Diagram Visual Grounding: Learning to See with Gestalt-Perceptual Attention

Xin Hu[1,2], **Lingling Zhang**[1,2*], **Jun Liu**[1,2], **Xinyu Zhang**[1,2], **Wenjun Wu**[1,2] and **Qianying Wang**[3]

[1] Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering,
School of Computer Science and Technology, Xi'an Jiaotong University, China
[2] National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, China
[3] Lenovo Research, Beijing, China
dr.huxin711@foxmail.com, {zhanglling, liukeen}@xjtu.edu.cn, zhibei1204@gmail.com,
nickjunwork@163.com, wangqya@lenovo.com

Abstract

Diagram visual grounding aims to capture the correlation between language expression and local objects in the diagram, and plays an important role in the applications like textbook question answering and cross-modal retrieval. Most diagrams consist of several colors and simple geometries. This results in sparse low-level visual features, which further aggravates the gap between low-level visual and high-level semantic features of diagrams. The phenomenon brings challenges to the diagram visual grounding. To solve the above issues, we propose a gestalt-perceptual attention model to align the diagram objects and language expressions. For low-level visual features, inspired by the gestalt that simulates human visual system, we build a gestalt-perception graph network to make up the features learned by the traditional backbone network. For high-level semantic features, we design a multi-modal context attention mechanism to facilitate the interaction between diagrams and language expressions, so as to enhance the semantics of diagrams. Finally, guided by diagram features and linguistic embedding, the target query is gradually decoded to generate the coordinates of the referred object. By conducting comprehensive experiments on diagrams and natural images, we demonstrate that the proposed model achieves superior performance over the competitors.

1 Introduction

Visual grounding is to localize the corresponding object in the image referred by an expression. Common expressions are either simple entities, such as *the cat*, or phrases, such as *the cup on the table*, or sentences like *the girl is holding a pen*. Visual grounding plays an important role in visual question and answering, visual dialogue, and other applications.

The visual grounding models have achieved ideal effects on natural images, while another kind of image with great research value has not received much attention, that is, the diagram. Diagrams widely exist in textbooks, technology forums, etc. Although there is a research on diagram-sentence matching [Hu *et al.*, 2021], this work only outputs the whole diagrams corresponding to the sentence. Diagram visual grounding mines the fine-grained objects by expressions. The interaction between objects and expressions can enhance the semantic, which is of great significance to the textbook question and answering, intelligent dialogue, etc.

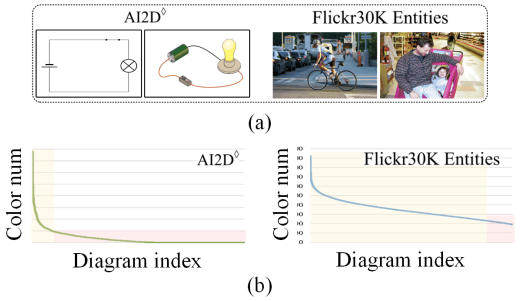

Figure 1: Comparison of visual characteristics between diagrams in AI2D◇ and natural images in Flickr30K Entities.

Considering the research gap on diagram visual grounding at present, we apply the latest model VLTVG [Yang *et al.*, 2022] to verify the performance on diagrams. The experimental results show that the accuracy of VLTVG on the natural image dataset Flickr30K Entities [Plummer *et al.*, 2015] achieves 79.18%, while that on the diagram dataset AI2D◇ is only 25.79%. The reasons can be analyzed from the following two aspects. **The low-level visual feature of diagram is sparse.** The diagram is drawn by experts in special fields, usually consisting of simple geometries and color blocks, and the low-level visual information is sparse. As shown in Figure 1(a), the natural image belongs to the realistic style, while the diagram focuses on the expression of knowledge. The diagram does not have a complex background compared to the natural image, while the color information is monotonous for the foreground objects that the model pays more attention to. The two line charts in Figure 1(b) represent the number of RGB colors contained in each image. The green line represents the statistic of diagrams in AI2D◇ and the blue line indicates that of natural images in Flickr30K Entities. It can be seen that the natural images are rich in colors and most im-

*Corresponding author: Lingling Zhang

ages contain a variety of colors. The color distribution of the diagram shows an obvious long-tail distribution. As a result, the diagram visual features are extremely sparse, and only using traditional backbone networks can not learn high-quality diagram representation. Therefore, how to effectively represent the foreground objects in the diagram is a problem. In addition, **the gap between low-level visual and high-level semantic features is exacerbated** by the sparse visual information and diverse forms of diagram expression. As shown in Figure 1(a), the two diagrams both represent *circuit*. The *bulb* in the first diagram is a circle with a fork, while the *bulb* in the second diagram is represented as a more vivid yellow object. This different form of visual objects brings semantic problem to the alignment of diagram and language expression in visual grounding task.

Gestalt is a psychological concept, which mainly includes some laws such as similarity, common fate, proximity, and continuity, and simulates the perception process of human eyes on visual elements. A previous work [Hu *et al.*, 2022] has proved that even if the visual information of the diagram is sparse, the key objects in the diagram can be effectively identified through the cooperation of multiple gestalt laws. Inspired by this, we focus on the visual and semantic representation of diagrams and propose a **G**estalt-**P**erceptual **A**ttention (GPA) model for diagram visual grounding task.

GPA mainly contains a diagram local learner and a spatial-semantic association module. The former enhances the local visual features of diagrams and the later promotes the association between diagram objects and language expression. Concretely, inspired by the powerful node association ability of graph networks [Zhang *et al.*, 2022b; Zhang *et al.*, 2022a], a gestalt-perception graph (GPG) network is built in the local learner. Each layer of GPG network has diagram patches as nodes, and the relationships between patches as edges. Gestalt laws are used to simulate the relationship between patches. Taking patch A and patch B as an example, that is, two nodes in the GPG are recorded as node_A and node_B respectively. If the spatial positions of A and B on the diagram are close, proximity law is used as a priori knowledge to enhance the location association between node_A and node_B, thus guiding the local representation of diagram patches by updating the GPG. In the association module, a multi-modal context attention is designed to improve the semantics of the diagram by the interaction with linguistic embedding. Finally, the target query predicts the referred box under the guidance of the diagram and language expression. Our main contributions are summarized as follows:

- As far as we know, the diagram visual grounding is studied for the first time. Because of the visual sparsity of diagrams and the gap between the visual features and semantic features of diagrams, we propose a gestalt-perceptual attention model for the novel task.
- We construct a gestalt-perception graph network to guide edge learning by gestalt laws and the representation of diagram patches is enhanced by information propagation between nodes. The multi-modal context attention mechanism is designed to associate visual representation of diagrams with linguistic embedding.
- We conduct experiments on a diagram dataset AI2D$^\diamond$ and a natural image dataset Flickr30K Entities. The experimental results indicate that the proposed GPA modal achieves the best accuracy in the diagram visual grounding task, and also obtains a comparable accuracy over the competitors in natural images.

2 Related Works

Gestalt Perception Theory. The theory [Wagemans *et al.*, 2012; Pomerantz *et al.*, 1977; Wertheimer, 1922; Desolneux *et al.*, 2004] aims to explain the process of human's overall cognition of objects, that the human visual system tends to perceive objects that are similar, close or connected without abrupt directional changes. They are characterized by the laws of proximity, continuity, common fate and so on. At present, there are several researches [Gnjatović *et al.*, 2022; Yan *et al.*, 2018; Xu *et al.*, 2019] on solving computer vision tasks with gestalt perception theory. For example, CogG [Yan *et al.*, 2021] model recognizes saliency in three phases, seeing, perceiving, and cogitating, mimicking human's perceptive and cognitive thinking of an image. [Gnjatović *et al.*, 2022] proposes an approach to traffic accident clustering and it is psychologically inspired to the extent that it introduces a clustering criterion based on the gestalt principle of proximity. In this work, we design an adaptive gestalt perception method with similarity, proximity and smoothness laws, for diagram visual grounding task.

Visual Grounding. The main idea of two-stage visual grounding [Wang and Specia, 2019; Hong *et al.*, 2019; Yang *et al.*, 2019; Wang *et al.*, 2019] is similar to RCNN-series method. In the first stage, several proposals are usually generated by the pretrained object detectors [Ren *et al.*, 2015; Redmon *et al.*, 2016]. In the second stage, score each proposal by calculating the correlation between proposal and the language expression, and select the top-ranked proposals as the final predictions [Liu *et al.*, 2021]. Recently, the one-stage approach has been widely studied due to its concise end-to-end architecture. TransVG [Deng *et al.*, 2021] is the first transformer-based framework for the visual grounding task and it avoids the manually-designed mechanisms to perform the query reasoning and multi-modal fusion. However, the visual-linguistic module in TransVG does not focus on the correspondence between the visual features and the target objects, which may affect the performance. VLTVG [Yang *et al.*, 2022] is also a transformer-based framework that directly learns the attention scores between the fine-grained visual features and target objects. Specifically, the visual encoding is guided by the semantics of the textual contexts, so that the VLTVG model more focuses on the discriminant region related to the language expression. [Li and Sigal, 2021] proposed a multi-task model, which leverages the transformer architecture and jointly learns the referring expression comprehension and segmentation tasks. In [Ye *et al.*, 2021] and [Liao *et al.*, 2020], the visual grounding task is regarded as a filtering-based reasoning process. Due to the sparse visual features and complex semantics of diagrams, we integrate gestalt perception into a one-stage architecture to conduct the diagram visual grounding task.

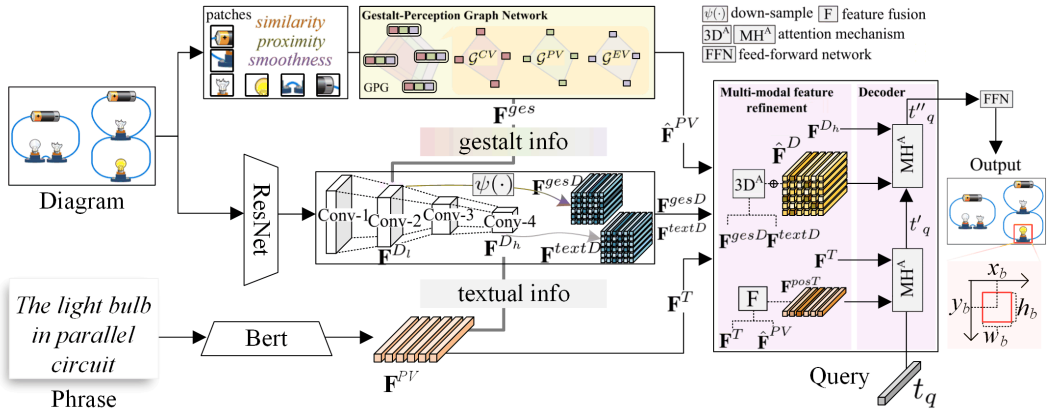

Figure 2: The overall architecture of GPA model for diagram visual grounding task.

3 Methodology

This section introduces the architecture of GPA model as shown in Figure 2. The goal is to give a language expression T and locate its referred region on the diagram D. For the language expression, we encode it as a sequence of linguistic embedding $\mathbf{F}^T \in \mathbb{R}^{K\times d}$ with BERT [Kenton and Toutanova, 2019]. For the diagram, we utilize the backbone network ResNet-50 [He *et al.*, 2016] to extract the low-level visual feature $\mathbf{F}^{D_l} \in \mathbb{R}^{h_l\times w_l\times c_l}$ from Conv-2 block and the high-level semantic feature $\mathbf{F}^{D_h} \in \mathbb{R}^{h_h\times w_h\times c_h}$ from Conv-4 block. We also divide the diagram into patches to facilitate the learning of local features with gestalt perception.

Specifically, GPA contains three key steps: 1) *Enhance the low-level visual features by gestalt laws.* The diagram patches, as the node of gestalt-perception graph (GPG) network, initialize GPG by extracting features from color, position and contour views; the node features in the GPG are updated based on similarity, proximity, and smoothness laws, and then the nodes are utilized to enhance the low-level diagram feature $\mathbf{F}^{D_l}$; 2) *Enrich the high-level semantic features by multi-modal context attention.* $\mathbf{F}^{D_h}$ and linguistic embedding $\mathbf{F}^T$ interact to enhance high-level semantic features of the diagram, and then refine the diagram features through 3D-attention. 3) *Target query decoder.* Guided by diagram features and linguistic embedding, the target query is gradually decoded to generate the coordinates of the predicted bounding box. Finally, the regression loss is computed with the ground-truth box to optimize the training of GPA model.

3.1 Improve Low-level Visual Feature with GPG

Gestalt-perception Graph Network

The gestalt laws simulate the process of human visual system and effectively identify the regions in the diagram with limited annotations. Taking diagram patches as input, the GPA model builds a gestalt-perception graph GPG with patches as nodes and the relationship between patches as edges. Under the guidance of gestalt laws, more meaningful diagram patch feature $\mathbf{F}^{ges}$ is output after the L-layer GPG network updating. The GPG consists of subgraphs $\mathcal{G}^{CV}$, $\mathcal{G}^{PV}$ and $\mathcal{G}^{EV}$ for color, position and contour views, respectively.

Notations	Explanations
$D \in \mathbb{R}^{H\times W\times C}$	Original diagram
$D^P = \{d_i^p \mid i = 1, \ldots, N\}$	Diagram patches
$\mathcal{G}^{CV} = (\mathcal{N}^{CV}, \mathcal{E}^{CV})$	Color view subgraph
$\mathcal{G}^{PV} = (\mathcal{N}^{PV}, \mathcal{E}^{PV})$	Position view subgraph
$\mathcal{G}^{EV} = (\mathcal{N}^{EV}, \mathcal{E}^{EV})$	Contour view subgraph
$\mathbf{F}^{CV} \in \mathbb{R}^{N\times 9}$	Node features for $\mathcal{N}^{CV}$ in $\mathcal{G}^{CV}$
$\mathbf{F}^{PV} \in \mathbb{R}^{N\times 4}$	Node features for $\mathcal{N}^{PV}$ in $\mathcal{G}^{PV}$
$\mathbf{F}^{EV} \in \mathbb{R}^{N\times(2\times\frac{W_0}{\sqrt{N}}+2\times\frac{H_0}{\sqrt{N}})}$	Node features for $\mathcal{N}^{EV}$ in $\mathcal{G}^{EV}$

Table 1: Important notations and explanations.

Node Features. In order to learn the local representations of the diagram, we divide the diagram D into N patches that is same as [Dosovitskiy *et al.*, 2020]. A patch represents a node in the GPG. The important notations and explanations are summarized in Table 1. $\mathbf{F}^{CV}$ is the color feature of node set $\mathcal{N}^{CV}$, representing the central moments [Stricker and Orengo, 1995] of patches. $\mathbf{F}^{PV}$ is the position feature of the node set $\mathcal{N}^{PV}$, which is composed of the coordinates of the top-left corner and the bottom-right corner of each patch. $\mathbf{F}^{EV}$ is the contour feature of the node set $\mathcal{N}^{EV}$, including the top, bottom, left, and right contours of each patch.

To facilitate the interaction between features of three visual views, $\mathbf{F}^{CV}$, $\mathbf{F}^{PV}$ and $\mathbf{F}^{EV}$ are mapped into the same feature space through the multi-layer perceptron MLP$(\cdot)$ network as shown in (1), where d_c, d_p and d_e are mapping dimensions and $||$ indicates the concatenating operator.

$$\begin{aligned}\hat{\mathbf{F}}^{CV} &= \text{MLP}^{CV}(\mathbf{F}^{CV}), \hat{\mathbf{F}}^{CV} \in \mathbb{R}^{N\times d_c},\\ \hat{\mathbf{F}}^{PV} &= \text{MLP}^{PV}(\mathbf{F}^{PV}), \hat{\mathbf{F}}^{PV} \in \mathbb{R}^{N\times d_p},\\ \hat{\mathbf{F}}^{EV} &= \text{MLP}^{EV}(\mathbf{F}^{EV}_{(t;b)} || \mathbf{F}^{EV}_{(l;r)}), \hat{\mathbf{F}}^{EV} \in \mathbb{R}^{N\times 4\times d_e}.\end{aligned} \quad (1)$$

Edge Matrices. Gestalt perception emphasizes the correlation between visual regions. That is, the human visual system tends to perceive regions that are similar, close or connected without abrupt directional changes as a perceptual whole object [Wagemans *et al.*, 2012]. Thereby, we encode the edges

in $\mathcal{G}^{CV}$, $\mathcal{G}^{PV}$ and $\mathcal{G}^{EV}$ with the assistant of similarity, proximity and smoothness laws, respectively.

- *Color Similarity.* $\mathcal{E}^{CV} \subseteq \mathcal{N}^{CV} \times \mathcal{N}^{CV}$ represents the color similarity between nodes. Given two node features $\hat{\mathbf{F}}_i^{CV}$ and $\hat{\mathbf{F}}_j^{CV}$, the weight of $\mathcal{E}_{ij}^{CV}$ is shown in (2), where $\text{sim}(\cdot)$ is a cosine similarity function.

$$\mathbf{A}_{ij}^{CV} = \text{sim}(\hat{\mathbf{F}}_i^{CV}, \hat{\mathbf{F}}_j^{CV}). \tag{2}$$

- *Position Proximity.* In order to measure the proximity of spatial positions, $\mathcal{E}^{PV} \subseteq \mathcal{N}^{PV} \times \mathcal{N}^{PV}$ is formulated to learn the positional relation between two patches. The weight of $\mathcal{E}_{ij}^{PV}$ is calculated by (3).

$$\mathbf{A}_{ij}^{PV} = 1.0 - \text{Norm}(\sqrt{\sum_{t=1}^{d_p}(\hat{\mathbf{F}}_{it}^{PV} - \hat{\mathbf{F}}_{jt}^{PV})^2}). \tag{3}$$

- *Contour Smoothness.* The law of contour smoothness states humans perceive objects as continuous in a smooth pattern. In order to judge whether two patches may belong to the same object, $\mathcal{E}^{EV} \subseteq \mathcal{N}^{EV} \times \mathcal{N}^{EV}$ determines the possibility of splicing two patches. The weight of $\mathcal{E}^{EV}$ is computed as follows, where $\hat{\mathbf{F}}_i^{EV_b}$ represents the bottom contour feature of patch d_i^P.

$$\begin{cases} \delta_1 = \text{sim}(\hat{\mathbf{F}}_i^{EV_b}, \hat{\mathbf{F}}_j^{EV_t}); \delta_2 = \text{sim}(\hat{\mathbf{F}}_i^{EV_t}, \hat{\mathbf{F}}_j^{EV_b}), \\ \delta_3 = \text{sim}(\hat{\mathbf{F}}_i^{EV_l}, \hat{\mathbf{F}}_j^{EV_r}); \delta_4 = \text{sim}(\hat{\mathbf{F}}_i^{EV_r}, \hat{\mathbf{F}}_j^{EV_l}), \\ \mathbf{A}_{ij}^{EV} = \max\{\delta_1, \delta_2, \delta_3, \delta_4\}. \end{cases} \tag{4}$$

GPG Network Updating

Each layer of the GPG network is stacked by the propagation block and the aggregation block as shown in Figure 3.

Propagation Block. The three subgraphs in GPG focus on the color, position and contour features of diagram patches, respectively. According to [Wagemans *et al.*, 2012], there is cooperation and confliction among multiple gestalt laws. For example, when human visual system perceives various objects, the perception order and contribution of similarity, proximity and smoothness laws are different. Therefore, an adaptive learning strategy is designed in GPG to propagate information between patches. For example, when judging the degree of association between two patches, we give priority to whether the spatial positions are close. Because all parts of a complete object are usually concentrated in a certain area, rather than scattered in the diagram, the position proximity is particularly important. GPG sets three adaptive factors α, β, and γ to control the contribution of similarity, proximity, and smoothness respectively. The propagation methods of $\mathcal{G}^{CV}$, $\mathcal{G}^{PV}$ and $\mathcal{G}^{EV}$ are similar. For $\mathcal{G}^{CV}$, the update of node features is shown in (5), where $\mathbf{A}^{(l)}$ denotes the adjacency matrix and $\mathbf{W}_C^{(l)}$ refers to the l-th layer parameter of $\mathcal{G}^{CV}$. $\mathbf{A}^{(l)}$ is calculated by adaptive weighting of three factors α, β, and γ in (6).

$$\begin{cases} \tilde{\mathbf{F}}^{CV(l)} = \Phi_C(\hat{\mathbf{F}}^{CV(l)}, \mathbf{A}^{(l)}), \\ \Phi_C(\hat{\mathbf{F}}^{CV(l)}, \mathbf{A}^{(l)}) = \text{ReLU}(\mathbf{A}^{(l)}\hat{\mathbf{F}}^{CV(l)}\mathbf{W}_C^{(l)}), \end{cases} \tag{5}$$

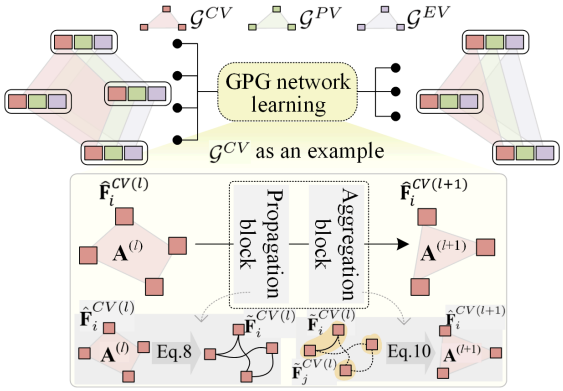

Figure 3: The GPG network learning process.

$$\mathbf{A}^{(l)} = \alpha \cdot \mathbf{A}^{CV(l)} + \beta \cdot \mathbf{A}^{PV(l)} + \gamma \cdot \mathbf{A}^{EV(l)}. \tag{6}$$

Aggregation Block. The propagation in GPG network promotes the correlation between patches, and some patches can be aggregated to form more meaningful objects. Therefore, we denote a learned assignment matrix [Ying *et al.*, 2018] at layer l as $\mathbf{S}^{(l)} \in \mathbb{R}^{N_l \times N_{l+1}}$, where N_l is the number of nodes at layer l. It provides a soft assignment of each node at layer l to the next layer l+1. Taking $\mathcal{G}_{CV}$ as an example, the node feature $\hat{\mathbf{F}}^{CV(l+1)}$ at layer l+1 is aggregated by (7). The node aggregation method of $\mathcal{G}_{PV}$ and $\mathcal{G}_{EV}$ is similar as that of $\mathcal{G}_{CV}$. GPG concatenates $\hat{\mathbf{F}}^{CV(L)}$, $\hat{\mathbf{F}}^{PV(L)}$ and $\hat{\mathbf{F}}^{EV(L)}$, and then applies a $\text{MLP}(\cdot)$ layer for feature projection.

$$\begin{cases} \hat{\mathbf{F}}^{CV(l+1)} = \mathbf{S}^{(l)^\top} \times \tilde{\mathbf{F}}^{CV(l)}, \\ \mathbf{F}^{ges} = \text{MLP}(\alpha\hat{\mathbf{F}}^{CV(L)} || \beta\hat{\mathbf{F}}^{PV(L)} || \gamma\hat{\mathbf{F}}^{EV(L)}), \end{cases} \tag{7}$$

Interaction between Visual and Gestalt Features

For how to enhance diagram representation with the help of gestalt perception, we have improved the approach in our pervious work, which simply interacts gestalt feature with diagram feature extracted from backbone network, without considering the feature difference of different blocks in backbone. As gestalt laws simulate the perception process of human visual system, our GPA model utilizes the gestalt feature $\mathbf{F}^{ges}$ to enhance low-level diagram feature $\mathbf{F}^{D_l}$, that is, the diagram feature based on the gestalt perception is generated with the multi-head attention module as shown in (8), where d_k is the projection channel dimension. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ represent the parameters of linear projection. Finally, a convolutional network $\psi(\cdot)$ acts on $\mathbf{F}^{gesD_l}$ to generate the final gestalt guided diagram feature $\mathbf{F}^{gesD}$. S indicates the Softmax function.

$$\mathbf{F}^{gesD} = \psi(\text{S}(\frac{(\mathbf{F}^{D_l}\mathbf{W}_Q)(\mathbf{F}^{ges}\mathbf{W}_K)^\top}{\sqrt{d_k}})(\mathbf{F}^{ges}\mathbf{W}_V)). \tag{8}$$

3.2 Enrich High-level Semantic Feature

Multi-modal Context Attention. The diagram contains a variety of contents in foreground, and the corresponding language expression emphasizes the local object. Therefore, the

interaction between the language expression and the visual information is conducive to making the GPA model easier to recognize and locate the referred object. Firstly, the correlation between high-level visual feature $\mathbf{F}^{D_h}$ and linguistic embedding $\mathbf{F}^T$ is calculated by multi-head attention as shown in (9). Here, the diagram feature $\mathbf{F}^{D_h}$ serves as the query and the linguistic embedding $\mathbf{F}^T$ acts as the key and value. The diagram feature guided by semantic contexts is $\mathbf{F}^{textD}$.

$$\mathbf{F}^{textD} = \mathrm{S}(\frac{(\mathbf{F}^{D_h}\mathbf{W}_Q)(\mathbf{F}^T\mathbf{W}_K)^\top}{\sqrt{d_k}})(\mathbf{F}^T\mathbf{W}_V). \tag{9}$$

Multi-modal Feature Refinement. A 3D-attention module acts on $\mathbf{F}^{gesD}$ and $\mathbf{F}^{textD}$, respectively. The process of 3D-attention is similar as [Woo *et al.*, 2018]. Taking $\mathbf{F}^{gesD}$ as an example, the first step is squeezing the spatial dimension of $\mathbf{F}^{gesD}$ by average-pooling and max-pooling simultaneously. Then a shared MLP$(\cdot)$ network is applied to the two pooling features. The channel attention map is generated after element-wise summation of pooling features and sigmoid operation. The second step is splicing the two pooling features along the channel dimension, and then applies a convolutional block and a sigmoid function to compute the spatial attention map $\mathbf{F}_a^{gesD}$. $\mathbf{F}_a^{textD}$ is calculated in the same way to $\mathbf{F}_a^{gesD}$, and is used to represent the attention of diagram features under the effect of linguistic embedding.

Subsequently, $\mathbf{F}_a^{gesD}$ for visual feature and $\mathbf{F}_a^{textD}$ for semantic feature are multiplied and then mapped through MLP$(\cdot)$ network to obtain the attention score denoted as Score_D. At this time, the diagram feature $\hat{\mathbf{F}}^D$ of multi-modal enhancement is shown in (10).

$$\begin{cases} \mathrm{Score}_D = \mathrm{MLP}(\mathbf{F}_a^{gesD} \otimes \mathbf{F}_a^{textD}), \\ \hat{\mathbf{F}}^D = \mathbf{F}^{gesD} + \mathbf{F}^{textD} + \mathrm{Score}_D \cdot \mathbf{F}^{D_h}. \end{cases} \tag{10}$$

In order to associate the language expression with the location in the diagram, the position view in GPG is used to enhance the embedding of language expression. Through the multiplication of linguistic embedding $\mathbf{F}^T$ and position feature $\hat{\mathbf{F}}^{PV}$, and then a softmax function is used to calculate the attention score of linguistic embedding on position feature. The enhanced linguistic embedding is termed as $\mathbf{F}^{posT}$.

3.3 Target Query Decoder

Spatial and Semantic Association

The decoding process aims to filter out the object feature referred to the language expression and generate position coordinates under the joint action of diagram feature and linguistic embedding. Based on the decoder structure of Transformer [Carion *et al.*, 2020], we employ a learnable target query $t_q \in \mathbb{R}^{1\times d}$ as the initial representation of the referred object. For each decoder layer, t_q is decoded by textual and visual information in turn. The final target query is directly used to predict the coordinates.

For the interaction of target query and linguistic embedding. ${t'}_q$ indicates that giving more informative semantics to the target query through the multi-head attention between t_q and $\mathbf{F}^{posT}$, where the t_q serves as the query and the $\mathbf{F}^{posT}$ acts as the key and value.

For the interaction of target query and visual feature. To accurately locate the referred object on the diagram, another multi-head attention module is used that taking ${t'}_q$ as a query to calculate its correlation with the diagram feature $\hat{\mathbf{F}}^D$. In this manner, the target feature ${t''}_q$ is produced to locate the referred object. Concretely, the GPA model makes the final prediction with a feed-forward neural network FFN$(\cdot)$, which is composed of a three-layer MLP$(\cdot)$ network with ReLU activation. For simplicity, the $a(\cdot)$ in (11) indicates the attention operation in multi-head attention module.

$$\begin{cases} {t'}_q = a(t_q, \mathbf{F}^{posT}) \cdot \mathbf{F}^{posT}, \\ {t''}_q = a({t'}_q, \hat{\mathbf{F}}^D) \cdot \hat{F}^D, \\ output = \mathrm{FFN}({t''}_q) = \mathrm{ReLU}(\mathrm{MLP}({t''}_q)). \end{cases} \tag{11}$$

Training Objective

Unlike most two-stage methods, which regard the location of target object as a ranking problem of candidate proposals, the GPA model directly projects the decoded target query representation into a 4-dimensional position feature, which consists of the center coordinates (x_b, y_b), the width w_b, and the height h_b of the predicted box. For diagram D and language expression T, assume that b denotes the prediction while $\hat{b}$ is the ground truth box, the training objective of our GPA is shown in (12). $\mathcal{L}_{giou}$ and $\mathcal{L}_{L_1}$ are the GIoU loss [Rezatofighi *et al.*, 2019] and L1 loss respecitively. λ_{giou} and λ_{L_1} are the balance factors between the two losses.

$$\mathcal{L} = \lambda_{giou} \cdot \mathcal{L}_{giou}(b, \hat{b}) + \lambda_{L_1} \cdot \mathcal{L}_{L_1}(b, \hat{b}). \tag{12}$$

4 Experiments

4.1 Datasets

We evaluate the GPA model on an AI2D$^\diamond$ dataset with diagrams and a Flickr30K Entities dataset with natural images.

AI2D$^\diamond$. As there is no research related to diagram visual grounding at present, we have annotated a novel dataset AI2D$^\diamond$ on the basis of original AI2D dataset to verify the performance of this task. AI2D is a dataset focusing on the scientific topic of primary and secondary schools, mainly composed of diagrams in the fields of biology, astronomy and geography. We draw a rectangular bounding box for each object in the diagram, and assign corresponding language expressions to all objects. AI2D$^\diamond$ is a more challenging dataset than Flickr30k Entities, containing only 2,038 diagrams and about 132k referred targets. We split AI2D$^\diamond$ into a train set with 1,634 diagrams and a test set with 404 diagrams.

Flickr30k Entities. In addition to diagram visual grounding, the framework of GPA model is also applicable to processing natural images. To this end, we select a benchmark Flickr30k Entities [Plummer *et al.*, 2015]. It has 31,783 images and 427k referred targets. We follow the same split as in the previous works [Deng *et al.*, 2021; Yang *et al.*, 2022] for training, validation, and testing.

Models	Backbone	Acc
ZSGNet [Sadhu *et al.*, 2019]	ResNet-50	15.31
ReSC-Large [Yang *et al.*, 2020]	DarkNet-53	23.93
TransVG [Deng *et al.*, 2021]	ResNet-50	12.50
VLTVG [Yang *et al.*, 2022]	ResNet-50	25.79
GPA(Ours)	ResNet-50	**27.67**
VLTVG [Yang *et al.*, 2022]	ResNet-101	28.11
GPA(Ours)	ResNet-101	**30.54**

Table 2: The accuracy (Acc %) results on AI2D$^{\diamond}$ test set.

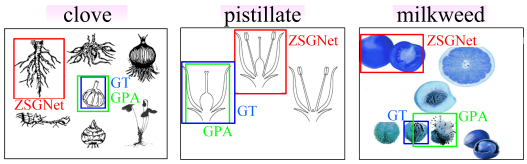

Figure 4: Case comparison between GPA and ZSGNet. "GT" represents the ground truth box.

4.2 Experimental Settings

GPA Implementation. Our model is implemented using PyTorch. For fair comparison, we resize the visual input into $640 \times 640 \times 3$ and follow the pervious works [Deng *et al.*, 2021; Yang *et al.*, 2022] to perform data augmentation. The maximum length of the language expression is set to 40. When extracting the linguistic embedding, we append the [CLS] and [SEP] tokens to the head and tail of the language expression, respectively. The convolutional neural network ResNet50 is used to extract the global visual feature, while the linguistic embedding is initialized with BERT.

Training and Evaluation. When training the GPA model, we use Adam for parameter optimization with an initial learning rate of 10^{-4}. We set the learning rate for visual backbone network and linguistic BERT to 10^{-5} and the weight decay is 10^{-4}. For comparison with the baseline models, we extend the training epochs to 90, and decay the learning rate by 10 after 60 epochs. In Eq. (10), we set 0.5 as the initial value of α, β, and γ. To avoid overfitting, we exploit dropout operation after the multi-head attention layer and the dropout rate is set to 0.1 by default. The evaluation of GPA model is in the same way as VLTVG [Yang *et al.*, 2022] and we set $\lambda_{giou} = 2$ and $\lambda_{L_1} = 5$.

4.3 Analyses of Diagram Visual Grounding

Performance Comparison. We compare the performance of the proposed GPA model with several state-of-the-art competitors. Table 2 shows that the GPA achieves the best accuracy. Specifically, the accuracy of GPA with ResNet-101 is 2.43% higher than the latest VLTVG model. In addition, there are some language expressions in the test set of AI2D$^{\diamond}$ dataset that have never appeared during training. ZSGNet, as a special model to deal with the unseen language expression in the reasoning, its accuracy of diagram visual grounding is only 15.31%. Figure 4 shows three cases of ZSGNet and GPA respectively. The *clove*, *pistillate*, and *milkweed* referred objects don't belong to these categories pre-defined in train set. For *clove* and *pistillate*, our GPA model accurately predicts the target boxes compared with ZSGNet. For *milkweed*, although the prediction by GPA is biased from the ground truth box, compared with the ZSGNet, GPA locates the region as similar as possible to the referred object.

Models	$Attn^G$	$Attn^T$	$Attn^{PV}$	Acc
BASE	-	-	-	17.33
GPA$^{\sharp 1}$		$\surd$		26.12
GPA$^{\sharp 2}$	$\surd$			26.26
GPA$^{\sharp 3}$	$\surd$	$\surd$		27.19
GPA	$\surd$	$\surd$	$\surd$	**27.67**

Table 3: The accuracy (Acc %) comparison of various ablation models. $Attn^G$ and $Attn^T$ indicate that 3D-attention is applied to $\mathbf{F}^{gesD}$ and $\mathbf{F}^{textD}$ respectively. $Attn^{PV}$ denotes the attention interaction between position view (PB) and linguistic embedding.

Ablation of Key Modules. To verify the contribution of each module in the GPA, we study ablation models and the differences between these versions are shown in Table 3. 1) BASE is a basic model that directly adopts gestalt-perception graph to guide diagram feature F^{gesD} and linguistic embedding to guide F^{textD} to predict the target object, and does not use any strategy to refine the diagram features. 2) GPA$^{\sharp 1}$ and GPA$^{\sharp 2}$ indicate that the 3D-attention mechanism of refining $F^{gesD}$and F^{textD} is adopted respectively. 3) GPA$^{\sharp 3}$ means that on the basis of BASE, more discriminative diagram features F^{gesD} and F^{textD} are simultaneously obtained by the 3D-attention mechanism. 4) Compared with GPA$^{\sharp 3}$, the proposed GPA model adopts gestalt position view to guide the linguistic embedding.

Table 3 shows that when gestalt feature or linguistic embedding is used separately to refine the diagram feature, GPA$^{\sharp 1}$ and GPA$^{\sharp 2}$ achieve comparable accuracy, 26.12% and 26.26% respectively. When $Attn^G$ and $Attn^T$ attention mechanisms are used simultaneously, that is, from the visual perspective of gestalt-perception and the semantic perspective of linguistic embedding to jointly enhance diagram representation, the accuracy of GPA$^{\sharp 3}$ has increased by 0.93% compared with GPA$^{\sharp 2}$. In order to effectively locate the language expression on the diagram, GPA adds the attention mechanism of gestalt position view (PV) and linguistic embedding on the basis of GPA$^{\sharp 3}$. The accuracy of GPA is 0.48% higher than that of GPA$^{\sharp 3}$.

Visualization of Diagram Feature Map. The visualization results of the diagram feature map generated by the multimodal attention are shown in Figure 5. The dark red area represents the target object that language expression refers to. On the left are some correct cases, from which we can see that the GPA model accurately aligns the language expressions, such as *the luminous bulb*, with the corresponding objects in the diagrams. There are some error cases in the dotted box. It is found that when there are multiple objects with similar appearance but different semantics in the diagram, our GPA cannot effectively distinguish the target object. Taking the first case as an example, the *tree* and *shrub* in the diagram are drawn by irregular curves that are approximately circular. When the language expression is *shrub*, the GPA model is incorrectly positioned to the *tree* in the diagram.

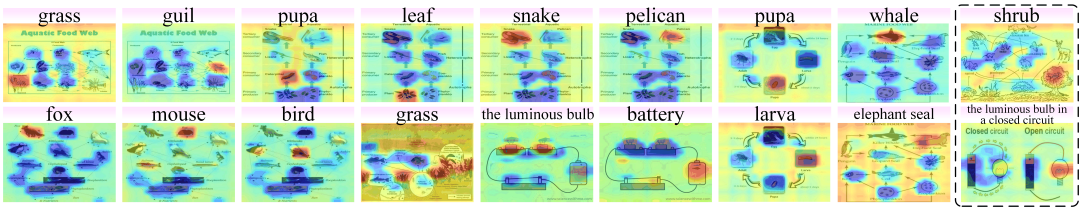

Figure 5: Visualization of diagram feature map guided by the language expression.

Models	Acc
GPG-C	24.25
GPG-P	25.69
GPG-E	25.74
GPG-CPE	25.32
GPG-CPE w/ Adaptive	**27.19** (↑ 1.87)

Table 4: The effect of gestalt laws on the accuracy (Acc %).

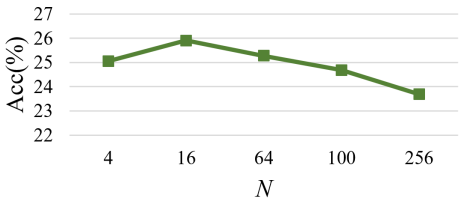

Figure 6: The effect of diagram patch number N.

In Table 4, GPG-C, GPG-P and GPG-E respectively indicate that only similarity, proximity and smoothness laws are used for GPG updating. GPG-CPE denotes that the three laws work together. Note that GPG-CPE simply adds the edge weights A^{CV}, A^{PV}, and A^{EV} of the three subgraphs $\mathcal{G}^{CV}$, $\mathcal{G}^{PV}$, and $\mathcal{G}^{EV}$. The experimental results show that when three gestalt laws are simply used at the same time, the accuracy of GPG-CPE is not much different from that of a single law. When α, β, and γ are used in an adaptive manner, the accuracy of GPG-CPE w/Adaptive is 1.87% higher than the GPG-CPE model, because the adaptive factors can effectively alleviate the bias problem of the gestalt laws.

Impact of Gestalt Laws in GPG. Adaptive gestalt factors α, β, and γ play key roles in the GPG network updating. As we all know, when human visual system perceives image patches, many gestalt laws work together. For example, if two patches have similar colors, but they are far away in the image, the probability that these two patches belong to the same object is low. For another example, if two patches are close to each other and can form a smooth contour, they may belong to the same object even if their colors are different. Inspired by this, GPG designs three adaptive factors to control the contribution of similarity, proximity and smoothness laws respectively.

Impact of Diagram Patch Number. To verify the impact of the number of diagram patches N, GPA is modified to only use N as a variable. N is set to 4, 16, 64, 100 and 256 respectively, and the experimental results are shown in Figure 6. GPA achieves the optimal result when $N = 16$. Subsequently, as N increases, performance gradually decreases.

Type	Models	Backbone	Acc
Two-stage	CITE	ResNet-101	61.33
	DDPN	ResNet-101	73.30
	Pseudo-Q	ResNet-101	60.41
One-stage	MultiG	PNASNet	69.19
	ZSGNet	ResNet-50	63.39
	ReSC	DarkNet-53	69.28
	TransVG	ResNet-50	78.47
	VLTVG	ResNet-50	**79.18**
One-stage	GPA(Ours)	ResNet-50	76.26

Table 5: The accuracy (Acc %) results on Flickr30K Entities test set.

4.4 Analyses of Natural Image Visual Grounding

The gestalt-perception graph network designed in the GPA model aims to simulate a series of gestalt laws that the human visual system follows when perceiving objects in the diagram. Of course, these laws are also applicable to human understanding of natural images. Therefore, we also conduct the experiments in natural image visual grounding to verify the effectiveness of GPA model.

Table 5 shows that the accuracy of the GPA model is 2.96% higher than that of the two-stage model DDPN [Yu *et al.*, 2018]. Compared with latest one-stage models, our GPA also achieves comparable result.

5 Conclusion and Future Work

In this paper, we conduct the diagram visual grounding with a gestalt-perceptual attention model GPA. Gestalt perception simulates human visual system, which can effectively identify key objects in images with limited samples. As the visual features of diagrams are sparse and there exist semantic gaps with language expressions, the GPA model designs a gestalt-perception graph network to learn the local visual representation of the diagram. In addition, there is a multimodal context attention mechanism to enhance the semantics of the diagram, thereby promoting accurate grounding on the diagram with the target language expression.

Although the GPA model reflects the effectiveness of the diagram visual grounding, we need to explore a variety of gestalt laws to deal with more complex diagrams in the future, not just the most conventional similarity, proximity and smoothness laws utilized in this work.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (62137002, 62192781, 62293553, 61937001, 62250066, and 62106190), Innovative Research Group of the National Natural Science Foundation of China (61721002), "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, CCF-Lenovo Blue Ocean Research Fund, Foundation of Key National Defense Science and Technology Laboratory (6142101210201), Natural Science Basic Research Program of Shaanxi (2023-JC-YB-293), the Youth Innovation Team of Shaanxi Universities, the Fundamental Research Funds for the Central Universities (xhj032021013-02), XJTU Teaching Reform Research Project "Acquisition Learning Based on Knowledge Forest".

References

[Carion *et al.*, 2020] 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.

[Deng *et al.*, 2021] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[Desolneux *et al.*, 2004] Agne Desolneux, Lionel Moisan, and Jean-Michel Morel. Gestalt theory and computer vision. In *Seeing, Thinking and Knowing*, pages 71–101. Springer, 2004.

[Dosovitskiy *et al.*, 2020] 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. In *International Conference on Learning Representations*, 2020.

[Gnjatović *et al.*, 2022] Milan Gnjatović, Ivan Košanin, Nemanja Maček, and Dušan Joksimović. Clustering of road traffic accidents as a gestalt problem. *Applied Sciences*, 12(9):4543, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hong *et al.*, 2019] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[Hu *et al.*, 2021] Xin Hu, Lingling Zhang, Jun Liu, Qinghua Zheng, and Jianlong Zhou. Fs-dsm: Few-shot diagram-sentence matching via cross-modal attention graph model. *IEEE Transactions on Image Processing*, 30:8102–8115, 2021.

[Hu *et al.*, 2022] Xin Hu, Lingling Zhang, Jun Liu, Jinfu Fan, Yang You, and Yaqiang Wu. Gptr: Gestalt-perception transformer for diagram object detection. *arXiv preprint arXiv:2212.14232*, 2022.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Li and Sigal, 2021] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34:19652–19664, 2021.

[Liao *et al.*, 2020] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.

[Liu *et al.*, 2021] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5612–5621, 2021.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[Pomerantz *et al.*, 1977] James R Pomerantz, Lawrence C Sager, and Robert J Stoever. Perception of wholes and of their component parts: some configural superiority effects. *Journal of Experimental Psychology: Human Perception and Performance*, 3(3):422, 1977.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[Ren *et al.*, 2015] 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.

[Rezatofighi *et al.*, 2019] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[Sadhu *et al.*, 2019] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4694–4703, 2019.

[Stricker and Orengo, 1995] Markus Andreas Stricker and Markus Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases III*, volume 2420, pages 381–392. SPiE, 1995.

[Wagemans *et al.*, 2012] Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A Van der Helm, and Cees Van Leeuwen. A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations. *Psychological Bulletin*, 138(6):1218, 2012.

[Wang and Specia, 2019] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4663–4672, 2019.

[Wang *et al.*, 2019] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019.

[Wertheimer, 1922] Max Wertheimer. Untersuchungen zur lehre von der gestalt. *Psychologische Forschung*, 1(1):47–58, 1922.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Xu *et al.*, 2019] Lijuan Xu, Zhihang Ji, Laura Dempere-Marco, Fan Wang, and Xiaopeng Hu. Gestalt-grouping based on path analysis for saliency detection. *Signal Processing: Image Communication*, 78:9–20, 2019.

[Yan *et al.*, 2018] Yijun Yan, Jinchang Ren, Genyun Sun, Huimin Zhao, Junwei Han, Xuelong Li, Stephen Marshall, and Jin Zhan. Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognition*, 79:65–78, 2018.

[Yan *et al.*, 2021] Ke Yan, Xiuying Wang, Jinman Kim, Wangmeng Zuo, and Dagan Feng. Deep cognitive gate: Resembling human cognition for saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4776–4792, 2021.

[Yang *et al.*, 2019] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019.

[Yang *et al.*, 2020] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020.

[Yang *et al.*, 2022] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508, 2022.

[Ye *et al.*, 2021] Jiabo Ye, Xin Lin, Liang He, Dingbang Li, and Qin Chen. One-stage visual grounding via semantic-aware feature filter. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1702–1711, 2021.

[Ying *et al.*, 2018] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in Neural Information Processing Systems*, 31, 2018.

[Yu *et al.*, 2018] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1114–1120, 2018.

[Zhang *et al.*, 2022a] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Zhihui Li, Lina Yao, and Alex Hauptmann. Tn-zstad: Transferable network for zero-shot temporal activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Zhang *et al.*, 2022b] Lingling Zhang, Shaowei Wang, Jun Liu, Qika Lin, Xiaojun Chang, Yaqiang Wu, and Qinghua Zheng. Mul-grn: Multi-level graph relation network for few-shot node classification. *IEEE Transactions on Knowledge and Data Engineering*, 2022.