

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

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

Abstract—Unmanned aerial vehicles (UAV) are increasingly deployed to assist humans in diverse tasks, where understanding human intentions is critical to effective collaboration. Referring expression comprehension (REC) links language to visual targets, allowing UAV to recognize human-intended targets of interest, thereby supporting subsequent actions. However, existing REC research is almost exclusively confined to ground-based scenarios, leaving aerial scenarios largely unexplored. In this paper, we formally define UAV-based REC as a new research problem and highlight its unique challenges, including abundant background interference, small target size, and complex referring relations. To enable systematic study, we introduce SkyFind, a large-scale dataset with one million high-quality target-expression pairs, providing a solid foundation. In addition, we propose AerialREC, a baseline framework that reduces background interference in UAV imagery by searching for a potential target region before localization. We establish benchmark results on SkyFind using ten representative REC methods and validate the effectiveness of the AerialREC framework. The dataset is publicly available at: <https://github.com/wangkunyu241/SkyFind>.

Index Terms—Unmanned Aerial Vehicles, Referring Expression Comprehension, Benchmark and Dataset.

1 INTRODUCTION

UNMANNED aerial vehicles (UAV), owing to their flexibility and versatility, have been widely deployed in various real-world scenarios to assist humans in diverse tasks [1]–[6]. To effectively support such missions, UAV require not only autonomous flight and navigation capabilities [7], [8] but also the ability to understand human intentions and objectives, which is essential for subsequent actions. For example, in security operations [9], UAV need to first recognize specific individual or object before proceeding with subsequent tracking and surveillance; in search-and-rescue missions [10], [11], UAV need to confirm the target of interest before carrying out close-range rescue operations or delivering supplies. These cases highlight that accurately understanding human intent is a prerequisite for mission success. Since language is the most direct and natural medium through which humans convey their intentions, it becomes a critical source of information for UAV to understand task objectives. Within this context, referring expression comprehension (REC) [12], which connects linguistic expressions with visual targets, emerges as an essential capability. By enabling UAV to ground natural language descriptions in specific targets within the visual scene, REC enhances their capacity to interpret human intent and perform subsequent tasks.

Despite significant progress in REC, current datasets, such as RefCOCO and RefCOCO+ [13], are primarily collected in everyday ground-level scenarios. These datasets are inadequate for supporting the development of REC in aerial scenarios. To bridge this gap and facilitate research in this realm, we introduce SkyFind, a large-scale dataset tailored to UAV-based REC, comprising one million high-

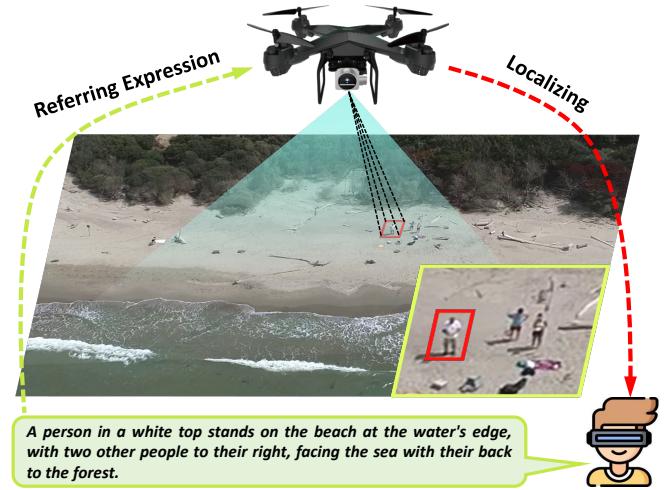


Fig. 1: Demo of UAV-based REC. A human provides a textual description of a specified target, and the UAV comprehends the expression and localizes the target accordingly.

quality target-expression pairs. As shown in Fig. 1, a human provides a free-form textual description, which the UAV comprehends and localizes the specified target in the UAV-captured image. The SkyFind dataset establishes a foundation for systematic investigation of REC in aerial scenarios. The examples of the dataset are shown in Fig. 2.

For the construction of the SkyFind dataset, we collect images from both publicly available UAV datasets and UAV data mined from the web, ensuring both scale and diversity. Based on these raw images, we annotate referring expressions together with the bounding boxes of the specified targets. Leveraging recent advancements in perception and understanding by large models, we initially utilize large models as assistants to pre-annotate the data, thereby reduc-

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.
E-mail: kuniuwang@mail.ustc.edu.cn

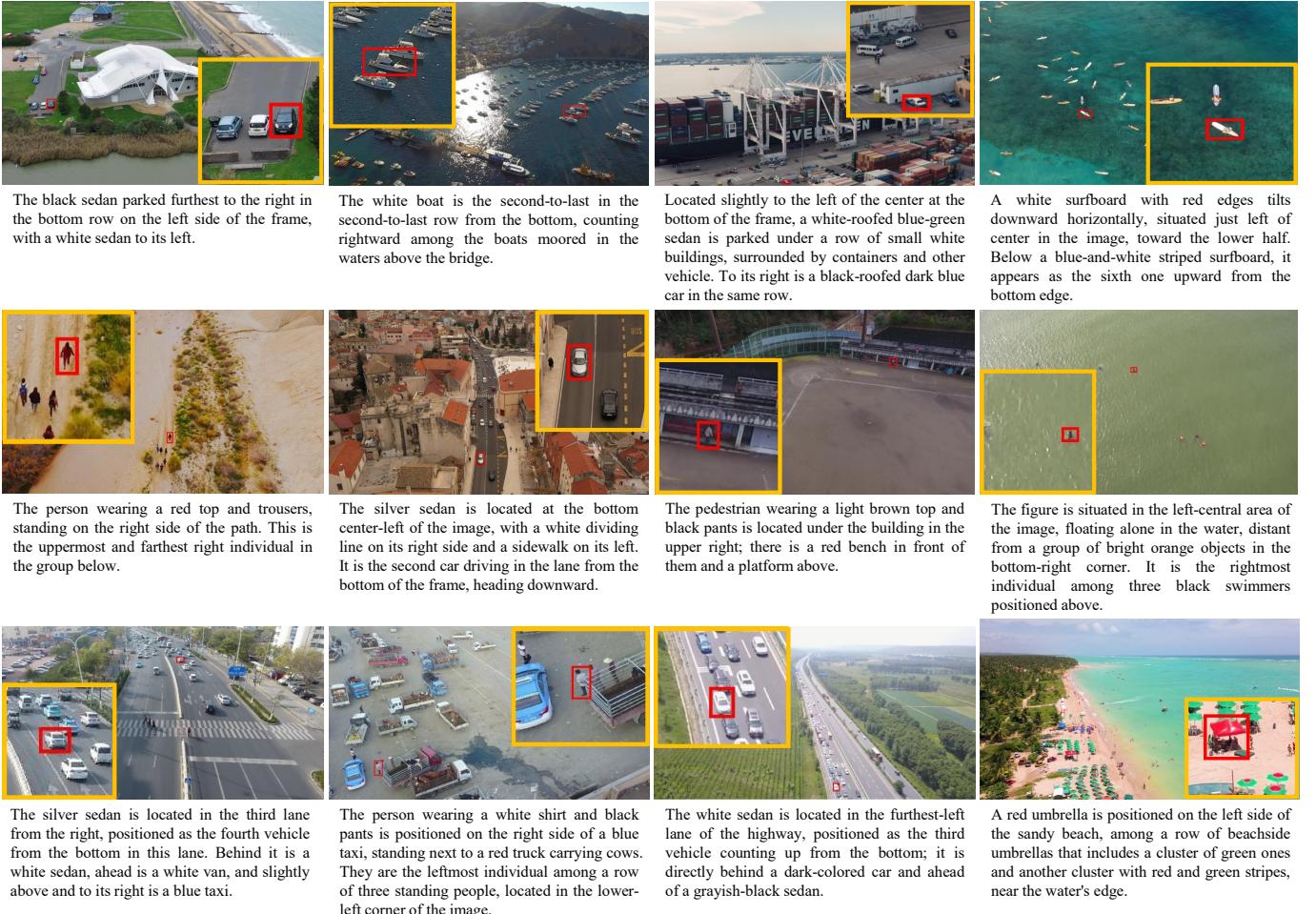


Fig. 2: A glance at the new large-scale SkyFind dataset, comprising over one million high-quality target-expression pairs.

ing the workload of manual annotation from scratch. Then, we manually refine and re-annotate the pre-annotations to ensure the high quality of the target-expression data. In addition to providing essential textual referring expressions, we also supply audio descriptions generated using text-to-speech software. This encourages explorations on understanding human instructions from speech, providing a more convenient mode of interaction.

Compared with general REC, UAV-based REC exhibits three distinct characteristics that introduce significant challenges: **(1) Abundant background interference.** The wide field of view in UAV imagery often leads to scenes with rich semantic content, containing numerous non-target entities that resemble the target. Such distracting entities in the background increase the difficulty of localization. **(2) Small target size.** UAV imagery typically follows a “large scene, small object” pattern, where the target occupies only a small fraction of the frame. Such target often has blurred boundary and weak texture, which increase the difficulty of recognition. **(3) Complex referring relations.** To identify a target in cluttered UAV imagery, referring expressions often incorporate fine-grained details, making them longer, characterized by more intricate reference relations, and consequently more difficult to comprehend. A more detailed comparison and discussion can be found in Section 3.1.

In addition to proposing the SkyFind dataset, we present

a baseline framework to address its challenges. Current REC methods suffer interference from non-target entities in UAV imagery that are semantically and spatially similar to the target, leading to confusion and degraded localization performance. To address this, we propose AerialREC, which formulates target localization as a two-step process. Specifically, we introduce a preliminary search step to initially identify a potential target region, providing a more focused area and explicitly reducing irrelevant background interference. Subsequently, we concentrate on this clearer region with less interference to precisely localize the target, thereby enhancing accuracy. Experiments conducted with two advanced REC methods demonstrate that our framework yields substantial performance gains on the SkyFind dataset, validating its effectiveness in UAV scenarios.

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

- We are the first to push the boundary of REC into aerial scenarios, formally introducing UAV-based REC as a new research problem and highlighting the unique challenges it poses compared with general REC.
- We construct SkyFind, a large-scale dataset with one million high-quality target-expression pairs, which is three orders of magnitude larger than existing counterparts, providing a solid foundation for systematic research on UAV-based REC.
- We propose AerialREC, a baseline framework that mi-

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 [14]	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 [15]	Life	26,711	54,822	85.7K	8.93	Free	480×583	8.93 %	✗	Manual
CLEVR-Ref+ [16]	Synthetic	85,000	492,727	998.7K	22.40	Free	480×320	2.70 %	✗	Simulator
Talk2Car [17]	Vehicle	9,217	10,519	11.9K	11.01	Free	1600×900	3.65 %	✗	Manual
Cops-Ref [18]	Life	75,299	148,712	148.7K	14.40	Template	521×432	13.10 %	✗	Manual
KB-Ref [19]	Life	16,917	43,284	43.2K	13.32	Free	500×413	18.20 %	✗	Manual
gRefCOCO [20]	Life	19,994	80,287	278.2K	13.22	Free	485×592	8.82 %	✗	Manual
SK-VG [21]	Movie	4,000	39,182	39.1K	4.46	Free	1601×848	12.00 %	✗	Manual
InDET [22]	Life	120,604	908,410	3.6M	6.20	Free	456×548	15.94 %	✗	LM
SkyFind	UAV	35599	352,910	1.0M	27.12	Free	1952×1155	0.76 %	✓	LM+Manual

gates interference from similar non-target entities in UAV imagery by introducing a preliminary search step, improving target localization accuracy.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of relevant literature. Section 3 elaborates on the uniqueness of UAV-based REC, explains the necessity for constructing the SkyFind dataset, and describes the construction process along with the statistical analysis. Section 4 presents our proposed baseline framework AerialREC. Section 5 establishes benchmark results on the SkyFind dataset and validates the effectiveness of the AerialREC framework. Finally, Section 6 concludes this work and discusses potential future research directions.

2 RELATED WORK

2.1 Referring Comprehension Datasets

In the deep learning era, benchmark datasets [23], [24] 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 [14] 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 [25], 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 [15], also originating from MSCOCO, stands out for its utilization of longer language expressions compared to its predecessors. CLEVR-Ref+ [16] emerges as a synthetic diagnostic dataset for REC, addressing bias issues, and facilitating the assessment of models' intermediate reasoning processes. Talk2Car [17] 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 [26] dataset. Cops-Ref [18] introduces intricate and compositional expressions, challenging models to demonstrate complex reasoning abilities beyond simple object recognition, attributes, and relations. KB-Ref [19] pushes the boundaries of REC

models by necessitating the incorporation of commonsense knowledge in identifying referent objects. It encourages models to explore information not only from images but also from external knowledge. gRefCOCO [20] extends the classic REC by allowing expressions to describe any number of target objects, including multi-target expressions and no-target expressions. SK-VG [21] 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 [22] proposes a data generation pipeline that relies on foundational models to generate instructions, paving the way for enhancing data scale. However, existing datasets primarily focus on ground-based scenarios, while aerial scenarios have been largely overlooked. This setting is in fact highly valuable, as it enables UAV to better understand human intentions and facilitate human-UAV interaction. To fill this gap, we formally introduce UAV-based REC and construct the million-scale SkyFind dataset, providing a foundation to advance research in this direction. Detailed comparisons with prior datasets are provided in Table 1.

Beyond the above, WebUAV-3M [27] presents a dataset with annotations akin to our work. However, WebUAV-3M is tailored for the UAV-based 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 target-expression annotations. In contrast, our proposed SkyFind dataset primarily focuses on the UAV-based REC task, comprising 1.0M pairs of target-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], [28]–[31] 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 [32]–[42] has recently

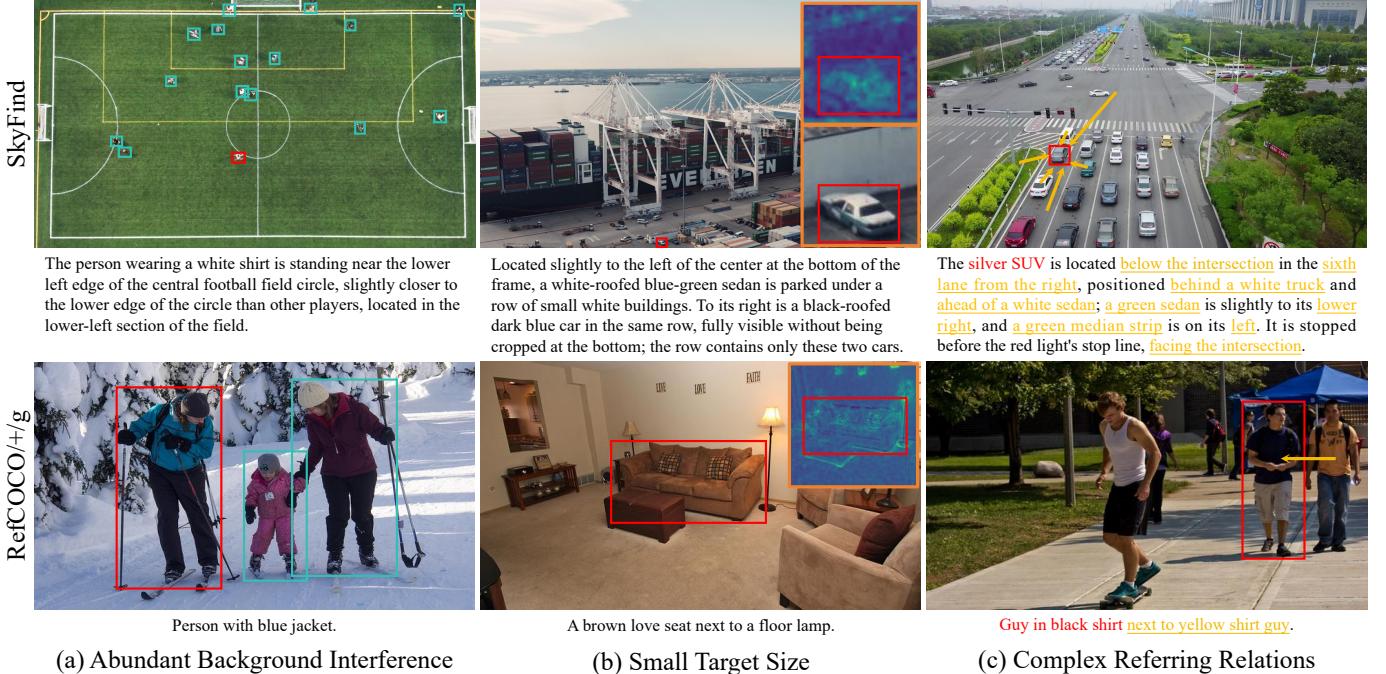


Fig. 3: Comparison between general REC and UAV-based REC, where UAV-based REC faces three unique challenges of abundant background interference, small object size, and complex referring relations.

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 [43]–[51]. For example, ReSC [43] introduces a recursive sub-query construction framework to enhance one-stage visual grounding, overcoming limitations in modeling long and complex queries. TransVG [45] 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 [47] represents the bounding box of the referent with a sequence of discrete coordinate tokens, which are predicted via a Transformer architecture. PolyFormer [50] formulates REC and RES tasks as a sequence-to-sequence prediction problem, generating sequential polygon vertices and bounding box corner points. Besides, recent advances in large-scale vision-and-language (V&L) pretraining have led to the emergence of powerful large models [52]–[63], which demonstrate strong generalization across a wide range of downstream V&L tasks, including REC. These models are trained with millions or even billions of image-text pairs and contain a large number of parameters, enabling them to achieve zero-shot performance on diverse benchmarks.

Despite progress, no existing method has been specifically developed for REC in aerial scenarios. Compared with general REC, UAV-based REC introduces three unique

challenges: abundant background interference, small target size, and complex referring relations. These challenges significantly increase the difficulty of REC and highlight the need for dedicated methods.

3 SKYFIND DATASET

3.1 Uniqueness

Compared with general REC that are typically performed in ground-based everyday scenarios, UAV-based REC demonstrates three distinctive characteristics that pose challenges:

Abundant Background Interference. Due to the wide field of view in UAV imagery, the captured scenes typically span large areas and contain abundant semantic information with diverse entity distributions. Consequently, numerous non-target objects often appear that resemble the referred target in category, appearance, or spatial location. Such entities exhibit strong referential or semantic similarity to the target, leading to confusion and interference, thereby degrading performance. For UAV-based REC, accurately localizing the referred target within such complex and distracted background constitutes a core challenge. As shown in Fig. 3 (a), compared with Refcoco/+/g, numerous non-target entities with semantic or referential similarity to the target appear in the image background of the SkyFind dataset, such as the person in a white T-shirt or the person near the edge of the central football field circle. This requires the model to perform fine-grained discrimination and precise localization within the scene.

Small Target Size. A notable characteristic of UAV imagery is the coexistence of large scenes with small targets. Due to the high imaging altitude, the referred target occupies only a minute portion of the entire image, leading to blurred boundaries and insufficient texture details. As a result, the

TABLE 2: Models trained on general REC datasets struggle with the SkyFind dataset, revealing the domain gap and the necessity of SkyFind dataset. ‘RefC’ denotes RefCOCO, RefCOCO+, and RefCOCOG.

Method	Train Data	SkyFind Test			
		IoU@0.5	Δ	IoU@mean	Δ
RefTR [44]	RefC	5.03		2.54	
	SkyFind Train	22.68	+17.65	13.46	+10.92
TransVG [45]	RefC	5.96		2.60	
	SkyFind Train	22.00	+16.04	11.90	+9.30
VGTR [48]	RefC	6.01		2.72	
	SkyFind Train	20.16	+14.15	10.55	+7.83
VLVTG [49]	RefC	5.70		2.24	
	SkyFind Train	23.52	+17.82	13.32	+11.08
SeqTR [47]	RefC	5.65		2.37	
	SkyFind Train	25.74	+20.09	12.57	+10.20
QRNet [46]	RefC	6.02		3.81	
	SkyFind Train	26.21	+20.19	11.22	+7.41
SimREC [51]	RefC	5.09		3.54	
	SkyFind Train	21.50	+16.41	12.15	+8.61
PolyFormer [50]	RefC	6.77		4.47	
	SkyFind Train	25.50	+18.73	16.44	+11.97

target’s feature is sparse and incomplete. Meanwhile, other semantically salient regions in the large scene compete for visual attention, further attenuating the prominence of the already inconspicuous target. As shown in Fig. 3 (b), compared with the large targets in Refcoco/+/_g, the small targets in the SkyFind dataset exhibit indistinct boundaries and blurred details at the feature level, such as vehicle contours and structural structures, thereby making precise localization of the referred target more difficult.

Complex Referring Relations. The complexity of UAV imagery naturally gives rise to highly intricate referring expressions. To accurately identify a target, the language often incorporates multi-level descriptions, including relative spatial relations between the target and reference objects, absolute spatial relations, categories, and attributes. Such expressions are typically lengthy and contain multiple semantic elements and dependencies, which makes them considerably more difficult to comprehend than short phrase-level expressions. As shown in Fig. 3 (c), the expression in the SkyFind dataset are more complex than that in Refcoco/+/_g, exemplified by references to intersection, lane, median strip, and multiple nearby vehicles used to specify the target. Such complexity underscores the challenge of fine-grained expression comprehension in UAV-based REC.

3.2 Necessity

Compared with general REC, UAV-based REC presents unique challenges, yet existing datasets (e.g., RefCOCO, RefCOCO+, and RefCOCOG, collectively referred to as RefC) are built from ground-level everyday scenes and thus cannot adequately support REC in aerial scenarios. To validate the necessity of introducing the customized large-scale SkyFind dataset, we train eight advanced REC methods separately on RefC and the SkyFind training set, and evaluate all models on the SkyFind test set. Note that SkyFind training and test sets are distributionally disjoint

TABLE 3: Introduction to the 14 publicly available UAV datasets. ‘Data’ represents the data type, and ‘Anno’ represents the annotation type.

Dataset	Task	Data	Anno	Target class
DroneVehicle [64]	Detect	Image	Box	Car, Bus, Truck, Van, Freight Car
SeaDronesSee [65]	Detect	Image	Box	Swimmer, Floater, Life Jacket, Boats
VisDrone2019 [66]	Detect	Image	Box	Pedestrian, Person, Car, Van, Bus, Truck, Motor, Bicycle, Awningtricycle, Tricycle
UAVDT [67]	Detect	Image	Box	Car, Vehicle, Truck, Bus
AU-AIR [68]	Detect	Video	Box	Person, Car, Van, Truck, Bike, Motorbike, Bus, Trailer
CAPRK [69]	Count	Video	Box	Car
MOBDrone [70]	Detect	Video	Box	Person, Boat, Wood, Life Buoy, Surfboard
Okutama-Action [71]	Detect	Video	Box	Human
Stanford Drone [72]	Track	Video	Box	Pedestrians, Bikers, Golf Carts, Cars, Buses, Skateboarders
Semantic Drone [73]	Segment	Image	Mask	Vegetation, Dirt, Gravel, Rocks, Water, Pool, Person, Dog, Car, Bicycle, Roof, Wall, Fence, Window, Door
UAvid [74]	Segment	Image	Mask	Building, Road, Low Vegetation, Tree, Car, Human, Clutter
UDD [75]	Segment	Image	Mask	Vegetation, Building, Free Space
UVSD [76]	Segment	Image	Mask	Vehicle
Aeroscapes [77]	Segment	Video	Mask	Sky, Road, Vegetation, Car, Obsfacle, Animal, Boat, Drone, Construction, Bike, Person

(see Section xx for details), thus avoiding potential distribution leakage. As shown in Table 2, models trained on RefC perform significantly worse on the SkyFind test set compared to those trained on the SkyFind training set. This substantial performance gap reflects a pronounced domain shift rooted in fundamental differences between ground-based and UAV-based scenarios, highlighting the limitations of existing datasets and the necessity of introducing the SkyFind dataset.

3.3 Dataset Construction

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 14 UAV datasets [64]–[77], as outlined in Table 3. The latter comprises UAV videos primarily obtained from YouTube under Creative Commons licenses¹, 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

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

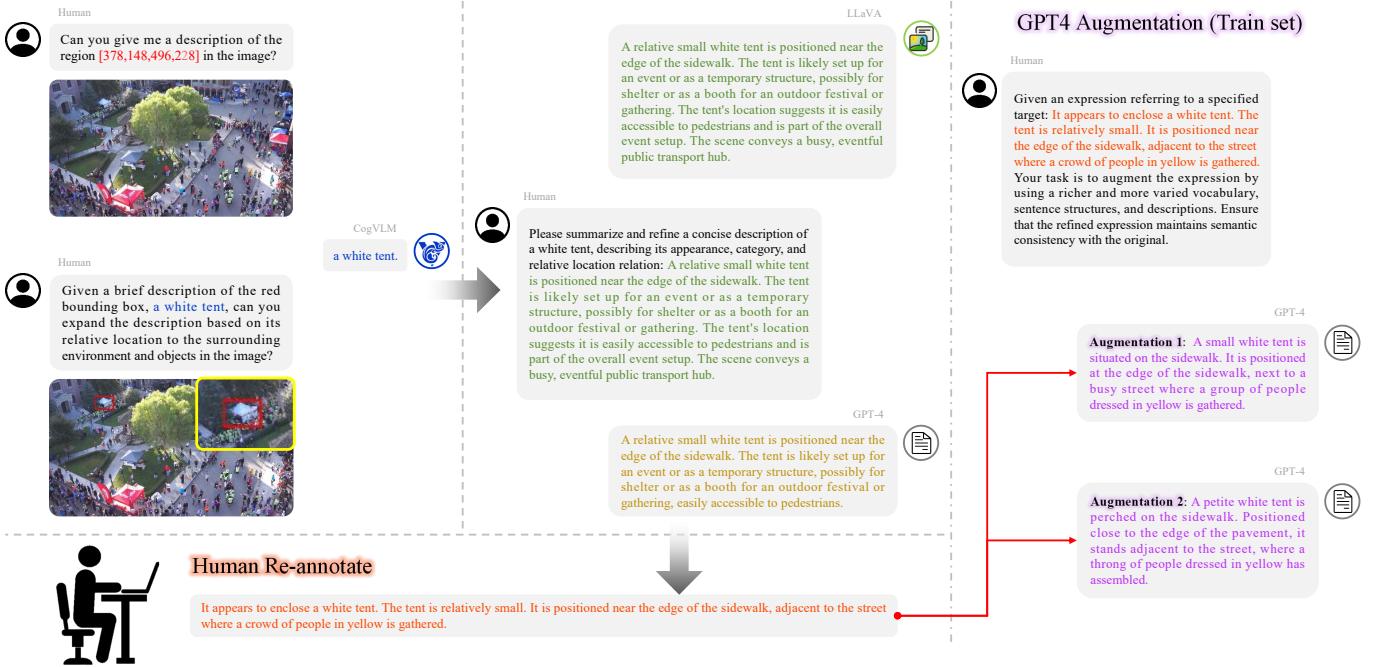


Fig. 4: Overview of the LM+Manual annotation pipeline, which integrates CogVLM [52] for **basic descriptions**, LLaVA [53] for **detailed descriptions**, GPT-4 [78] for **concise descriptions**, human input for **re-annotation**, and GPT-4 for **training set expression augmentation**.

valuable target, such as landscapes or fields, to ensure the dataset’s efficacy.

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 35,599 and 352,910 high-quality images and objects.

Data 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 annotation pipeline is shown in Fig. 4.

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 [52] 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 [53]. 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 [79], 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 [78] 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. Therefore, we manually re-annotate and refine all the pre-annotations to ensure the high quality of the target-expression data. Note that for all processes involving human, including screening images, annotating boxes, and re-annotating pre-annotations, 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 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 work-

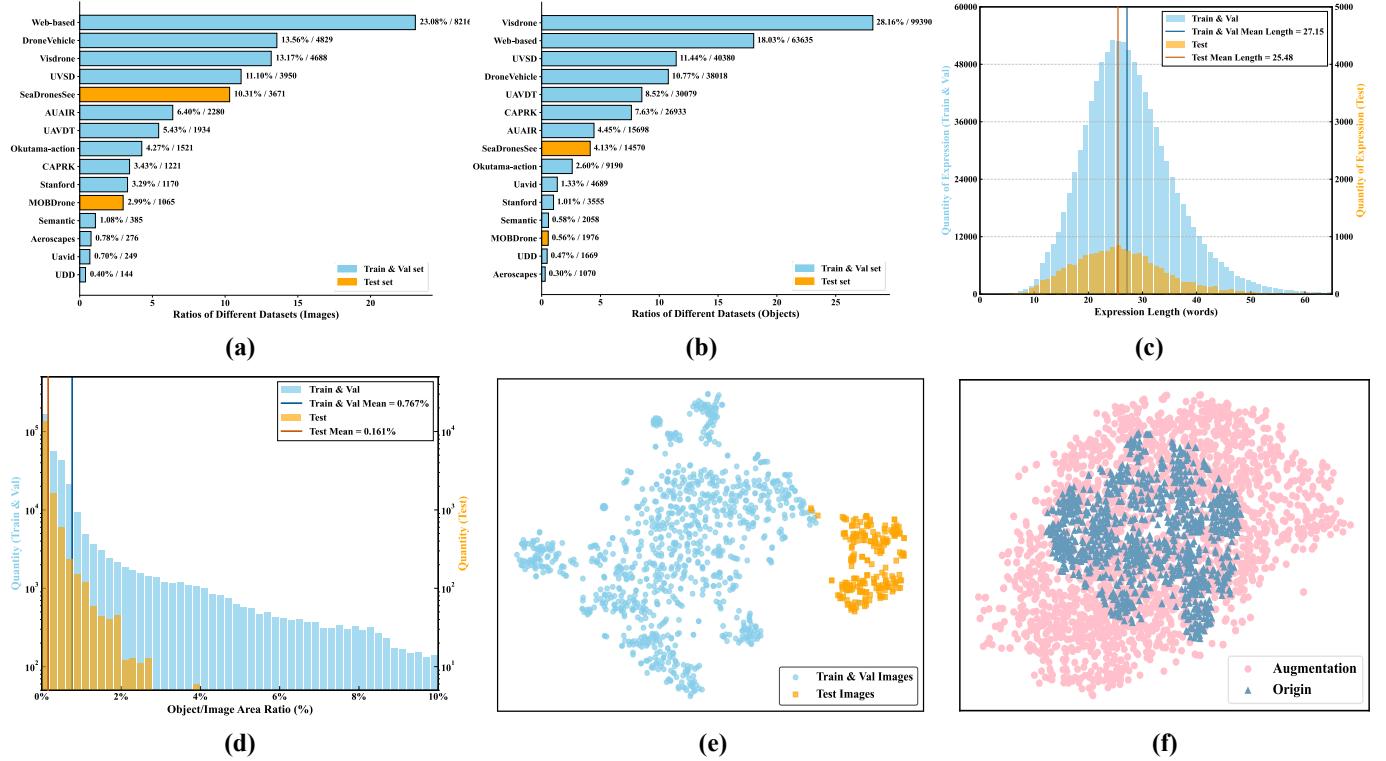


Fig. 5: Statistical analysis of the SkyFind dataset: (a) (b) Proportions of different data sources, by image count and annotated object count. (c) Distribution of expression lengths. (d) Distribution of the average object-to-image area ratio. (e) T-SNE visualization of images from training/validation and test sets. (f) T-SNE visualization of origin and GPT-4 augmented expressions in training set.

flow, 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.

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 AI³ as our audio generation tool, integrating diverse tones, timbres, accents, and voice effects throughout the generation process to deliver comprehensive audio descriptions.

3.4 Dataset Split

We divide the SkyFind dataset into training, validation, and test sets. To ensure distributional difference between the training/validation sets and the test set, which is crucial for meaningful evaluation, we perform the split based on data sources. Specifically, our dataset is constructed from web-crawled aerial images and 14 publicly available datasets. Among these, only SeaDronesSee and MoBDrone focus on maritime scenarios, which exhibit clear scene-level distinctions from the other sources. Therefore, we assign all target-expression pairs originating from these two datasets to the test set, resulting in 16,546 samples. The remaining data are used for training and validation. From these, we

randomly sample 5,000 instances for validation, and use the rest (331,364 samples) for training.

In addition to ensuring source-level diversity, we further enrich the training data on the language side. To enhance the density and coverage of the linguistic space, we employ GPT-4 [78] to augment the original annotations with semantically equivalent but lexically and syntactically more diverse descriptions. This augmentation exposes the model to a broader range of linguistic variations during training, thereby improving its ability to consistently interpret diverse expressions and leading to greater robustness at test time. As shown in Fig. 4, We prompt GPT-4 with instructions 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 origin expression. Finally, we obtain 999,092 diverse training target-expression pairs.

3.5 Dataset Analysis

Our SkyFind dataset consists of 1,015,638 target-expression pairs across 35,599 images and 352,910 objects. The average length of the expressions is 27.12 words, and the vocabulary size is 11,934. We now present a more detailed statistical analysis as below.

Fig. 5 (a) and (b) show the proportions of different data sources in the SkyFind dataset, measured respectively by

2. <https://www.robots.ox.ac.uk/~vgg/software/via/>
 3. <https://cloud.google.com/text-to-speech>



Fig. 6: Word cloud of major referring targets, illustrating the dataset's broad coverage and diversity.

Target	Pre-annotation	Re-annotation
	<p>A white sedan parked near the center of the image, in front of a blue truck and slightly to the left of a white car, on the paved area between two rows of apartment buildings.</p>	<p>A white sedan partially covered by leaves, located on the left in a row of cars beneath a blue truck, with an empty parking space separating it from a red car on the right.</p>
	<p>The car is a small, dark-colored vehicle located in the parking lot near the center of the scene. It is positioned close to the row of palm trees lining the edge of the marina, to the left of a curved walkway, and directly adjacent to another parked black car.</p>	<p>The white sedan is the seventh car from the right in the topmost row of the parking lot, flanked by a grayish-black sedan on its left and a white sedan with a black roof on its right.</p>

Fig. 7: Examples of pre-annotated expressions generated by large models compared to those re-annotated by humans.

the number of images and the number of annotated objects. Fig. 5 (c) reports the expression length distributions of the training & validation and test sets, with average lengths of 27.15 and 25.48 words, respectively. Fig. 5 (d) presents the distributions of the object-to-image area ratio, where the averages are 0.767% for the training & validation sets and 0.161% for the test set. Fig. 5 (e) provides a T-SNE visualization of image features, extracted using ResNet-50 from randomly sampled 1,000 training & validation images and 200 test images. The distinct clusters confirm the distributional difference between training/validation and test sets, validating our split strategy. Fig. 5 (f) illustrates a T-SNE visualization of randomly sampled 1,000 original expressions of training set and their 2,000 GPT-4-augmented counterparts, with features extracted by BERT. The augmented data expands the distribution, reflecting greater lexical and structural diversity while preserving semantic consistency. Finally, Fig. 6 depicts a word cloud of major described objects, highlighting the dataset's broad coverage and diversity.

3.6 Discussions

Large Models as Pre-Annotation Assistants. In our annotation pipeline, we employ large models (LM) to generate initial annotations, which are then refined by human annotators, leading to a substantial reduction in human annotation cost. The effectiveness of this strategy stems from the strong generalization ability that LM acquire through large-scale multi-modal or language pretraining. Nevertheless, there remains a non-negligible domain gap between general-purpose LM and UAV-specific tasks. As a result,

while model outputs often provide useful reference cues, they may also contain errors and noise.

Accordingly, we do not directly adopt the raw outputs of large models as final annotations, since their accuracy is insufficient for direct use. Instead, they serve as a starting point for human annotation. The informative parts can be leveraged by annotators to reduce workload. As shown in Fig. 7, although LM-generated pre-annotations may contain noise, they typically provide useful content that can be revised, thereby making the human annotation process more efficient. This paradigm shifts the human annotation task from “thinking and authoring” to “judging and editing,” which substantially lowers cognitive load and reduces annotation time. A comparison between annotating from scratch and refining pre-annotations shows that the latter achieves significant time savings, reducing the average annotation time by 38.8% (from 85 seconds to 52 seconds). The gain in efficiency mainly arises from this change in workflow, while the final annotation quality remains unaffected.

We do not fine-tune LMs for UAV-specific tasks for two main reasons. First, closed-source models such as GPT-4 do not provide fine-tuning access. Second, UAV datasets are relatively small compared to the scale required for LM pretraining, making fine-tuning prone to overfitting and loss of generalization, which may even reduce the amount of useful information contained in pre-annotations, while also introducing considerable computational overhead.

In summary, the “LM-based pre-annotation + human refinement” paradigm ensures final annotation quality while markedly reducing human cost, and avoids the access, overfitting, and computational issues of fine-tuning, thus offering strong practical value.

Privacy and ethical concerns. Our dataset is constructed in strict accordance with legal and ethical requirements. Images obtained from publicly available UAV datasets are properly cited, while those collected from the internet are explicitly marked with Creative Commons licenses to ensure lawful use and sharing. The image content is carefully controlled to include only aerial views of natural scenes and public areas, excluding restricted sites and sensitive facilities to avoid potential security risks. During the annotation process, we follow a non-identifiable principle, ensuring that no sensitive information linked to personal identity is included; annotations are restricted to general categories necessary for the intended research tasks. Furthermore, the dataset will be released under an academic use agreement that requires strict compliance with privacy and ethical standards and explicitly prohibits any use for surveillance or privacy-intrusive purposes.

4 AERIALREC BASELINE FRAMEWORK

Compared with general REC, UAV-based REC poses unique challenges. In UAV-captured scenes, images are typically acquired from high-altitude viewpoints, resulting in broad fields of view with complex backgrounds that contain numerous objects semantically similar to the target. Such conditions make localization particularly challenging, as the model needs to discriminate the referred target from a large number of distractors. Consequently, existing methods that

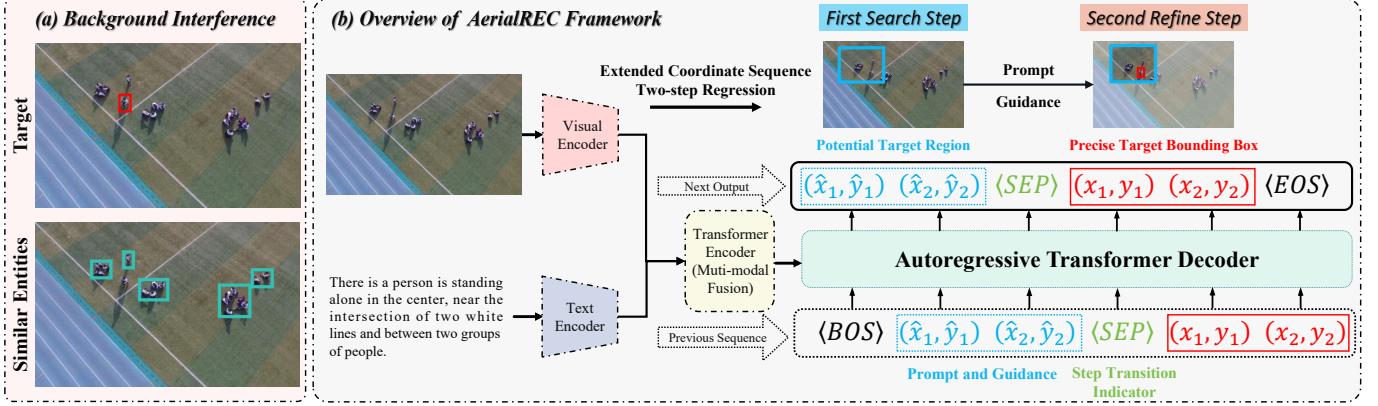


Fig. 8: (a) UAV-based REC with high-altitude views and complex backgrounds, where similar non-target entities hinder target localization. (b) Overview of the AerialREC baseline, which introduces a search step before localization to filter background interference.

directly regress the target box are prone to background interference, leading to degraded REC performance.

To address this challenge, we propose a baseline framework termed AerialREC, which formulates target localization as a two-step process, as shown in Fig. 8. Unlike existing methods that directly regress the precise target bounding box, our framework first introduces a search step, where the model learns to identify a potential target region. This step aims to delineate an approximate area of interest, effectively filtering out substantial irrelevant background interference and providing a more focused region for subsequent analysis. In the second step, the model then performs precise localization within this clearer and less cluttered region, thereby achieving more accurate predictions. Benefiting from the guidance of the potential target region identified in the first step, the target localization is carried out within a more focused area with reduced background interference, thereby leading to improved localization accuracy.

To validate the effectiveness of the proposed AerialREC framework, we instantiate it on recent sequence-to-sequence (seq2seq) REC methods, which represent a mainstream paradigm in current REC research. Specifically, these methods cast REC as a sequence modeling problem, where the coordinates of the target bounding box corners are generated in an autoregressive manner. A typical seq2seq REC model 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:

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

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.

In our framework, the sequence regression process for target localization is extended into a two-step formulation.

The extended coordinate sequence is defined as:

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

In the search step, we first regress the coordinates of a potential target region $\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^2$, aiming to filter out irrelevant background and provide a more focused candidate region:

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

where AR denotes the autoregressive transformer decoder. The potential region is not required to precisely align with the ground-truth bounding box but should encompass the target. To generate its label, we randomly sample from enlarged intervals during training:

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

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

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

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

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

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. In the next step, we regress the precise bounding box $\{(x_i, y_i)\}_{i=1}^2$. By nature, the auto-regressive transformer decoder generates each token conditioned on the previously predicted ones. This property naturally aligns with our two-step framework: the precise bounding box is regressed based on the potential target region predicted in the first step, which serves as a prompt and guidance for the subsequent prediction. In this way, the model focuses on a clearer sub-region with reduced background interference, thereby improving localization accuracy.

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

where $\langle \text{SEP} \rangle$ is a newly added special token that indicates the transition between the two steps. During inference, the coordinates obtained from the second step are used as the final target bounding box.

Theoretical Justification. Beyond the intuitive explanation, we provide a theoretical perspective from the viewpoint

of optimization to understand the effectiveness of the proposed two-stage framework.

Instead of directly regressing the precise target bounding box (PTBB), we introduce an intermediate objective: identifying a potential target region (PTR). Supervision is provided by coarse boxes b_c sampled from an expanded region $\mathcal{B}(b^*)$ of the ground-truth b^* . The two-stage losses are defined as:

$$\mathcal{L}_{\text{PTBB}}(b) = d(b, b^*), \quad \mathcal{L}_{\text{PTR}}(b) = \mathbb{E}_{b_c \sim \mathcal{B}(b^*)} d(b, b_c), \quad (10)$$

where $d(\cdot, \cdot)$ denotes a distance metric. Since b_c is sampled from $\mathcal{B}(b^*)$, this is equivalent to optimizing over a distributional perturbation around b^* . Let this induced distribution be denoted as $\nu(\epsilon)$, such that $b_c = b^* + \epsilon$, $\epsilon \sim \nu$. The PTR loss can then be expressed as:

$$\mathcal{L}_{\text{PTR}}(b) = \int d(b, b^* + \epsilon) \nu(\epsilon) d\epsilon = (\mathcal{L}_{\text{PTBB}} * \nu)(b), \quad (11)$$

where “ $*$ ” denotes convolution. Since convolution with a distribution corresponds to taking a weighted average of shifted copies of the original function, it naturally smooths out sharp variations. This formulation shows that \mathcal{L}_{PTR} is essentially a smoothed version of $\mathcal{L}_{\text{PTBB}}$: the sharp unimodal structure of $\mathcal{L}_{\text{PTBB}}$ centered at b^* is expanded into a broader and smoother basin of attraction. The smoothing effect provides informative gradients over a wider region, mitigating the risk of suboptimal local minima.

Based on this, the proposed two-stage framework can be interpreted through the lens of the continuation method [80], [81]. In general, continuation introduces a family of objective functions that gradually transition from a simplified surrogate to the exact target objective:

$$\mathcal{L}_\lambda(b) = (1 - \lambda) \cdot \mathcal{L}_{\text{coarse}}(b) + \lambda \cdot \mathcal{L}_{\text{fine}}(b), \quad \lambda \in [0, 1], \quad (12)$$

where λ controls the degree of transition. The benefit of this strategy lies in its ability to optimize a smoother and easier objective at the early stage of training, thereby alleviating the risk of being trapped in poor local minima and providing a stable convergence path for optimization. In our framework, the first step leverages coarse supervision to optimize toward the vicinity of the ground truth, while the second stage employs fine supervision to refine predictions for accurate localization. Throughout this process, the smoother loss landscape induced by coarse supervision helps steer optimization into the correct region, whereas fine supervision ensures precise alignment with the target.

Furthermore, our framework is conceptually consistent with the principle of Curriculum Learning (CL) [82], [83]. CL advocates training in an easy-to-difficult order, starting with simplified objectives and gradually progressing to more challenging ones. In our case, we adopt a similar philosophy by guiding the model to learn from a simpler objective (PTR) to a more complex one (PTBB), which helps improve performance on the final task.

Discussion. Regarding the use of random sampling rather than fixed values for generating PTR labels. From the optimization perspective discussed earlier, the surrogate objective induced by fixed values remains a single deterministic target and therefore does not alleviate the intrinsic difficulty of optimization; the loss landscape around that target stays

steep and narrow. By contrast, random sampling injects distributional perturbations around the ground truth and turns the objective from a point-wise one into distributional risk minimization, which is equivalent to convolving the original loss with the perturbation distribution. This convolution smooths the loss surface, lowers the optimization difficulty, and yields a more stable training trajectory.

Regarding whether decomposing localization into more steps can improve performance, this should be discussed in the context of the auto-regressive structure. In training, teacher forcing ensures that each decoding step receives ground-truth input. At inference, however, student forcing requires the model to condition on its own previous predictions, making later steps dependent on the accuracy of earlier ones. As the number of steps increases, prediction errors will be propagated and amplified, impairing localization accuracy. Our two-step design provides a practical trade-off, and introducing additional steps is unlikely to yield further benefits and may instead degrade performance due to compounding errors.

5 EXPERIMENTS

5.1 Implementation Details

In the process of generating pre-annotations with the assistance of large models, we utilize CogVLM [52] (CogVLM-grounding-generalist-v1.1-17B) for basic descriptions, LLaVA [53] (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

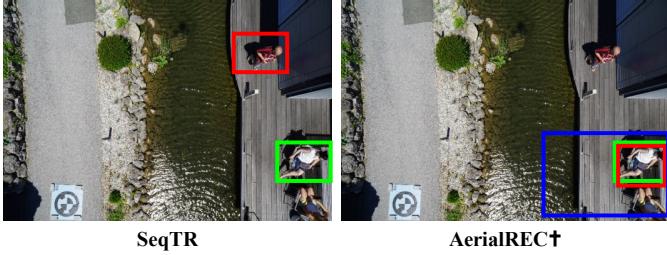
On the proposed SkyFind dataset, we select ten representative and publicly available REC methods to establish benchmark results, including FAOA [35], RSC [43], RefTR [44], TransVG [45], VGTR [48], VLVTG [49], SeqTR [47], QRNet [46], SimREC [51], and PolyFormer [50]. We train and evaluate all the above methods on the SkyFind dataset using the official implementations and training configurations. When pretraining is involved, we adopt the officially released pretrained weights. As shown in Table 4, PolyFormer and SeqTR achieve the top two overall performances. On the in-domain validation set, their IoU@0.5 scores reach 42.84 and 37.94, respectively; on the cross-domain test set, the scores reach to 31.01 and 25.74. Nevertheless, compared with their results on general REC benchmarks such as RefCOCO, RefCOCOg, and RefCOCO+, these numbers remain considerably lower. This demonstrates that UAV scenarios introduce substantial challenges to existing REC methods, highlighting the necessity of designing customized approaches tailored to this domain.

We further validate the effectiveness of the proposed AerialREC baseline framework in tackling the challenge of abundant background interference in UAV-based REC. The framework is instantiated on SeqTR and PolyFormer, two

TABLE 4: Benchmark results of 10 representative REC methods on our SkyFind validation and test sets. For the proposed AerialREC framework, we evaluate its effectiveness using two recent seq2seq REC methods, SeqTR and PolyFormer, denoted as AerialREC \dagger and AerialREC \diamond , respectively. The gray-highlighted rows present the comparison between the original performance and the results after incorporating our AerialREC framework, making the differences clearer.

Method	Visual Feature	Text Feature	SkyFind Val		SkyFind Test		Average	
			IoU@0.5	IoU@mean	IoU@0.5	IoU@mean	IoU@0.5	IoU@mean
FAOA [35]	DarkNet-53	BERT	23.10	10.21	13.08	6.91	18.09	8.56
RSC [43]	DarkNet-53	BERT	29.61	16.38	17.59	9.01	23.60	12.70
RefTR [44]	ResNet-101	BERT	31.22	15.80	22.68	13.46	26.95	14.63
TransVG [45]	ResNet-101	BERT	35.49	20.56	22.00	11.90	28.75	16.23
VGTR [48]	ResNet-101	LSTM	35.30	21.99	20.16	10.55	27.73	16.27
VLVTC [49]	ResNet-101	BERT	30.29	14.40	23.52	13.32	26.91	13.86
SeqTR [47]	DarkNet-53	GRU	37.49	24.45	25.74	12.57	31.62	18.51
QRNet [46]	Swin-S	BERT	33.90	22.10	26.21	11.22	30.06	16.66
SimREC [51]	CSPDarkNet-53	LSTM	31.17	22.09	21.50	12.15	26.34	17.12
PolyFormer [50]	Swin-B	BERT	42.84	25.50	31.01	16.44	36.93	20.97
AerialREC \dagger	DarkNet-53	GRU	42.95	28.06	32.13	18.00	37.54	23.03
AerialREC \diamond	Swin-B	LSTM	45.21	29.78	38.13	20.38	41.67	25.08

Expression: A person sitting on a wooden deck near the water's edge with legs crossed, wearing a white top and dark shorts.



Expression: The person wearing an orange shirt and black pants, standing in the top-left corner of the image.



Fig. 9: Qualitative comparison of methods with vs. without the proposed AerialREC framework. Green boxes denote the ground truth, blue boxes represent the potential target region predicted by AerialREC in the initial search step, and red boxes indicate the localization results.

representative sequence-to-sequence REC models, denoted as AerialREC \dagger and AerialREC \diamond , respectively. Experimental results show that, compared with SeqTR, AerialREC \dagger improves IoU@0.5 by 5.35 and 3.61 and IoU@mean by 6.39 and 5.43 on the validation and test sets. Similarly, compared with PolyFormer, AerialREC \diamond achieves gains of 2.37 and 4.28 in IoU@0.5 and 7.12 and 3.94 in IoU@mean on the validation and test sets. These results clearly shows that the proposed two-step localization framework effectively improves target localization accuracy in complex UAV-captured scenes with strong background interference. Moreover, the qualitative comparisons in Fig. 9 further reveal the mechanism of our framework. In the first step, the model leverages high-level referential cues in the expression to narrow down a coarse region. In the second step, fine-grained details are used to distinguish the target and accurately localize it. This two-stage process effectively alleviates confusion caused by salient and similar non-target entities in UAV imagery.

5.3 Ablation Studies

Framework Design. As shown in Table 5, We conduct a systematic ablation study to assess the effectiveness of key components in the two-step localization framework, including the introduction of the special token $\langle\text{SEP}\rangle$, the form of supervision applied in the first-step search, and whether extending localization to more steps yields additional benefits. Our analysis shows that random sampling (RS) of the

potential target region during the first-step supervision is essential for the effectiveness of the two-step framework:

$$(\hat{x}_1, \hat{y}_1) \sim \text{RS}(R_1), \quad (\hat{x}_2, \hat{y}_2) \sim \text{RS}(R_2). \quad (13)$$

In contrast, fixed-value strategies do not yield noticeable improvements. For example, the innermost case, where the potential region exactly matches the ground-truth box:

$$(\hat{x}_1, \hat{y}_1) = (x_1, y_1), \quad (\hat{x}_2, \hat{y}_2) = (x_2, y_2), \quad (14)$$

or the outermost case, where the region is deterministically expanded to its maximum:

$$(\hat{x}_1, \hat{y}_1) = (\max(0, x_1 - \alpha x_{\text{left}}), \max(0, y_1 - \alpha y_{\text{left}})), \quad (15)$$

$$(\hat{x}_2, \hat{y}_2) = (\min(w, x_2 + \alpha x_{\text{right}}), \min(h, y_2 + \alpha y_{\text{right}})), \quad (16)$$

do not bring significant gains. This observation is consistent with our theoretical analysis: using fixed values essentially reduces the task to point-wise regression, which neither smooths the loss nor eases optimization. In contrast, random sampling transforms the objective into distributional risk minimization, equivalent to convolving the original loss with the perturbation distribution. This operation smooths the loss surface, lowers optimization difficulty, and yields a more stable training trajectory. Moreover, introducing the special token $\langle\text{SEP}\rangle$ between the two regression steps further boosts performance by explicitly distinguishing the two stages and providing a clear modeling cue. Finally,

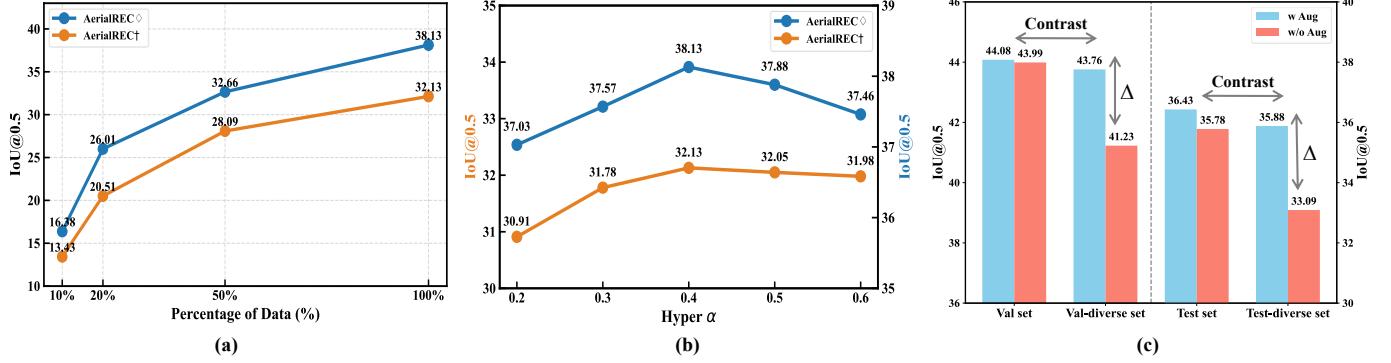


Fig. 10: (a) Effect of data scale (x-axis: percentage of training data). (b) Effect of hyper-parameter α on localization performance. (c) Effect of GPT-4 training augmentation on robustness to diverse expressions.



Origin: The person wearing a red life jacket on the far right of the boat, standing under the red sunshade and holding onto the rectangular silver structure on the boat.

Version 1: At the far right end of the boat sits a person in a red safety jacket, just below the red shade, keeping one hand on the rectangular silver structure.

Version 2: The person positioned at the far-right edge of the boat, under the red cover, holding the silver rectangular structure, and wearing a red life vest.

Fig. 11: Failure cases of the model trained without GPT-4 augmentation, where variations in expression style (e.g., word order or word choice) mislead the model and result in incorrect localization.

TABLE 5: Effect of components in constructing the extended regressed coordinate sequence. ‘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	SkyFind Test	
				IoU@0.5	IoU@mean
AerialREC \dagger	✓		—	25.74	12.57
			RandomSample	31.01	17.64
			Fixed (inner)	26.06	12.60
	✓	✓	Fixed (outer)	26.78	13.01
	✓	✓	RandomSample	32.13	18.00
	✓	✓	RandomSample	31.12	16.44
AerialREC \diamond	✓		—	28.74	15.61
			RandomSample	31.01	16.44
			Fixed (inner)	37.56	19.88
	✓	✓	Fixed (outer)	32.11	16.81
	✓	✓	RandomSample	33.89	17.02
	✓	✓	RandomSample	38.13	20.38
Three-step	✓	✓	RandomSample	38.01	19.55
	✓	✓	RandomSample	34.31	17.10

extending localization beyond two steps with repeated enlargement and random sampling does not lead to additional benefits. Instead, performance degrades due to the autoregressive nature of the framework, where student-forcing during inference accumulates and amplifies errors, thereby reducing localization accuracy.

Data Scale. In this work, we construct a large-scale dataset, SkyFind, which contains over one million target expressions. To analyze the effect of dataset scale, we conduct

experiments using subsets of 10%, 20%, 50%, and the full dataset, and report the corresponding performance on the test set. As shown in Fig. 10 (a), model performance consistently improves as the dataset size increases, but the marginal gains gradually diminish and approach saturation. This indicates that, while further expansion may still bring additional improvements, the current scale of SkyFind is already sufficient to support effective training and evaluation.

Hyper-parameter α . We investigate the impact of varying the hyper-parameter α on model performance. As shown in Fig. 10 (b), a smaller α forces the model to search within a relatively small area, which does not substantially alleviate the 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.

Impact of GPT-4 Training Augmentation. We employ GPT-4 to augment the expressions in the training set, exposing the model to a broader range of linguistic variations during training and thereby enhancing its robustness in understanding diverse expressions at test time. To verify the effectiveness of this augmentation, we design two training settings: using only the original training set and using the augmented training set. To ensure a fair comparison with the same number of training iterations, in the augmented setting we sample one expression per target from the pool of original and augmented expressions, so that the effective training set size matches that of the original-only setting.

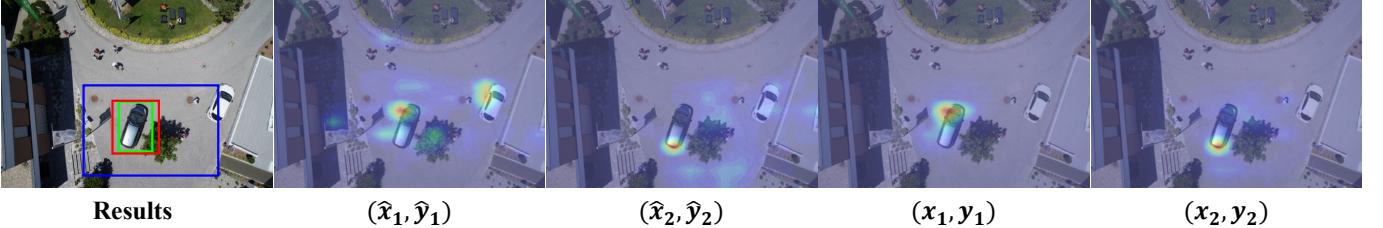
Expression: A silver sedan parked near the bottom-left corner of the image, close to some greenery.

Fig. 12: The cross-attention maps of the decoder when generating each new vertex token, using AerialREC \diamond 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.

TABLE 6: Zero-shot performance of 12 large-scale vision-language pretrained large models on the SkyFind test set.

Method	Visual Feature	Text Feature	SkyFind Test	
			IoU@0.5	IoU@mean
OFA [84]	ResNet-152	—	13.01	6.14
OBE-PEACE [59]	—	—	11.03	6.56
Grounding-DINO [61]	Swin-B	BERT	15.16	8.25
UNINEXT [56]	ConvNeXt-L	BERT	12.70	7.40
Sphinx [57]	Mixed	LLaMA	13.54	10.22
Shikra [58]	OpenCLIP-L	Vicuna	14.95	11.85
Qwen-VL [85]	OpenCLIP-G	Qwen	17.11	12.78
MiniGPT-V2 [60]	EVA-CLIP-G	LLaMA	18.80	13.64
Ferret [55]	CLIP-L	Vicuna	16.02	11.63
GLEE [54]	EVA02-CLIP-L	CLIP	14.09	12.11
LLaVA [53]	OpenCLIP-L	Vicuna	19.12	15.75
CogVLM [52]	EVA02-CLIP-E	LLaMA	23.67	16.99

For evaluation, we report performance on the validation and test sets, and additionally construct val-diverse and test-diverse by manually writing two extra expressions for each sample to simulate different expression styles. We then evaluate both models on val, val-diverse, test, and test-diverse (see Fig. 10 (c)). Experimental results show that, compared with performance on val and test, the model trained on the original set exhibits a larger drop on val-diverse and test-diverse, whereas the model trained with GPT-4 augmentation remains substantially robust under such variations. We further conduct a qualitative analysis of failure cases (see Fig. 11). Differences in expression style, such as altering word order (e.g., placing the spatial relation to “the boat” at the beginning) or word choice (e.g., replacing sit with position), directly mislead the model trained without augmentation. In the first case, the mention of “boat” at the beginning draws the model’s attention prematurely to the wrong entity, causing it to ignore the subsequent discriminative description. In the second case, replacing sit with position weakens the specificity of the expression, leading the model to confuse the target with another person. In contrast, the augmented model remains robust to such variations, verifying both the effectiveness and necessity of GPT-4 augmentation.

Visualization of Attention Map. Here, we visualize the cross-attention maps (averaged across all layers and heads) during the regression of each new vertex token, as shown in Fig. 12. In the search step, the model predicts the potential target region, $(\hat{x}_i, \hat{y}_i)_{i=1}^2$, with attention primarily

distributed over the broader area containing the specified car. In the subsequent refine step, guided by the narrowed candidate region, the model produces more accurate predictions, $(x_i, y_i)_{i=1}^2$, with attention now concentrated on the black car. These visualizations demonstrate that beyond the output results, the model internally learns a two-stage mechanism, first focusing on a coarse region and then refining to precisely localize the target.

Evaluation on Large Models. Recent advances in large-scale vision-language pretraining have given rise to a series of powerful large models with strong generalization ability across diverse multi-modal tasks. These models are equipped with highly capable visual and textual feature extractors and are trained on hundreds of millions to billions of image-text pairs, thereby achieving remarkable zero-shot performance on many established benchmarks, including REC. Motivated by this, we extend our evaluation to 12 representative vision-language models: OFA [84], ONE-PEACE [59], Grounding-DINO [61], UNINEXT [56], Sphinx [57], Shikra [58], Qwen-VL [85], MiniGPT-V2 [60], Ferret [55], GLEE [54], LLaVA [53], and CogVLM [52]. We perform zero-shot evaluation on the SkyFind test set to examine whether their generalization ability transfers to the newly introduced UAV-based REC setting, following the official evaluation codes and released weights. As reported in Table 6, the results are unsatisfactory, with most models achieving IoU@0.5 scores below 20. This outcome underscores the distinctive challenges inherent to UAV-based REC and highlights the limitations of existing large models in this setting. At the same time, it also shows the value of SkyFind dataset: incorporating SkyFind into the pretraining process has the potential to enhance the generalization ability of large models in aerial scenarios.

6 CONCLUSION AND FUTURE WORK

Conclusion. In this paper, we formally introduce UAV-based referring expression comprehension (REC) as a new research problem and analyze its unique challenges, including abundant background interference, small target size, and complex referring relations. To support systematic research in this direction, we construct SkyFind, a large-scale dataset with one million high-quality target-expression pairs. Furthermore, we propose AerialREC, a baseline framework that addresses the inherent difficulties of UAV-based REC by introducing a two-step localization process

to reduce background interference. Comprehensive experiments on SkyFind establish the benchmark and validate the effectiveness of AerialREC, while also revealing substantial room for future improvement. We hope that SkyFind and AerialREC can serve as strong foundations and catalysts for advancing research in human-UAV interaction.

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 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 [86], [87]. 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 [88]–[90]. 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 framework represents a starting point for advancing REC methods in aerial scenarios. Building on this, several promising research directions emerge. A 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. E. Wang, "Aerial vision-and-dialog navigation," *arXiv preprint arXiv:2205.12219*, 2022.
- [8] J. Lee, T. Miyanishi, S. Kurita, K. Sakamoto, D. Azuma, Y. Matsuo, and N. Inoue, "Citynav: Language-goal aerial navigation dataset with geographic information," *arXiv preprint arXiv:2406.14240*, 2024.
- [9] K. S. Lee, M. Ovinis, T. Nagarajan, R. Seulin, and O. Morel, "Autonomous patrol and surveillance system using unmanned aerial vehicles," in *2015 IEEE 15th International Conference on Environment and Electrical Engineering (EEEIC)*. IEEE, 2015, pp. 1291–1297.
- [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] Y. Qiao, C. Deng, and Q. Wu, "Referring expression comprehension: A survey of methods and datasets," *IEEE Transactions on Multimedia*, vol. 23, pp. 4426–4440, 2020.
- [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] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014, pp. 787–798.
- [15] 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.
- [16] R. Liu, C. Liu, Y. Bai, and A. L. Yuille, "Clevr-ref+: Diagnosing visual reasoning with referring expressions," in *CVPR*, 2019, pp. 4185–4194.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] S. He, H. Ding, C. Liu, and X. Jiang, "Grec: Generalized referring expression comprehension," *arXiv preprint arXiv:2308.16182*, 2023.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Lioung, 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.
- [27] 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.

- [28] 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.
- [29] 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.
- [30] J. Liu, L. Wang, and M.-H. Yang, "Referring expression generation and comprehension via attributes," in *ICCV*, 2017, pp. 4856–4864.
- [31] R. Luo and G. Shakhnarovich, "Comprehension-guided referring expressions," in *CVPR*, 2017, pp. 7102–7111.
- [32] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *ICCV*, 2019, pp. 4694–4703.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," in *NeurIPS*, pp. 19 652–19 664, 2021.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," in *ICME*, 2022.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvilm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.
- [53] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024, pp. 26 296–26 306.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," in *NeurIPS*, pp. 6616–6628, 2020.
- [64] 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.
- [65] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "Seadroneesee: A maritime benchmark for detecting humans in open water," in *WACV*, 2022, pp. 2260–2270.
- [66] 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.
- [67] 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.
- [68] 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.
- [69] 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.
- [70] 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.
- [71] 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.
- [72] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *ECCV*, 2016, pp. 549–565.
- [73] I. team, "Semantic Drone Dataset," <http://dronedataset.icg.tugraz.at>.
- [74] 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.
- [75] 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.

- [76] 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.
- [77] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *WACV*, 2018, pp. 1499–1508.
- [78] 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.
- [79] A. Shvedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," in *ICCV*, 2023, pp. 11987–11997.
- [80] E. L. Allgower and K. Georg, *Numerical continuation methods: an introduction*. Springer Science & Business Media, 2012, vol. 13.
- [81] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.
- [82] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.
- [83] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *International Journal of Computer Vision*, vol. 130, no. 6, pp. 1526–1565, 2022.
- [84] 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. 23318–23340.
- [85] 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.
- [86] 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.
- [87] 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. 15435–15459, 2022.
- [88] 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.
- [89] 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.
- [90] 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 machine learning and embodied AI.



Guanbo Wu is pursuing a bachelor's degree at the School of Information Science and Technology of the University of Science and Technology of China.



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, TNNLS, 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.