SkyFind: A Large-Scale Benchmark Unveiling Referring Expression Comprehension for UAV

Kunyu Wang†, Xingbo Wang†, Kean Liu†, Xin Lu, Chengjie Ge, Wei Zhai, Xueyang Fu and Zheng-Jun Zha*

Abstract—The ability of unmanned aerial vehicles (UAV) to comprehend and follow human instructions is pivotal, enhancing both usability and efficiency. To propel research in human-UAV interaction, we introduce SkyFind, a new task that accurately localizes targets specified by human instructions within UAV-captured images, aiming to provide a more efficient and user-friendly UAV experience. To support and implement the SkyFind task, we construct a large-scale SkyFind dataset comprising one million high-quality object-expression pairs. We source images from both publicly available UAV datasets and a diverse range of UAV data mined from the internet to ensure scale and diversity. We employ large models as assistants for pre-annotation to reduce the annotation workload from scratch, followed by manual re-annotation to ensure high-quality labeling. In addition, we propose a baseline method, AerialREC, which tackles the unique challenges of complex background interference in the SkyFind task by initially narrowing down the potential target area. We conduct extensive experiments with various representative REC methods and our baseline, establishing an initial benchmark and verifying the effectiveness of our AerialREC.

Index Terms—Unmanned Aerial Vehicles, Referring Expression Comprehension, Human-UAV Interaction, Benchmark.

1 Introduction

Unmanned Aerial Vehicles (UAV) equipped with cameras serve a myriad of roles in human society, owing to their versatility and efficiency across various tasks [1]–[6]. As UAV become increasingly integral to human endeavors, their ability to understand human language and execute instructions assumes paramount importance for societal advancement. This seamless human-UAV interaction not only enhances user experience but also optimizes the efficiency of human-led tasks involving UAV.

Despite significant advancements in human-UAV interaction [7]–[9], a critical yet often overlooked task remains: the precise localization of human-specified targets within UAV-captured images. This capability is particularly useful in diverse applications, such as UAV search and rescue operations [10], [11]. UAV capture thousands of images across vast disaster areas. Given the large-scale scenes and the sheer volume of data, locating survivors or critical assets within these images is a highly challenging and time-consuming task for rescue personnel. By leveraging this capability, rescue teams can be effectively assisted in locating specific targets based on free-form instructions, thereby reducing manual workload, enhancing efficiency, and expediting response in rescue efforts. Therefore, this unexplored area presents a pivotal opportunity for further research and development in human-UAV interaction.

To facilitate research in this realm and bridge this gap, we introduce SkyFind, a UAV-based referring expression comprehension task. As depicted in Fig. 1, the users provide a free-form textual description, which the UAV comprehend and localize the specified target, then respond by indicating the location. By incorporating more details in free-form description, such as spatial relationships, or object attributes, potential ambiguities of specified target can be alleviated. Consequently, SkyFind allows users to retrieve objects in UAV-captured images using natural language instructions, realizing human-UAV interaction and offering a more user-friendly experience.

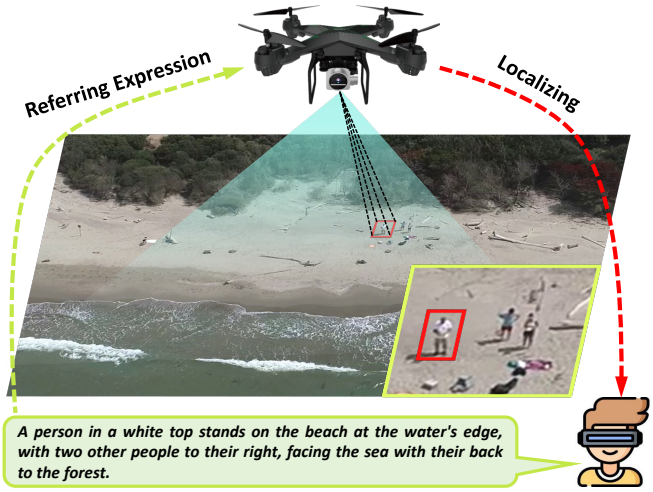

Fig. 1: Demo of the SkyFind task. Users provide textual descriptions of a specified target, and the UAV responds by indicating the specified target's location.

To support and implement the SkyFind task, we construct a new large-scale SkyFind dataset comprising one million high-quality object-expression pairs. For image data sourcing, we both collect publicly available UAV datasets and mine more web-based UAV data from the internet to ensure scale and diversity. During the annotation process, leveraging recent advancements in perception and under-

The authors are with the School of Information Science and Technology and MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, 230026, China.
† *Joint first authors.* * *Corresponding author.*

TABLE 1: Comparative analysis with existing REC datasets. 'Avg Length' denotes average expression length, 'Avg Img Res' denotes average image resolution, 'Avg O/I Ratio' denotes average object area to image area ratio, 'Anno' denotes annotation method, 'LM' denotes large models.

Dataset	Aera	Img	Obj	Expr	Avg Length	Expr Type	Avg Img Res	Avg O/I Ratio	Audio	Anno
ReferIt [12]	Life	19,894	96,654	130.5K	3.46	Free	485×592	14.65 %	✗	Manual
RefCOCO [13]	Life	26,711	50,000	142.2K	3.61	Free	480×583	9.05 %	✗	Manual
RefCOCO+ [13]	Life	19,992	49,856	141.5K	3.53	Free	485×592	8.82 %	✗	Manual
RefCOCOg [14]	Life	26,711	54,822	85.7K	8.93	Free	480×583	8.93 %	✗	Manual
CLEVR-Ref+ [15]	Synthetic	85,000	492,727	998.7K	22.40	Free	480×320	2.70 %	✗	Simulator
Talk2Car [16]	Vehicle	9,217	10,519	11.9K	11.01	Free	1600×900	3.65 %	✗	Manual
Cops-Ref [17]	Life	75,299	148,712	148.7K	14.40	Template	521×432	13.10 %	✗	Manual
KB-Ref [18]	Life	16,917	43,284	43.2K	13.32	Free	500×413	18.20 %	✗	Manual
gRefCOCO [19]	Life	19,994	80,287	278.2K	13.22	Free	485×592	8.82 %	✗	Manual
SK-VG [20]	Movie	4,000	39,182	39.1K	4.46	Free	1601×848	12.00 %	✗	Manual
InDET [21]	Life	120,604	908,410	3.6M	6.20	Free	456×548	15.94 %	✗	LM
SkyFind	UAV	34,548	346,635	1.0M	32.45	Free	2502×1468	0.48 %	✓	LM+Manual

standing by large models, we initially utilize large models as assistants to pre-annotate the data, thereby reducing the workload of manual annotation from scratch. Then, we manually refine and re-annotate the pre-annotations to ensure the high quality of the object-expression data. In addition to providing essential textual referring expressions, we also supply audio descriptions generated using text-to-speech software. This encourages explorations that integrate visual, linguistic, and audio features, thereby further enhancing the human-UAV interaction.

Compared to general REC, the SkyFind task poses notable challenges due to the complex and interfering backgrounds in UAV-captured images. Taken from high altitudes, these images offer expansive fields of view featuring diverse landscapes and objects rich in semantic content. This complexity greatly interferes precise target localization hidden within the intricate scenes. To tackle these challenges, we introduce AerialREC, a dedicated baseline method that incorporates an intermediate step to initially narrow down the potential target area, thereby reducing irrelevant background interference and providing a more focused area. Subsequently, guided by this narrowed area, we concentrate on a clearer sub-area with less interference to precisely localize the target location, thereby enhancing accuracy. We conduct extensive experiments with various representative REC methods and our AerialREC, establishing the initial benchmark and verifying the effectiveness of our baseline. We expect that the SkyFind dataset and our AerialREC baseline will significantly advance future research endeavors related to the SkyFind task.

Overall, the contributions of this paper are three-fold:

- We propose a new SkyFind task for precisely localizing specified targets in UAV-captured images, pushing general REC task into challenging yet practical UAV-based scenarios, advancing human-UAV interaction research and bridging the research gap.
- We propose a large-scale SkyFind dataset comprising one million object-expression annotations—three orders of magnitude larger than other comparable dataset, sourced from 13 publicly available UAV datasets and web-mined UAV data to ensure diversity, and supplemented with audio description annotations, making the first significant effort to support research on the SkyFind task.
- We propose AerialREC, a dedicated baseline method to tackle the unique challenges of complex and interfering backgrounds posed by the SkyFind task. Furthermore, we establish initial benchmark results that provides a solid foundation for future studies in this field.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of relevant literature. Section 3 describes the construction of the SkyFind dataset and provides a detailed statistical analysis of the dataset's characteristics. Section 4 presents our proposed baseline method, AerialREC. Section 5 offers the initial benchmark on our dataset, along with ablation studies and exploratory experiments. Finally, Section 6 concludes our work, discussing some possible future research directions.

2 Related Work

2.1 Referring Comprehension Datasets

In the deep learning era, benchmark datasets [22], [23] have become the critical infrastructure for the computer vision research community. Thanks to the publicly referring expression comprehension (REC) datasets, the REC task have evidenced notable progresses. ReferIt [12] is the first dataset comprising natural language expressions referring to objects in real-world scenes, pioneering the development of REC in diverse contexts. Subsequently, RefCOCO and RefCOCO+ [13] are introduced, both derived from the MSCOCO dataset [24], accompanied by concise phrase descriptions. RefCOCO imposes no restrictions on language expressions, whereas RefCOCO+ prioritizes purely appearance-based descriptions, prohibiting the use of location words. RefCOCOg [14], also originating from MSCOCO, stands out for its utilization of longer language expressions compared to its predecessors. CLEVR-Ref+ [15] emerges as a synthetic diagnostic dataset for REC, addressing bias issues, and facilitating the assessment of models' intermediate reasoning processes. Talk2Car [16] emerges as the first object referral dataset meticulously tailored for self-driving cars, providing natural language commands for actions related to urban street scene objects, built upon the nuScenes [25] dataset. Cops-Ref [17] introduces intricate and compositional expressions, challenging models to demonstrate complex reasoning abilities beyond simple object recognition, attributes, and relations. KB-Ref [18] pushes the boundaries of REC models by necessitating the incorporation of commonsense

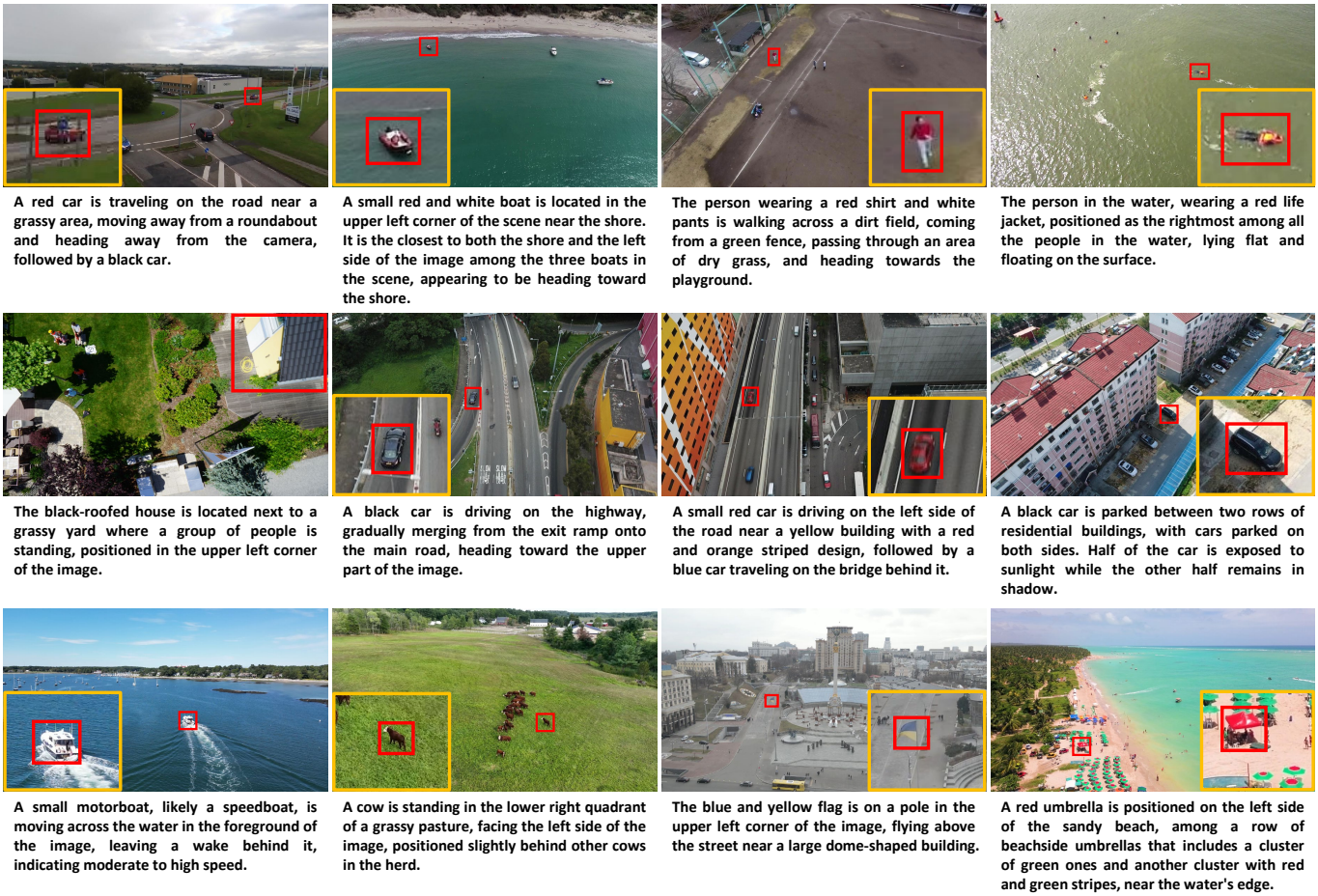

Fig. 2: A glance at the new large-scale SkyFind dataset, comprising over one million high-quality object-expression pairs.

knowledge in identifying referent objects. It encourages models to explore information not only from images but also from external knowledge. gRefCOCO [19] extends the classic REC by allowing expressions to describe any number of target objects, including multi-target expressions and no-target expressions. SK-VG [20] incorporates reasoning over scene knowledge, i.e., long-form text-based stories, alongside image content and referring expressions, necessitating models to process image, scene knowledge, query triples for comprehensive understanding. InDET [21] proposes a data generation pipeline that relies on foundational models to generate instructions, paving the way for enhancing data scale. However, these works primarily focus on ground-based scenarios, overlooking UAV-based scenarios. In this paper, we introduce the SkyFind dataset to address this research gap. Due to the nature of UAV-captured scenes, the SkyFind dataset features high shooting angles, high image resolution, complex and cluttered backgrounds, relatively small object sizes, and detailed referring expressions. In addition, our annotations utilize free-form descriptions and a semi-automatic labeling method to reduce workload, along with supplementary audio descriptions. Detailed comparisons with prior datasets are provided in Table 1. Some qualitative examples are shown in Fig. 2.

Beyond the above, WebUAV-3M [26] presents a dataset with annotations akin to our work. However, it's noteworthy that WebUAV-3M is tailored for the UAV tracking task, i.e., tracking targets in UAV videos. For each tracked object in the videos, it supplements this by furnishing a language specification, totaling approximately 4.5K object-expression annotations. In contrast, our proposed SkyFind dataset primarily focuses on the UAV-based REC task, comprising 1.0M pairs of object-expression annotations. The quantity, diversity, and comprehensiveness of both objects and expressions in our dataset far exceed WebUAV-3M.

2.2 Referring Comprehension Methods

REC predicts a bounding box that accurately encompasses the target object in an image based on a given referring expression. Early works often follow a two-stage pipeline. Specifically, two-stage models [13], [27]–[30] first detect the salient regions of an image and then treat the REC task as a region-expression ranking problem. Despite their considerable success, these two-stage methods exhibit significant drawbacks in terms of model efficiency and generalization. To address these issues, one-stage REC [31]–[41] has recently become a popular research direction. By eliminating the region detection and image-text matching steps inherent in multi-stage modeling, one-stage models significantly reduce inference time. However, they often exhibit sub-optimal REC performance compared to two-stage methods, primarily due to their limited reasoning

capabilities. In response, recent advancements have focused on enhancing the reasoning capabilities of one-stage REC. Various novel multi-modal networks have been proposed to improve the performance of one-stage REC models [42]–[50]. For example, ReSC [42] introduces a recursive sub-query construction framework to enhance one-stage visual grounding, overcoming limitations in modeling long and complex queries. TransVG [44] offers an elegant perspective for capturing intra- and inter-modality context uniformly, formulating visual grounding as a direct coordinates regression problem. More Recently, researchers have redefined REC as a sequence prediction task, leading to the development of several novel sequence-to-sequence frameworks. For example, SeqTR [46] represents the bounding box of the referent with a sequence of discrete coordinate tokens, which are predicted via a Transformer architecture. PolyFormer [49] formulates REC and REC tasks as a sequence-to-sequence prediction problem, generating sequential polygon vertices and bounding box corner points.

In addition to the aforementioned specialist REC models, recent advances in large-scale vision-and-language (V&L) pre-training have propelled generalist models [51]–[62] to achieve new state-of-the-art performance in REC. These models are pre-trained with millions or billions of image-text pairs and an extensive number of parameters, allowing them to excel across various downstream V&L tasks, including REC. For example, CogVLM [51], trained on publicly available image-text pairs from datasets totaling about 1.5 billion images, attains SOTA performance across 17 classic cross-modal benchmarks. LLaVA [52], highlighting its data and training efficiency, leverages only 1.2M publicly available data, yet achieves SOTA results across 11 benchmarks. In this paper, we first establish benchmark results by evaluating both representative specialist REC methods and powerful generalist methods on the proposed task and dataset. Subsequently, considering the challenges posed by the SkyFind task, we propose AerialREC, a baseline method that introduces an initial search step to narrow the potential target area, eliminating irrelevant background interference. This approach yields substantial improvements and provides a start point for future research.

3 Construction of SkyFind Dataset

3.1 Data Collection

The SkyFind dataset is primarily sourced from two channels: publicly available UAV datasets and UAV videos downloaded from the internet. The former comprises 13 UAV datasets [63]–[76], as outlined in Table 2. The latter comprises UAV videos primarily obtained from YouTube under Creative Commons licenses [1], utilizing keywords such as aerial video, aerial photography, and drone photography. We have collectively downloaded approximately 20k raw videos. For video-based UAV datasets and web-based UAV videos, we initially transform them into image data by extracting frames from videos and computing the P-Hash of images, thus reducing data redundancy. Subsequently, all obtained image data is converted to the HSV color space, and images with average brightness below or above a specific threshold are eliminated, thereby conducting initial screening for anomalous images. Finally, we conduct a manual screening of images, considering the following criteria: (1) We exclude ambiguous images, such as those with motion blur, or low resolution, as they fail to provide clear and abundant semantic information. (2) We exclude images not taken from the perspective of the UAV, such as overhead or high-angle shots, to maintain the dataset's UAV attributes. (3) We exclude images lacking valuable target, such as landscapes or fields, to ensure the dataset's efficacy.

1. https://creativecommons.org/licenses/

TABLE 2: Introduction to the 13 publicly available UAV datasets. 'Data' represents the data type, and 'Anno' represents the annotation type.

Dataset	Task	Data	Anno	Target class
DroneVehicle [63]	Detect	Image	Box	Car, Bus, Truck, Van, Freight Car
SeaDroneSea [64]	Detect	Image	Box	Swimmer, Floater, Life Jacket, Boats
VisDrone2019 [65]	Detect	Image	Box	Pedestrian, Person, Car, Van, Bus, Truck, Motor, Bicycle, Awningtricycle, Tricycle
AU-AIR [66]	Detect	Video	Box	Person, Car, Van, Truck, Bike, Motorbike, Bus, Trailer
CAPRK [67]	Count	Video	Box	Car
MOBDrone [68]	Detect	Video	Box	Person, Boat, Wood, Life Buoy, Surfboard
Okutama-Action [69]	Detect	Video	Box	Human
Stanford Drone [70]	Track	Video	Box	Pedestrians, Bikers, Golf Carts, Cars, Buses, Skateboarders
Semantic Drone [72]	Segment	Image	Mask	Vegetation, Dirt, Gravel, Rocks, Water, Pool, Person, Dog, Car, Bicycle, Roof, Wall, Fence, Window, Door
UAVid [73]	Segment	Image	Mask	Building, Road, Low Vegetation, Tree, Car, Human, Clutter
UDD [74]	Segment	Image	Mask	Vegetation, Building, Free Space
UVSD [75]	Segment	Image	Mask	Vehicle
Aeroscapes [76]	Segment	Video	Mask	Sky, Road, Vegetation, Car, Obstacle, Animal, Boat, Drone, Construction, Bike, Person

Based on the images above, we further extract objects. For images sourced from existing datasets, we rely on the provided object annotations. These annotations primarily include bounding box and mask annotations. To convert mask-based annotations into box-based annotations, we initially process the mask by category, encoding each pixel label corresponding to the target class. Then, we identify the contour points of connected regions for each class in the mask image. These contour points provide the necessary information to determine the bounding box coordinates. By sorting the horizontal and vertical coordinates of the contour points, we can extract the minimum and maximum values, thus standardizing all annotations into box-based formats. Additionally, we perform box filtering to remove any outliers, excluding those with top-left coordinates less than or equal to the bottom-right coordinates. For images sourced from the internet, we extract objects through manual annotation. The principle of annotation is to ensure that the objects possess clear semantics, descriptive value, and can be explicitly described. Finally, we obtain 34,538 and 346,635 high-quality images and objects.

Note that for all processes involving manual intervention, including screening, annotation, and re-annotation, we follow a three-step "Process, Verify, Re-process" workflow. We first divide the annotation team into three sub-teams and partition the data requiring manual intervention into three

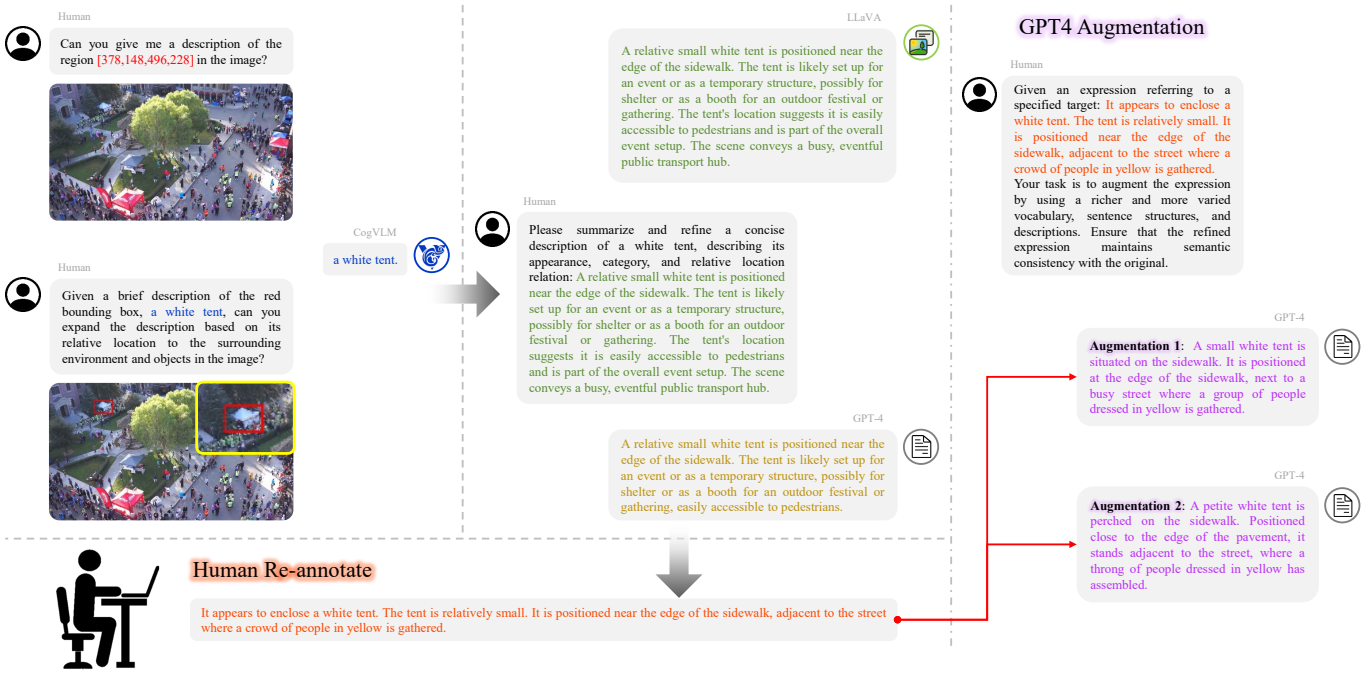

Fig. 3: The pipeline of the assisted manual annotation process, which includes CogVLM [51] for generating basic descriptions, LLaVA [52] for detailed descriptions, GPT-4 [77] for concise descriptions, human input for re-annotation, and GPT-4 for augmentation.

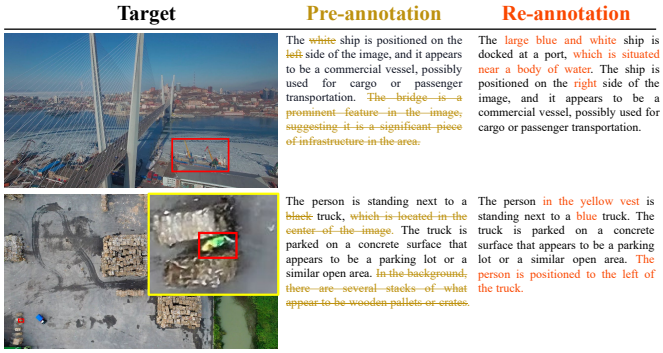

Fig. 4: Examples of pre-annotated expressions generated by large models compared to those re-annotated by humans.

parts. For each portion of the data, sub-team 1 conducts the initial processing, sub-team 2 performs verification, and sub-team 3 re-processes any data with issues. By following this workflow, the three sub-teams take turns processing the three portions of the data. For the annotation tool, we use the VGG Image Annotator (VIA) [2].

3.2 Assisted-Manual Annotation

The significant advancements demonstrated by large models have underscored their remarkable efficacy in proficiently executing a wide array of tasks. This includes tasks such as comprehension and reasoning for large language models, and grounding and captioning for large vision-language models. Motivated by these developments, we initially utilize large models as assistants to generate expressions for each object as pre-annotations, thereby reducing the workload of manual annotation from scratch. The pipeline of the annotation process is shown in Fig. 3.

Specifically, given the bounding box of an object, we first employ various box-to-caption task templates to generate prompts that include the bounding box information for each object. For example, "***Give me a description of the region <bbox> in the picture.***" These prompts are then input into CogVLM [51] to obtain a basic description for the specified object in the image. Next, we enhance these basic descriptions with the relative location of the target using LLaVA [52]. This facilitates more detailed referencing. For example, "***Given a brief description of the red bounding box, <basic description>, can you expand the description based on its relative location to the surrounding environment and objects in the image?***" Inspired by [78], we augment the images inputted to LLaVA with a visual prompt, outlining the target with a red bounding box to direct LLaVA attention to the specified object. Upon obtaining a detailed description, we utilize GPT-4 [77] to meticulously summarize and refine it, aiming to enhance the overall quality of the detailed description. We prompt GPT-4 to succinctly summarize, resulting in a concise description, such as: "***Please summarize and refine a concise description of <basic description>, describing its appearance, category, and relative location relation: <detailed description>.***"

However, the pre-annotations generated by large models may contain noise. A qualitative comparison between the pre-annotated expressions generated by the large models and those re-annotated by humans is provided in Fig. 4. We find that the pre-annotations still have a gap in accurately and concisely referring to the target. Therefore, we manually re-annotate and refine all the pre-annotations to ensure the high quality of the object-expression data. Furthermore, to generate diverse expressions and enrich the dataset with

2. https://www.robots.ox.ac.uk/~vgg/software/via/

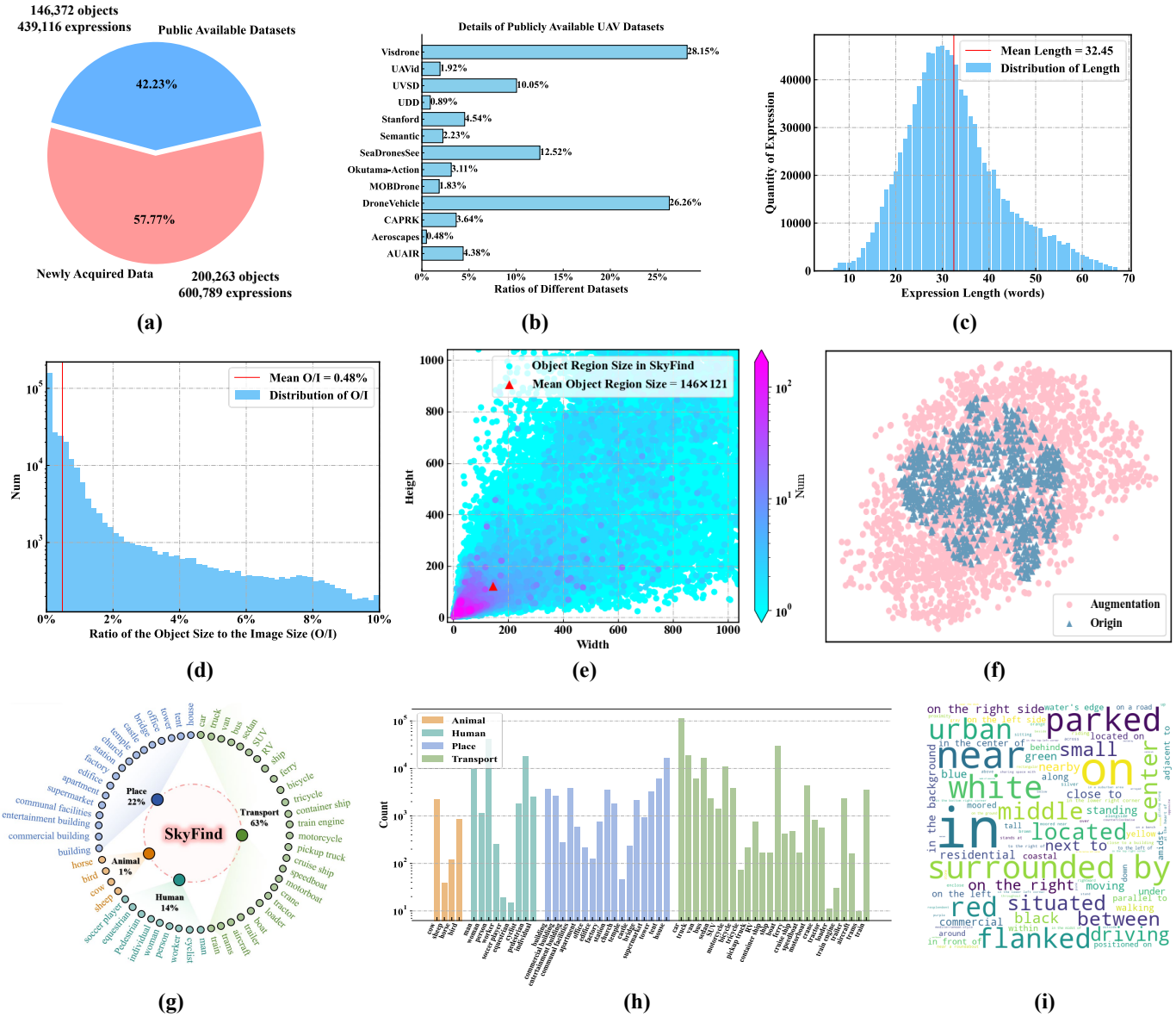

Fig. 5: Statistical analysis of the SkyFind dataset: (a) Proportion of newly acquired data from the internet and data from publicly available datasets. (b) Proportion of data from each of the 13 publicly available datasets. (c) Distribution of expression lengths. (d) Distribution of the average object-to-image area ratio. (e) Distribution of object dimensions (width and height). (f) T-SNE visualization of human re-annotated and GPT-4 augmented expressions. (g) Classification of specified objects, showing 4 super-classes and 56 classes. (h) Data counts for the 4 super-classes and 56 classes. (i) Word cloud of object attributes and relationships within SkyFind. For visual clarity, the negligible portions of the long-tail distribution have been excluded.

varied vocabulary, sentence structures, and descriptions, we use GPT-4 [77] to augment the re-annotations above. We prompt GPT-4 such as: "***Given an expression referring to a specified target: <re-annotation>. Your task is to augment the expression by using a richer and more varied vocabulary, sentence structures, and descriptions. Ensure that the refined expression maintains semantic consistency with the original.***" We generate two augmented expressions for each concise expression. Finally, we obtain 1,039,905 high-quality and diverse object-expression pairs.

In addition to the essential textual referring expressions, we also utilize text-to-speech software to produce corresponding audio files. This aims to encourage and facilitate explorations that integrate visual, linguistic, and audio features, thereby further enhancing human-UAV interaction and user experience. We utilize Google TTS AI [3] as our audio generation tool, integrating diverse tones, timbres, accents, and voice effects throughout the generation process to deliver comprehensive audio descriptions.

3.3 Dataset Split

The SkyFind dataset primarily consists of data from 13 publicly available UAV datasets and newly acquired UAV data

3. https://cloud.google.com/text-to-speech

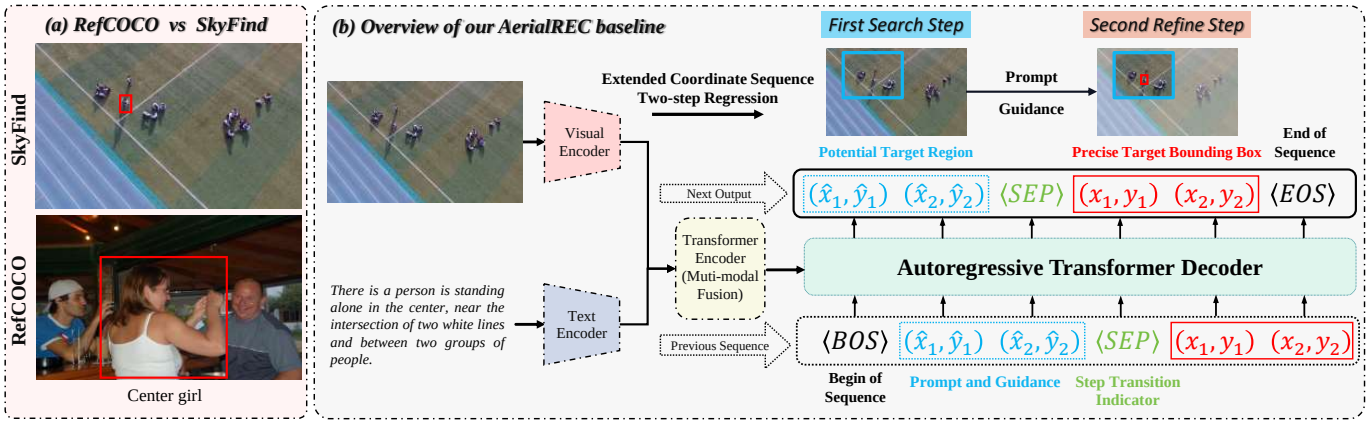

Fig. 6: (a) Comparison between general REC (RefCOCO) and our SkyFind. (b) Overview of our AerialREC baseline.

from the internet. We perform two types of train/test splits on the SkyFind dataset: intra-domain split and cross-domain split. Firstly, the intra-domain split involves randomly partitioning 5% of the SkyFind dataset, ensuring that all data domains are represented in both the training and testing sets, denoted as **TestA**. Secondly, the cross-domain split involves selecting data from the Semantic Drone dataset as the test set, while the remaining 12 publicly available datasets and newly acquired data constitute the train set. This division ensures non-overlapping data distributions between the train and test sets, facilitating the evaluation of the model's generalization performance, referred to as **TestB**. We conduct experiments on both of these splits.

3.4 Dataset Analysis

Our constructed AerialREC dataset consists of 1,039,905 object-expression pairs across 34,538 images and 346,635 objects, encompassing 4 super-classes and 56 classes of objects. The average length of the expressions is 32.45 words, and the vocabulary size is 10,097. We now present a more detailed statistical analysis.

Fig. 5 (a) shows the proportions of objects and expressions derived from newly acquired internet data and 13 publicly available datasets within the SkyFind dataset. Fig. 5 (b) further breaks down the data proportions from each of these 13 datasets. Fig. 5 (c) depicts the distribution of expression lengths, with an average of 32.45 words. Fig. 5 (d) illustrates the distribution of the average object-to-image area ratio, with an average ratio of 0.48%. Fig. 5 (e) presents the distribution of object dimensions (width and height), with an average size of 146 $\times$ 123 pixels. Fig. 5 (f) shows a T-SNE visualization of human re-annotated expressions and those augmented by GPT-4. We randomly sampled 1,000 original expressions and 2,000 corresponding augmented expressions, extracted features using BERT, and applied T-SNE for dimensionality reduction. The augmented data expands beyond the original distribution, indicating greater diversity in vocabulary and structure without significant semantic changes. Fig. 5 (g) provides statistics for objects specified by expressions, showing 4 super-classes and 56 classes, with their proportions highlighted. Fig. 5 (h) details the data counts for each of the 56 classes. Fig. 5 (i) displays a word cloud representing the terms describing object attributes and relationships in the SkyFind dataset.

4 AerialREC: A Baseline Approach

The SkyFind task presents unique challenges arising from the heightened complexity and interference within the UAV-captured scenes. The high-altitude perspective of UAV captures broad fields of view, encompassing diverse landscapes, numerous objects, and rich semantics. This results in highly complex and varied backgrounds, leading to significant interference that makes it challenging for current REC methods to accurately localize specified targets hidden within the intricate UAV-captured images, highlighting the difficulty of the SkyFind task, as shown in Fig. 6 (a).

To address the above challenges, we propose a new baseline method, AerialREC, which introduces an initial search step to first narrow down the potential target region, thereby eliminating irrelevant background interference. In the search step, we aim to identify the potential target region within complex UAV-captured scenes without the need for precise localization. This step preliminarily delineates the approximate target area, filtering out substantial irrelevant background interference and providing a more focused area for subsequent analysis. Next, in the refine step, guided by the narrowed potential region, we concentrate on a relatively clear sub-region with reduced interference and perform precise localization to pinpoint the exact target location. By employing our method, we decompose a challenging task into two steps. Compared to directly localizing the target, our method simplifies the process, thereby enhancing accuracy.

We implement our idea based on the recent seq2seq REC framework [46], [49] due to its simplicity and effectiveness, as shown in Fig. 6 (b). This framework formulates the REC task as a sequence modeling problem, generating the coordinates of the target bounding box top-left and bottom-right corners in an auto-regressive manner. It consists of four main components: a vision encoder that extracts visual features f_v from images, a text encoder that extracts textual features f_t from expressions, a transformer encoder that fuses these visual and textual features, generating the multi-modal features f_m, and a transformer decoder that successively regresses the entire coordinate sequence, with

each prediction conditioned on the multi-modal features f_m and the previously regressed coordinates. The regressed coordinate sequence is formulated as:

$$\left[\langle \text{BOS} \rangle, \{(x_i, y_i)\}_{i=1}^{2}, \langle \text{EOS} \rangle\right], \tag{1}$$

where $\{(x_i, y_i)\}_{i=1}^{2}$ represent the coordinates of the target bounding box top-left and bottom-right corners, $\langle \text{BOS} \rangle$ and $\langle \text{EOS} \rangle$ are special tokens to indicate the beginning and end of the sequence.

To address the challenges posed by complex and interfering backgrounds in the SkyFind task, we propose a two-step extension to the sequence regression process: search and refine. The extended regressed coordinate sequence can be formulated as:

$$\left[\langle \text{BOS} \rangle, \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{2}, \langle \text{SEP} \rangle, \{(x_i, y_i)\}_{i=1}^{2}, \langle \text{EOS} \rangle\right]. \tag{2}$$

In the search step, we initially regress the coordinates of the potential target region top-left and bottom-right corners $\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{2}$, aiming to filter out substantial irrelevant background interference, providing a more focused area:

$$[\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{2}] = \text{AR}([\langle \text{BOS} \rangle], f_m), \tag{3}$$

where AR denotes the auto-regressive transformer decoder. The potential target region does not need to precisely align with the target bounding box but should encompass the specified target. We formulate this potential target region label by randomly sampling from the enlarged intervals during training:

$$x_{\text{left}} = w - (x_2 - x_1), \quad y_{\text{left}} = h - (y_2 - y_1), \tag{4}$$

$$R_1 = [\max(0, x_1 - \alpha x_{\text{left}}), x_1] \times [\max(0, y_1 - \alpha y_{\text{left}}), y_1], \tag{5}$$

$$R_2 = [x_2, \min(w, x_2 + \alpha x_{\text{left}})] \times [y_2, \min(h, y_2 + \alpha y_{\text{left}})], \tag{6}$$

$$(\hat{x}_1, \hat{y}_1) \sim \text{RandomSample}(R_1), \tag{7}$$

$$(\hat{x}_2, \hat{y}_2) \sim \text{RandomSample}(R_2), \tag{8}$$

where $\{R_i\}_{i=1}^{2}$ denote the enlarged intervals, w and h denote the width and height of the images, α is a hyper-parameter that controls the size of the enlarged intervals. When α is zero, it provides precise alignment with no margin. As α increases towards one, the intervals enlarge while still remaining within the image bounds.

In the refine step, we further regress the coordinates of the precise target bounding box top-left and bottom-right corners $\{(x_i, y_i)\}_{i=1}^{2}$. Leveraging the auto-regressive nature of the transformer decoder, the regression of the next target bounding box $\{(x_i, y_i)\}_{i=1}^{2}$ is conditioned on the preceding potential target region $\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{2}$. This provides prompt and guidance, helping the decoder to focus on a clearer sub-region with reduced interference, thereby improving localization accuracy:

$$[\{(x_i, y_i)\}_{i=1}^{2}] = \text{AR}([\langle \text{BOS} \rangle, \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{2}, \langle \text{SEP} \rangle], f_m), \tag{9}$$

where $\langle \text{SEP} \rangle$ is a newly added special token that serves as an indicator to help the decoder understand the transition between two steps. During inference, we also employ the two-step regression, using the coordinates obtained from the refine step as the bounding box for the specified target.

5 Experiments

5.1 Implementation Details

In the process of generating pre-annotations with the assistance of large models, we utilize CogVLM [51] (CogVLM-grounding-generalist-v1.1-17B) for basic descriptions, LLaVA [52] (LLaVA-1.5-13B) for detailed descriptions, and GPT-4 (gpt-4-turbo) for concise descriptions and augmented expressions. For evaluation metrics, we use the standard IoU@0.5. Additionally, we introduce IoU@mean, a weighted average calculated across intervals from IoU@0.5 to IoU@0.9 in steps of 0.1, to better highlight the performance of precise target localization. The hyper-parameter α is set to 0.4 for our AerialREC baseline.

5.2 Benchmark Results

We categorize our evaluation methods into two distinct groups: specialists and generalists. Specialists refer to methods explicitly designed for visual grounding. We have selected 10 representative and open-sourced REC methods, namely FAOA [34], RSC [42], RefTR [43], TransVG [44], VGTR [47], VLVTG [48], SeqTR [46], QRNet [45], SimREC [50], PolyFormer [49]. We train and test the above specialist methods on our SkyFind dataset. We use the official code and training settings, following the best-performing training procedures. When pre-training is involved, we utilize the officially provided pre-trained weights. Moreover, with the rapid development of large models, a range of generalist models has emerged. These models are pre-trained on extensive datasets and exhibit proficiency not only in visual grounding task but also in various other domains. Consequently, our evaluation extends to encompass 12 influential generalist models: OFA [79], ONE-PEACE [58], Grounding-DINO [60], UNINEXT [55], Sphinx [56], Shikra [57], Qwen-VL [80], Qwen-VL [80], MiniGPT-V2 [59], Ferret [54], GLEE [53], LLaVA [52], CogVLM [51]. We conduct zero-shot evaluation to test the above generalist methods without further training, utilizing the official provided test code, procedures, prompts and weights.

The experimental results are shown in Table 3. We use IoU@0.5 as the primary metric for ranking the methods. Among all the trained specialists, SeqTR performs best on the intra-domain TestA, achieving 21.86% on IoU@0.5 and 13.98% on IoU@mean, slightly outperforming the second-place PolyFormer by 0.02% and 0.12%, and the third-place SimREC by 0.69% and 1.59%. This indicates that the top three methods perform similarly. On the cross-domain TestB, PolyFormer outperforms the others, surpassing the second-place SeqTR by 10.58% and 3.24%, and the third-place QRNet by 11.76% and 4.06%, demonstrating PolyFormer's superior generalization performance. Based on average metrics, the top three methods are PolyFormer, SeqTR, and QRNet. Among all the generalists tested in the zero-shot setting, CogVLM achieves the best performance on average metrics, with 21.09% on IoU@0.5 and 13.71% on IoU@mean, outperforming the second-place MiniGPT-V2 by 3.84% and 2.21%, and the third-place LLaVA by 3.86% and 3.55%. We observe that stronger visual and textual features in generalists lead to better generalization. However, despite extensive pre-training on large-scale general data and the large number of parameters, these generalists do not achieve

TABLE 3: Benchmark results of 10 specialist and 12 generalist REC methods on our SkyFind intra-domain TestA and cross-domain TestB sets, evaluated under both training and zero-shot settings. For our proposed AerialREC, we implement AerialREC based on the recent seq2seq REC methods, SeqTR and PolyFormer, referred to as AerialREC† and AerialREC♢, respectively. The gray-highlighted sections showcase the comparison between the original baseline performance and the results after integrating our AerialREC method, making it easier to compare the differences.

Type	Method	Year	Visual Feature	Text Feature	TestA		TestB		Average	
					IoU@0.5	IoU@mean	IoU@0.5	IoU@mean	IoU@0.5	IoU@mean
Specialist (Train)	FAOA [34]	2019	DarkNet-53	BERT	13.01	7.03	13.28	5.81	13.15	6.42
	RSC [42]	2020	DarkNet-53	BERT	14.71	8.26	14.59	6.37	14.65	7.32
	RefTR [43]	2021	ResNet-101	BERT	15.66	8.81	12.68	6.64	14.17	7.73
	TransVG [44]	2021	ResNet-101	BERT	15.49	8.56	12.11	5.12	13.80	6.84
	VGTR [47]	2022	ResNet-101	LSTM	15.13	9.68	14.61	6.77	14.87	8.23
	VLVTG [48]	2022	ResNet-101	BERT	20.29	10.48	15.52	10.41	17.91	10.45
	SeqTR [46]	2022	DarkNet-53	GRU	21.86	13.98	20.89	12.04	21.38	13.01
	QRNet [45]	2022	Swin-S	BERT	20.39	12.01	19.71	11.22	20.05	11.62
	SimREC [50]	2023	CSPDarkNet-53	LSTM	21.17	12.39	18.51	12.15	19.84	12.27
	PolyFormer [49]	2023	Swin-B	BERT	21.84	13.86	31.47	15.28	26.66	14.57
	AerialREC†	2024	DarkNet-53	GRU	**24.80**	**16.86**	**25.31**	**13.59**	**25.06**	**15.23**
	AerialREC♢	2024	Swin-B	BERT	**25.21**	**16.78**	**37.61**	**18.84**	**31.41**	**17.81**
Generalist (Zero-shot)	OFA [79]	2022	ResNet-152	—	12.61	8.06	12.54	7.30	12.58	7.68
	ONE-PEACE [58]	2023	—	—	12.88	6.26	10.19	5.95	11.54	6.11
	Grounding-DINO [60]	2023	Swin-B	BERT	14.36	9.01	13.66	8.39	14.01	8.70
	UNINEXT [55]	2023	ConvNeXt-L	BERT	12.78	8.44	13.94	8.04	13.36	8.24
	Sphinx [56]	2023	Mixed	LLaMA	15.69	11.05	16.16	9.83	15.93	10.44
	Shikra [57]	2023	OpenCLIP-L	Vicuna	13.92	9.50	12.52	7.20	13.22	8.35
	Qwen-VL [80]	2023	OpenCLIP-G	Qwen	15.17	10.88	13.97	8.85	14.57	9.87
	MiniGPT-V2 [59]	2023	EVA-CLIP-G	LLaMA	18.18	12.46	16.31	10.53	17.25	11.50
	Ferret [54]	2024	CLIP-L	Vicuna	17.52	10.14	15.72	6.48	16.62	8.31
	GLEE [53]	2024	EVA02-CLIP-L	CLIP	15.11	10.01	14.66	7.32	14.89	8.67
	LLaVA [52]	2024	OpenCLIP-L	Vicuna	18.72	11.85	15.73	8.47	17.23	10.16
	CogVLM [51]	2024	EVA02-CLIP-E	LLaMA	22.07	14.93	20.10	12.48	21.09	13.71

TABLE 4: Effect of components in constructing the extended regressed coordinate sequence, including two-step sequence extension, the special token ⟨SEP⟩, and strategies for formulating the potential target region. 'Fixed (inner)' uses the innermost value to match the target box precisely, while 'Fixed (outer)' extends the region to its maximum range.

Ablation	Two-step	⟨SEP⟩	Strategy	TestA		TestB		Average	
				IoU@0.5	IoU@mean	IoU@0.5	IoU@mean	IoU@0.5	IoU@mean
AerialREC†			—	21.86	13.98	20.89	12.04	21.38	13.01
	✓		RandomSample	23.91	16.45	22.98	13.73	23.45	15.09
	✓	✓	Fixed (inner)	22.00	13.95	20.31	10.94	21.16	12.45
	✓	✓	Fixed (outer)	23.12	14.96	20.89	11.33	22.01	13.15
	✓	✓	RandomSample	24.80	16.86	25.31	13.59	25.06	15.23
AerialREC♢			—	21.84	13.86	31.47	15.28	26.66	14.57
	✓		RandomSample	25.21	16.48	37.15	18.59	31.18	17.54
	✓	✓	Fixed (inner)	22.18	13.81	33.71	15.60	27.95	14.71
	✓	✓	Fixed (outer)	21.86	13.77	33.19	16.09	27.53	14.93
	✓	✓	RandomSample	25.21	16.78	37.61	18.84	31.41	17.81

satisfactory generalization in UAV-specific scenarios. Overall, there remains significant room for improvement for both specialists and generalists on the SkyFind dataset.

To further improve performance on the SkyFind dataset, we proposed a dedicated method called AerialREC to address the unique challenges posed by the heightened complexity and interference within UAV-captured scenes. We implement our AerialREC based on the recent seq2seq REC methods, SeqTR and PolyFormer, which we refer to as AerialREC† and AerialREC♢, respectively. Compared to SeqTR, AerialREC† shows improvements of 2.94% and 2.97% in IoU@0.5 and IoU@mean, respectively, on the intra-domain TestA, and 4.42% and 1.55% on IoU@0.5 and IoU@mean on the cross-domain TestB, with an average improvement of 3.68% and 2.26%. Similarly, compared to PolyFormer, AerialREC♢ achieves improvements of 3.37% and 2.92% in IoU@0.5 and IoU@mean on TestA, and 6.14%and 3.56% on TestB, leading to an average improvement of 4.75% and 3.24%. These results validate the effectiveness of our proposed method.

5.3 Ablation Studies

Component Ananlysis. We study the impact of various components in constructing the extended regressed coor-

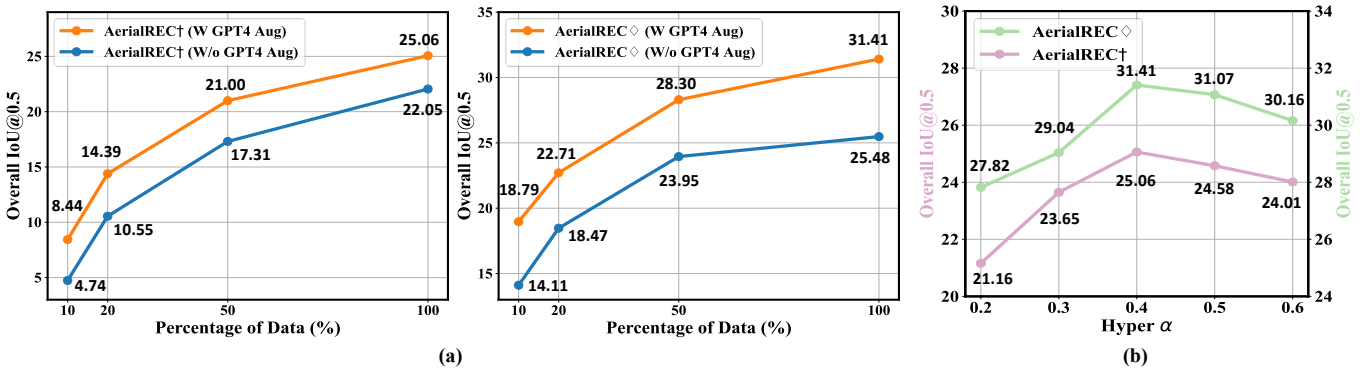

Fig. 7: (a) Impact of data scale, showing the percentage of data (x-axis) and the effect of GPT-4 data augmentation (yellow vs. blue curves). (b) Impact of different hyper-parameter α settings. We provide the Average IoU@0.5 metric for reference.

TABLE 5: Models trained on general REC datasets struggle with the SkyFind task, highlighting the importance and necessity of the SkyFind dataset. 'RefC' represents RefCOCO, RefCOCO+, and RefCOCOg.

Method	Train Data	SkyFind TestB			
		IoU@0.5	Δ	IoU@mean	Δ
VLVTG	RefC	5.71	+9.81	2.26	+8.15
	SkyFind	15.52		10.41	
SeqTR	RefC	5.85	+15.04	2.07	+9.97
	SkyFind	20.89		12.04	
QRNet	RefC	6.37	+13.34	2.81	+8.41
	SkyFind	19.71		11.22	
SimREC	RefC	5.89	+12.62	2.45	+9.70
	SkyFind	18.51		12.15	
PolyFormer	RefC	6.80	+24.67	2.41	+12.87
	SkyFind	31.47		15.28	

dinate sequence, including the two-step sequence extension, the newly added special token ⟨SEP⟩, and different strategies for formulating the potential target region. These strategies involve: (1) randomly sampling the potential target region from the enlarged intervals, i.e.,

$$(\hat{x}_1, \hat{y}_1) \sim \text{RS}(R_1), \quad (\hat{x}_2, \hat{y}_2) \sim \text{RS}(R_2), \tag{10}$$

(2) using the fixed innermost value where the potential target region precisely matches the target box without enlargement, i.e.,

$$(\hat{x}_1, \hat{y}_1) = (x_1, y_1), \quad (\hat{x}_2, \hat{y}_2) = (x_2, y_2), \tag{11}$$

and (3) using the fixed outermost value to extend the region to its maximum range, i.e.,

$$(\hat{x}_1, \hat{y}_1) = (\max(0, x_1 - \alpha x_{\text{left}}), \max(0, y_1 - \alpha y_{\text{left}})), \tag{12}$$

$$(\hat{x}_2, \hat{y}_2) = (\min(w, x_2 + \alpha x_{\text{left}}), \min(h, y_2 + \alpha y_{\text{left}})). \tag{13}$$

The results are shown in Table 4. We find that the two-step regression approach results in a significant performance improvement. The newly added special token also provides the model with valuable cues. In terms of formulation strategy, random sampling within the intervals proves to be more effective than using fixed values. Regressing the same coordinates twice does not yield substantial benefits and may even introduce negative effects. Additionally, using the outermost value of the enlarged intervals can impose a fixed prior, potentially confusing the model.

Dataset Necessity. To demonstrate the necessity and effectiveness of our SkyFind dataset for the proposed SkyFind task, we compare the performance of previous SOTA methods (i.e., VLVTG, SeqTR, QRNet, SimREC, and PolyFormer) when trained on general REC datasets versus our SkyFind dataset. 'RefC' denotes RefCOCO, RefCOCO+, and RefCOCOg. We choose to compare performance on SkyFind testB because the cross-domain test setting provides a more fair comparison. As shown in Table 5, there is a significant gap between general REC and SkyFind. Models trained on general REC datasets struggle to effectively handle SkyFind, with similar performance across different methods. However, when models are trained on our SkyFind dataset and then tested cross-domain, they exhibit substantial performance improvements in both IoU@0.5 and IoU@mean metrics compared to those trained on general REC data, confirming the necessity of the SkyFind dataset.

Data Scale. We investigate the impact of data scale on model performance by training and testing with 10%, 20%, 50%, and 100% of the data. As shown in Fig. 7 (a), model performance decreases at smaller scales but steadily improves as the scale increases. However, the rate of improvement diminishes, indicating a trend toward saturation. This suggests that further increasing the data scale may continue to boost performance, albeit at a reduced rate. Additionally, we examine the effect of data augmentation using GPT-4. In the figure, the yellow curve represents augmented data, while the blue curve represents non-augmented data. The results demonstrate that GPT-4 augmentation enhances data diversity, enriches data points, and consequently improves model performance.

Hyper-parameter α. We investigate the impact of varying the hyper-parameter α on model performance. As shown in Fig. 7 (b), a smaller α forces the model to search within a relatively small area, which does not effectively reduce task difficulty. Conversely, a larger α results in overly broad search areas, which fails to effectively prompt the second refinement step for precise localization. An α value of 0.4 yields the optimal performance.

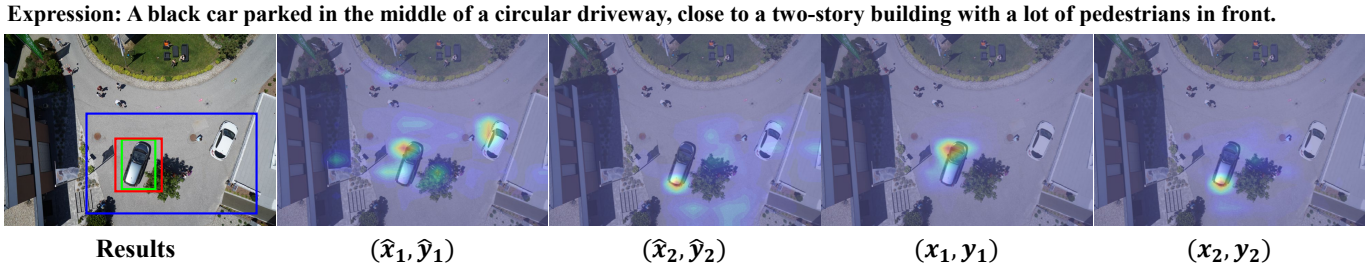

Fig. 8: The cross-attention maps of the decoder when generating each new vertex token, using AerialREC♢ for illustration. The green box denotes the ground truth, the blue box indicates the potential target region $(\hat{x}_i, \hat{y}_i)_{i=1}^{2}$ predicted in the initial search step, and the red box shows the precise target bounding box $(x_i, y_i)_{i=1}^{2}$ predicted in the refine step.

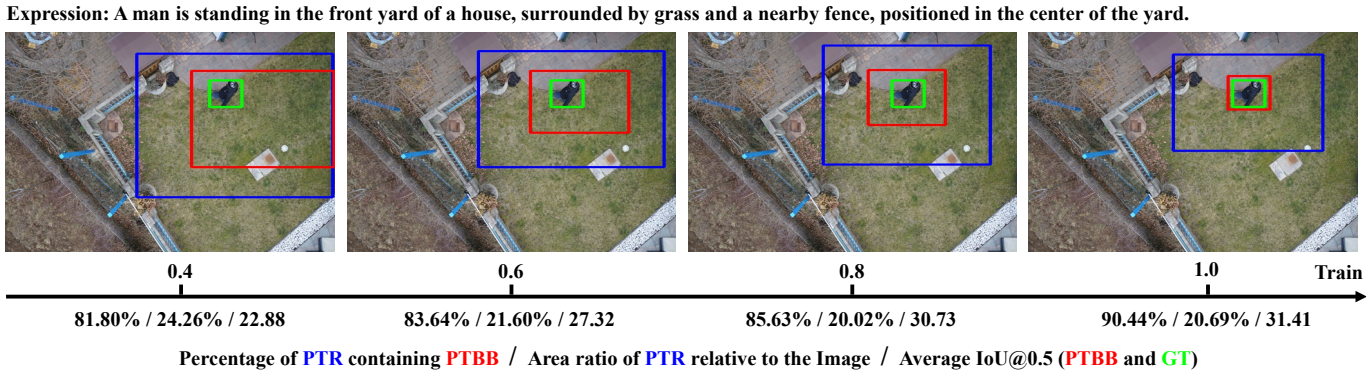

Fig. 9: Visualization of the two-step localization results at different training stages. The green box denotes the ground truth (GT), the blue box indicates the potential target region (PTR) from the initial search step, and the red box shows the precise target bounding box (PTBB) from the refine step. The three metrics below the x-axis represent the percentage of PTBB fully contained within PTR, the average area ratio of the PTR relative to the entire image, and the average IoU@0.5. We use AerialREC♢ for illustration.

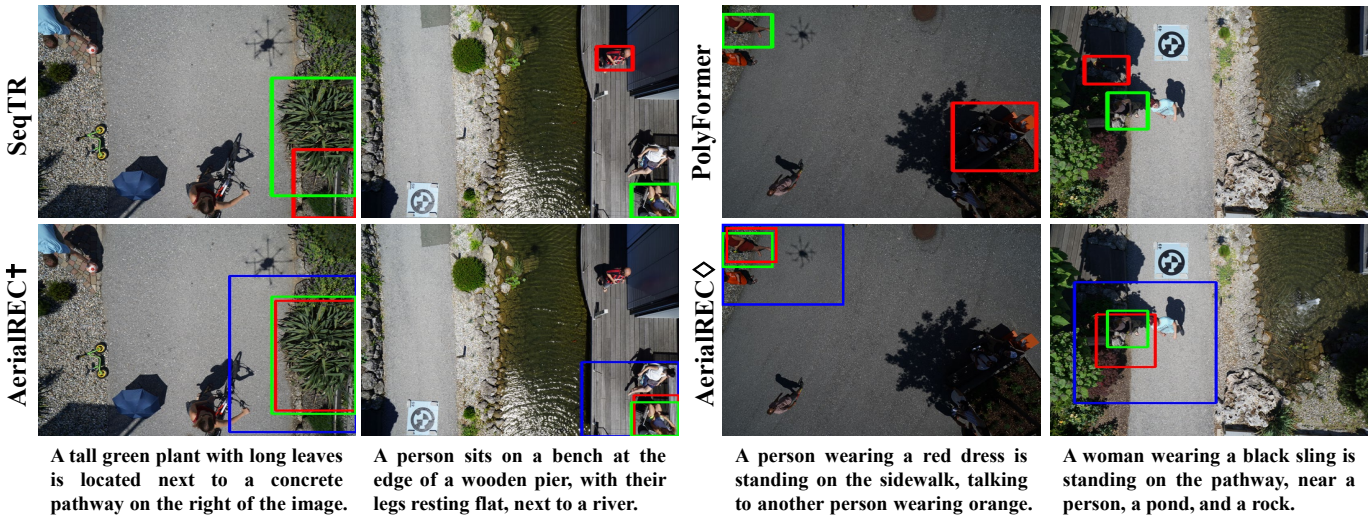

Fig. 10: Comparison of results between the baseline methods and AerialREC, including SeqTR vs. AerialREC† and PolyFormer vs. AerialREC♢. Green boxes indicate the ground truth, blue boxes show the potential bounding box predicted by AerialREC in the initial search step, and red boxes depict the final localization results predicted by different methods.

5.4 Visualization Results

Cross-attention Map. Here, we visualize the cross-attention maps (averaged across all layers and heads) during the model's regression of each new vertex token, as shown in Fig. 8. In the search step, the model predicts the potential target region, $(\hat{x}_i, \hat{y}_i)_{i=1}^{2}$, with attention primarily focused on the area of the image containing two cars. In the subsequent refine step, guided by the narrowed potential target region, the model achieves more precise localization, predicting $(x_i, y_i)_{i=1}^{2}$, with attention now concentrated on the black car.

Learning Process. We visualize the two-step localization results at different stages of training, i.e., 40%, 60%, 80%, and 100% of the total epochs or iterations. We show the potential target region (PTR) from the first search step, marked with blue boxes, and the precise target bounding box (PTBB) from the subsequent refine step, marked with red boxes. The ground truth (GT) is indicated in green. We provide three metrics evaluated on the test sets for each training stage: the percentage of PTBB fully contained within PTR, the average reduction in interference area across the image achieved by PTR, and the average IoU@0.5. As training progresses, the model learns to predict the PTBB within the PTR. The coverage ratio increases from 81.80% to 90.44%, and the area ratio of the PTR relative to the entire image decreases from 24.26% to 20.69%, gradually excluding more interference regions from the image. Consequently, the model's localization accuracy enhances from 22.88 to 31.41 in terms of average IoU@0.5.

Prediction Visualization. We present visualization results for both the baselines and our methods, including SeqTR and AerialREC†, as well as PolyFormer and AerialREC♢, as illustrated in Fig. 10. Compared to the baseline methods, AerialREC achieves superior accuracy in target localization within UAV scenarios. In the first search step, AerialREC identifies an approximate region containing the specified target based on the given expression, effectively reducing background interference. In the subsequent refine step, it precisely locates the target within this region. In contrast, baseline methods often experience mislocalization due to interference from complex scenes.

6 Conclusion and Future Work

Conclusion. In this paper, we introduce a novel SkyFind task, aimed at advancing research in the field of human-UAV interaction. The SkyFind task has broad application prospects in various scenarios, including search and rescue, environmental monitoring, security patrols, geographic information collection, and precision agriculture. It can significantly enhance work efficiency, reduce manual workload, and provide a user-friendly experience. Furthermore, to support and realize the SkyFind task, we construct a large-scale SkyFind dataset containing one million data points and propose a dedicated AerialREC baseline to address the unique challenges posed by complex and cluttered backgrounds, thereby promoting further research and development in this area. We establish the initial benchmark on the SkyFind dataset. Experimental results validate the effectiveness of the proposed AerialREC baseline, while also highlighting significant potential for further improvement.

Future Work. Building on the existing work presented in this paper, there are several promising directions for future research, primarily from the perspectives of both dataset and method. We highlight them as follows:

(1) Video-based SkyFind Dataset. The proposed SkyFind dataset is built on image data, which offers several advantages for UAV platforms. First, image-based processing is well-suited for UAV edge devices due to common challenges such as limited storage capacity, restricted bandwidth, and high energy consumption [81], [82]. Real-time processing of high-resolution video streams can heavily strain these resources, whereas image-based approaches make more efficient use of them. Second, image-based processing is better aligned with UAV tasks that require rapid response and inference [83]–[85]. In urgent scenarios, swift decision-making is paramount, and image-based processing provides immediate analysis, enabling UAV to promptly execute critical missions. However, video-based processing holds potential advantages in certain contexts. When ample resources such as storage, bandwidth, and energy are available, video-based approaches offer richer temporal information and a more comprehensive context for the SkyFind task. Thus, future work may explore efficient methods to construct a video-based dataset while maintaining temporal and semantic coherence in the referring expressions.

(2) Tailored REC Methods for UAV. The introduction of AerialREC as a baseline marks the starting point for a broader exploration of REC methods specifically designed for UAV applications. Several promising directions for future research emerge. One promising direction is the integration of additional modalities, such as audio, to enrich interactions between users and UAV. Combining audio with visual data can provide nuanced descriptions, disambiguate complex scenes, and improve the system's ability to comprehend instructions. Multi-modal REC methods that seamlessly fuse audio, visual, and other sensory inputs could significantly enhance the precision and reliability of UAV localization. Another critical area is boosting the computational efficiency of REC methods. Given the resource constraints of UAV platforms, such as limited processing power, energy, and bandwidth, developing lightweight, high-performance models is essential for real-time localization and decision-making. Future research should focus on optimizing algorithms to balance accuracy with speed and resource efficiency. Finally, multi-turn interactions between users and UAV offer a powerful approach to refining REC task. In complex environments, a single referring expression may not suffice. Enabling UAV to engage in dialogues-asking clarifying questions or seeking additional information could reduce errors and improve understanding, further enhancing accuracy and usability.

References

[1] J. Del Cerro, C. Cruz Ulloa, A. Barrientos, and J. de León Rivas, "Unmanned aerial vehicles in agriculture: A survey," *Agronomy*, vol. 11, no. 2, p. 203, 2021.

[2] S. A. H. Mohsan, N. Q. H. Othman, Y. Li, M. H. Alsharif, and M. A. Khan, "Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends," *Intelligent Service Robotics*, vol. 16, no. 1, pp. 109–137, 2023.

[3] K. Wang, X. Fu, Y. Huang, C. Cao, G. Shi, and Z.-J. Zha, "Generalized uav object detection via frequency domain disentanglement," in *CVPR*, 2023, pp. 1064–1073.

[4] K. Wang, X. Fu, C. Ge, C. Cao, and Z.-J. Zha, "Towards generalized uav object detection: A novel perspective from frequency domain disentanglement," *International Journal of Computer Vision*, pp. 1–29, 2024.

[5] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, and J. Han, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] L. Li, X. Yao, G. Cheng, and J. Han, "Aifs-dataset for few-shot aerial image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[7] Y. Fan, W. Chen, T. Jiang, C. Zhou, Y. Zhang, and X. Wang, "Aerial vision-and-dialog navigation," in *ACL*, 2023, pp. 3043–3061.

[8] S. Liu, H. Zhang, Y. Qi, P. Wang, Y. Zhang, and Q. Wu, "Aerialvln: Vision-and-language navigation for uavs," in *ICCV*, 2023, pp. 15 384–15 394.

[9] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *CVPR*, 2021, pp. 16 266–16 275.

[10] I. Martinez-Alpiste, G. Golcarenarenji, Q. Wang, and J. M. Alcaraz-Calero, "Search and rescue operation using uavs: A case study," *Expert Systems with Applications*, vol. 178, p. 114937, 2021.

[11] M. Lyu, Y. Zhao, C. Huang, and H. Huang, "Unmanned aerial vehicles for search and rescue: A survey," *Remote Sensing*, vol. 15, no. 13, p. 3266, 2023.

[12] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014, pp. 787–798.

[13] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *ECCV*, 2016, pp. 69–85.

[14] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016, pp. 11–20.

[15] R. Liu, C. Liu, Y. Bai, and A. L. Yuille, "Clevr-ref+: Diagnosing visual reasoning with referring expressions," in *CVPR*, 2019, pp. 4185–4194.

[16] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M. F. Moens, "Talk2car: Taking control of your self-driving car," in *EMNLP*, 2019, pp. 2088–2098.

[17] Z. Chen, P. Wang, L. Ma, K.-Y. K. Wong, and Q. Wu, "Cops-ref: A new dataset and task on compositional referring expression comprehension," in *CVPR*, 2020, pp. 10 086–10 095.

[18] P. Wang, D. Liu, H. Li, and Q. Wu, "Give me something to eat: referring expression comprehension with commonsense knowledge," in *ACM MM*, 2020, pp. 28–36.

[19] S. He, H. Ding, C. Liu, and X. Jiang, "Grec: Generalized referring expression comprehension," *arXiv preprint arXiv:2308.16182*, 2023.

[20] Z. Chen, R. Zhang, Y. Song, X. Wan, and G. Li, "Advancing visual grounding with scene knowledge: Benchmark and method," in *CVPR*, 2023, pp. 15 039–15 049.

[21] R. Dang, J. Feng, H. Zhang, G. Chongjian, L. Song, G. Lijun, C. Liu, Q. Chen, F. Zhu, R. Zhao *et al.*, "Instructdet: Diversifying referring object detection with generalized instructions," in *ICLR*, 2024.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[25] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020, pp. 11 621–11 631.

[26] C. Zhang, G. Huang, L. Liu, S. Huang, Y. Yang, X. Wan, S. Ge, and D. Tao, "Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9186–9205, 2022.

[27] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. v. d. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *CVPR*, 2019, pp. 1960–1968.

[28] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *CVPR*, 2017, pp. 1115–1124.

[29] J. Liu, L. Wang, and M.-H. Yang, "Referring expression generation and comprehension via attributes," in *ICCV*, 2017, pp. 4856–4864.

[30] R. Luo and G. Shakhnarovich, "Comprehension-guided referring expressions," in *CVPR*, 2017, pp. 7102–7111.

[31] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *ICCV*, 2019, pp. 4694–4703.

[32] B. Huang, D. Lian, W. Luo, and S. Gao, "Look before you leap: Learning landmark features for one-stage visual grounding," in *CVPR*, 2021, pp. 16 888–16 897.

[33] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *CVPR*, 2020, pp. 10 034–10 043.

[34] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *ICCV*, 2019, pp. 4683–4693.

[35] H. Jiang, Y. Lin, D. Han, S. Song, and G. Huang, "Pseudo-q: Generating pseudo language queries for visual grounding," in *CVPR*, 2022, pp. 15 513–15 523.

[36] Y. X. Chng, H. Zheng, Y. Han, X. Qiu, and G. Huang, "Mask grounding for referring image segmentation," in *CVPR*, 2024, pp. 26 573–26 583.

[37] W. Tang, L. Li, X. Liu, L. Jin, J. Tang, and Z. Li, "Context disentangling and prototype inheriting for robust visual grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[38] X. Liu, L. Li, S. Wang, Z.-J. Zha, Z. Li, Q. Tian, and Q. Huang, "Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3003–3018, 2022.

[39] K. Li, J. Li, D. Guo, X. Yang, and M. Wang, "Transformer-based visual grounding with cross-modality interaction," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–19, 2023.

[40] M. Ni, Y. Zhang, K. Feng, X. Li, Y. Guo, and W. Zuo, "Ref-diff: Zero-shot referring image segmentation with generative models," *arXiv preprint arXiv:2308.16777*, 2023.

[41] B. Chen, Z. Hu, Z. Ji, J. Bai, and W. Zuo, "Position-aware contrastive alignment for referring image segmentation," *arXiv preprint arXiv:2212.13419*, 2022.

[42] Z. Yang, T. Chen, L. Wang, and J. Luo, "Improving one-stage visual grounding by recursive sub-query construction," in *ECCV*, 2020, pp. 387–404.

[43] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," *NeurIPS*, pp. 19 652–19 664, 2021.

[44] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "Transvg: End-to-end visual grounding with transformers," in *ICCV*, 2021, pp. 1769–1779.

[45] J. Ye, J. Tian, M. Yan, X. Yang, X. Wang, J. Zhang, L. He, and X. Lin, "Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding," in *CVPR*, 2022, pp. 15 502–15 512.

[46] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, and R. Ji, "Seqtr: A simple yet universal network for visual grounding," in *ECCV*, 2022, pp. 598–615.

[47] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," in *ICME*, 2022.

[48] L. Yang, Y. Xu, C. Yuan, W. Liu, B. Li, and W. Hu, "Improving visual grounding with visual-linguistic verification and iterative reasoning," in *CVPR*, 2022, pp. 9499–9508.

[49] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha, "Polyformer: Referring image segmentation as sequential polygon generation," in *CVPR*, 2023, pp. 18 653–18 663.

[50] G. Luo, Y. Zhou, J. Sun, X. Sun, and R. Ji, "A survivor in the era of large-scale pretraining: An empirical study of one-stage referring expression comprehension," *IEEE Transactions on Multimedia*, 2023.

[51] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.
[52] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024, pp. 26 296–26 306.
[53] J. Wu, Y. Jiang, Q. Liu, Z. Yuan, X. Bai, and S. Bai, "General object foundation model for images and videos at scale," in *CVPR*, 2024, pp. 3783–3795.
[54] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," in *ICLR*, 2024.
[55] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu, "Universal instance perception as object discovery and retrieval," in *CVPR*, 2023, pp. 15 325–15 336.
[56] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen *et al.*, "Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models," *arXiv preprint arXiv:2311.07575*, 2023.
[57] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.
[58] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou, "One-peace: Exploring one general representation model toward unlimited modalities," *arXiv preprint arXiv:2305.11172*, 2023.
[59] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
[60] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[61] Z. Yang, Z. Gan, J. Wang, X. Hu, F. Ahmed, Z. Liu, Y. Lu, and L. Wang, "Unitab: Unifying text and box outputs for grounded vision-language modeling," in *ECCV*, 2022, pp. 521–539.
[62] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," *NeurIPS*, pp. 6616–6628, 2020.
[63] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022.
[64] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "Seadronessee: A maritime benchmark for detecting humans in open water," in *WACV*, 2022, pp. 2260–2270.
[65] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.
[66] I. Bozcan and E. Kayacan, "Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *ICRA*, 2020, pp. 8504–8510.
[67] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *ICCV*, 2017, pp. 4145–4153.
[68] D. Cafarelli, L. Ciampi, L. Vadicamo, C. Gennaro, A. Berton, M. Paterni, C. Benvenuti, M. Passera, and F. Falchi, "Mobdrone: A drone video dataset for man overboard rescue," in *ICIP*, 2022, pp. 633–644.
[69] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *CVPRW*, 2017, pp. 28–35.
[70] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *ECCV*, 2016, pp. 549–565.
[71] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *ECCV*, 2018, pp. 370–386.
[72] I. team, "Semantic Drone Dataset," http://dronedataset.icg.tugraz.at.
[73] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.
[74] Y. Chen, Y. Wang, P. Lu, Y. Chen, and G. Wang, "Large-scale structure from motion with semantic constraints of aerial images," in *PRCV*, 2018, pp. 347–359.
[75] W. Zhang, C. Liu, F. Chang, and Y. Song, "Multi-scale and occlusion aware network for vehicle detection and segmentation on uav aerial images," *Remote Sensing*, vol. 12, no. 11, p. 1760, 2020.
[76] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *WACV*, 2018, pp. 1499–1508.
[77] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[78] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," in *ICCV*, 2023, pp. 11 987–11 997.
[79] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *ICML*, 2022, pp. 23 318–23 340.
[80] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.
[81] K. Telli, O. Kraa, Y. Himeur, A. Ouamane, M. Boumehraz, S. Atalla, and W. Mansoor, "A comprehensive review of recent research trends on unmanned aerial vehicles (uavs)," *Systems*, vol. 11, no. 8, p. 400, 2023.
[82] P. McEnroe, S. Wang, and M. Liyanage, "A survey on the convergence of edge computing and ai for uavs: Opportunities and challenges," *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 15 435–15 459, 2022.
[83] D. Erdos, A. Erdos, and S. E. Watkins, "An experimental uav system for search and rescue challenge," *IEEE Aerospace and Electronic Systems Magazine*, vol. 28, no. 5, pp. 32–37, 2013.
[84] C. Kanellakis and G. Nikolakopoulos, "Survey on computer vision for uavs: Current developments and trends," *Journal of Intelligent & Robotic Systems*, vol. 87, pp. 141–168, 2017.
[85] S. H. Alsamhi, A. V. Shvetsov, S. Kumar, S. V. Shvetsova, M. A. Alhartomi, A. Hawbani, N. S. Rajput, S. Srivastava, A. Saif, and V. O. Nyangaresi, "Uav computing-assisted search and rescue mission framework for disaster and harsh environment mitigation," *Drones*, vol. 6, no. 7, p. 154, 2022.

Kunyu Wang is currently pursuing the Ph.D. degree with the National Engineering Laboratory for Brain-inspired Intelligence Technology and Application (NEL-BITA), University of Science and Technology of China, Hefei, China. His research interests include unmanned aerial vehicles and computer vision.

Xingbo Wang is currently pursuing a Master's degree at the University of Science and Technology of China. He completed his undergraduate studies at the Harbin Institute of Technology (Shenzhen) from 2019 to 2023, majoring in Automation. Wang's primary research interests lie in the fields of computer vision and machine learning.

Kean Liu received the B.S. from Hefei University of Technology in 2024 and is currently pursuing an M.S. at the University of Science and Technology of China. His research interests include computer vision and machine learning.

Xin Lu is currently pursuing an M.S. degree at the University of Science and Technology of China, involved in research projects at the National Engineering Laboratory for Brain-inspired Intelligence Technology and Application (NEL-BITA). He received a B.S. degree in Vehicle Engineering from Wuhan University of Technology (2023) and is interested in computer vision and intelligent vehicles.

Chengjie Ge is currently pursuing the Ph.D. degree with the National Engineering Laboratory for Brain-inspired Intelligence Technology and Application (NEL-BITA) at the University of Science and Technology of China, Hefei, China. His research interests include event cameras and image processing.

Wei Zhai received the Ph.D. degree with the University of Science and Technology of China (USTC), Hefei, China, in 2022. He is now a Associate Research Fellow with the School of Information Science and Technology, University of Science and Technology of China. He is also a member of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). His research interests mainly include computer vision and deep learning. He has published more than 30 papers in these areas with a series of publications on top journals and conferences, such as T-PAMI, IJCV, T-IP, T-NNLS, T-MM, CVPR, ICCV, ECCV, NeurIPS, AAAI, and IJCAI. Dr. Zhai was a recipient of AAAI Distinguished Paper Award.

Xueyang Fu (Member, IEEE) received the PhD degree in signal and information processing from Xiamen University, in 2018. He was a Visiting Scholar with Columbia University, sponsored by the China Scholarship Council, from 2016 to 2017. He is currently an associate professor with the School of Information Science and Technology, University of Science and Technology of China. He is also a member of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). His research interests include machine learning and image processing.

Zheng-Jun Zha (Member, IEEE) received the BE and PhD degrees from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively. He is currently a full professor with the School of Information Science and Technology, University of Science and Technology of China, and the Executive Director of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). He has authored or coauthored more than 200 papers in his research field with a series of publications on top journals and conferences, which include multimedia analysis and understanding, computer vision, pattern recognition, and also brain-inspired intelligence. He was the recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and Best Student Paper Award in ACM Multimedia etc. He is an associated editor for IEEE Transactions on Circuits and Systems for Video Technology and ACM Transactions on Multimedia Computing, Communications, and Applications.