Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241075

HiFormer: Hierarchical transformer for grounded situation recognition

Yulin Pan

Huitong School, Shenzhen, China

panbrian2077@gmail.com

Abstract. The prevalence of monitoring video is critical to public safety, but existing Object Detection and Action Recognition models are overwhelmed by camera operators, unable to identify relevant events. In light of this, Grounding Situation Recognition (GSR) provides a practical solution to recognize the events in a surveillance video. That is, GSR can identify the noun entities (e.g., humans) and their actions (e.g., driving), and provide grounding frames for involved entities. **Compared with Action Recognition and Object Detection, GSR is more in line with human cognitive habits, better allowing law enforcement agencies to understand the predictions. However, the crucial issue with most existing frameworks is the neglect of verb ambiguity, that is, superficially similar verbs but have distinct meanings (e.g. buying v.s. giving)**. Many existing works propose a two-stage model, which first blindly predicts the verb, and then uses this verb information to predict semantic roles. These frameworks ignore the importance of noun information during verb prediction, making them susceptible to misidentifications. **To address this problem and better discern between ambiguous verbs, we propose HiFormer, a novel hierarchical transformer framework with direct and comprehensive consideration of similar verbs for each image, to more accurately identify the salient verb, semantic roles, and the grounding frames. Compared with the state-of-the-art models in Grounded Situation Recognition (SituFormer and CoFormer), HiFormer shows an advantage of over 35% and 20% on the Top-1 and Top-5 verb accuracy respectively, as well as 13% on the Top-1 Noun accuracy.**

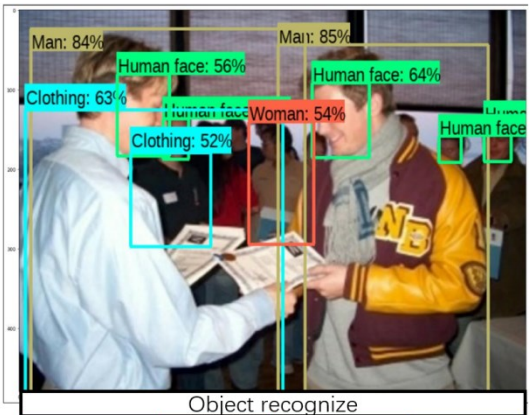
Keywords: Grounded Situation Recognition, Transformer, Deep Learning.

1. Introduction

The rapidly developing surveillance technology is significantly improving our lives. At the end of 2019, the number of surveillance cameras in the world exceeded 770 million [1], which is approximated to have a market size of 69.1 billion dollars by 2026 [2]. Despite the high coverage of surveillance cameras, human activity detections still rampage in places with a less sufficient police force, even directly under high-resolution cameras. Now, as the field of Artificial Intelligence and Computer Vision develops, new solutions such as Biometric Identification [3,4,5], Object Detection/Tracking [6,7,8], Crowd Density Analysis [9,10,11,12,13], and Action Recognition [14,15,16,17,18,19] have come to light, which in theory could automatically detect objects or actions. *However, this seemingly useful technology remains confined to the laboratories as a consequence of deficiencies: 1) The narrow scope of the detection process (e.g., action-only or object-only) cripples the model accuracy, for it requires a comprehensive*

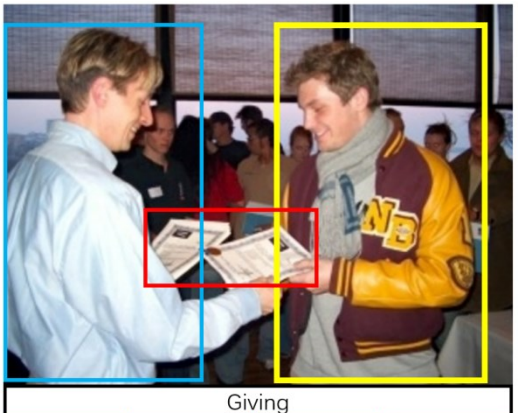
DOI: 10.54254/2755-2721/49/20241075

consideration of multiple factors to determine the nature of a situation. 2) Although these models could determine what action or object is in the image, they could not recognize the details or causes of the entire event. As shown in Figure 1, the Object Detection model monotonously determines the probability for each noun entity but makes no effort to recognize the action taking place.



Object recognize

Human face	Woman	Clothing	Man
64%	54%	63%	85%
56%		52%	84%
51%			
45%			



Giving

Agent	Goods	Giver	Place
man	certificate	man	classroom

Figure 1. Comparison between Object Detection (left) and Grounded Situation Recognition (right).

To tackle the existing problems, Grounded Situation Recognition (GSR) [20], first proposed by Pratt et al., aims to recognize images following the human cognitive pattern. Unlike the Object Detection process shown in Figure 1, which mechanically predicts the likelihood of each noun entity, the task of GSR is to understand the given image from an event-based perspective, which is to identify the involved noun entities, their mutual interactions as well as the relative and absolute locations of the noun entities. In the case of Figure 1, GSR recognizes not only the two men and the certificate, but also the apparent action of giving, as well as the classroom where the action is taking place. This comprehensive image analysis makes GSR models more accurate, significantly reducing misidentifications. Furthermore, in addition to the traditional Situation Recognition, GSR also makes a grounding frame prediction for all relevant noun entities based on the semantic information of the salient action, the noun entities, and their mutual relations. This provides the model with locations of noun entities within the image, thereby benefiting many downstream tasks such as multimedia understanding [21,22,23] and information retrieval [24,25,26]. In summary, Grounded Situation Recognition aims to predict the salient verb, noun entities, and grounded frames when analyzing an image by utilizing the crisscrossing semantic relations.

DOI: 10.54254/2755-2721/49/20241075

Figure 2. Instance of Verb Ambiguity.

However, almost all existing works in this field neglect the existence of verb ambiguity (similar-looking verbs with different meanings). For instance, "providing," "buying," and "giving" have drastically different meanings despite looking quite similar (as shown in Figure 2). Most works in this field use identical two-stage frameworks, which first predict the verb, and then use this information to predict the nouns. With the complete ignorance of noun information during verb prediction, these models could not discern between many similar verbs, whose only difference is with their semantic roles.

Recent works (such as SituFormer [27] and GSRFormer [28]) try to tackle both of the above problems by proposing a three-stage transformer framework. On the firsthand, they try to alternately update verbs and semantic roles by 1) predicting the verb in stage 1; 2) predicting semantic roles in stage 2 based on the verb; 3) refining the verb features in the third stage with the newly acquired semantic roles. On the other hand, when processing an image with the salient verb found, SituFormer tries to consider the other highly similar verbs. However, the approaches to both problems are still largely ineffective: 1) the interconnecting semantic relations between verbs and nouns remain underexploited, for the refinement process lacks repetition; 2) the problem of verb ambiguity still remains unsolved, for many of the computed similar verbs could still have drastically different meanings due to their different semantic roles. Therefore, without considering these semantic roles, the actually similar verbs remain neglected. Furthermore, they have overcomplicated this task by adding redundant and inherently ineffective modules.

To address this problem, we propose HiFormer, a novel transformer-based GSR model. HiFormer takes advantage of its hierarchical internal structure and directly considers all similarities during the decision-making process, significantly reducing the chance of misidentifying, and leading to more reliable and accurate predictions for GSR. We uniquely contribute to the task of Grounded Situation Recognition in the following three aspects:

1. We unravel the problems of previous works, pointing out their neglect of verb ambiguity. The lack of macroscopic analysis regarding the similarity between verbs causes their accuracy to re-main gloomy. Despite repeated attempts at improvement, their current performances are still pessimistic, for they oversimplify the retrieval process of similar verbs and ignore the significance of semantic roles.

2. We propose a novel hierarchical transformer framework, taking advantage of its internal hierarchy to explicitly consider all similarities for each image. This approach effectively reduces the misidentifications caused by verb ambiguity, directly confronting the most crucial issue of the GSR task.

3. We achieve stage-of-the-art performance accuracy on the challenging SWiG benchmark, far surpassing all the previous works by over *35%, 21%* on the Top-1, Top- 5 verb accuracy, and over *13%* on the Top-1 noun accuracy.

Our proposed HiFormer can be applied to the surveillance camera network to alert the local authorities and medical centers in the event of criminal activity or risky behavior. Due to the rapid

DOI: 10.54254/2755-2721/49/20241075

improvement in the economy and the development of technology both in developed and developing countries, surveillance technology will become increasingly widespread. Therefore, our Hiformer could deter and drastically reduce criminal activity, and in the end, improve the life quality of the general public as a whole.

2. Related Work

This section will briefly introduce some previous essential works in Transformers, Action Recogni-tion, and Grounded Situation Recognition.

2.1. Progression of Action Recognition

According to an analysis conducted by Gella et al. [29], Action Recognition is the task of recognizing the activity taking place in pictures and videos. Most existing works can be classified into four categories: 1) Action Classification; 2) Human-Object Interaction detection; 3) Visual Verb Sense Disambiguation; 4) Visual Semantic Role Labeling. In the beginning, Action Recognition took the form of Action Classification based on small-scale data sets [30,31,32,33], which laid the foundation numerous later works. However, this kind of classification is problematic in two folds. Firstly, the methods could not be extended to large-scale data sets. Secondly, these works all assume a singular verb label for each image, ignoring the fact that multiple activities could take place in the images simultaneously. In response to these problems, Human-Object Interaction detection was proposed [34], which solved both of the above problems. However, more deficiencies have come to light regarding HOI detection. Firstly, it neglects the existence of multiple meanings of the same verb. For example, the verb “take” means accepting, acquiring, and carrying. Secondly, it ignores the importance of noun information within the verb prediction process, which is often the only difference between similar verbs such as “riding a bicycle” or “riding a horse”. These observations have led to many arguments regarding how actions should be analyzed on the level of verb senses. Later, Gella et al. [36] proposed a new task of visual Verb Sense Disambiguation, where each image is annotated with verb sense labels. However, although this task handles the ambiguity of verbs, it neither identifies nor localizes the noun entities within the images. Some recent works [37,38] solve this deficiency by proposing Visual Semantic Role Labeling, which not only predicts and identifies the semantic roles but also provides grounding frames to localize these roles.

2.2. Transformer in Grounded Situation Recognition

The Transformer Framework [39] was first proposed by Vaswani et al. to tackle problems in Natural Language processing. Its built-in attention mechanism allows it to easily model the long-range dependencies between words and phrases without laboriously stacking up multiple layers. It is significantly more efficient than conventional Convolutional Neural Networks. Furthermore, due to the inherent structure of the attention block, transformers are much more parallelizable compared to Recurrent Neural Networks. Later, this model became widely used and modified in Natural Language Processing. Modifications on the transformer framework such as Lightweight variants [40,41,42], recurrent transformers [43,44] and hierarchical transformers [45,46,47], etc. According to a survey by Khan et al. [48] many works have been trying to implement this successful model in Computer Vision. Its minimal need for inductive biases and ability to tackle long-range dependencies makes it much more suitable for the task than conventional convolutional neural networks. Furthermore, the robust design of the transformer model makes it competent in various sub-areas, such as video, image, and audio, without any laborious modifications. Therefore, the transformer model is becoming increasingly popular in Computer Vision.

DOI: 10.54254/2755-2721/49/20241075

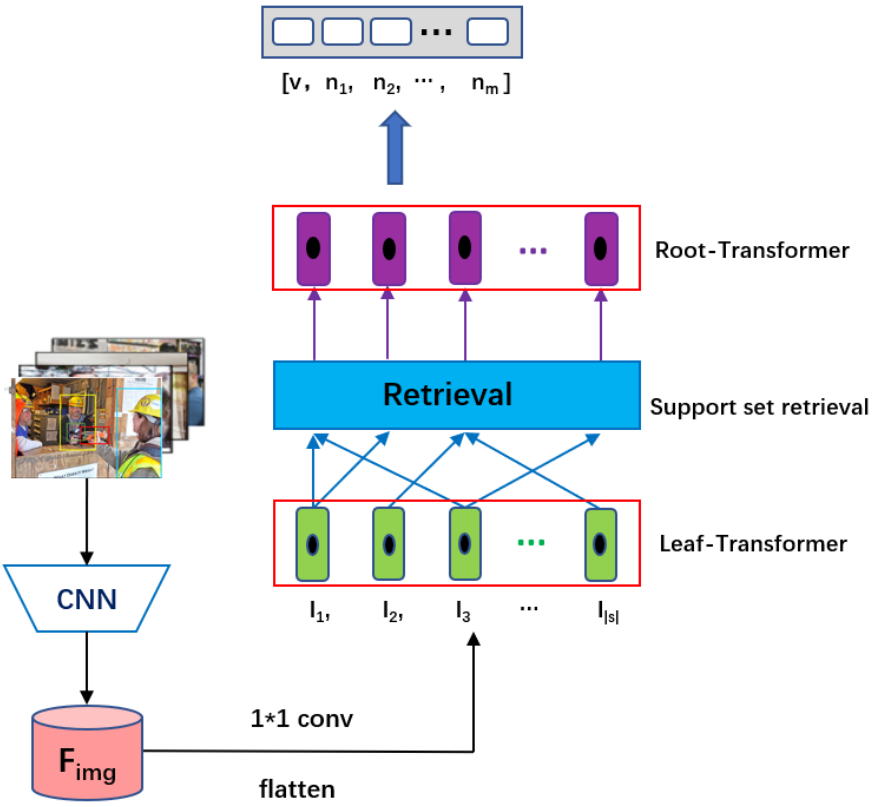


Figure 3. Overall Architecture of Hierarchical Transformer.

In the Specific field of Grounded Situation Recognition, the Transformer model was first proposed by Cho et al. in GSRTR [49], by replacing the object-centric queries in DETR [50] with semantic role queries. More recently, Wei et al. proposed SituFormer [27], a two-stage model that predicts the verb and nouns separately using two transformer-based detectors, to improve the performance.

3. Hierarchical Transformer Framework

This section will further elaborate on the Hierarchical Transformer Framework (HiFormer).

3.1. Overview of HiFormer

To address the above problems, we reshape the transformer framework in Grounded Situation Recognition by proposing a renewed learning framework named HiFormer. As shown in Figure 3, HiFormer is a two-stage transformer-based model that directly tackles the problem of verb ambiguity. It computes and considers the similar images for each training set image and thoroughly exploits the semantic verb- noun relations. In the first stage, the Leaf Transformer (TRM_{leaf}) learns the preliminary representation for each image. Then, the support image set for each image is computed during the KNN retrieval pro- cess, which serves as a transition to the Root Transformer. Finally, the Root Transformer utilizes the support image sets to refine the representations of each image.

Formally, HiFormer can be represented by Eq. 1-3 as,

$$\mathrm{TRM}_{\mathrm{leaf}}\left(R^{(0)}, P^{(0)} \mid I\right) \tag{1}$$

$$\mathrm{KNN}\left(\left\{R^{(0)}, P^{(0)}\right\}_{\mathrm{sim}} \mid \left\{R^{(0)}, P^{(0)}\right\}_{\mathrm{all}}\right) \tag{2}$$

$$\mathrm{TRM}_{\mathrm{root}}\left(\left\{R^{(t)}, P^{(t)}\right\}_{\mathrm{sim}} \mid R^{(t+1)}, P^{(t+1)}\right) \tag{3}$$

DOI: 10.54254/2755-2721/49/20241075

where the three equations respectively denote the working procedures of the Leaf Transformer, KNN retrieval process, and Root Transformer. Firstly, the Leaf Transformer learns the role tokens $R^{(0)}$ and the preliminary representations $P^{(0)}$ of each image, using the image features I extracted by the CNN backbone. Afterwards, we compute a support image set $\{R^{(0)}, P^{(0)}\}$ sim of size k for each image in the retrieval process. Finally, the Root Transformer refines the preliminary representations of the original images with their support verb sets.

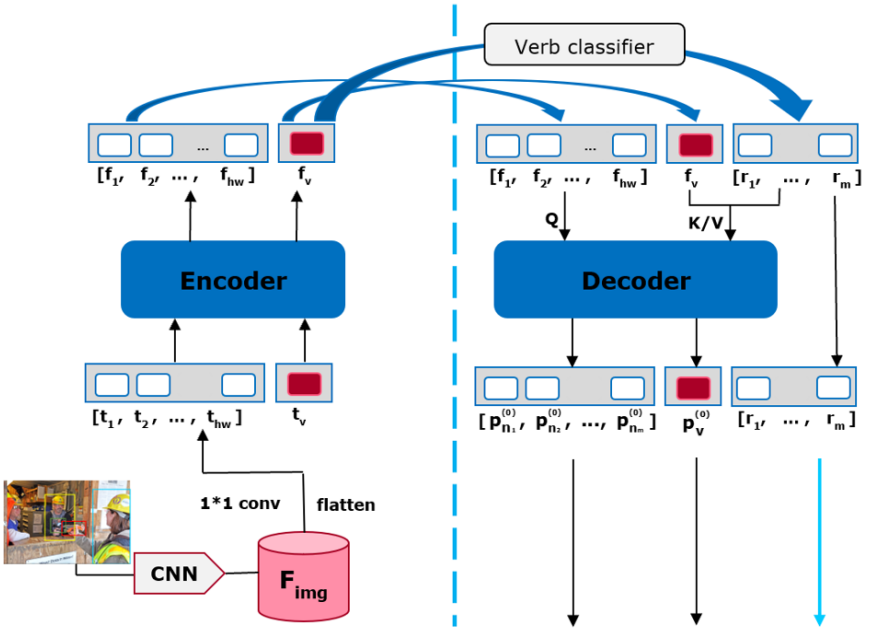


Figure 4. Leaf Transformer.

3.2. Leaf Transformer

Leaf Transformer is trained to independently learn the preliminary representations of salient verbs and their corresponding semantic roles, consisting of an encoder and a decoder.

As shown in Figure 4, the Leaf Transformer consists of an Encoder and a Decoder. The Leaf Transformer Encoder is designed to learn the representation of the salient verb and predict the preliminary verb category. At the same time, the Leaf Transformer Decoder is devised to learn the corresponding semantic role representation based on the salient verb.

3.2.1. Representation of Salient Verb

As shown in Figure 4, the CNN backbone first extracts a feature map $F_{img} \in \mathbb{R}^{c\times h\times w}$ from the im-age, with is transformed into a sequence of image features $[f_1, f_2, \ldots, f_{h-w}]$ by a 1×1 convolutional layer and a flattening operator, where each element $f_i \in \mathbb{R}^d$ represents the features of a single pixel. Inspired by ViT [51], we initialize a vector of learnable verb tokens $t_v \in \mathbb{R}^d$ to represent the salient verb. Then, the verb tokens and image features are encoded with positional embedding E_{pos} to take positional information into account. Finally, the encoded image features are fed to the Leaf Transformer Encoder, equipped with a multi-head self-attention module. The detailed working procedure of the Leaf Transformer Encoder stands as below.

$$[e_v, e_1, e_2, \ldots, e_{h\times w}] = \text{MSA}\underbrace{\left([t_v, f_1, f_2, \ldots, f_{h\times w}] \oplus E_{\text{pos}}\right)}_{\text{query/key/value}} \tag{4}$$

The output can be divided in to two parts: 1) verb embedding $e_v \in \mathbb{R}^{1\times d}$; 2) image embedding $e_{1\ldots hw} \in \mathbb{R}^{hw\times d}$. Which will later serve as input for the decoder.

DOI: 10.54254/2755-2721/49/20241075

3.2.2. Representation of Semantic Roles

Before entering the decoder, a verb classifier determines the preliminary verb category v based on the verb embedding acquired by the encoder, from which we fetch the corresponding semantic roles and initialize them to a role embedding vector $[r_1, \ldots, r_m]$. Where m is the number of roles, and each element $r_i \in \mathbb{R}^d$ represents the embedding for a single role.

After acquiring the verb, semantic role, and image embedding, we feed them to a transformer decoder module to further learn the preliminary representations of the verb p_v and semantic roles $[p_{n_1}, \ldots, p_{n_m}]$ as shown below:

$$P_v = [p_v, p_{n_1}, \ldots, p_{n_m}] = \text{MHA}(\underbrace{[e_{1\ldots hw}] \oplus E_{pos}}_{\text{query}}, \underbrace{[e_v, r_1, \ldots, r_m]}_{\text{key/value}}), \tag{5}$$

where $p_v, p_{n_1}, \ldots, p_{n_m}$ are all vectors of real numbers with size $1 \times d$.

Note that the Leaf Transformer Decoder is equipped with a multi-head cross-attention block, where the image embedding with position encoding serves as the query for the Attention Mechanism, and the concatenated vector of verb features and role features serves as both the key and value.

3.3. Retrieval Process

In this stage, we compute the support image set for each image (as shown in Figure 5), which consists of k image with the highest cosine similarity to the said image.

3.3.1. Computation of the Support Image Set

Before the actual computation begins, we split the Training/Validation/Test datasets into multiple segments of around 10000 images to guarantee the program efficiency under the high complexity of the pairwise similarity calculation. Next, for each segment in the dataset, we calculate the pairwise cosine similarity of all the images within using a brute-force method with $\mathcal{O}(n^2 \cdot d)$ time complexity. Finally, for each image, we find k images in the segment with the highest cosine similarity to it, and then save this information in a hash-table form.

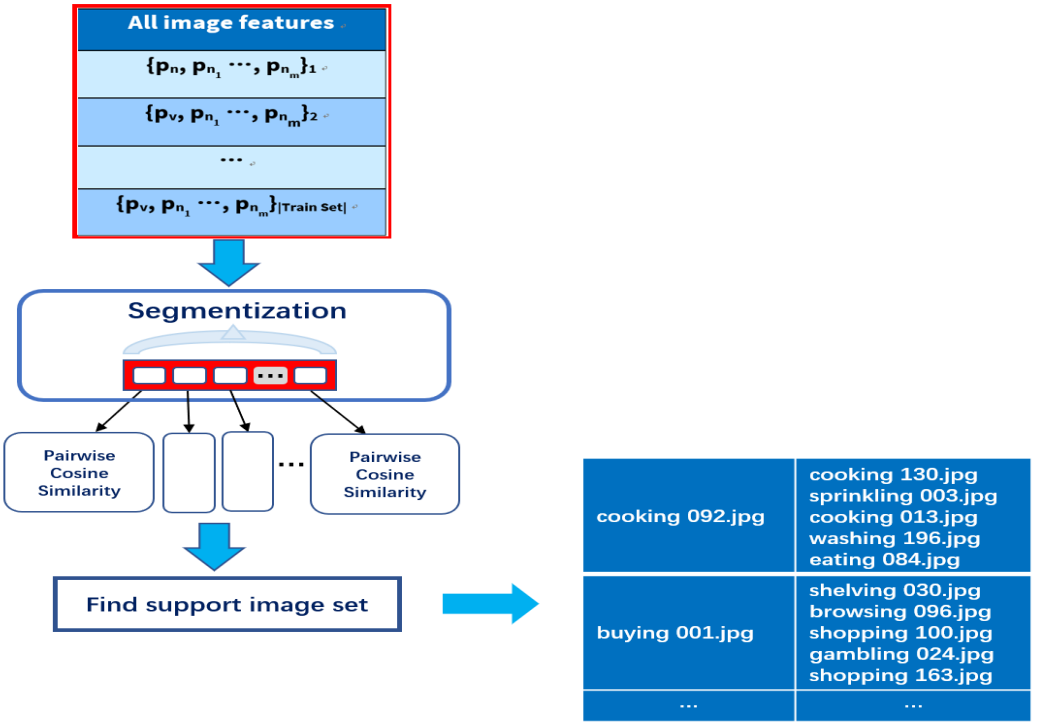


Figure 5. Retrieval Process.

DOI: 10.54254/2755-2721/49/20241075

3.3.2. *Retrieval of the Support Image Set*

Using the verb and role representations computed by the Leaf Transformer, we retrieve a support set $\{P\}_{\text{sim}}$ of k images with the highest similarity to the current image in order to consider all the possibilities and prevent misidentification of similar verbs.

The Retrieval process is as follows:

$$\{P\}_{sim} = \underset{P_{I_j}}{\text{argtop} - \text{K}} \in IS\left(P, P_j\right), \tag{6}$$

$$\text{S}\left(P, P_j\right) = \frac{1}{m}\sum_{i=1}^{m} \text{CosineSimilarity}\left(p_{n_i}^{(0)} \in P, p_{n_i}^{(0)} \in P_j\right), \tag{7}$$

where I is the current image segment. Note that in the actual training or evaluation processes, the support image set of each image is already computed. Therefore, the retrieval process could be done in $\mathcal{O}(1)$ time complexity using the preprocessed hash-table.

3.4. *Root Transformer*

In this stage, the Root Transformer is trained to refine the verb and role features in an iterative and alternating way. Before the main procedure begins, we first use the pre-trained CNN backbone, Leaf Transformer, and hash-table to extract the preliminary representations from the raw images. Recognize that the Leaf Transformer is already trained to its full extent in the preliminary stage and does not participate in loss calculation or backward propagation in this stage. Then, the four main steps stand as follows: 1) retrieval of the support image set; 2) computation of neural messages; 3) refinement of semantic role features using previously acquired verb information; 4) refinement of verb features using previously acquired semantic role information. In this stage, the above steps are repeated five times, in which (t) denotes the messages and representations in the t-th iteration.

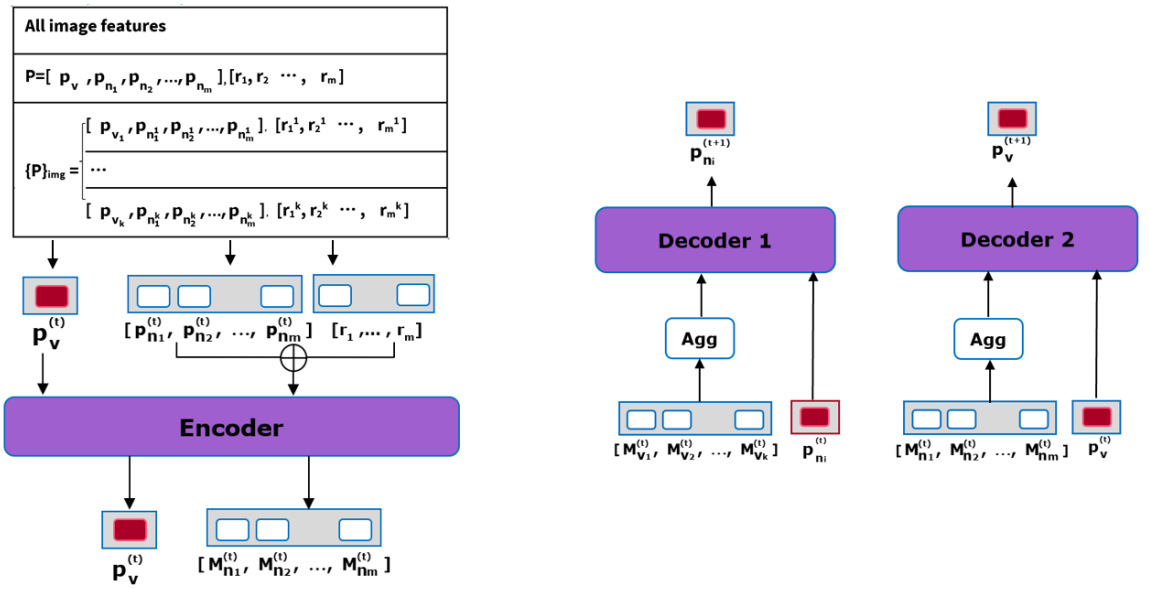


Figure 6. Root Transformer.

3.4.1. *Computation of Neural Messages*

To allow the decoder to simultaneously consider all the images in the support image set, we use the Root Transformer Encoder to compute a compacted semantic message, following the neural message passing mechanism.

$$\left[M_v^{(t)}, M_{n_1}^{(t)}, \ldots, M_{n_m}^{(t)}\right] = \text{MHA}\left(p_v^{(t)}, p_{n_1}^{(t)} \oplus r_1, p_{n_2}^{(t)} \oplus r_2, \ldots n_{n_m}^{(t)} \oplus r_m\right), \tag{8}$$

DOI: 10.54254/2755-2721/49/20241075

In which $M_v^{(t)} \in \mathbb{R}^{1\times d}$ denotes the verb message, and $M_{n_{1\ldots m}}^{(t)} \in \mathbb{R}^{m\times d}$ denotes the m semantic role messages. Unlike the Leaf Transformer Encoder, whose structure is highly similar to those of the Vanilla Transformer [39], the Root Transformer Encoder here consists only of a single multi-head self-attention module.

3.4.2. Refinement of Semantic Role Messages

For each vector of role representations $p_{v\to n_i}^{(t)}$ of the original image, we utilize the semantic relations between it and the $K+1$ verb representation vectors to compute an update message $M_{v_{\text{all}}}^{(t)} \in \mathbb{R}^{1\times d}$. More specifically, we aggregate $M_v^{(t)}, M_{v_1}^{(t)}, M_{v_2}^{(t)}, \ldots, M_{v_K}^{(t)}$ on their second dimension using a Fully Connected Network, which corresponds to the Agg module in Figure 6. The update message and the role embedding vector are then fed to the transformer sub-layer consisting of two Layer Normalization modules with a Feedforward Network in between, which will update the role feature $p_{n_i}^{(t)}$ to $p_{n_i}^{(t+1)}$.

$$M_{v_{all}}^{(t)} = \text{Agg}\left(M_v^{(t)}, M_{v_1}^{(t)}, M_{v_2}^{(t)}, \ldots, M_{v_K}^{(t)}\right) \tag{9}$$

$$p_{n_i}^{(t+1)} = \text{MLP}\left(M_{v_{all}}^{(t)} \oplus p_{n_i}^{(t)}\right) \tag{10}$$

3.4.3. Refinement of Verb Messages

Similarly, the verb representations of the original image are refined by the semantic role representations. We first aggregate $M_{v\to n_{1\ldots m}}^{(t)}$ like the previous section to compute the update message. Then, we feed it to another transformer sub-layer, along with the to-be updated verb features, to conduct the refinement process below.

$$M_{n_{\text{all}}}^{(t)} = \text{Agg}\left(M_{v\to n_1}^{(t)}, M_{v\to n_2}^{(t)}, \ldots, M_{v\to n_1}^{(t)}\right) \tag{11}$$

$$p_v^{(t+1)} = \text{MLP}\left(M_{n_{\text{all}}}^{(t)} \oplus p_v^{(t)}\right) \tag{12}$$

3.5. Training Objectives

3.5.1. Preliminary Stage

In the preliminary stage, we calculate the loss functions for the three main outputs: 1) the preliminary verb category predicted by the verb classifier after the encoder module; 2) the preliminary verb representations produced by the decoder; 3) the preliminary noun representations produced by the decoder, where the first output represents the encoder, and the second and third outputs represent the decoder. Although the preliminary verb category does not participate in the future stages, we still impose a loss function for it since it plays a crucial role in the input of the Leaf Transformer Decoder.

The details of the loss functions are as below:

$$L_{verb_{e1}} = L_{\text{CE}}(v^{gt}, v_1), \tag{13}$$

$$L_{verb_{e2}} = L_{\text{CE}}(v^{gt}, v_2), \tag{14}$$

$$L_{\text{noune}} = \sum_{i=1}^{m}\left[L_{\text{CE}}\left(n_i^{gt}, n_i\right)\right), \tag{15}$$

$$L_{bbox} = L_{\text{CE}}\left(b_i^{gt}, b_i\right), \tag{16}$$

where v_1 is the verb category predicted by the verb classifier between the encoder and decoder in the preliminary stage; $v_2, [n_{1\ldots m}]$ and $[b_{1\ldots m}]$ are the verb, noun, and bounding box predictions respectively, based on the preliminary representations acquired at the end of the preliminary stage. The purpose of the three loss functions are: 1) $L_{verb_{el}}$ helps optimize the preliminary verb category v_1;2)

DOI: 10.54254/2755-2721/49/20241075

$L_{verb_{e2}}$ assist optimization of the verb prediction v_2 based on the preliminary representations of the salient verb. 3) L_{noun} quantifies the loss of noun prediction $[n_{1\ldots m}]$ based on the preliminary representation of semantic roles.

3.5.2. Refinement Stage

Like the Leaf Transformer-decoder, we have to optimize the verb, semantic roles, and bounding boxes via cross-entropy loss functions. Same as the second step of the encoder, the detail of the decoder loss calculation stands as:

$$L_{\text{verb}d} = L_{\text{CE}}(v^{gt}, v), \tag{17}$$

$$L_{\text{noun}d} = \sum_{i=1}^{m} L_{\text{CE}}\left(n_i^{gt}, n_i\right), \tag{18}$$

$$L_{bbox} = L_{\text{CE}}\left(b_i^{gt}, b_i\right), \tag{19}$$

where $v, [n_{1\ldots m}]$ and $[b_{1\ldots m}]$ are the verb, noun, and bounding box predictions based on their corresponding refined representations. The respective purposes of these loss functions are identical to those above.

3.6. Process of Evaluation and Inference

During non-training processes such as evaluation, our framework produces the result straightforwardly. The evaluation process consists of five simple steps: 1) the CNN backbone extracts the image features from the raw images; 2) the Leaf Transformer produces the preliminary verb and semantic role representations of the images; 3) we retrieve the support image sets of the images from the precomputed hash table (note that the support image sets of validation/test image are also precomputed); 4) the Root Transformer refines the verb and semantic role representations; 5) the verb category, noun category, and bounding box predictors use the well-refined representations to produce the final outputs of Grounded Situation Recognition.

In addition, the custom image inference process is highly similar to the evaluation process above. However, since we cannot precompute the support image set for inference images during training, the inference process is slightly different from the evaluation process in the third step. Instead of using the precomputed hash table, we manually find its support image set in the training set by calculating its Cosine Similarity with every image in the set.

4. Experiments

4.1. Dataset

We use the most dominant dataset in Grounded Situation Recognition, the SWiG benchmark, to train and evaluate HiFormer. SWiG builds upon the imSitu dataset while retaining the original images and the frame annotations. SWiG provides additional grounding frames for each image's visible semantic roles. There are 126,102 images, 504 verb classes, and 190 semantic role classes, where each verb is followed by 1 to 6 corresponding semantic roles. For each image, three sets of annotations exist made by different annotators. We split the Training/Validation/Test datasets into sets with sizes of *75* K/*25* K/*25* K, respectively, following the official dataset split.

DOI: 10.54254/2755-2721/49/20241075

Table 1. Performance comparison with state-of-the-art models on the test set.

Dataset	Models	Top-1 Verb					Top-5 Verb					Ground Truth Verb			
		verb	value	val-all	grnd	grnd-all	verb	value	val-all	grnd	grnd-all	value	val-all	grnd	grnd-all
Test	Traditional Situation Recognition models														
	CRF [52]	32.34	24.64	14.19	-	-	58.88	42.76	22.55	–	–	65.66	28.96	-	-
	CRF + DataAug [53]	34.12	26.45	15.51	-	-	62.59	46.88	25.46	–	–	70.44	34.38	-	-
	VGG+RNN [54]	35.90	27.45	16.36	-	-	63.08	46.88	26.06	–	–	70.27	35.25	-	-
	FC-Graph [55]	36.72	27.52	19.25	-	-	61.90	45.39	29.96	–	–	69.16	41.36	-	-
	CAQ [56]	38.19	30.23	18.47	-	-	65.05	50.21	28.93	–	–	73.41	38.52	-	-
	Kernel GraphNet [57]	43.27	35.41	19.38	-	-	68.72	55.62	30.29	–	–	72.92	42.35	-	-
	Grounded Situation Recognition models														
	ISL [20]	39.36	30.09	18.62	22.73	7.72	65.51	50.16	28.47	36.60	11.56	72.42	37.10	52.19	14.58
	JSL [20]	39.94	31.44	18.87	24.86	9.66	67.60	51.88	29.39	40.60	14.72	73.21	37.82	56.57	18.45
	GSRTR [49]	40.63	32.15	19.28	25.49	10.10	69.81	54.13	31.01	42.50	15.88	74.11	39.00	57.45	19.67
	SituFormer [27]	44.20	35.24	21.86	29.22	13.41	71.21	55.75	33.27	46.00	20.10	75.85	42.13	61.89	24.89
	CoFormer [1]	44.66	35.98	22.22	29.05	12.21	73.31	57.76	33.98	46.25	18.37	75.95	41.87	60.11	22.12
	GSRFormer[3]	49.42	40.42	26.25	34.41	18.14	74.42	59.81	36.78	49.41	24.31	79.79	46.63	64.95	28.20
	HiFormer (Ours)	79.29	49.20	17.35	35.83	7.33	94.06	57.01	19.95	42.13	8.95	59.73	20.87	44.50	9.69

4.2. Performance Comparison with State-of-the-Art Models

For HiFormer, we use the evaluation metric for Grounded Situation Recognition proposed by Pratt et al. 1., which stands as below: 1) verb: accuracy of the verb prediction; 2) value: accuracy of prediction a single semantic role; 3) val-all: the accuracy of correctly predicting all the semantic roles in an image simultaneously; 4) grnd: the accuracy of single bounding box predictions; 5): the accuracy of correctly predicting all the bounding boxes at once. We deem a bounding box prediction as correct if the IoU between it and the ground truth bounding box is above 0.5.

We further implement the above metrics under three different settings: 1) Top-1-Verb: only calculate the accuracy of the top-1 verb, its corresponding semantic roles, and grounding frames; 2) Top-5-Verb: calculate the accuracy of the top-5 verbs, its semantic roles and; 3) Ground-Truth-Verb: the ground truth verb is known before the prediction, so only the accuracy of roles and frames are calculated. Note that in the first two settings, the role and bounding box predictions are automatically considered incorrect if the top- k verbs do not include the ground truth verb.

DOI: 10.54254/2755-2721/49/20241075

Table 2. Performance comparison with state-of-the-art models on the validation set.

Dataset	Models	Top-1 Verb					Top-5 Verb					Ground Truth Verb			
		verb	value	val-all	grnd	grnd-all	verb	value	val-all	grnd	grnd-all	value	val-all	grnd	grnd-all
Test	Traditional Situation Recognition models														
	CRF [53]	32.34	24.64	14.19	-	-	58.88	42.7650.21	22.55	-	-	65.66	28.96	-	-
	CRF + DataAug [54]	34.12	26.45	15.51	-	-	62.59	46.88	25.46	-	-	70.44	34.38	-	-
	VGG+RNN [55]	35.90	27.45	16.36	-	-	63.08	46.88	26.06	-	-	70.27	35.25	-	-
	FC-Graph [56]	36.72	27.52	19.25	-	-	61.90	45.39	29.96	-	-	69.16	41.36	-	-
	CAQ [57]	38.19	30.23	18.47	-	-	65.05	50.21	28.93	-	-	73.41	38.52	-	-
	Kernel GraphNet [58]	43.27	35.41	19.38	-	-	68.72	50.21	30.29	-	-	72.92	42.35	-	-
	Grounded Situation Recognition models														
	ISL [23]	39.36	30.09	18.62	22.73	7.72	65.51	50.16	28.47	36.60	11.56	72.42	37.10	52.19	14.58
	JSL [23]	39.94	31.44	18.87	24.86	9.66	67.60	51.88	29.39	40.60	14.72	73.21	37.82	56.57	18.45
	GSRTR [50]	40.63	32.15	19.28	25.49	10.10	69.81	54.13	31.01	42.50	15.88	74.11	39.00	57.45	19.67
	SituFormer [27]	44.20	35.24	21.86	29.22	13.41	71.21	55.75	33.27	46.00	20.10	75.85	42.13	61.89	24.89
	CoFormer [58]	44.66	35.98	22.22	29.05	12.21	73.31	57.76	33.98	46.25	18.37	75.95	41.87	60.11	22.12
	GSRformer [28]	49.52	40.64	26.21	34.20	18.02	74.21	59.78	36.98	49.23	23.45	79.77	45.65	64.87	27.90
	HiFormer (Ours)	79.29	49.20	17.35	35.83	7.33	94.06	57.01	19.95	42.13	8.95	59.73	20.87	44.50	9.69

4.3. Performance Comparison with State-of-the-Art Models

As shown in Table 1 and 2, HiFormer achieves state-of-the-art verb and noun accuracy under the top-1 and top-5 verbs. Compared to the current best-performing GSR model CoFormer [12], the improvement in the verb prediction accuracy range from *21*% under the Top- 5 Verb to *35*% under the Top- 1 verb. Fur-thermore, HiFormer improves the noun and bounding box accuracy by *14*% and *6*%, respectively, under the Top-1 verb. However, our model shows some deficiency under the Ground Truth Verb metric, as well as the value-all and grnd-all accuracy (in which a prediction is only counted as correct if all nouns or grounding boxes are predicted correctly). Despite the model deficiency, the astounding improvement in the verb prediction accuracy demonstrates the effectiveness of our framework in solving verb ambiguity.

5. Conclusion

We propose a novel two-stage hierarchical transformer framework, in which we simultaneously consider all similarities for each image instance. With this improved framework, HiFormer outperforms all state-of-the-art models regarding verb and noun accuracy. Compared to the current two best-performing models, CoFormer [58] and SituFormer [27], HiFormer prevails by over 35% on the top-1 verb accuracy, 13% on the top-1 noun accuracy and 21% on the top-5 verb accuracy. Regardless, some limitations of HiFormer lie with the bounding box prediction and the accuracy under the Ground-Truth-Verb, which we intend to explore further in the future. Our hierarchical framework provides a foundation for future Grounded Situation Recognition works in solving the bottleneck for many downstream applications such as E-commerce [59,60,61,62,63], Intelligent Transportation [64,65,66], etc. We believe our work will contribute to moving Grounded Situation Recognition out of the laboratories and implementing it in the surveillance network to improve people's lives.

References

[1] Paul Bischoff. Surveillance camera statistics: which cities have the most cctv cameras? 2

[2] Video surveillance market by offering (hardware (camera, storage device, monitor), software (video analytics, vms), service (vsaas)), system (ip,analog, hybrid), vertical and geography (north america, europe, apac,row) - global forecast to 2027. 2

DOI: 10.54254/2755-2721/49/20241075

[3] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. IEEE transactions on pattern analysis and machine intelligence, 27(3):328–340, 2005. 2

[4] SiyuHuang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. Gnas: A greedy neural architecture search method for multi-attribute learning. In Proceedings of the 26th ACM international conference on Multimedia, pages 2049–2057, 2018. 2

[5] Xu Bao, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Wangmeng Xiang, Jingdong Sun, Hanbing Liu, Wei Liu, Bin Luo, Yifeng Geng, et al. Keyposs: Plug-and-play facial landmark detection through gps-inspired true-range multilateration. In Proceedings of the 31st ACM International Conference on Multimedia, 2023. 2

[6] 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. 2

[7] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015. 2

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440,2015. 2

[9] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander Hauptmann. Improving the learning of multi-column convolutional neural network for crowd counting. In Proceedings of the 27th ACM International Conference on Multimedia, 2019. 2

[10] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G Hauptmann. Learning spatial awareness to improve crowd counting. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6152–6161, 2019. 2

[11] Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G Hauptmann. Rethinking spatial invariance of convolutional networks for object counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19638–19648, 2022. 2

[12] SiyuHuang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. Stacked pooling for boosting scale invariance of crowd counting. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2578–2582. IEEE, 2020. 2

[13] Ji Zhang, Zhi-Qi Cheng, Xiao Wu, Wei Li, and Jian-Jun Qiao. Crossnet: Boosting crowd counting with localization. In Proceedings of the 30th ACM International Conference on Multimedia, pages 6436–6444, 2022. 2

[14] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. In arXiv preprint arXiv:2211.09590, 2022. 2

[15] Yuxuan Zhou, Zhi-Qi Cheng, Jun-Yan He, Bin Luo, Yifeng Geng, Xuansong Xie, and Margret Keuper. Overcoming topology agnosticism: Enhancing skeleton-based action recognition through redefined skeletal topology awareness. In arXiv preprint arXiv:2305.11468, 2023. 2

[16] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. Db-lstm: Densely-connected bi-directionallstm for human action recognition. Neurocomputing, 444:319–331, 2021. 2

[17] Hanbing Liu, Jun-Yan He, Zhi-Qi Cheng, Wangmeng Xiang, Qize Yang, Wenhao Chai, Gaoang Wang, Xu Bao, Bin Luo, Yifeng Geng, et al. Posynda: Multi-hypothesis pose synthesis domain adaptation for robust 3d human pose estimation. In Proceedings of the 31st ACM International Conference on Multimedia, 2023. 2

[18] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. Vireo@ trecvid 2017: Video-to-text, ad-hoc video search and video hyperlinking. In 2017 TREC Video Retrieval Evaluation (TRECVID2017), 2017. 2

DOI: 10.54254/2755-2721/49/20241075

[19] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In Proceedings of the 26th ACM international conference on Multimedia, pages 383–391, 2018. 2

[20] Luca Weihs Ali Farhadi Sarah Pratt, Mark Yatskar and Aniruddha Kembhavi. Grounded situation recognition. In European Conference on Computer Vision, pages Springer, 314–332, 2020. 2,11,12

[21] Mark Yatskar Ram Nevatia Aniruddha Kembhavi Arka Sadhu, Tanmay Gupta. Visual semantic role labeling for video understanding. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5589–5600. IEEE, CVF, 2021. 2

[22] Lluis Castrejon Paul Vicol, Makarand Tapaswi and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In IEEE Conference on Computer Vision and Pattern Recognition, pages 8581–8590. IEEE, 2018. 2

[23] Angelina Wang Dora Zhao and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In IEEE/CVF International Conference on Computer Vision, pages 14830–14840. IEEE, CVF, 2021. 2

[24] Yang Liu Zhi-Qi Cheng, Xiao Wu and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In IEEE Conference on Computer Vision and Pattern Recognition, page 4048–4056. IEEE, 2017. 2

[25] Jerome Revaud Albert Gordo, Jon Almazán and Diane Larlus. Deep image retrieval: Learning global repre- sentations for image search. In European Conference on Computer Vision, pages Springer, 241–257, 2016. 2

[26] Jack Sim Tobias Weyand Hyeonwoo Noh, Andre Araujo and Bohyung Han. Large-scale image retrieval with attentive deep local features. In IEEE International Conference on Computer Vision, page 3456–3465. IEEE, 2017. 2

[27] WeiJi Xiaoyu Yue Tat-Seng Chua Meng Wei, Long Chen. Rethinking the two-stage framework for grounded situation recognition. arXiv:2112.05375, 2021. 1,3,5,11,12

[28] Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In Proceedings of the 30th ACM International Conference on Multimedia, pages 3272–3281, 2022. 1,3,11,12

[29] Frank Keller Spandana Gella. An analysis of action recognition datasets for language and vision tasks. arXiv:1704.07129, 2017. 4

[30] Nazli Ikizler, R Gokberk Cinbis, and Pinar Duygulu. Human action recognition with line and flow histograms. In 2008 19th International Conference on Pattern Recognition, pages 1–4. IEEE, 2008. 4

[31] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE transactionson pattern analysis and machine intelligence, 31(10):1775–1789, 2009. 4

[32] Bangpeng Yao and LiFei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 9–16. IEEE, 2010. 4

[33] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy LaiLin, Leonidas Guibas, and LiFei-Fei. Human action recognition by learning bases of action attributes and parts. In 2011 International conference on computer vision, pages 1331–1338. IEEE, 2011. 4

[34] Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. Tuhoi: Trento universal human object interaction dataset. In Proceedings of the Third Workshop on Vision and Language, pages 17–24, 2014. 4

[35] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In Proceedings of the IEEE international conference on computer vision, pages 1017–1025, 2015. 4

DOI: 10.54254/2755-2721/49/20241075

[36] Frank Keller Spandana Gella, Mirella Lapata. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. arXiv:1603.09188, 2016. 4
[37] Jitendra Malik Saurabh Gupta. Visual semantic role labeling. arXiv:1505.04474, 2015. 4
[38] Luke Zettlemoyer Mark Yatskar and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA. IEEE, 2016. 4
[39] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. arXiv:1706.03762, 2017. 4,9
[40] Ji Lin Yujun Lin Song Han Zhanghao Wu, Zhijian Liu. Lite transformer with long-short range attention. arXiv:2004.11886, 2020. 4
[41] Yiming Yang Quoc V. Le Zihang Dai, Guokun Lai. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. arXiv:2006.03236, 2020. 4
[42] Srinivasan Iyer Luke Zettlemoyer Hannaneh Hajishirzi Sachin Mehta, Marjan Ghazvininejad. Delight: Deep and light-weight transformer. arXiv:2008.00623, 2020. 4
[43] Yiming Yang Jaime Carbonell Quoc V. Le Ruslan Salakhutdinov Zihang Dai, Zhilin Yang. Transformerxl: Attentive language models beyond a fixed-length contextzihang dai, zhilin yang, yiming yang, jaime carbonell, quoc v. le, ruslan salakhutdinov. arXiv:1901.02860, 2019. 5
[44] Kun Qian Jing Gu Alborz Geramifard Zhou Yu Qingyang Wu, Zhenzhong Lan. Memformer: A memory- augmented transformer for sequence modeling. arXiv:2010.06891, 2020. 5
[45] Ming Zhou Xingxing Zhang, Furu Wei. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. arXiv:1905.06566, 2019. 5
[46] Xiaonan Li Xipeng Qiu Hang Yan, Bocao Deng. Tener: Adapting transformer encoder for named entity recognition. arXiv:1911.04474, 2019. 5
[47] Enhua Wu Jianyuan Guo Chunjing Xu Yunhe Wang Kai Han, An Xiao. Transformer in transformer. arXiv:2103.00112, 2021. 5
[48] Munawar Hayat Syed Waqas Zamir Fahad Shahbaz Khan Mubarak Shah Salman Khan, Muzammal Naseer. Transformers in vision: A survey. arXiv:2101.01169, 2021. 5
[49] Hyeonjun Lee Suha Kwak Junhyeong Cho, Youngseok Yoon. Grounded situation recognition with transformers. arXiv:2111.10135, 2021. 5,11,12
[50] Gabriel Synnaeve Nicolas Usunier Alexander Kirillov Sergey Zagoruyko Nicolas Carion, Francisco Massa. End-to-end object detection with transformers. arXiv:2005.12872, 2020. 5
[51] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020. 7
[52] Luke Zettlemoyer Mark Yatskar and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In IEEE conference on computer vision and pattern recognition, page 5534–5542. IEEE, 2016. 11,12
[53] Luke Zettlemoyer Mark Yatskar, Vicente Ordonez and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition. In IEEE conference on computer vision and pattern recognition, page 7196–7205. IEEE, 2017. 11,12
[54] Arun Mallya and Svetlana Lazebnik. Recurrent models for situation recognition. In IEEE International Conference on Computer Vision, page 455–463. IEEE, 2017. 11,12
[55] Renjie Liao Jiaya Jia Raquel Urtasun Ruiyu Li, Makarand Tapaswi and Sanja Fidler. Situation recognition with graph neural networks. In IEEE International Conference on Computer Vision, page 4173–4182. IEEE, 2017. 11,12
[56] Ngai-Man Cheung Thilini Cooray and Wei Lu. Attention-based context aware reasoning for situation recog- nition. In IEEE/CVF International Conference on Computer Vision, page 4736–4745. IEEE,CVF, 2020. 11, 12

DOI: 10.54254/2755-2721/49/20241075

[57] Mohammed Suhail and Leonid Sigal. Mixture-kernel graph attention network for situation recognition. In IEEE/CVF International Conference on Computer Vision, page 10363–10372. IEEE, CVF, 2019. 11,12

[58] Suha Kwak Junhyeong Cho, Youngseok Yoon. Collaborative transformers for grounded situation recognition. arXiv:2203.16518, 2022. 1,11,12

[59] Zhi-Qi Cheng, Yang Liu, Xiao Wu, and Xian-Sheng Hua. Video ecommerce: Towards online video adver- tising. In Proceedings of the 24th ACM international conference on Multimedia, pages 1365–1374, 2016. 13

[60] Zhi-Qi Cheng, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. On the selection of anchors and targets for video hyperlinking. In Proceedings of the 2017 acm on international conference on multimedia retrieval, pages287–293, 2017. 13

[61] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4048–4056, 2017. 13

[62] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video ecommerce++: Toward large scale online video advertising. IEEE transactions on multimedia, 19(6):1170–1183, 2017. 13

[63] Guang-Lu Sun, Zhi-Qi Cheng, Xiao Wu, and Qiang Peng. Personalized clothing recommendation combining user social circle and fashion style consistency. Multimedia Tools and Applications, 77:17731–17754, 2018.13

[64] Jin-Peng Lan, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Xu Bao, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Procontext: Exploring progressive context transformer for tracking. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2022. 13

[65] Chenyang Li, Zhi-Qi Cheng, Jun-Yan He, PengyuLi, Bin Luo, Hanyuan Chen, Yifeng Geng, Jin-Peng Lan, and Xuansong Xie. Longshortnet: Exploring temporal and semantic features fusion in streaming percep- tion. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2022. 13

[66] Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Wangmeng Xiang, Binghui Chen, Bin Luo, Yifeng Geng, and Xuansong Xie. Damo-streamnet: Optimizing streaming perception in autonomous driving. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence, 2023. 13

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

DOI: 10.54254/2755-2721/49/20241075

Appendices

Appendices A. Internal Structure of Our Transformer Modules

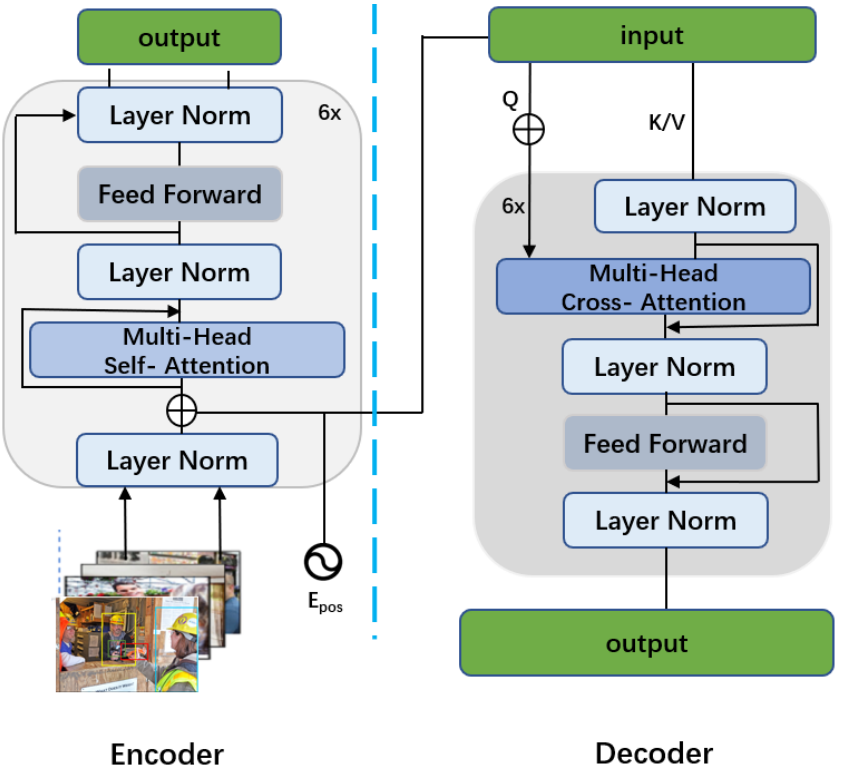


Figure A1. Internal Structure of Leaf Transformer.

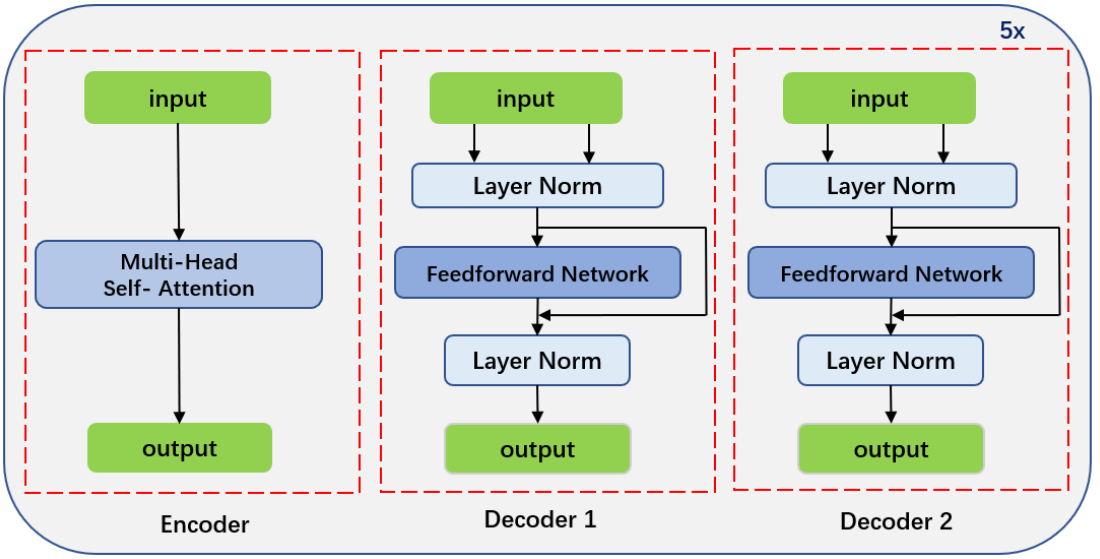


Figure A2. Internal Structure of Root Transformer.