcal{L}_{app}$ and the action knowledge by $\mathcal{L}_{action}$ with the knowledge transfer by $\mathcal{L}_{clc}$, we can directly utilize the suitable transferred appearance knowledge ($w^n_g$, $v^n_g$) and action knowledge ($w^v_g$, $v^v_g$) to localize the target segment boundary without any further training. Specifically, during the inference, given some videos $V_h \in \mathbb{R}^{N_g}$ and a query set $Q_g \in \mathbb{R}^{N_g}$, we first extract their appearance feature $f^n_{h,i,k}$, verb feature $f^v_{h,i,k}$, noun feature $f^{n,g,j}$, and verb feature $f^{v,g,j}$, respectively. Then, based on the collected appearance knowledge and action knowledge modules, we can obtain the corresponding consensus appearance and action features. We compute the cosine similarity matrix between consensus appearance features and video appearance features, and also calculate the cosine similarity matrix between consensus action features and video action features.

After that, following (Liu et al., 2022c), we select the best-matched clip with the highest scores (sum of appearance similarity and action similarity) in the target video as the target segment center. We add the left/right frames into the segment if the ratio of their scores to the frame score of the closest segment boundary is less than a threshold $\gamma$. We repeat this step until no frame can be added. In this way, we can generate the final segment boundary.
Table 1: Performance comparison with state-of-the-art methods. ‘Type’ refers to supervision level, FS: fully-supervised setting, WS: weakly-supervised setting, US: unsupervised setting.

<table>
<thead>
<tr>
<th>Method</th>
<th>ActivityNet Captions</th>
<th>Charades-STA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IoU=0.1</td>
<td>IoU=0.3</td>
</tr>
<tr>
<td>CTRL (Gao et al., 2017)</td>
<td>FS</td>
<td>49.10</td>
</tr>
<tr>
<td>2D-TAN (Zhang et al., 2020b)</td>
<td>FS</td>
<td>-</td>
</tr>
<tr>
<td>LGI (Mun et al., 2020)</td>
<td>FS</td>
<td>-</td>
</tr>
<tr>
<td>VSLNet (Zhang et al., 2020a)</td>
<td>FS</td>
<td>-</td>
</tr>
<tr>
<td>VCA (Wang et al., 2021c)</td>
<td>WS</td>
<td>73.73</td>
</tr>
<tr>
<td>RTBPN (Zhang et al., 2020c)</td>
<td>WS</td>
<td>74.47</td>
</tr>
<tr>
<td>CTF (Chen et al., 2020b)</td>
<td>WS</td>
<td>74.20</td>
</tr>
<tr>
<td>MARN (Song et al., 2020)</td>
<td>WS</td>
<td>-</td>
</tr>
<tr>
<td>SCN (Lin et al., 2020)</td>
<td>WS</td>
<td>74.48</td>
</tr>
<tr>
<td>BAR (Wu et al., 2020)</td>
<td>WS</td>
<td>-</td>
</tr>
<tr>
<td>CCL (Zhang et al., 2020d)</td>
<td>WS</td>
<td>-</td>
</tr>
<tr>
<td>LCNet (Yang et al., 2021)</td>
<td>WS</td>
<td>78.58</td>
</tr>
<tr>
<td>CNM (Zheng et al., 2022)</td>
<td>WS</td>
<td>78.13</td>
</tr>
<tr>
<td>WSTAN (Wang et al., 2021b)</td>
<td>WS</td>
<td>79.78</td>
</tr>
<tr>
<td>DSCNet (Liu et al., 2022c)</td>
<td>US</td>
<td>-</td>
</tr>
<tr>
<td>Our CMKT</td>
<td>US</td>
<td>73.35</td>
</tr>
</tbody>
</table>

4 Experiment

4.1 Datasets and Evaluation Metrics

ActivityNet Captions. From ActivityNet v1.3 (Heilbron et al., 2015; Lan et al., 2022), ActivityNet Captions contains 20,000 YouTube videos and 100,000 language queries. On average, a video is 2 minutes and a query has about 13.5 words. Following the public split (Gao et al., 2017), we utilize 17,031 video-query pairs for testing.

Charades-STA. Built upon the Charades dataset (Sigurdsson et al., 2016), Charades-STA contains 16,128 video-sentence pairs. The average video length is 0.5 minutes. The language annotations are generated by sentence decomposition and keyword matching with manual checks. Following (Gao et al., 2017), we remove 12,408 training pairs and utilize the others for testing.

Evaluation metrics. Following previous works (Mun et al., 2020; Zhang et al., 2020a), we use “IoU=m” for evaluation, which denotes the percentage of queries having at least one result whose Intersection over Union (IoU) with ground truth is larger than m. In our experiments, m ∈ {0.3, 0.5, 0.7} for Charades-STA and m ∈ {0.1, 0.3, 0.5, 0.7} for ActivityNet Captions. Also, we utilize mean IoU (mIoU) as the averaged temporal IoU between the predicted boundary and the ground-truth one.

4.2 Implementation details

For the appearance knowledge module, we utilize all the words on the Visual Genome dataset, so the noun number of region-level semantics is 27,801. The feature dimension $d$ is set to 512 for all the features. For the action knowledge module, we use the pre-trained C3D (Tran et al., 2015) network to extract the single-action and multi-action information on the Kinetics dataset. Also, we use all the verbs on the Kinetics dataset. To make a fair comparison with previous works (Gao et al., 2017; Zhang et al., 2020b), we also utilize the pre-trained C3D model to extract video features and employ the Glove model to obtain word embeddings. As some videos are too long, we set the length of video feature sequences to 128 for Charades-STA and 256 for ActivityNet Captions, respectively. We fix the query length to 10 in Charades-STA and 20 in ActivityNet Captions. Threshold $\gamma$ is set to 0.9 on Charades-STA and 0.8 on ActivityNet Captions. All the experiments are implemented by PyTorch.

4.3 Comparison with State-of-the-Arts

We conduct performance comparison with three categories of state-of-the-art TSG methods: (i) Fully-supervised (FS) setting (Gao et al., 2017; Zhang et al., 2020b; Mun et al., 2020; Zhang et al., 2020a); (ii) Weakly-supervised (WS) setting (Chen et al., 2020b; Lin et al., 2020; Wang et al., 2021b; Wu et al., 2020; Song et al., 2020; Zhang et al., 2020d; Huang et al., 2021; Wang et al., 2021c; Yang et al., 2021; Zhang et al., 2020c; Zheng et al., 2022); (iii) Unsupervised (US) setting (Liu et al., 2022c).

Table 1 reports the performance comparison. Without any supervision, our CMKT still obtains the competitive performance to WS and US methods on the ActivityNet Captions dataset. Compared with the US method DSCNet, our CMKT improves the performance by 3.40% and 3.12% in ActivityNet Captions.
Table 2: Main ablation study, where we remove each key component to investigate its effectiveness. “AK” denotes “appearance knowledge”, “MK” denotes “action knowledge”, “KT” denotes “knowledge transfer”.

<table>
<thead>
<tr>
<th>Model</th>
<th>AK</th>
<th>MK</th>
<th>KT</th>
<th>ActivityNet Captions</th>
<th>Charades-STA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IoU=0.3</td>
<td>mIoU</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IoU=0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IoU=0.7</td>
<td></td>
</tr>
<tr>
<td>CMKT(a)</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>45.81</td>
<td>24.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.52</td>
<td>31.36</td>
</tr>
<tr>
<td>CMKT(b)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>47.50</td>
<td>26.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.37</td>
<td>32.71</td>
</tr>
<tr>
<td>CMKT(c)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>48.02</td>
<td>27.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.96</td>
<td>34.05</td>
</tr>
<tr>
<td>CMKT(Full)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>50.69</td>
<td>31.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.42</td>
<td>38.79</td>
</tr>
</tbody>
</table>

Table 3: Necessity of average pooling on ActivityNet Captions, where “MIP” means min-pooling, “MAP” means max-pooling, and “A VP” means average pooling.

<table>
<thead>
<tr>
<th>Module</th>
<th>Changes</th>
<th>IoU=0.3</th>
<th>IoU=0.5</th>
<th>IoU=0.7</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance</td>
<td>MIP</td>
<td>47.80</td>
<td>29.08</td>
<td>13.76</td>
<td>35.70</td>
</tr>
<tr>
<td></td>
<td>MAP</td>
<td>48.03</td>
<td>30.72</td>
<td>14.45</td>
<td>36.68</td>
</tr>
<tr>
<td></td>
<td>AVP</td>
<td>50.69</td>
<td>31.28</td>
<td>16.42</td>
<td>38.79</td>
</tr>
<tr>
<td>Action</td>
<td>MIP</td>
<td>49.27</td>
<td>29.26</td>
<td>16.20</td>
<td>36.61</td>
</tr>
<tr>
<td></td>
<td>MAP</td>
<td>48.36</td>
<td>28.72</td>
<td>13.43</td>
<td>35.53</td>
</tr>
<tr>
<td></td>
<td>AVP</td>
<td>50.69</td>
<td>31.28</td>
<td>16.42</td>
<td>38.79</td>
</tr>
</tbody>
</table>

4.4 Ablation study

Main ablation studies. To examine the effectiveness of each module, we conduct the main ablation studies in Table 2. CMKT(a) is the baseline model, which only utilizes pre-trained appearance and action features for inference without fine-tuning knowledge. CMKT(b) removes appearance knowledge module, and directly use the action knowledge to locate the target segment. In CMKT(c), we only use the appearance information to obtain the segment boundary by removing the action information. CMKT(Full) is our full CMKT model. CMKT(Full) performs better than all the ablation models, demonstrating the effectiveness of appearance and action knowledge for understanding the long yet complex sentence in TSG.

Influence on different types of action knowledge. To investigate the contribution of single- and multi-action videos to our final consensus action features, we conduct the corresponding ablations in Table 4 to evaluate the significance of single-action and multi-action knowledge. Obviously, both single-action and multi-action features contribute to our action knowledge module.

Investigation on different types of knowledge transfer. To evaluate the importance of appearance and action knowledge, we conduct corresponding experiments. As shown in Table 5, each knowledge contributes a lot to transferring the knowledge into our task. By jointing appearance and action knowledge, the performance can further boost.

Effect of the copy-paste strategy. To analyze the (MAP) on both two knowledge collection modules, showing that AVP is suitable for integrating the information among a single modality for representing the same semantic as the other modality.

Effect of the copy-paste strategy. To analyze the...
performance of our copy-paste strategy, an ablation experiment is conducted in Table 6, where “w/ video connection” means that we direct connect two videos for data synthesis. Thus, the copy-paste strategy is more effective to synthesize the multi-action videos for better grounding performance.

**Analysis on the hyper-parameters.** As shown in Figure 3, we investigate the impact of four hyper-parameters: object number ($K$); paste rate ($\mu$); positive parameter ($\beta$); evaluation metrics (IoU). (i) With larger $K$, our proposed CMKT performs better. To balance the performance and the computation cost of object detection, we choose $K = 8$. (ii) All the variants obtain the best performance when the paste rate $\mu$ is set to 0.8. (iii) Our CMKT obtains the best performance when $\beta = 0.9$. (iv) CMKT(Full) obtains the best performance on all IoUs, demonstrating the effectiveness of each module. Thus, we set $K = 8$, $\mu = 0.8$ and $\beta = 0.9$ in all the experiments.

### 4.5 Qualitative Results

As shown in Figure 4, we visualize the localization results. Our CMKT can predict more precise segment boundaries than fully-supervised model 2D-TAN and weakly-supervised method WSTAN.

![Figure 3: Effect of parameters on ActivityNet Captions.](image1)

![Figure 4: Qualitative results sampled from ActivityNet Captions (top) and Charades-STA (bottom).](image2)

### 5 Conclusion

In this paper, we address the challenging yet practical unsupervised TSG task from a brand-new perspective of knowledge transfer. Without further grounding training, we can directly utilize the general knowledge from other cross-modal tasks to guide to match the video and query for retrieving the target segment. To our best knowledge, we make the first attempt to transfer the knowledge from other multi-modal topics to our TSG task. Experimental results validate the effectiveness of CMKT, outperforming the existing unsupervised method by a large margin.
6 Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) under Grant No. 61972448.

7 Limitations

This work analyzes an interesting yet challenging problem of how to transfer the annotation knowledge from other cheap cross-modal annotations to the complicated and challenging unsupervised TSG task. Although we can transfer the knowledge from other cheap multi-modal datasets, the contexts of these datasets are not always helpful for our grounding task. However, our proposed CMKT still provides a brand-new and interesting idea, cross-modal knowledge transfer, for promoting the development of these areas. Therefore, a more contextual and generalizable way to transfer cross-modal knowledge between different topics is a promising future direction.

References


Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In 2015 IEEE conference on computer vision and pattern recognition (CVPR), pages 961–970. IEEE.


Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023b. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023, pages 2499–2508. ACM.


8733