Semantics Guided Contrastive Learning of Transformers for Zero-shot Temporal Activity Detection

Sayak Nag*, Orpaz Goldstein†, Amit K. Roy-Chowdhury*
*University of California, Riverside, USA, †Amazon, USA
{snag@ece, amitrc@ece.}ucr.edu, orpgol@cs.ucla.edu

Abstract

Zero-shot temporal activity detection (ZSTAD) is the problem of simultaneous temporal localization and classification of activity segments that are previously unseen during training. This is achieved by transferring the knowledge learned from semantically-related seen activities. This ability to reason about unseen concepts without supervision makes ZSTAD very promising for applications where the acquisition of annotated training videos is difficult. In this paper, we design a transformer-based framework titled TranZAD, which streamlines the detection of unseen activities by casting ZSTAD as a direct set-prediction problem, removing the need for hand-crafted designs and manual post-processing. We show how a semantic information-guided contrastive learning strategy can effectively train TranZAD for the zero-shot setting, enabling the efficient transfer of knowledge from the seen to the unseen activities. To reduce confusion between unseen activities and unrelated background information in videos, we introduce a more efficient method of computing the background class embedding by dynamically adapting it as part of the end-to-end learning. Additionally, unlike existing work on ZSTAD, we do not assume the knowledge of which classes are unseen during training and use the visual and semantic information of only the seen classes for the knowledge transfer. This makes TranZAD more viable for practical scenarios, which we evaluate by conducting extensive experiments on Thumos'14 and Charades.

1. Introduction

With video content growing rapidly on the internet [1], automated indexing and analysis of video data have taken a pivotal position in information retrieval studies. In recent years, deep learning based temporal activity detection (TAD) has emerged as a solution for automating the retrieval of pertinent activities in long untrimmed videos [67, 15, 16, 57, 9, 60]. However, most of these methods need to be trained with heavy supervision to achieve good performance. In real-world applications, it is often quite difficult and expensive to acquire well-annotated video samples that exhaust all possible activity classes, which makes existing TAD frameworks prone to misclassifying activity instances that are previously unseen during training. Therefore, there is a growing need to develop methods that can learn with limited supervision.

One such approach is zero-shot learning (ZSL), where training and testing data come from disjoint sets of classes sharing some semantic relation. The goal is to transfer knowledge learned from the detection of seen classes to the detection of unseen classes, which is accomplished by exploiting some common prior information such as hand-crafted attributes or semantic label embeddings. The ability to generalize to unseen concepts without heavy supervision makes ZSL very attractive for applications like video analysis on the edge, where the lower computation power of edge devices makes large-scale supervised learning infeasible.

Existing ZSL studies have largely focused on image data with zero-shot classification/recognition (ZSR) being the most popular [35, 34, 63, 10] followed by zero-shot object detection (ZSD) [41, 42, 2]. The limited work done on videos has largely focused on extending ZSR for activity classification in short-trimmed video clips [39, 8, 7, 28, 58]. However, in real-world settings, web videos are long and untrimmed, containing multiple action instances, making TAD a much more challenging problem than simple activity recognition [49]. In this paper, we address the problem of TAD in the ZSL setup, formally called Zero-Shot Temporal Activity Detection (ZSTAD), whereby the knowledge learned from modeling the spatio-temporal dynamics of seen activities is transferred to the detection of unseen activities.

Recently, [62] addressed this task by introducing a modified version of the popular RC3D framework [57], where semantic embeddings are used for the metric-based classification of temporal region proposals. RC3D, being a two-stage detector, relies heavily on several hand-crafted components such as manually designed anchor sets and non-maximal suppression (NMS) to improve performance [5]. Due to the dynamic nature of video data, designing anchor sets that cover all ground-truth instances is very challenging

[49], which is further amplified for the zero-shot case, where unseen class activities have no supervision for manually tuning the anchor-set design. Consequently, this affects the quality of the unseen class proposals [61]. Additionally, post-processing steps like NMS increase inference time, making current ZSTAD frameworks like [62] unsuitable for applications such as edge computing, where low latency inference is crucial. Another critical point is [62] assumes the availability of the unseen class semantic embeddings during the training phase itself, which they use to construct a super-class classification loss that significantly boosts their zero-shot detection performance. However, the information provided by the unseen class semantic embeddings, although not as rich as visual information, enables the framework of [62] to gauge which classes are unseen during training. This renders the learning model of [62] impractical for real-world scenarios where retrieval systems may not have any prior context about the unseen class distribution during the training phase and are introduced to them only during inference.

To overcome the challenges of two-stage detectors [45, 19, 57], in recent years, DETR [5] and its variants [69, 54, 25, 49] have introduced transformer-based set-prediction models that streamlines localization, by-passing the need for proposal generation and its hand-crafted components and consequently achieving faster inference. However, these models have been introduced for fully supervised learning and are not fit for the ZSL setup. In this work, we cast ZSTAD as a set-prediction problem and introduce a transformer-based zero-shot activity detector titled TranZAD. *TranZAD leverages the multi-headed attention of the transformer [52] along with the prior information from semantic word embeddings to transfer knowledge from the seen activities to the detection of semantically related unseen activities. Additionally, unlike [62], we only use the semantic information of the seen classes for model training, thus developing a more practical retrieval framework.*

An illustration of our setup is shown in Fig. 1. During the training phase, videos containing only seen class activities are available. In order to transfer the knowledge learned from seen activities to the detection of unseen activities, TranZAD learns to associate the visual features of an activity with its corresponding class-specific semantic embedding. This is accomplished via a contrastive learning strategy, whereby the transformer model learns to map the *visual features* of the activities in a video to the *semantic feature space* which are then contrasted with the seen class semantic embeddings for metric-based classification. In this way, TranZAD learns a consistent visual-semantic mapping, enabling it to transfer the association knowledge learned from the seen activities to semantically related the unseen ones.

ZSTAD being a localization problem, necessitates distinguishing the background information in videos from the pertinent activities. This is enabled by assigning a robust

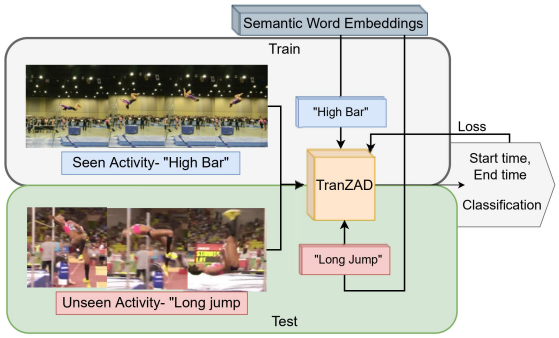

Figure 1: **High-level illustration of ZSTAD using TranZAD.** TranZAD is trained on videos containing only seen class activity instances. Prior information from the semantic word embeddings, of the activity labels, are utilized to transfer knowledge from the seen to the unseen classes. At inference, TranZAD utilizes the unseen class semantic embeddings to perform zero-shot detection of the unseen activity segments.

semantic embedding to the background class to prevent confusion with the unseen activity classes. Zhang et al. [62] achieved this by solving an optimization problem whereby a fixed representation is assigned to the background class which is least similar to all the label embeddings (seen and unseen). Their background embedding, derived solely from the semantic information of the classes, fails to incorporate their corresponding visual information, rendering it ineffective in modeling complex and diverse background information. Moreover, this approach of [62] again relies on the impractical assumption that the unseen classes are known apriori during training. We rectify this by learning the background embedding in a joint end-to-end manner, enabling it to model more complex background information.

While the contrastive learning based classification enables effective visual-semantic mapping, it is also necessary to ensure that visual features of an activity remain consistent across different videos. This is motivated by the concept of *temporal coherence* [33, 44], whereby the features corresponding to an activity should be focused on the discriminative aspects, such as gait, and ignore background nuances such as illumination and occlusion across different videos, as well as, different time segments of the same video. To enable this we apply a supervised-contrastive loss [26, 48] on the intermediate visual representations produced by the transformer, and show it's efficacy in boosting the zero-shot detection performance of unseen classes.

Main contributions. To the best of our knowledge, *this is the first work to utilize a transformer-based set-prediction framework to address ZSTAD, where the framework is trained using a semantic information guided contrastive learning strategy.* The main characteristics of our proposed solution are as follows:

1. We frame ZSTAD as a set-prediction problem and introduce a transformer-based setup, TranZAD, for direct detection

of unseen activities, removing the need for hand-crafted components and consequently achieving faster inference.
2. We introduce a novel approach for obtaining the background label embedding, which enables the modeling of diverse and complex background scenes.
3. This the first study to explore semantic information guided contrastive learning of a multi-headed attention model to address the challenging problem of ZSTAD.
4. Compared to the existing state-of-the-art, TranZAD does not rely on the explicit knowledge of which classes are unseen and still achieves superior or comparable performance, validated by experiments on two popular ZSTAD datasets THUMOS'14 [22], and Charades [47].

2. Related Work

Temporal Activity Detection. Temporal activity detection (TAD) is the study of simultaneous classification and temporal localization of multiple action instances in long untrimmed videos. Current state-of-the-art TAD methods are primarily two-stage detectors involving temporal proposal generation followed by action classification [67, 15, 16, 57, 9, 60]. The performance of these methods is attributed to fully supervised training on large-scale annotated data, which makes them fail to detect unseen activities during inference and require re-training with heavy supervision on the new activities. This is often difficult to acquire in the real world. Furthermore, the proposal generation mechanism relies on hand-crafted anchor placements [3, 20, 15] or manually-tuned boundary matching mechanisms [67, 30, 29], as well as NMS-based post-processing [49] which increases inference time.

Zero-Shot Learning. Zero-shot learning is the study of generalizing to previously unseen classes by transferring knowledge from semantically related seen classes. Unlike the simple classification problem of zero-shot recognition (ZSR), zero-shot detection (ZSD) is much more challenging as it focuses on the joint localization and classification of previously unseen instances [42]. Existing literature on both ZSD and ZSR have largely focused on image data [10, 27, 35, 63, 41, 56, 64, 42, 65, 32, 53, 42, 2, 68, 40, 66, 4]. The limited work done on videos [39, 8, 7, 28, 58, 37] are mostly focused on extending ZSR for classification of short-trimmed video clips. Recently, Zhang *et al.* [62] attempted to address this by introducing a modified RC3D framework, which maps temporal region proposals to the semantic space and compares them with Word2Vec [18] embeddings for metric-based classification. However, their framework suffers from the challenges associated with two-stage detectors like RC3D [57] and similar to many previous ZSD studies on images [42, 40], assume the knowledge of which classes are unseen during training and use the semantic context of these unseen classes to boost ZSD performance. This does not reflect many practical scenarios where any information about the unseen classes may not be available during model training. Additionally, Zhang et al. [62], like many prior ZSD studies [42, 41, 40], assign a fixed representation to the background embedding, which is ineffective in modeling complex background information of video data.

Transformers in Vision. The success of transformers in NLP tasks [52] has inspired several computer vision applications, such as image recognition [43, 13, 14], image generation [36], object detection [5, 69, 61, 54, 25] and also video understanding [49, 17]. Recently, Carion *et al.* [5] introduced DETR, a transformer-based set-prediction method for object detection in images, removing the need for hand-crafted designs and manual-post processing. However, as shown by Tan *et al.* [49], directly extending the setup of [5] to videos is problematic, owing to the inherent slowness of video features which makes the traditional transformer encoder prone to over-smoothing the video representations leading to a reduction in their discriminability. Tan *et al.* [49] addressed this by substituting the transformer encoder with a multi-layered perceptron (MLP). We use this insight of [49] in our experiments. However, unlike [49], which addresses activity proposal generation in the fully supervised setup, our framework focuses on the direct set-based prediction of unseen activities in a zero-shot setting.

Contrastive Learning Contrastive learning focuses on learning representations that maximize the alignment of similar instances. The utilization of contrastive losses has led to significant performance gains in self-supervised representation learning [21, 55, 51, 12, 11]. Recently, many studies have extended the batched contrastive loss to the supervised setting [26, 48, 59]. In this paper, we utilize supervised contrastive learning in two ways, 1) to build an effective visual-semantic mapping relationship for performing semantics-guided classification, and 2) to enable consistency of the activity visual features across different temporal segments within the same video, as well as, across different videos.

3. Methodology

3.1. Problem Description

In ZSTAD, the task is to perform joint classification and temporal localization of activity categories that are previously unseen during training. Therefore, given $\mathcal{C}_s$ seen activity classes and $\mathcal{C}_u$ unseen activity classes, the training data set $(\mathcal{X}_{c_s}, \mathcal{Y}_{c_s}) = \{(\mathbf{x}_{c_s,i}, y_{c_s,i})\}_{i=1}^{N_s}$ is composed of N_s untrimmed videos each containing temporal annotations from only the seen activity classes, and the testing set $\mathcal{X}_{c_u} = \{\mathbf{x}_{c_u,j}\}_{j=1}^{N_u}$ is comprised of N_u videos, with each containing at least one activity from the unseen classes. The seen and unseen are semantically related, and we exploit this relationship to guide the training of our framework.

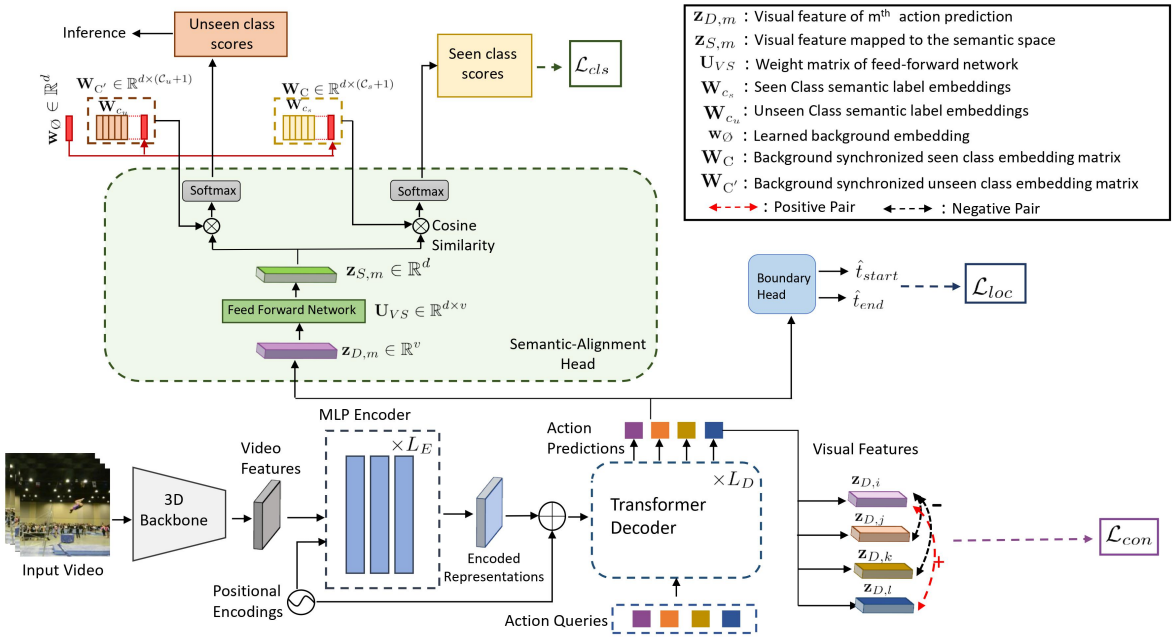

Figure 2: **Overview of TranZAD**. TranZAD addressses ZSTAD as a direct set-prediction problem, and has four components, 1) a multi-layered perceptron (MLP) encoder (with L_E layers) for transforming the extracted video features into a compact representation, 2) a transformer decoder (with L_D layers) for generating the visual features of all the activities in a video, 3) a semantic alignment head which decodes the class labels by contrasting the semantic embeddings with the visual features, and 4) a boundary head for obtaining the temporal coordinates.

3.2. Overview of Method

In order to transfer knowledge from the seen to the unseen classes, we leverage the semantic word embeddings of the textual descriptors of each activity category obtained using unsupervised vector embedding models such as Word2Vec [18] and GLoVE [38]. These embeddings provide a measure of the semantic relationship between the seen and unseen classes [42, 4, 62]. The seen and unseen class semantic embeddings are denoted as, $\mathbf{W}_{c_s} = \{\mathbf{w}_{c_s,i}\}_{i=1}^{\mathcal{C}_s} \in \mathbb{R}^{d\times\mathcal{C}_s}$ and $\mathbf{W}_{c_u} = \{\mathbf{w}_{c_u,j}\}_{j=1}^{\mathcal{C}_u} \in \mathbb{R}^{d\times\mathcal{C}_u}$ respectively. The background class embedding is denoted as $\mathbf{w}_{\emptyset}$, which is used to distinguish the pertinent activity classes from the background information in videos. Unlike existing ZSTAD work [62] we do not assign a fixed representation to $\mathbf{w}_{\emptyset}$ but model it as a learnable network parameter (For eg, in pytorch it is formulated as `nn.Embedding()` the weights of which are made to simulate the background embedding). During training only, $\mathbf{W}_{c_s}$ is available, which along with $\mathbf{w}_{\emptyset}$ is used to establish an effective visual-semantic mapping for the seen activities, which in turn enables the transfer of knowledge to the detection of semantically-related unseen activities.

The schematic representation of TranZAD is shown in Fig. 2. For each video we obtain detections in a sliding window manner, dividing the video into T overlapping segments $\{\hat{\mathbf{x}}_i\}_{i=1}^{T}$, where T depends on the temporal window, overlap ratio, and video duration. A 3D convolutional network is used to extract short-term spatio-temporal features of each temporal segment $\hat{\mathbf{x}}_i$, which along with fixed positional encodings [52] are provided as input to the multi-layered perceptron (MLP) based encoder for obtaining a compact representation of $\hat{\mathbf{x}}_i$. As discussed earlier, the traditional transformer encoder [5] is substituted by an MLP since the inherent slowness of video features makes the former prone to over-smoothing the video representations in-turn diminishing their discriminability [49]. The encoder representations and M learned query encodings called *action queries* are passed to the transformer decoder which utilizes multi-headed attention [52] to aggregate long-term temporal information from the encoder representations into the action queries and transforms them into a set of M *action predictions*, each of which represent the visual features of the activities in $\hat{\mathbf{x}}_i$. The *action predictions* are parallelly decoded into their respective classes and temporal coordinates (start and end times), using the semantic-alignment head and the boundary head, respectively.

The training of the framework is guided by three losses $\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$ and $\mathcal{L}_{con}$. $\mathcal{L}_{loc}$ is the temporal localization loss applied on the predicted coordinates generated by

the boundary head. On the other hand $\mathcal{L}_{cls}$ and $\mathcal{L}_{con}$ are supervised contrastive losses, where the former is the main classification loss associated with the semantic-alignment head whereby the visual features are first mapped to the semantic space and contrasted with the $\mathbf{W}_s$ and $\mathbf{w}_\emptyset$. The latter on the other hand, is applied on the intermediate visual features of the decoder and ensures that the visual features of each activity are consistent throughout the training set irrespective of background nuances in different videos. The individual modules and losses are discussed in detail below.

3.3. Network Architecture

Video Feature Extraction. The video features can be extracted using any 3D convolutional network [50, 6] as the backbone. Therefore, for each video segment $\hat{\mathbf{x}}_i$ in a given sliding window t, the extracted features are denoted as $\mathbf{f}(\hat{\mathbf{x}}_i)$ with a set temporal length of l_t.

Feature Encoder. The encoder is designed as an MLP, which takes in the feature representation of a video segment $\mathbf{f}(\hat{\mathbf{x}})$ along with fixed positional encodings, $pos(\mathbf{f}(\hat{\mathbf{x}}))$ and transforms it into a compact representation [52, 5, 49]. Formally the output of the encoder with L_E layers is given as follows,

$$\mathbf{Z}_E = \sum_{j=1}^{L_E} \mathbf{U}_{(E,j)}^T (\mathbf{f}(\hat{\mathbf{x}}) + pos(\mathbf{f}(\hat{\mathbf{x}}))) \tag{1}$$

where, $\mathbf{U}_{(E,j)}$ is the weight matrix of the j^{th} layer.

Transformer Decoder. We use the standard transformer [52] decoder in our framework. It takes as input the encoded video representation, $\mathbf{Z}_E$ and M *action queries* $q \in \mathbb{R}^{v \times M}$. The transformer decoder leverages multiple encoder-decoder and stacked multi-head attention to model the long term spatio-temporal relationship between all the activities in a video clip [49, 5]. In this way, the decoder learns all inter-dependencies in a pair-wise manner and refines the action queries q into a set of M action predictions $\mathbf{Z}_D \in \mathbb{R}^{v \times M}$. Thus the output $\mathbf{Z}_D$ is a collection of the visual features of all the activities in a video clip, which are then decoded to their respective class labels and temporal coordinates by the detection heads. To ensure that $\mathbf{Z}_D$ is generalizable for zero-shot detection, it is necessary to infuse information from the semantic embeddings in the decoding process. This is done using the semantic alignment head as described below.

Detection Heads. The following detection heads are used to independently decode the M action predictions into their class labels and temporal coordinates.

1. **Semantic Alignment Head**: The primary purpose of this head is to learn the relationship between the visual and semantic features of the seen activities by establishing an effective visual-semantic mapping during the training phase. As shown in Fig. 2, this is accomplished by first mapping $\mathbf{Z}_D$ to the semantic space using a feed-forward network with weights $\mathbf{U}_{VS} \in \mathbb{R}^{v \times d}$ to obtain $\mathbf{Z}_S$, where $\mathbf{Z}_S = \mathbf{U}_{VS}^T \mathbf{Z}_D$. Simultaneously, $\mathbf{w}_\emptyset$ is concatenated with $\mathbf{W}_{c_s}$ to get a background synchronized seen-class embedding matrix $\mathbf{W}_\mathrm{C} = [\mathbf{W}_{c_s}; \mathbf{w}_\emptyset] \in \mathbb{R}^{d \times (\mathcal{C}_s + 1)}$. Therefore, the classification score for the m^{th} action prediction is obtained as follows,

$$p_{m,\mathrm{c}}(\mathrm{c}|\mathbf{z}_{S,m}) = \frac{exp(\hat{\mathbf{w}}_\mathrm{c}^T \hat{\mathbf{z}}_{S,m} / \tau_{cls})}{\sum_{\mathrm{c}=1}^{\mathrm{C}} exp(\hat{\mathbf{w}}_\mathrm{c}^T \hat{\mathbf{z}}_{S,m} / \tau_{cls})} \tag{2}$$

where $\hat{\mathbf{w}}_\mathrm{c}$ and $\hat{\mathbf{z}}_{S,m}$ are the l_2 normalized feature vectors of the c^{th} semantic embedding and the m^{th} mapped visual feature respectively, $\mathrm{C} = \mathcal{C}_s + 1$, and τ_{cls} is a learnable temperature parameter for scaling the cosine similarity. The p.m.f. over all C classes is $\mathbf{p}_m = [p_{m,1}, p_{m,2}, ..., p_{m,\mathrm{C}}]$. Compared to assigning a fixed value to τ_{cls}, we observe a more discriminative visual-semantic feature alignment by learning it.

2. **Boundary Head**: The boundary head is a simple feed-forward network with two output nodes that takes as input the activity visual features $\mathbf{Z}_D$ and outputs their individual temporal coordinates $\hat{\mathbf{t}}_m = (\hat{t}_{start}, \hat{t}_{end})_m$.

3.4. Loss Functions

Set based label assignment. The optimal bipartite matching between a set of N_g ground truth instances and the set of M activity detections is obtained using the Hungarian matching algorithm as shown in [5, 49]. The matched ground truth label of the m^{th} detection is given as $\sigma(m)$. If a detection does not match any ground instance then it is assigned the background class i.e. $\sigma(m) = \emptyset$.

Visual-Semantic Contrastive Loss. This is the classification loss associated with the semantic-alignment head and is modeled as follows,

$$\mathcal{L}_{cls} = -\frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{\mathrm{C}} \mathbf{1}^c \log \frac{exp(\hat{\mathbf{w}}_\mathrm{c}^T \hat{\mathbf{z}}_{S,m} / \tau_{cls})}{\sum_{\mathrm{c}=1}^{\mathrm{C}} exp(\hat{\mathbf{w}}_\mathrm{c}^T \hat{\mathbf{z}}_{S,m} / \tau_{cls})} \tag{3}$$

where $\mathbf{1}^c$ is the one-hot vector corresponding to the c^{th} class. Since the classification score is obtained by contrasting the visual and semantic information of the seen classes, by minimizing $\mathcal{L}_{cls}$, TranZAD effectively learns to associate the visual features of a seen activity with it's corresponding semantic concept. This visual-semantic consistency enables TranZAD to detect previously unseen activities that are semantically-related with the seen ones.

Localization Loss. The temporal localization loss of the boundary head is the following regression loss,

$$\mathcal{L}_{loc} = \frac{1}{N_g} \sum_{m, \sigma(m) \neq \emptyset}^{M} \lambda_a \cdot L_{tbox}(\mathbf{t}_{\sigma(m)}, \hat{\mathbf{t}}_m) + \lambda_b \cdot L_{gtIoU}(\mathbf{t}_{\sigma(m)}, \hat{\mathbf{t}}_m) \tag{4}$$

where, $\hat{\mathbf{t}}_m$ and $\mathbf{t}_{\sigma(m)}$ are the detected and matched ground-truth temporal coordinates respectively, L_{tbox} and L_{gtIoU} are the l_1 and generalized temporal IoU losses of [49].

Visual Consistency Loss. While the visual-semantic contrastive loss brings consistency between the visual and semantic concepts of each activity, it is also necessary to ensure that the distribution of the visual features for each activity remains temporally coherent. This means that for each activity, its visual features should remain consistent across different temporal segments of the same video as well as across different videos. We accomplish this by leveraging supervised contrastive learning [26] on the intermediate visual features $\mathbf{Z}_D$ generated by the transformer decoder. Therefore for each positive pair of detected features $(\mathbf{z}_{D,i}, \mathbf{z}^+_{D,j})$ the consistency loss is as follows,

$$l(\mathbf{z}_{D,i}, \mathbf{z}^+_{D,j}) = -\log \frac{exp(\frac{\hat{\mathbf{z}}_i^T \hat{\mathbf{z}}_j^+}{\tau_{con}})}{\sum_{\substack{k=1 \\ \sigma(k)\neq\emptyset}}^{M_{p_i}} exp(\frac{\hat{\mathbf{z}}_i^T \hat{\mathbf{z}}_k^+}{\tau_{con}}) + \sum_{\substack{k=1 \\ \sigma(k)\neq\emptyset}}^{M_{n_i}} exp(\frac{\hat{\mathbf{z}}_i^T \hat{\mathbf{z}}_k^-}{\tau_{con}})} \quad (5)$$

where, a pair is considered positive if their matched ground-truth classes are the same i.e. $y_{\sigma(i)} = y_{\sigma(j)}$, $\hat{\mathbf{z}}_i$ is the l_2 normalized feature of $\mathbf{z}_{D,i}$, M_{p_i} and M_{n_i} are the number of positive and negative pairs w.r.t. i, and τ_{con} is a fixed temperature parameter as used in [26]. The total visual consistency loss over all pairs is formulated as follows,

$$\mathcal{L}_{con} = \sum_{i=1,\sigma(i)\neq\emptyset}^{M} \frac{1}{M_{p_i}} \sum_{j=1,\sigma(j)\neq\emptyset}^{M_{p_i}} l(\mathbf{z}_{D,i}, \mathbf{z}^+_{D,j}) \quad (6)$$

Minimizing $\mathcal{L}_{con}$ enforces the transformer to focus on the discriminative aspects of each activity and ignore background nuances resulting in consistent visual features for each activity across different videos. It must be noted that the background visual features are excluded from the computation of $\mathcal{L}_{con}$. This is because the number of predictions matched to the ground truth classes is sparser than background predictions, and so using the background class features leads to an overwhelming influx of irrelevant information to $\mathcal{L}_{con}$, causing distortion in the distribution of the visual feature space.

3.5. Training and Inference

Training. During training only the seen class visual and semantic information is available in $\mathcal{X}_{c_s}$ and $\mathbf{W}_{c_s}$ respectively. For each $\mathbf{x}_i$ in $\mathcal{X}_s$ the detections are obtained in a sliding window manner, where each temporal segment and $\mathbf{W}_{c_s}$ are forward passed to the model and the losses described above are computed. The entire framework is trained end-to-end by backpropagating over the following loss,

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda_{con} \cdot \mathcal{L}_{con} \quad (7)$$

where, $\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{loc}$ and λ_{con} is a hyperparameter that controls the trade-off between the detection and the visual-consistency loss. Since $\mathbf{w}_\emptyset$ dynamically updates itself as a part of this end-to-end learning, it is able to incorporate both visual and semantic information to model a more generalizable background embedding for ZSTAD.

Inference. During the testing phase, the seen and unseen class activity segments are detected separately for each video in $\mathcal{X}_{c_u}$. In either case, prior to the computation of the classification score, $\mathbf{w}_\emptyset$ is concatenated with $\mathbf{W}_{c_s}$ or $\mathbf{W}_{c_u}$ to obtain the background synchronized seen and unseen class embedding matrices given as $\mathbf{W}_{\mathrm{C}}$ and $\mathbf{W}_{\mathrm{C}'}$, respectively (Fig. 2). The predicted unseen activity segments are used to evaluate the zero-shot detection performance of TranZAD.

Table 1: Zero-shot temporal activity detection performance on 8 unseen classes of Thumos'14 in terms of mAP (%) at different tIoU thresholds. The best results for each backbone are highlighted in bold.

		tIoU				
	Backbone	0.1	0.2	0.3	0.4	0.5
RC3D+SE	C3D	13.96	12.61	10.81	7.91	5.11
RC3D+CONSE	C3D	14.16	12.54	10.93	8.02	5.29
ZS_RC3D	C3D	21.34	16.98	15.01	11.12	9.15
TranZAD-G	C3D	21.59	**20.61**	19.14	**16.37**	**12.84**
TranZAD-W	C3D	**22.27**	20.58	**19.40**	15.93	12.36
TranZAD-G	I3D	**24.33**	**22.51**	**20.04**	**17.69**	**14.17**
TranZAD-W	I3D	23.31	21.54	19.48	17.21	13.84

4. Experiments

4.1. Experimental Setup

Baselines. We compare the performance of TranZAD with the modified RC3D [57] framework of Zhang *et al.* [62], which is currently the only work that addresses ZSTAD. We refer to this baseline as ZS-RC3D and compare with two additional baselines designed by [62] called RC3D-ConSE and RC3D-SE, which combine the vanilla RC3D framework with ZSR methods of [35] and [58] respectively.

Datasets. We perform experiments on two popular activity detection datasets Thumos'14 [22] and Charades [47], which have also been used for ZSTAD [62]. For a fair comparison the class and train-test splits are kept the same as [62] for both datasets.

- **Thumos'14**: This dataset has temporal annotations for 20 activity classes and 200 validation, and 213 test videos. Following [62], 12 activities are selected as seen classes and 8 are selected as unseen, with the 200 untrimmed validation videos being used for training, and the 213 test videos being used for testing.

- **Charades**: This dataset comprises of 9848 videos of 157 daily indoor activities collected using Amazon Mechanical Trunk. Following [62], we consider 120 activities as seen classes and the remaining 37 as

Table 2: Thumos'14 per-unseen class AP (%) at tIoU $= 0.5$. The best results for each backbone are higlighted in bold.

	Backbone	**Baseball Pitch**	**Cricket Bowling**	**Diving**	**Hammer Throw**	**Long Jump**	**Shotput**	**Soccer Penalty**	**Tennis Swing**
R-C3D+SE [62]	C3D	2.23	3.09	3.13	9.21	12.15	3.42	3.38	4.29
R-C3D+ConSE [62]	C3D	2.21	3.07	3.23	9.53	12.54	3.56	3.46	4.72
ZS-RC3D [62]	C3D	4.34	4.87	5.03	18.12	20.78	7.06	**6.03**	6.93
TranZAD-W	C3D	**5.16**	6.35	14.23	**21.49**	**27.55**	11.73	5.09	7.30
TranZAD-G	C3D	5.08	**8.58**	**15.03**	21.26	27.41	**12.89**	5.14	**7.33**
TranZAD-W	I3D	5.39	**8.10**	**15.91**	**18.60**	32.65	17.06	5.28	7.73
TranZAD-G	I3D	**6.44**	7.79	15.26	18.28	**34.59**	**17.54**	5.64	**7.82**

Table 3: Zero-shot detection performance on 37 unseen classes of Charades in terms of mAP (%) of [46]. Overall best results are highlighted in bold.

	Backbone	mAP
RC3D+SE	C3D	9.17
RC3D+CONSE	C3D	9.84
ZSRC3D	C3D	**13.23**
TranZAD-G	C3D	13.14
TranZAD-W	C3D	13.05
TranZAD-G	I3D	**13.56**
TranZAD-W	I3D	**13.21**

the unseen classes, with the training and testing set comprised of 7985 videos and 1863 videos, respectively.

Additional details of these datasets and their class-splits are provided in the supplementary material.

Semantic Embeddings. We experiment with both GloVE [38] and Word2Vec [18] embeddings, each with a dimension size of 300. The semantic embedding of each activity class is obtained by averaging the representations of all the words describing that class. Unlike the simple activity captions of Thumos'14, which are tags like 'Basketball Dunk' the activity captions of Charades are gerund phrases, such as 'Someone is Eating' and so we follow [62] and remove some of the prepositions and quantifiers from each caption before obtaining the final embedding. The GloVE model is referred to as TranZAD-G and the Word2Vec one as TranZAD-W.

Implementation. The feature extractor can be any 3D convolutional backbone and we show results using both C3D [50] and I3D [6] features, pretrained on Sports-1M [23] and Kinetics [24], respectively. The sliding temporal window is set to 500 and 250 frames for Thumos'14 and Charades, respectively. The overlap ratio is set to 0.75 during training and 0.5 during testing for both datasets. In the ZSL setting, the training data must not contain any instances from the unseen classes. Therefore, following the same principle as [62], we remove any window segment with activities belonging to the unseen classes in the training videos. The temporal length of the video features l_T is set to 100 for Thumos'14 and 50 for Charades. The number of action queries M is set to 32 for Thumos'14 and 8 for Charades. In all our experiments,

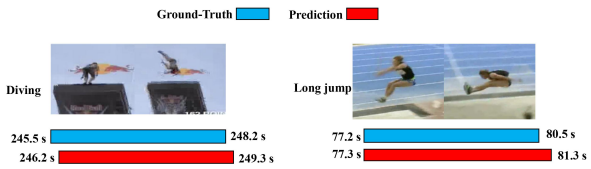

Figure 3: Qualitative results of TranZAD on Thumos'14.

we use fixed *sine* positional encodings [52].

λ_a and λ_b are set to 5 and 2 following [49], λ_{con} is set to 0.01. The learnable scaling parameter τ_{cls} is initialized with 0.1, and the scalar τ_{con} is fixed at 0.05 for all experiments. Training is conducted for 100 epochs using the AdamW [31] optimizer, with a batch size of 64 and a learning rate of 10^{-4} which is dropped by a factor of 10 after 70 epochs. Additional implementation details are listed in the supplementary.

4.2. Comparative Results

Results on Thumos'14: The results on Thumos'14 are shown in Table 1, reported in terms of mean average precision (mAP) at tIoU thresholds [0.1,0.5]. The performance of the baselines is taken directly from their paper [62]. With both C3D and I3D backbones, TranZAD outperforms ZS-RC3D and the other baselines. Specifically for tIoU $= 0.5$, the best results of TranZAD with the C3D features achieve $> 3\%$ increase in mAP and with the I3D features, TranZAD achieves $> 5\%$ increase. The performance of TranZAD-W and TranZAD-G are nearly identical, and the slight improvement of the latter is due to the better representation of the GloVE embeddings over the Word2VEC ones. Table 2 shows the per unseen class average performance (AP) at tIoU $= 0.5$, and it can be observed that TranZAD outperforms ZS-RC3D on the majority of the unseen classes. This shows that our anchor-free learning with transformers and an adaptive background embedding helps to overcome the challenges of a two-stage detector like ZS-RC3D, resulting in more true positive detections. Visualizations of some qualitative results are shown in Fig. 3.

Results on Charades: For Charades, as per common practice to performance is computed in terms of Sigurdsson *et al.*'s [46] mAP metric and is reported in 3. Overall, TranZAD achieves comparable performance using both C3D and I3D features. On average, there is about 79% temporal overlap

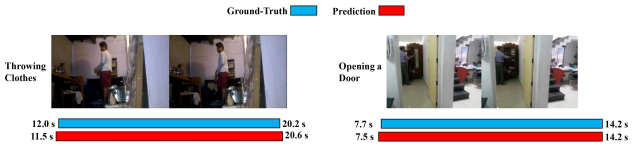

Figure 4: Qualitative results of TranZAD on Charades.

Table 4: Comparison of our framework with ZS-RC3D w/o super-class classification loss. The 3D backbone is C3D for both methods. For Thumos'14 the metric is mAP@tIoU=0.5 and for Charades its the mAP of [46].

	ZS-RC3D-L_{sc}	TranZAD-W
Thumos'14	8.25	**12.36**
Charades	11.72	**13.05**

between activities in the Charades dataset compared to 8% in Thumos'14, which makes Charades a much more challenging dataset for ZSTAD. Also, the number of pertinent activities in each video of Charades is much sparse compared to Thumos'14, so we use fewer action queries for Charades. Qualitative results are shown in Fig. 4. It must be noted that the performance of ZS-RC3D receives a significant boost by using unseen class semantic context during training, which is only possible if the model knows which classes are unseen beforehand. We analyze this as follows.

Removing the need for unseen class context. A significant boost in the performance of ZS-RC3D [62] can be attributed to using context information about the unseen classes during training. This means that ZS-RC3D is aware of which classes are unseen apriori, which is not practical. Specifically, ZS-RC3D utilizes the unseen class semantic embeddings to compute a background embedding as well as super-classes the latter being used as a supervisory signal for a max-margin classification loss L_{sc}. As shown in Table 4 the performance of ZS-RC3D drops when L_{sc} is not used (even though it's $\mathbf{w}_{\phi}$ is obtained using seen+unseen class embeddings), with the drop being most significant for the Charades dataset. In contrast, our TranZAD framework does not use any information from the unseen classes during training (visual or semantic). Despite that it achieves better or comparable performance to ZS-RC3D and is more applicable for practical scenarios.

4.3. Ablation Studies

Impact of learned background embedding. We study the effectiveness of our learned $\mathbf{w}_{\emptyset}$ by comparing with the strategy of Zhang *et al.* [62] for obtaining $\mathbf{w}_{\emptyset}$. Therefore, we remove $\mathbf{w}_{\emptyset}$ as a learnable network parameter and assign it a fixed representation obtained following [62]. As seen in the third row of Table 5 and 6, the performance of TranZAD degrades when we assign $\mathbf{w}_{\emptyset}$ the fixed representation of [62]. This shows that it is more important to learn $\mathbf{w}_{\emptyset}$, so that it incorporates both semantic and visual cues to model more diverse background information in videos.

Table 5: **Effectiveness of learned background embedding and visual consistency loss on Thumos'14**. Results are with I3D features in terms of mAP@tIoU $=0.5$.

$\mathcal{L}_{con}$	learned $\mathbf{w}_{\emptyset}$	TranZAD-G	TranZAD-W
×	×	12.58	11.12
×	✓	12.79	13.18
✓	×	12.76	12.89
✓	✓	**14.17**	**13.84**

Table 6: **Effectiveness of learned background embedding and visual consistency loss on Charades**. Results are reported with I3D features in terms of mAP of [46].

$\mathcal{L}_{con}$	learned $\mathbf{w}_{\emptyset}$	TranZAD-G	TranZAD-W
×	×	12.09	12.01
×	✓	12.17	12.52
✓	×	12.31	12.24
✓	✓	**13.56**	**13.21**

Impact of visual consistency loss. As shown in Table 5, the addition of $\mathcal{L}_{con}$ enables increased consistency of the visual features of each activity across different videos. This consequently boosts the zero-shot detection performance of TranZAD by effectively enhancing the discriminability of the visual features. From the second row of Table 5 and 6, it can be observed that, for both Word2Vec and GLoVE embeddings, the performance declines when $\mathcal{L}_{con}$ is not included i.e. $\lambda_{con}=0$. A sensitivity analysis of λ_{con} is given in the supplementary.

5. Conclusion

In this paper, we introduce a transformer-based setup called TranZAD to address the challenging task of ZSTAD. TranZAD streamlines the detection of unseen activities by performing direct set-based prediction, removing hand-crafted designs, and consequently achieving faster inference. We show how the visual and semantic information of *only* the seen classes can be used to train TranZAD via a contrastive learning strategy enabling improved reasoning of previously unseen activities. We also propose an adaptive approach for modeling the background class embedding enabling greater distinguishability of unseen classes from the background information in videos. Experimental analysis on Thumos'14 and Charades establishes TranZAD as a new baseline for the modeling and search of semantic relationships in videos under data and computational scarcity.

Acknowledgement. This paper is partially supported by ONR grants N00014-19-1-2264 and N00014-18-1-2252.

References

[1] https://socialblade.com/youtube/.

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[3] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association, 2019.

[4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Xiaojun Chang, Yi Yang, Alexander Hauptmann, Eric P Xing, and Yao-Liang Yu. Semantic concept discovery for large-scale zero-shot event detection. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[8] Xiaojun Chang, Yi Yang, Guodong Long, Chengqi Zhang, and Alexander Hauptmann. Dynamic concept composition for zero-example event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[9] Xiaojun Chang, Yao-Liang Yu, Yi Yang, and Eric P Xing. Semantic pooling for complex event analysis in untrimmed videos. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1617–1632, 2016.

[10] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5327–5336, 2016.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[12] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[13] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE international conference on computer vision*, pages 3628–3636, 2017.

[16] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*, 2017.

[17] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[18] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1914–1923, 2016.

[21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[22] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[24] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[25] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[27] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3174–3183, 2017.

[28] Zhihui Li, Lina Yao, Xiaojun Chang, Kun Zhan, Jiande Sun, and Huaxiang Zhang. Zero-shot event detection via event-adaptive concept relevance mining. *Pattern Recognition*, 88:595–603, 2019.

[29] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action

proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019.

[30] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2188–2196, 2018.

[33] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *ECCV*, 2016.

[34] Li Niu, Jianfei Cai, Ashok Veeraraghavan, and Liqing Zhang. Zero-shot learning via category-specific visual-semantic mapping and label refinement. *IEEE Transactions on Image Processing*, 28(2):965–979, 2018.

[35] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[36] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[37] Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Text-based temporal localization of novel events. 2022.

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[39] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2833–2842, 2017.

[40] Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11932–11939, 2020.

[41] Shafin Rahman, Salman Khan, and Fatih Porikli. A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. *IEEE Transactions on Image Processing*, 27(11):5652–5667, 2018.

[42] Shafin Rahman, Salman H Khan, and Fatih Porikli. Zero-shot object detection: Joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128(12):2979–2999, 2020.

[43] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] Dripta S. Raychaudhuri and Amit K. Roy-Chowdhury. Exploiting temporal coherence for self-supervised one-shot video re-identification. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII*, page 258–274, Berlin, Heidelberg, 2020. Springer-Verlag.

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[46] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017.

[47] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[48] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021.

[49] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13526–13535, 2021.

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[51] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Vinay Kumar Verma, Dhanajit Brahma, and Piyush Rai. Meta-learning for generalized zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6062–6069, 2020.

[54] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[55] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[56] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[57] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection.

In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.

[58] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015.

[59] Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-guided contrastive network for zero-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[60] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7103, 2019.

[61] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning. *CoRR*, abs/2103.11731, 2021.

[62] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander Hauptmann. Zstad: Zero-shot temporal activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2020.

[63] Lingling Zhang, Jun Liu, Minnan Luo, Xiaojun Chang, and Qinghua Zheng. Deep semisupervised zero-shot learning with maximum mean discrepancy. *Neural Computation*, 30(5):1426–1447, 2018.

[64] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015.

[65] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.

[66] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang. Gtnet: Generative transfer network for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12967–12974, 2020.

[67] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[68] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020.

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.