Panoptic Perception: A Novel Task and Fine-grained Dataset for Universal Remote Sensing Image Interpretation

Danpei Zhao, *Member, IEEE,* Bo Yuan, Ziqiang Chen, Tian Li, Zhuoran Liu, Wentao Li, and Yue Gao

***Abstract*—Current remote-sensing interpretation models often focus on a single task such as detection, segmentation, or caption. However, the task-specific designed models are unattainable to achieve the comprehensive multi-level interpretation of images. The field also lacks support for multi-task joint interpretation datasets. In this paper, we propose Panoptic Perception, a novel task and a new fine-grained dataset (FineGrip) to achieve a more thorough and universal interpretation for RSIs. The new task, 1) integrates pixel-level, instance-level, and image-level information for universal image perception, 2) captures image information from coarse to fine granularity, achieving deeper scene understanding and description, and 3) enables various independent tasks to complement and enhance each other through multi-task learning. By emphasizing multi-task interactions and the consistency of perception results, this task enables the simultaneous processing of fine-grained foreground instance segmentation, background semantic segmentation, and global fine-grained image captioning. Concretely, the FineGrip dataset includes 2,649 remote sensing images, 12,054 fine-grained instance segmentation masks belonging to 20 foreground things categories, 7,599 background semantic masks for 5 stuff classes and 13,245 captioning sentences. Furthermore, we propose a joint optimization-based panoptic perception model. Experimental results on FineGrip demonstrate the feasibility of the panoptic perception task and the beneficial effect of multi-task joint optimization on individual tasks. The dataset will be publicly available [1].**

***Index Terms*—remote sensing images, panoptic perception, fine-grained interpretation, benchmark dataset, multi-task learning**

I. Introduction

REMOTE sensing image (RSI) interpretation has witnessed a rapid development tendency in multiple tasks, including image classification [1], object detection [2], semantic segmentation [3], instance segmentation [4], image caption generation [5], etc. However, these tasks only cover single task interpretation. Particularly, object detection focuses on locating and classifying individual objects within images. Instance segmentation identifies refined contours. Semantic segmentation aims to assign a class label to every pixel. While image caption generation attempts to provide a holistic understanding of RSIs. However, models of these tasks are usually designed independently and overlook the rich semantics and contextual relationships in RSIs. However, in the field of remote sensing observation, a very common and practical expectation is to achieve multi-level, fine-grained, perceptive interpretation of RSIs.

Recently, new efforts have emerged to promote a more comprehensive RSI interpretation. For example, panoptic segmentation [6] combines the advantages of instance and semantic segmentation, enabling simultaneous foreground instance masking and background pixel classification. However, the datasets and research on RSI panoptic segmentation [7] are scarce. However, panoptic segmentation still focuses on pixel-level and instance-level interpretation. Another frontier research focus, i.e., fine-grained object recognition [8] emerges as a pivotal task that focuses on identifying specific subcategories of target objects. However, these tasks can not cope with multi-modal interpretation covering from pixel-level to image-level, lacking the comprehensive perception capacity and universal interpretation models across multi-modal tasks.

In this paper, we introduce Panoptic Perception, a novel task for comprehensive remote sensing image interpretation along with a new benchmark dataset FineGrip. As shown in Fig. 1, panoptic perception can concurrently handle various subtasks across multi-level interpretation, including fine-grained instance segmentation of foreground instances, semantic segmentation for background areas, and image caption generation. This innovative task diverges from conventional tasks by not solely focusing on individual interpretation levels but leverages mutually reinforcing and interactive optimization. The collaborative processing of multiple tasks necessitates that the model comprehensively understand the global contextual relationships and semantic information at different levels. This, in turn, amplifies the model capability to extract and utilize the abundant information in RSIs. The proposed panoptic perception integrates pixel-level, instance-level, and image-level understanding to construct a universal interpretation framework.

We construct a novel benchmark dataset named FineGrip to support the development of the new task. It comprises 2,649 remote sensing images, with fine-grained aircraft instance segmentation annotations, diverse background semantics and fine-grained sentence description annotations. To the best of our knowledge, this is the first dataset integrating fine-grained

This work was supported by the National Natural Science Foundation of China under Grant 62271018. (Corresponding author: Danpei Zhao.)

Danpei Zhao, Bo Yuan, Tian Li are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and also with Tianmushan Laboratory, Hangzhou 311115, China (e-mail: zhaodanpei@buaa.edu.cn, yuanbobuaa@buaa.edu.cn, lit@buaa.edu.cn).

Ziqiang Chen, Zhuoran Liu, Wentao Li and Yue Gao are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: chenziqiang@buaa.edu.cn, liuzhuoran@buaa.edu.cn, canoe@buaa.edu.cn, bjlguniversity@163.com).

[1]https://ybio.github.io/FineGrip/

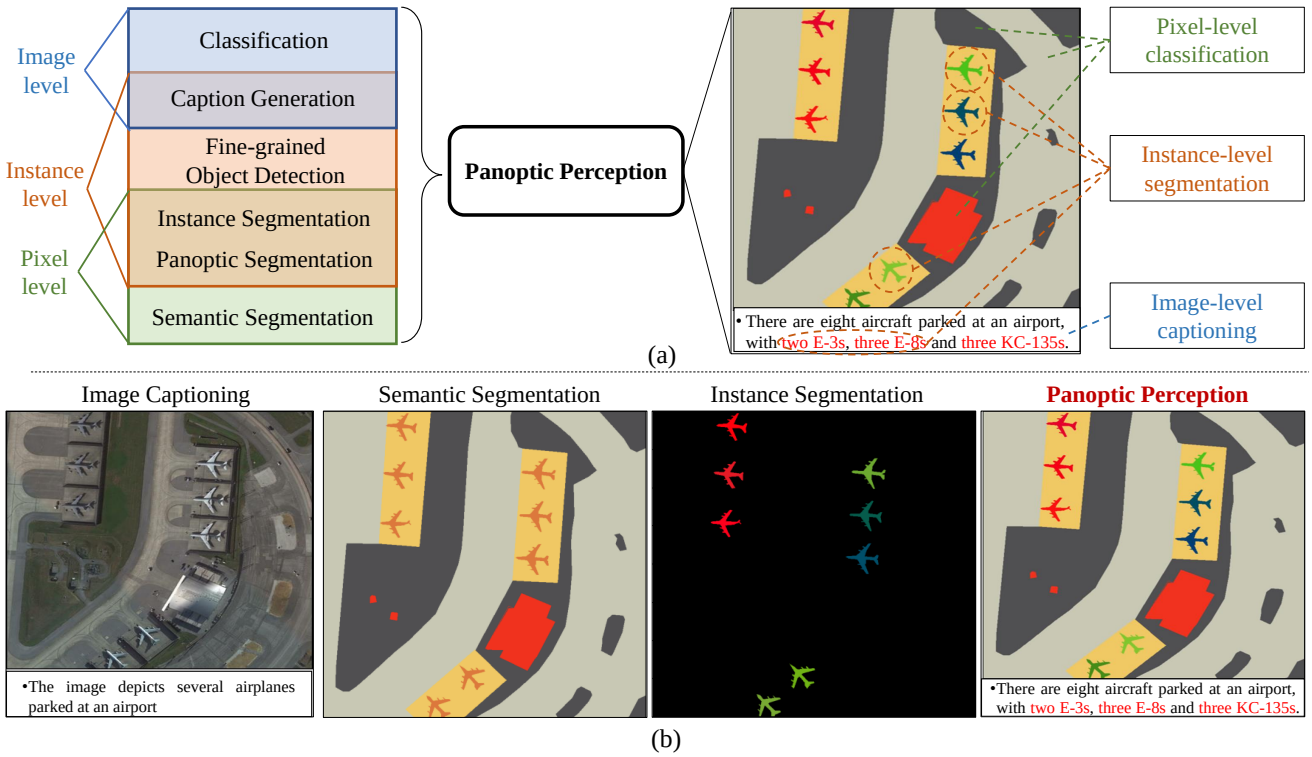

Fig. 1. The proposed fine-grained Panoptic Perception task. (a) The definition and scope of panoptic perception Task. (b) Qualitative comparison of panoptic perception and other interpretation tasks.

detection, instance segmentation, semantic segmentation, and fine-grained image caption annotations for RSIs. Additionally, we utilized the Segment Anything Model (SAM) to construct a semi-automatic segmentation annotation system. It fully leverages the robust zero-shot transfer capabilities of SAM, significantly enhancing the annotation efficiency of foreground segmentation.

To validate the feasibility of the proposed panoptic perception and the effectiveness of the dataset, we propose an end-to-end panoptic perception baseline model. The experimental results corroborated the viability of panoptic perception tasks and the beneficial impact of multi-task joint optimization on individual task enhancement.

The main contributions of this paper are summarized as follows:

- We propose panoptic perception, a novel and comprehensive interpretation task that synchronously performs pixel-level, instance-level and image-level perception.
- We present FineGrip, which to our best knowledge, serves as the first fine-grained panoptic perception dataset to support the multi-task interpretation of RSIs.
- A new semi-automatic annotation system is designed to enhance annotation efficiency by combining the zero-shot generalization ability of foundation models with the fitting capability of supervised models.
- We introduce an end-to-end panoptic perception baseline method that offers referential research directions for subsequent studies.

II. Related Works

A. Remote Sensing Datasets for Various Interpretation Tasks

Currently, a variety of task settings are employed in the interpretation of RSIs. These tasks range from pixel-level, instance-level to image-level, and low-level details to high-level semantics, providing a multi-faceted understanding of RSIs.

Semantic segmentation is a pixel-level interpretation task that aims to assign a category label to each pixel. There are multiple RSI semantic segmentation datasets constructed for various applications such as water area extraction [9], cloud and fog detection [10], [11], land cover and vegetation classification [12]–[15], road segmentation [16], [17], and disaster monitoring [18], etc. Semantic segmentation has achieved admirable accomplishments in these scenarios. However, the inability to distinguish between foreground instances limits its effectiveness in tasks focusing on specific foreground objects.

Instance-level interpretation tasks include object detection and instance segmentation. The former aims to locate and identify objects of interest. There are numerous datasets constructed for detecting typical objects such as airplanes, ships, oil tanks, bridges, and vehicles [2], [19]–[21]. The emerging fine-grained object detection tasks [8], [22] require recognizing subtle distinctions among sub-categories of objects. It goes beyond merely identifying a general category like an *airplane* and extends to specifying the exact type, such as *Boeing 747*. Instance segmentation [23], [24] provides an additional pixel-level mask for each identified object, enriching the understanding of the exact contour of objects. The localization, identifi-

cation, and segmentation of typical remote sensing targets are crucial in applications like disaster loss assessment [25] and military surveillance [26]. However, a challenge emerges from the sparse distribution of foreground objects in RSIs. Relying solely on the foreground object features could result in the significant neglect of invaluable background information.

The goal of image caption generation for RSIs is to perceive the overall situation and automatically generate text that describes the content of the images. In contrast to the prevalent focus on delineating actions and behaviors in natural images, datasets for RSI captioning [5], [15], [27] prioritize the recognition of surface attributes, geographical formations, and the interrelationships among geophysical entities, etc. It provides a concise overview of the image content. However, its inherent generality and ambiguity may cause models to overlook critical details such as the exact number, size, and type of targets.

Despite the specific significant progress, these tasks remain focused on specific scenarios and fail to fully extract and utilize the wealth of rich information contained in RSIs. A more thorough understanding of RSIs necessitates multi-modal and multi-task interpretation.

B. Panoptic Segmentation for RSIs

Panoptic segmentation [6] is a generalized segmentation task that offers a unified framework to address background semantic segmentation (Stuff) and foreground instance segmentation(Things). Specifically, it aims to assign a category label to every pixel in an image and distinguish different instances of the same category. Standard benchmark datasets for panoptic segmentation tasks include COCO-Panoptic [28] and Cityscapes [29]. However, panoptic segmentation in RSIs faces the following challenges. First, there is a lack of high-quality fine-grained datasets. PASTIS [30] proposes an open-source dataset for panoptic segmentation of RSIs. However, the restricted image size of 128×128 limits its application. BSB aerial dataset [7] contains fine-grained instance annotations only for building targets. On the other hand, currently most methods are carried out on a few categories with high differentiation such as [31], [32], and there is a lack of fine-grained labels with foreground instances. In addition, current panoptic segmentation methods lack the ability of multi-modal interpretation.

In this paper, we extend RSI interpretation from single-task to multi-modal panoptic perception. A new panoptic perception dataset is proposed, which supports multi-modal tasks including pixel-level classification, instance-level segmentation and image captioning.

C. Efficient Data Annotation

As the foundation for training robust segmentation models, accurate annotations offer benchmarks against model predictions and guide optimization. Over the years, high-quality annotated datasets have notably advanced RSI segmentation.

Traditional manual annotation tools, such as Labelme [33], achieve high precision by allowing users to delineate target contours with polygons. When it comes to annotating complex-shaped targets, these tools prove labor-intensive. With the expanding volume of RSI datasets, Improving annotation efficiency becomes a challenge.

The zero-shot learning capabilities of foundation models make them highly effective for annotating massive datasets. For example, the Segment Anything Model (SAM) [34], with its ability to effectively segment specific areas based on provided box prompts, inspires the development of numerous semi-automatic annotation systems. ISAT [35] integrates Labelme with SAM, facilitating interactive semi-automatic segmentation annotation. It enables users to craft and modify boundaries, which are taken as prompts into SAM to achieve target-specific segmentation. Semantic SAM [36] emerges as the first open framework employing SAM for semantic segmentation tasks. It smoothly enhances performance without fine-tuning the weights of SAM. Grounded-Segment-Anything fuses SAM with Grounding-DINO [37], delivering zero-shot detection results through textual input. The output bounding boxes then serve as prompts for instance segmentation.

The lack of familiarity with typical remote sensing targets poses challenges for the use of SAM in RSI annotation. While SAMRS [38] effectively converts bounding boxes into masks using original box annotations as prompts, It constrains category expansion because of the reliance on existing detection annotations. Meanwhile, fine-tuning SAM can address domain discrepancies but needs a large amount of training data and incurs significant computational costs. We aim to leverage SAM for quality annotations with limited resources and fewer images. Thus, instead of fine-tuning its parameters, we use SAM for inference and employ its results to train a more efficient segmentation model for enhanced outcomes.

III. Task Definition

In this paper, we propose a fine-grained, unified framework to simultaneously achieve pixel-level, instance-level, and image-level interpretation of RSIs. As illustrated in Fig. 1, beyond, our proposed task exceeds traditional single-task and necessitates models to extract more comprehensive contextual features and enables the joint interpretation of multiple tasks at various levels:

1) At the image level, the task demands the model to generate a concise overview of the entire content of the image and output this overview using natural language.

2) At the instance level, the model identifies the fine-grained categories of all foreground objects, distinguishes between different instances within the same category, and predicts an accurate contour for each instance. The task also demands the model to specify the number and specific categories of all foreground instances in its descriptive sentences.

3) At the pixel level, the task necessitates that each pixel in the image be assigned a distinct category of either foreground or background. Moreover, pixels associated with different foreground instances must be assigned a unique identifier.

Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we define a set of words $Wds = \{wd_1, wd_2, ..., wd_W\}$ and a set of categories $C^P = \{c_1, c_2, ..., c_C\}$, where W, C is the total number of words and categories, respectively. C^P can be further divided into

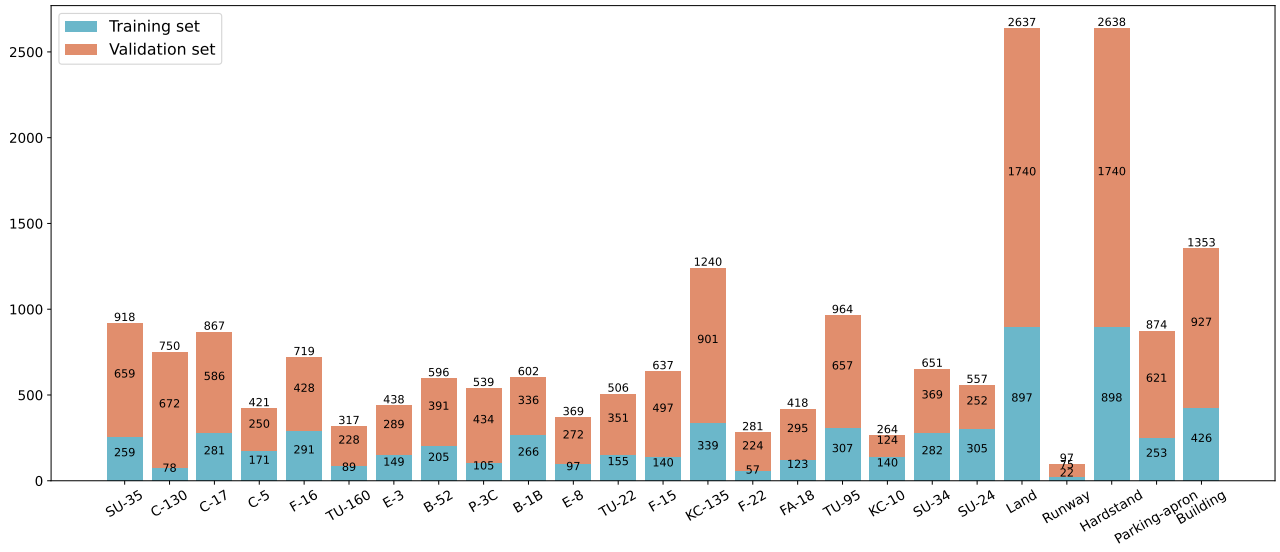

Fig. 2. Number of per-category masks in the FineGrip dataset across training and validation sets.

a foreground category (things) set C^{Th} and a background category (stuff) set C^{St}, where $C^{Th} \cap C^{St} = \emptyset$. The objectives of the fine-grained panoptic perception task are formally defined as follows:

1) For any given pixel (x, y) in the image, the model is required to predict both the category and the instance id of the pixel, denoted as $(c_{x,y}, id_{x,y})$. All pixels within the same instance should share identical category and number identifications. When a pixel belongs to a stuff category, the predicted instance id should be $\emptyset$.

2) Considering a maximum sentence length L, the model should generate a descriptive sentence for the image, represented as $\{w_1, w_2, ..., w_L | w_i \in Wds\}$. This sentence must include information about the number and types of foreground objects in the image.

Fine-grained panoptic perception demands the consistency of perception results across sub-tasks. As shown in Fig. 1(a), the caption concerning the quantities and types of foreground instances should align with the segmentation results.

Unlike a simple combination of multiple independent tasks, fine-grained panoptic perception emphasizes the complementarity and joint optimization among sub-tasks to achieve integrated image perception. Semantic segmentation provides detailed information about the background areas of RSIs, which aids in detecting and masking foreground instances. Meanwhile, fine-grained instance segmentation of foreground classes yields information on the number, category, and relative position of targets, which serve as critical materials for fine-grained caption generation. Moreover, accurate descriptive sentences can also help achieve more precise segmentation due to the constraint of consistency in perception results.

For the segmentation sub-task, following [6], we employ Panoptic Quality (PQ) to assess the performance and utilize PQ^{th} and PQ^{st} to measure the segmentation quality on *things* and *stuff* classes. Moreover, Recognition Quality (RQ) and Segmentation Quality (SQ) are applied to analyze the recognition and segmentation performance. As for image caption generation, we use the BLEU [39] metric to examine the overall description quality. Details of these evaluation metrics are introduced in section V.

IV. Dataset Construction

Existing datasets originate from diverse sources and are usually annotated for specific tasks, making them inadequate for the unified perception task defined in this paper. Therefore, we develop a fine-grained panoptic perception benchmark dataset based on a novel semi-automatic annotation system. In this section, we will provide a detailed introduction to the benchmark dataset and the semi-automatic segmentation annotation system.

A. The FineGrip Dataset

To facilitate the research on the new task, we constructed a Fine-Grained Panoptic Perception (FineGrip) Dataset encompassing 20 fine-grained foreground aircraft categories and 5 background stuff categories. The sample images in the proposed FineGrip are mainly derived from MAR20 [40] but with further cleaning and replenishment. As seen in Table I, the original MAR20 only contains bounding box annotations. However, we extend the annotations to support fine-grained instance segmentation, semantic segmentation, and image caption generation tasks. Concretely, FineGrip comprises 2,649 remote sensing images, 12,054 instance segmentation masks across 20 *things* categories, and 7,599 background semantic masks across 5 *stuff* categories. And 13,245 sentences with fine-grained category indications. The *things* categories include SU-35, C-130, C-17, C-5, F-16, TU-160, E-3, B-52, P-3C, B-1B, E-8, TU-22, F-15, KC-135, F-22, FA-18, TU-95, KC-10, SU-34, and SU-24. For representative convenience, the categories are respectively represented by $A1 \sim A20$. While the background *stuff* categories comprise *Land*, *Runway*, *Hardstand*, *Parking-apron* and *Building*. Fig. 2 illustrates

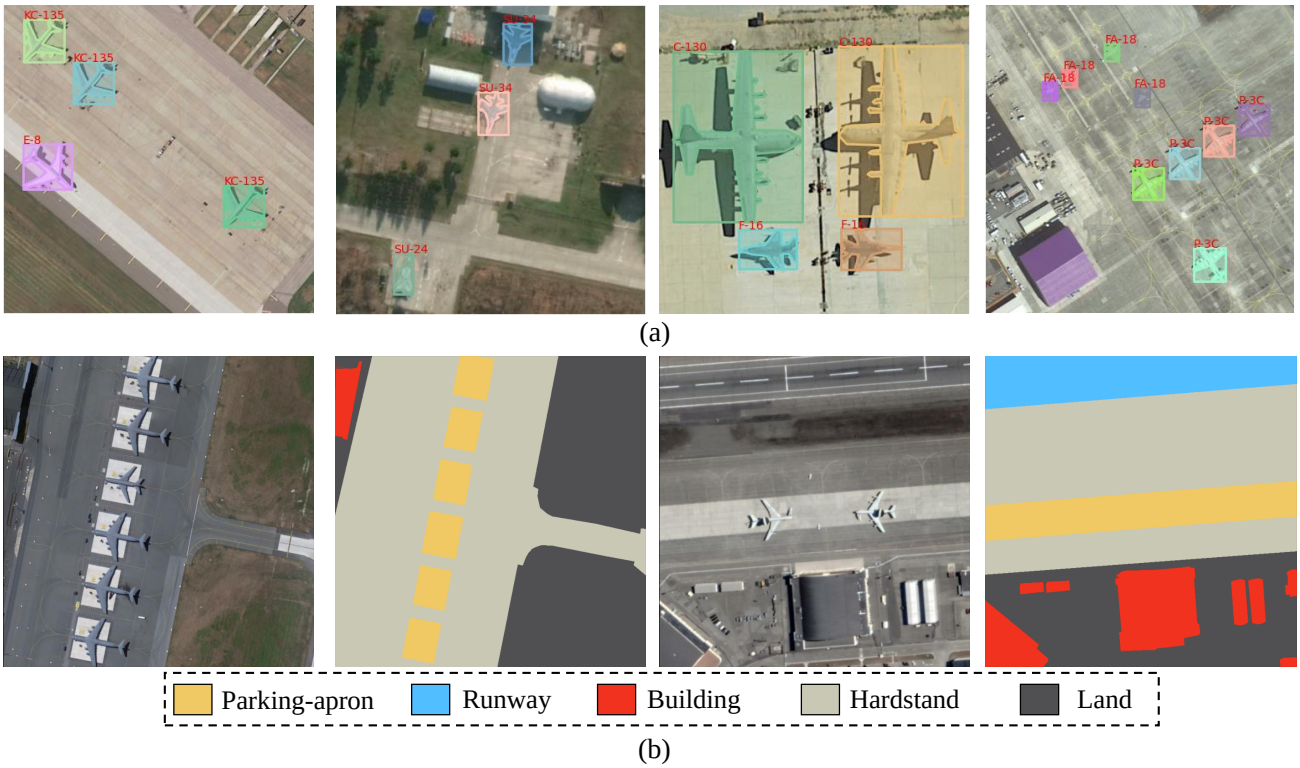

Fig. 3. Examples of Annotations in FineGrip. (a) Foreground fine-grained instance segmentation annotations.(b) Background stuff semantic segmentation annotations.

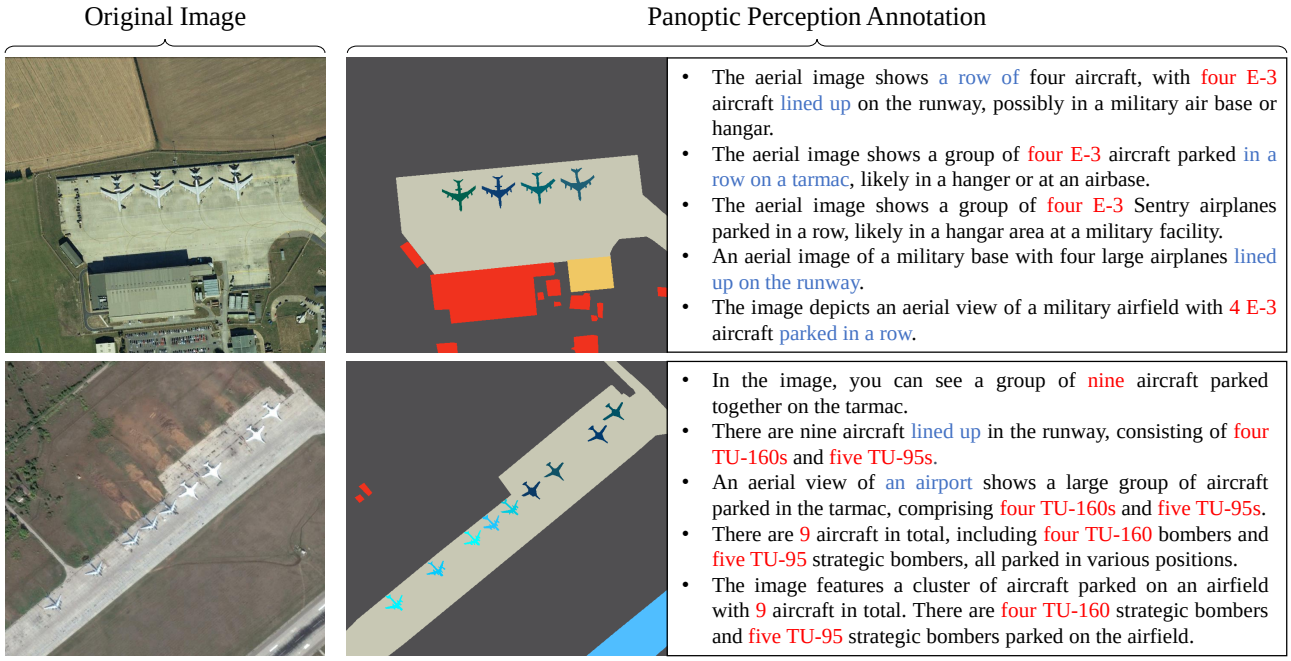

Fig. 4. Examples of Per-Image Annotations from FineGrip. In the captions, red highlights detail the fine-grained categories and quantities of aircraft instances, while blue highlights indicate relationships between instances.

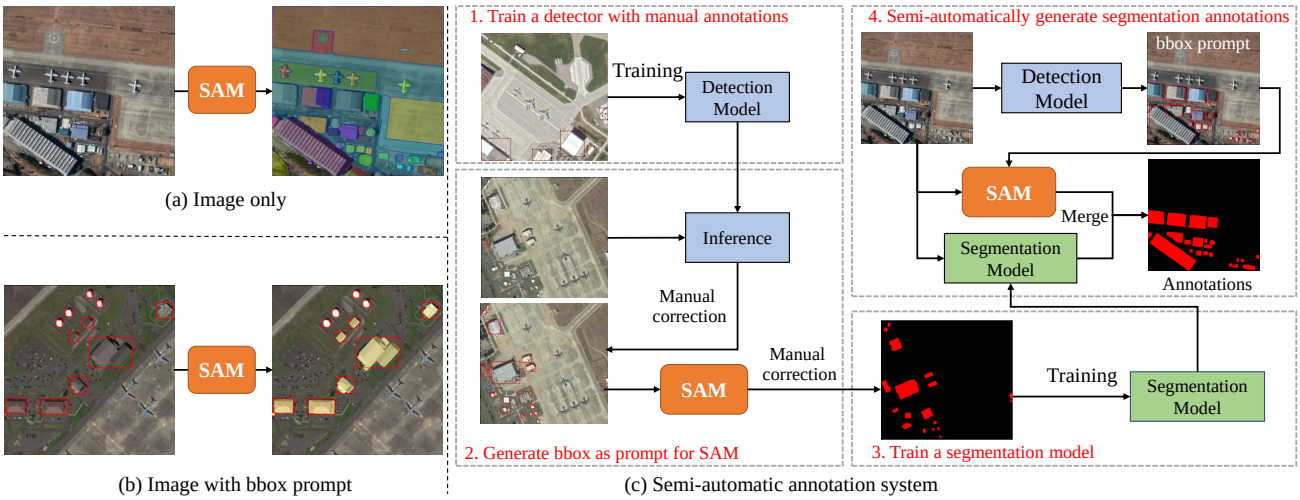

Fig. 5. Examples of Per-Image Complete Annotations from FineGrip. In the captions, red highlights detail the fine-grained categories and quantities of aircraft instances, while blue highlights indicate relationships between instances.

TABLE I
THE PROPOSED FINEGRIP DATASET.

Dataset	Class num.	Sample num.	Bbox	Masks	Caption	Fine grained	Cross modal
MAR20 [40]	20	3842	✓	×	×	✓	×
BSB [7]	15	3400	✓	✓	×	×	×
FineGrip	25	2649	✓	✓	✓	✓	✓

the number of segmentation masks for each category in the training/testing sets. The dataset contains 901 images for training and 1,748 images for testing.

The FineGrip dataset focuses on airport scenes, where various aircraft models are our primary foreground categories of interest. We annotated 20 fine-grained aircraft targets with instance-specific segmentation. Fig. 3 showcases some examples of FineGrip. In the context of background (*stuff*) categories, we prioritize areas closely related to aircraft targets. We define *Runway* as a long straight road marked with lines. A *Parking-apron* is a notable region where aircraft are parked for extended periods. *Hardstand* refers to areas, other than the previously mentioned two kinds, where aircraft can taxi. Buildings are categorized as background rather than foreground because we are not concerned with the specific instance segmentation of buildings in this scene. We merely need to identify which areas in the image pertain to buildings. Moreover, unlike aircraft targets, buildings often present a closely interconnected structure, complicating instance-level differentiation. While *Land* designates areas in the image other than the foreground aircraft targets and the mentioned stuff categories. Fig. 3 shows some examples of segmentation annotations for *things* and *stuff* categories.

As for the fine-grained image caption task, we place emphasis on the precise number and models of the foreground targets, as well as the relationships among the physical features. We generate five captions from five different annotators for each image to foster caption diversity. Ultimately, by integrating fine-grained instance segmentation, background semantic segmentation, and fine-grained caption annotations, we establish the FineGrip dataset. Fig. 4 displays some complete annotation examples from FineGrip.

Compared to conventional interpretation tasks and the recently proposed RSI panoptic segmentation dataset, our FineGrip exhibits significant characteristics in the following aspects:

- ***Abundant fine-grained semantic categories.*** The proposed FineGrip covers 20 fine-grained foreground categories and 5 densely labeled background categories. The samples from different categories possess diverse semantics, wide terrain scenes and complex semantic relations, etc. Besides, it also meets the practical challenges of small inter-class differences and large intra-class differences.
- ***Broader granularity of caption sentences.*** The captioning annotations span from general to specific granularity, delivering a comprehensive view of images. It is also fine-grained and consistent with the pixel-level annotations. In addition, the complex semantic relations are depicted in detail to achieve human-like perception from a global perspective. It gives a general overview of the image and pinpoints the exact count and model of primary targets.
- ***Affinity exploration of foreground-background relationships.*** In FineGrip, the foreground and background categories share a close relationship. For example, aircraft is predominantly parked in *Parking apron* or *Hardstand* areas, but rarely seen in *Land* areas. Moreover, building zones are often separated by hardstand areas. Such objective factors suggest that the panoptic perception model should consider these semantic relations, i.e., the foreground recognition and the background segmentation have the potential of mutual boosting.
- ***Synergized multi-tasking.*** Coordinating the instance segmentation and image caption tasks, which both identify the target quantity and sub-category, can mutually enhance their performance.

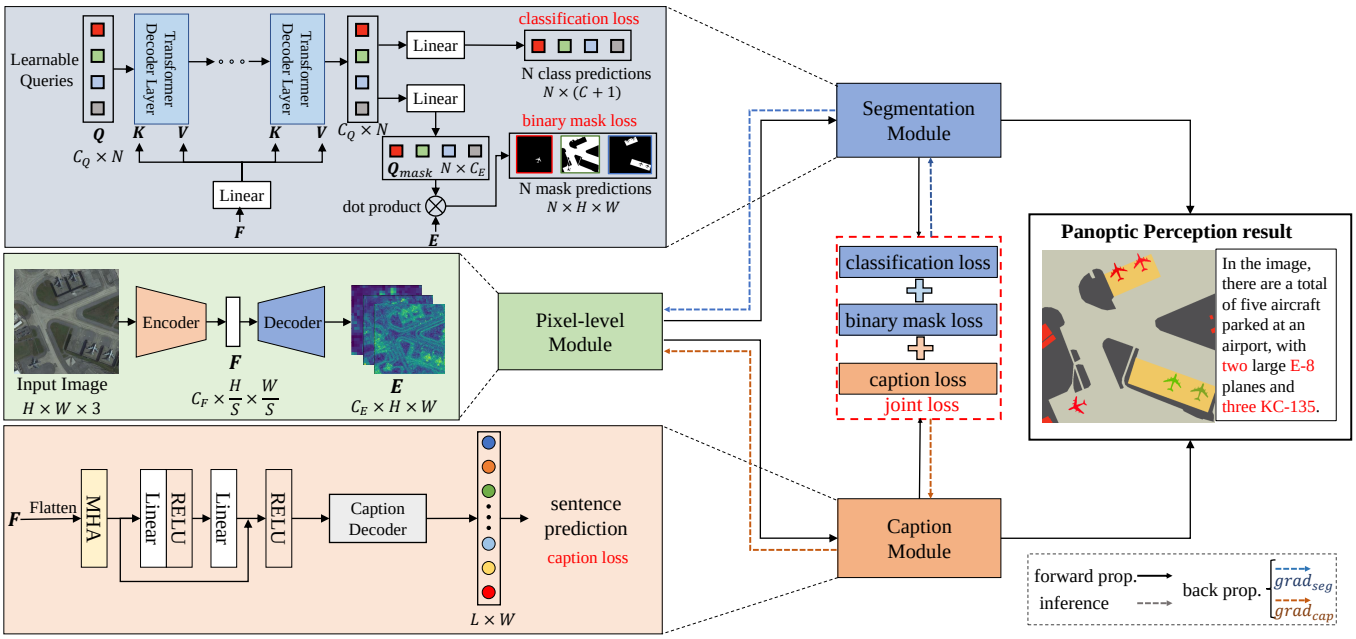

Fig. 6. Overall architecture of our proposed panoptic perception model. $grad_{seg}$ and $grad_{cap}$ indicate the gradient calculated from the segmentation loss and the caption loss.

B. Semi-Automatic Annotation System

Due to its robust generalization capabilities, SAM is widely applied in the segmentation of natural images. When provided with high-quality prompts, annotation systems based on SAM can generate high-quality unlabeled segmentation masks. Moreover, fine-tuning SAM with a small amount of annotated data can yield good performance in various downstream segmentation tasks.

However, SAM does not achieve comparable results for RSIs. As shown in Fig. 5, there are primarily two approaches using SAM for RSI segmentation annotations: directly inputting the image (a) and utilizing manually annotated bounding boxes as prompts (b). However, method (a) struggles with the substantial domain difference between natural images and RSIs. Meanwhile, method (b) does not eliminate the manual effort required for bounding box annotations.

The key to efficiently annotating RSIs with SAM is fully leveraging its zero-shot segmentation capabilities while compensating for its lack of RSI-specific knowledge. To address these challenges and improve annotation efficiency, we design a novel semi-automatic annotation system based on SAM, supplemented by a supervised detection and segmentation model, as shown in Fig. 5(c).

In the beginning, we manually annotate bounding boxes for a small set of images to train a detector. To ensure the quality of annotations for unseen images, bounding boxes generated by the detection model undergo manual checks. Subsequently, the predicted box results serve as prompts and are fed into SAM to segment certain targets in the image. We will train a supervised segmentation model after refining these segmentation results.

Note that the described process is iterative, which means the detection and segmentation results from the current step directly feed into the training data for the next step.

To annotate unseen images, we first employ the detection model to obtain box prompts. Then, both SAM and the trained segmentation model are used to predict segmentation results. We achieve the final segmentation annotations by merging results from SAM and our trained model. In practice, simply obtaining intersection areas can effectively combine the segmentation results.

V. Joint Optimization-Based Panoptic Perception Method

As shown in Fig. 6, we propose a simple baseline method to perform end-to-end fine-grained panoptic perception. It achieves multi-task, multi-stage information interaction by optimizing a joint loss function. The model consists of three parts: the pixel-level module, the panoptic segmentation module, and the image caption module. They are responsible for encoding image features, predicting object masks, and generating captions, respectively. The features derived from the pixel-level module are utilized for both mask prediction and caption generation. Therefore, during optimization, the gradient calculations from the loss functions of both segmentation and captioning tasks will contribute to updating the pixel-level module parameters.

A. Pixel-level Module

The pixel-level module primarily consists of an image encoder and a decoder. Consider an image of size $H \times W$. The image encoder produces a down-sampled feature F of size $C_F \times \frac{H}{S} \times \frac{W}{S}$, where S is the output stride and C_F is the number of channels of the encoded feature. The decoder then upsamples F to obtain per-pixel feature embeddings

$E \in \mathbb{R}^{C_E \times H \times W}$, which are used for subsequent mask predictions. We apply ResNet-50 [41] as the image encoder and Transformer Encoder [42] with convolutional layers as the image decoder. Specifically, we feed the flattened F into the Transformer decoder, reshape the obtained embeddings back to hxw, and then use deconvolution layers for upsampling. By default, we set $S = 256$, $C_F = 256$, and $C_E = 256$.

B. Segmentation Module

We regard both instance and semantic segmentation as mask classification problems and handle them with a Transformer-based architecture. Initially, N learnable queries $Q \in \mathbb{R}^{C_Q \times N}$ are initialized, where C_Q is the dimension of the queries. The features F obtained from the pixel-level module are used as keys (K) and values (V). We use a standard Transformer decoder to iteratively update Q. Similar to DETR, we will save the results of each decoder layer.

A typical Transformer decoder layer computation consists of three parts: self-attention on Q, cross-attention between Q, K, and V, and a feed-forward neural network. Readers are referred to [42] for more details. Note that we did not employ masked attention as there is no temporal relationship among the queries.

Through interactions with other queries and the image encoding features, the query can learn the features of various targets and their positional information within the image. Subsequently, we use these information-rich queries for mask classification and generation.

In the mask classification branch, the encoded query undergoes a linear transformation to yield a classification result of $N \times (C+1)$, where C is the total number of foreground and background categories. An additional category $\emptyset$ represents *no object*.

In the mask generation branch, the query is projected into mask embeddings $\mathbf{Q}_{\text{mask}} \in \mathbb{R}^{N \times C_E}$, which has the same channel dimensions as the per-pixel feature embeddings. Subsequently, the dot product is performed between the i-th mask embedding and the matrix $\mathbf{E}$, followed by applying a sigmoid function to generate the i-th mask prediction result.

Referring to [43], we adopt the Hungarian matching [44] to generate a one-to-one mapping between the mask prediction results and the ground truth. Consider the predictions

$$r^{\text{pred}} = \left\{(p_i, m_i) \mid p_i \in \mathbb{R}^{C+1}, m_i \in [0,1]^{H \times W}\right\}_{i=1}^{N} \tag{1}$$

and the M Ground Truth (GT) masks

$$r^{\text{gt}} = \left\{(c_j^{gt}, m_j^{gt}) \mid c_j^{gt} \in \{1, 2, ..., C\}, m_j^{gt} \in [0,1]^{H \times W}\right\}_{j=1}^{M}, \tag{2}$$

We denote $i = \sigma(j)$ as the mapping from r_j^{gt} to r_i^{pred}, and use

$$-p_i(c_j^{\text{gt}}) + \mathcal{L}_{mask}(m_i, m_j^{gt}) \tag{3}$$

as the edge weight in the bipartite graph for the Hungarian matching, where $\mathcal{L}_{mask}$ is the binary mask loss. Normally, we assume $N \gg M$. As a result, some prediction results will be matched to the *no object* category $\emptyset$. We then pad the M GT masks to match the prediction size N using $\emptyset$.

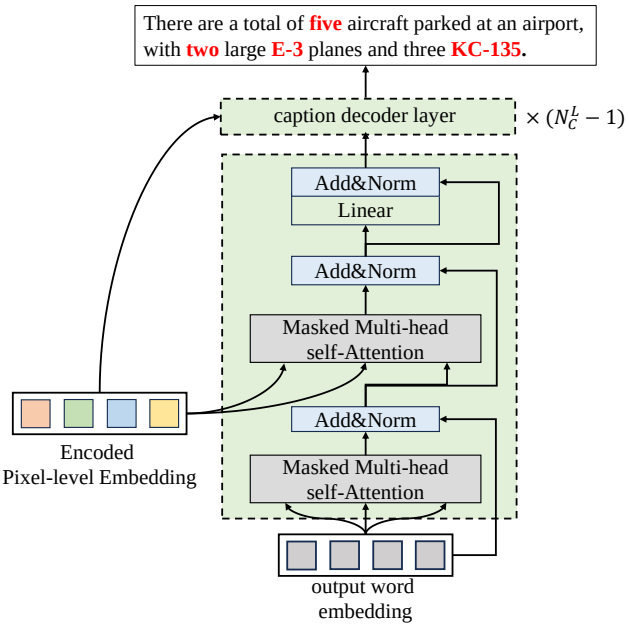

Fig. 7. The model-agnostic task-interactive caption module.

giving a matching σ, the mask prediction loss consists of a cross-entropy loss and a binary mask loss for each $pred - gt$ pair:

$$\mathcal{L}_{seg}(r, r^{gt}) = \sum_{j=1}^{N} [-\log p_{\sigma(j)}(c_j^{gt}) + \mathbb{1}_{c_j^{gt} \neq \emptyset} L_{mask}(m_{\sigma(j)}, m_j^{gt})] \tag{4}$$

Particularly, we apply Dice Loss [45] for the binary mask loss.

C. Caption module

To achieve end-to-end collaborative joint training, we propose a model-agnostic caption module that utilises a transformer decoder, which extracts more abstract semantic information from image features and outputs a sequence of words for description. As depicted in Fig. 6, the encoded image features F obtained from the Pixel-level module:

$$F' = \text{MHA}(F, F, F) \tag{5}$$

$$\hat{F} = \text{Linear}(\text{ReLU}(\text{Linear}(F'))) + F' \tag{6}$$

$$\hat{F} = \text{ReLU}(\hat{F}) \tag{7}$$

As depicted in Fig. 7, we use a model-agnostic Transformer-based decoder to perform caption generation.

Following [42], the Transformer-based-decoder used for sentence prediction consists of multiple decoder layers, each having a masked self-attention mechanism, a cross-attention mechanism, and a feed-forward neural network. We initialize a word representation $X_0 \in \mathbb{R}^{L \times C_F}$ with the START identifier as queries and use T as keys and values for the cross-attention. Similarly, we use a linear classifier for word prediction and compute the loss function using cross-entropy. The number of the decoder layers is denoted as N_C^L.

The output embedding of the decoder is denoted as h_t, and the probability distribution over W words is computed as:

$$p_t = \text{Softmax}(\text{Linear}(h_t)),\ p_t \in [0,1]^W \tag{8}$$

Considering the word predictions $\{p_t\}_{t=1}^{L}$ and the GT captions $\{wd_t|wd_t \in \{1,2,...,W\}\}_{t=1}^{L}$, we compute the caption generation loss using cross-entropy as:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{N_C^L} \log p_t(wd_t) \tag{9}$$

D. Overall Objective

To jointly optimize segmentation and caption branch, the total loss function is the weighted sum of $\mathcal{L}_{seg}$ and $\mathcal{L}_{cap}$:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{cap} \tag{10}$$

where λ is the weight to control the contribution of the caption module.

VI. Experiments

In this section, we validate the effectiveness of our proposed unified baseline in handling fine-grained panoptic segmentation tasks through experiments on the FineGrip dataset.

A. Evaluation Metrics

Referencing [6], we utilize the Panoptic Quality (PQ) to uniformly measure the performance of instance segmentation and semantic segmentation:

$$\text{PQ} = \frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p,g)}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|} \tag{11}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. A pair $(p,g) \in$ TP denotes a match between the predicted mask and the ground truth. Further, the PQ can be decomposed into Segmentation Quality (SQ) and Recognition Quality (RQ) to separately measure the performance of segmentation and recognition:

$$\text{SQ} = \frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p,g)}{|\text{TP}|} \tag{12}$$

$$\text{RQ} = \frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|} \tag{13}$$

For a more comprehensive evaluation, we compute PQ, RQ, and SQ scores separately for the *Things* and *Stuff* classes.

For the image caption task, Bilingual Evaluation Understudy (BLEU) [39] is used to evaluate the quality of generated texts. Specifically, BLEU-4 evaluates the overlap of n-grams between the generated sentences and the reference text for $n = 1, 2, 3, 4$.

The BLEU score is defined as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right) \tag{14}$$

Where:

$$p_n = \frac{\text{Count of matching n-grams}}{\text{Count of total n-grams in generated text}} \tag{15}$$

w_n is the weight for n-grams, typically set to $\frac{1}{4}$ for BLEU-4. The brevity penalty, BP, is defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases} \tag{16}$$

where c is the length of the generated translation and r is the effective reference corpus length.

The BLEU score ranges from 0 to 1, with 1 indicating a perfect match with the reference text. Before computing the BLEU-4 scores, we used beam search to obtain descriptive sentences. The beam sizes employed in the experiments were 3 and 5.

B. Implementation Details

All experiments are conducted on two NVIDIA RTX 4090 GPUs. We employ ResNet-50 [41] pre-trained on ImageNet [46] as the backbone for the image encoder in the pixel-level module. The total number of training epochs are 75 for all experiments. The learning rate was initialized as 0.0001 and decay to 0.00001 through multi-step decay. The input image resolution is 800×800 and the training batchsize is set to 4. The number of learnable queries N in the segmentation module is set to 200. The code implementation is based on MMDetection [47].

C. Quantitative Analysis

We evaluate the proposed panoptic perception model on multi-task and multi-modal interpretation. As seen in Table II, we conduct the proposed model with traditional panoptic segmentation and caption generation tasks. For panoptic segmentation, the PQ, SQ, RQ for all classes, *thing* classes and *stuff* classes are computed respectively. While for captioning task, two kinds of beam sizes are reported.

Segmentation performance. We embed the proposed model-agnostic panoptic perception method into several baselines including MaskFormer [43] and Mask2Former [49], of which the models support joint optimization with end-to-end training. Compared to the single-task segmentation model, our panoptic model achieves 0.7% and 1.3% PQ improvement, respectively. Further we respectively explore the segmentation performance on *thing* and *stuff* classes. On the one hand, ours w/ MaskFormer achieves effective improvement on *stuff* classes. On the other hand, ours w/ Mask2Former achieves solid advancements on *thing* classes. We argue that cross-modal joint training can boost divergent single-task performance.

Captioning performance. Regarding caption generation, we compare two baseline models including SAT [50] and MLAT [51]. We evaluate the performance of a variant of SAT without attention mechanisms and a variant of MLAT without LSTM. For fair comparisons, we conduct panoptic perception using both the LSTM decoder (ours w/ SAT) and the Transformer decoder (ours w/ MaskFormer and ours w/ Mask2Former) for the caption module. The quantitative results indicate that joint optimization of the panoptic perception model significantly enhances the performance of caption

TABLE II
COMPARISONS BETWEEN OUR PROPOSED PANOPTIC PERCEPTION METHOD AND OTHER INTERPRETATION APPROACHES ON FINEGRIP. THE BEST RESULTS ARE MARKED IN BOLD.

Task	Method	All			Things			Stuff			BLEU score	
		PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{st}	SQ^{St}	RQ^{St}	beam size=3	beam size=5
Panoptic Segmentation	Panoptic FPN [48]	54.1	78.6	68.3	55.8	79.0	**70.7**	47.1	77.1	58.8	-	-
	MaskFormer [43]	50.2	79.4	63.1	49.9	78.8	63.3	51.6	81.8	62.5	-	-
	Mask2Former [49]	55.2	**81.0**	67.9	54.8	80.5	67.9	57.0	83.0	68.0	-	-
Caption Generation	SAT [50]	-	-	-	-	-	-	-	-	-	30.4	29.4
	SAT w/o attention	-	-	-	-	-	-	-	-	-	36.2	35.9
	MLAT [51] w/o LSTM	-	-	-	-	-	-	-	-	-	35.9	35.8
	MLAT [51]	-	-	-	-	-	-	-	-	-	41.2	40.3
Panoptic Perception	Ours w/ SAT	49.6	79.5	62.1	48.5	78.7	61.4	53.8	**82.6**	64.7	**45.2**	**44.0**
	Ours w/ MaskFormer	50.9	79.3	63.8	49.8	78.6	63.2	55.1	82.1	66.1	42.4	41.7
	Ours w/ Mask2Former	**56.5**	80.9	**69.6**	**56.3**	**80.6**	69.7	**57.3**	82.2	**68.9**	42.3	41.5

generation. Concretely, our model with LSTM decoder significantly outperforms original captioning models SAT [50] and gains an improvement of 14.8% (beam-size=3) and 14.6% (beam-size=5) in BLEU score. While the Transformer-decoder-based panoptic perception also achieves favourable improvements on both segmentation and captioning tasks.

Cross-modal sub-tasks interaction. We demonstrate that joint optimization across multi-modal tasks within an intact model can advance each single-task performance. As shown in Table II, the proposed panoptic perception methods support model-agnostic combinations with current segmentation models and achieve pixel-wise classification, instance-wise segmentation and image captioning concurrently. And compared to the single segmentation or captioning task, the panoptic perception achieve higher performance on both tasks. To our knowledge, cross-modal interactive optimization can advance the overall performance. The experimental results further validate this perspective. Particularly, ours w/ Mask2Former achieves 1.3% and 1.7% improvement in PQ and RQ over all classes to the original Mask2Former. While in the fine-grained *things* classes, the metrics are also significantly improved by 1.5% and 1.8&, respectively.

Fig. 8 illustrates the qualitative comparisons of the panoptic perception results with separate training approach. In the first row, our panoptic perception model can accurately perceive the categories and quantities of the foreground airplane targets in both segmentation and description branches. While in the second row, the foreground and background semantics are both recognized with fine-grained annotations. In contrast, the predictions of the separately trained model are coarser in both segmentation and captioning tasks. Particularly in the captioning task, the model trained independently struggles to generate accurate fine-grained captions due to the lack of additional information from the segmentation task.

To intuitively demonstrate the impact of joint optimization on model behavior, we visualize the feature saliency maps for SAT and our panoptic perception model w/ SAT, w/ MaskFormer and w/ Mask2Former, respectively. We select the output feature map of the last layer of the ResNet-50 backbone with a shape of $2048 \times 25 \times 25$. After being averaged over channels and normalized, these feature maps are overlaid onto the original input images. As illustrated in Fig. 9, the SAT model pays more attention to the overall image while overlooking the foreground semantics. This is because the caption generation task requires the model to generalize over the entire image, which prevents the model from focusing on specific targets. On the contrary, ours w/ SAT focus on both foreground and background semantics. The Transformer decoder based panoptic perception methods also strike a balance by emphasizing foreground instances and maintaining a focus on the overall image. This validates the effectiveness of a unified multi-task framework and joint optimization strategy in enhancing a more comprehensive interpretation of RSIs.

D. Ablation Study

Collaborative Optimization. The proposed joint optimization-based panoptic perception model leverages both the segmentation branch and captioning branch with task-interactive optimization. Thus the weight contribution within multi-modal tasks has a vital impact on the joint optimization and multi-task learning problem. In Eqn. 10, λ was used to balance the contribution of multi-modal branches during training. As shown in Table III, we respectively compute the segmentation performance and the captioning score. The weight to balance the contribution of sub-tasks is not a fixed value. Concretely, regarding MaskFormer as the baseline, it achieves the highest 50.9% PQ when $\lambda = 2.0$, while the best BLEU score arrives with $\lambda = 0.7$. We infer that a lower λ prevents the caption module from providing sufficient instance-level information, while a higher λ can overwhelm the impact of the few instance-related words with the major non-instance parts of the sentence. While ours w/ Mask2Former achieves the highest PQ at $\lambda = 1.0$. To our knowledge, the gradient contributions from different sub-tasks should maintain balanced or tilted towards vital tasks, which have been investigated in multi-task learning problems.

Captioning Decoder. In our proposed panoptic perception model, we develop a universal Transformer decoder for captioning task, which is parallel and interactive with the Transformer decoder in the segmentation branch. Table IV displays the panoptic perception results of our model with

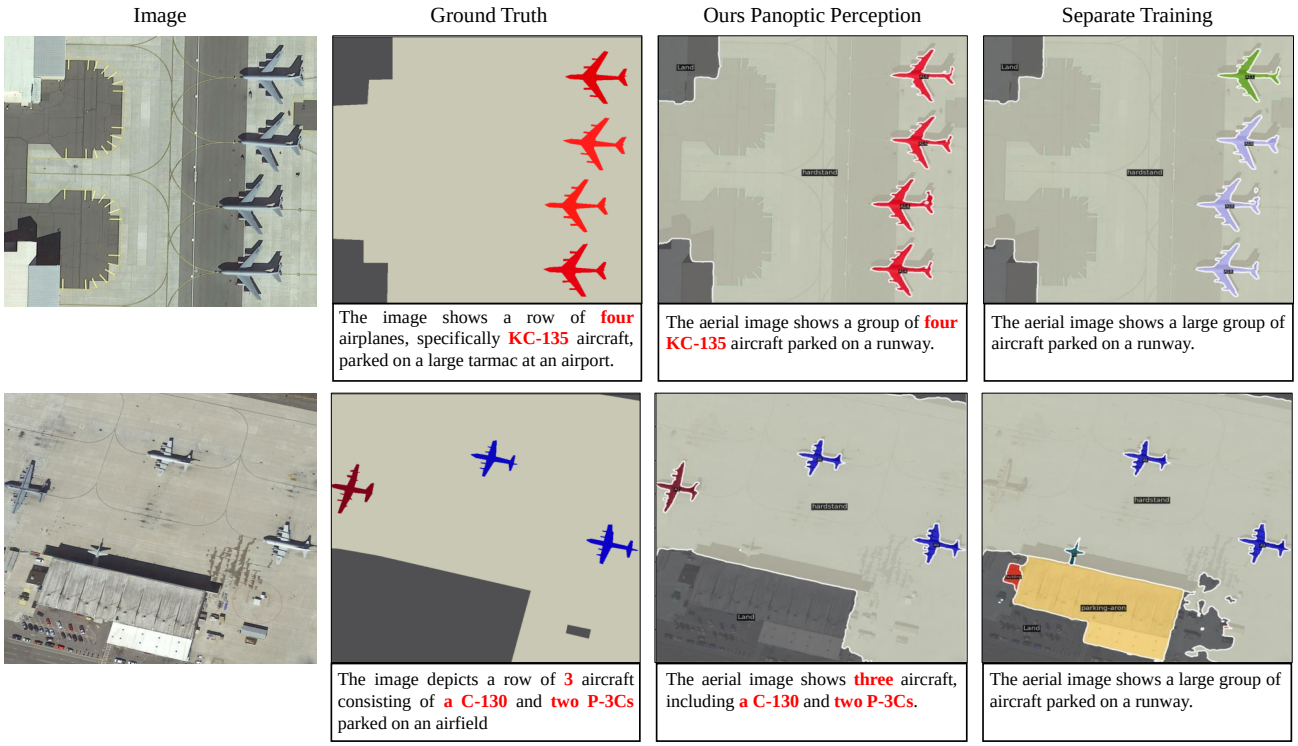

Fig. 8. Panoptic perception results from the jointly trained and separately trained models. The first column displays the ground-truth mask and one of the five caption sentences for each image.

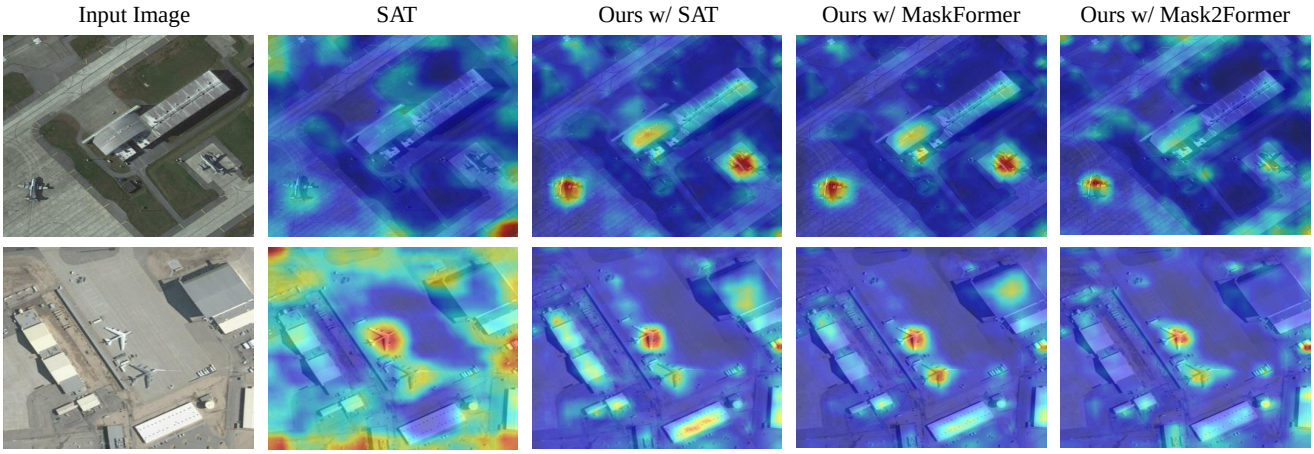

Fig. 9. Saliency maps of features from the last layer of the ResNet-50 backbone for different models. The regions with colors closer to red indicate areas of higher feature response intensity.

a Transformer decoder at different layer counts. Regarding Mask2Former as the base model, We observe that segmentation performance improves with increased model complexity, reaching 56.5% PQ with three decoder layers. However, The caption module performs best when applying two decoder layers. It achieves a 43.4% BLEU score at a beam size of 3, surpassing the model based on the SAT decoder.

There are two reasons for the significant difference in the impact of decoder layer counts on segmentation and captioning performance. On the one hand, substantial model complexity presents challenges in achieving efficient optimization, thereby hindering the caption performance. On the other hand, a more complex structure possesses stronger feature representation capabilities. It can encourage the backbone to learn feature representations containing richer information, ultimately leading to improved performance in the segmentation branch.

Model Attributes. We evaluate the proposed panoptic perception model attributes from parameters, FLOPs and inference FPS. The results are reported in Table V. In addition, we also analyze the category-wise improvements of the proposed panoptic perception model to the base segmentation model. Table VI shows the gain on PQ of the proposed model, especially on the challenging fine-grained foreground semantics, proving the effectiveness of boosting single-task interpretation.

TABLE III
THE IMPACT OF λ.

Model	λ	All			BLEU score	
		PQ	SQ	RQ	bs=3	bs=5
Ours w/ MaskFormer	0.5	49.4	**79.5**	61.9	43.2	42.1
	0.7	50.1	79.4	62.9	**43.8**	**43.2**
	1.0	50.2	79.2	63.1	43.1	42.6
	2.0	**50.9**	79.3	**63.8**	42.4	41.7
	3.0	50.7	79.4	63.5	42.6	42.2
Ours w/ Mask2Former	0.5	54.0	81.0	66.3	41.8	40.8
	0.7	54.4	**81.1**	66.6	41.3	40.7
	1.0	**56.5**	80.9	**69.6**	42.3	**41.5**
	2.0	54.1	81.0	66.6	40.4	40.8
	3.0	54.0	81.0	66.4	**42.8**	41.4

TABLE IV
THE IMPACT OF THE DECODER LAYER NUMBER N_{cap}^{dec} IN THE CAPTION MODULE.

Model	N_{cap}^{dec}	All			BLEU score	
		PQ	SQ	RQ	bs=3	bs=5
Ours w/ MaskFormer	6	50.5	79.1	63.6	41.6	41.5
	3	**50.9**	79.3	**63.8**	42.4	41.7
	2	50.6	**79.4**	63.5	**42.7**	**42.6**
	1	49.3	79.3	62.0	42.1	36.5
Oursw w/ Mask2Former	6	53.8	**81.2**	66.1	41.3	40.8
	3	**56.5**	80.9	**69.6**	42.3	41.5
	2	53.4	80.9	65.7	**43.4**	**42.7**
	1	54.6	81.1	67.2	42.8	42.3

VII. CONCLUSION

In this paper, we develop a novel fine-grained panoptic perception task to achieve a more comprehensive unified interpretation of RSIs. The main contributions include: 1) We introduce FineGrip, a novel fine-grained panoptic perception dataset. 2) We propose an end-to-end model-agnostic panoptic perception method that integrates multi-task joint optimization covering pixel-level classification, instance-level segmentation and image-level captioning within an intact model. 3) An efficient semi-automatic annotation system that provides scalability for segmentation annotations. Experimental results demonstrate the effectiveness of the proposed unified architecture for panoptic perception. Particularly, the feasibility of joint optimization across multiple tasks by interactive learning was validated. However, the consistency across multiple subtasks is still a challenging but promising research priority. Our future work will further expand the proposed dataset and improve the task-interactive facilitation of the panoptic perception model.

TABLE V
THE PARAMETERS, FLOPS AND INFERENCE FPS OF THE PROPOSED PANOPTIC PERCEPTION MODELS.

Model	Param. (M)	FLOPs (G)	FPS
MaskFormer	45.0	126	8.5
Ours w/ MaskFormer	52.1	135	4.8
Mask2Former	44.0	160	6.8
Ours w/ Mask2Former	46.5	172	6.2

REFERENCES

[1] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International journal of Remote sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[2] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS journal of photogrammetry and remote sensing*, vol. 159, pp. 296–307, 2020.

[3] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.

[4] H. Su, S. Wei, M. Yan, C. Wang, J. Shi, and X. Zhang, "Object detection and instance segmentation in remote sensing imagery based on precise mask r-cnn," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 1454–1457.

[5] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[6] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[7] O. L. F. de Carvalho, O. A. de Carvalho Júnior, C. R. e. Silva, A. O. de Albuquerque, N. C. Santana, D. L. Borges, R. A. T. Gomes, and R. F. Guimarães, "Panoptic segmentation meets remote sensing," *Remote Sensing*, vol. 14, no. 4, p. 965, 2022.

[8] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022.

[9] E. Fernandes, P. Wildemberg, and J. dos Santos, "Water tanks and swimming pools detection in satellite images: Exploiting shallow and deep-based strategies," in *Anais do XVI Workshop de Visão Computacional*. SBC, 2020, pp. 117–122.

[10] S. Mohajerani and P. Saeedi, "Cloud-net: An end-to-end cloud detection algorithm for landsat 8 imagery," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, July 2019, pp. 1029–1032.

[11] X. Roynard, J.-E. Deschaud, and F. Goulette, "Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification," *The International Journal of Robotics Research*, vol. 37, no. 6, pp. 545–557, 2018.

[12] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[13] H. Alemohammad and K. Booth, "Landcovernet: A global benchmark land cover classification training dataset," *arXiv preprint arXiv:2012.03111*, 2020.

[14] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," *arXiv preprint arXiv:2110.08733*, 2021.

[15] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[16] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2043–2056, 2018.

[17] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv preprint arXiv:1807.01232*, 2018.

[18] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89 644–89 654, 2021.

[19] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[20] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018.

TABLE VI
QUANTITATIVE COMPARISON IN CATEGORY-WISE PQ (%).

Method	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	Land	Run.	Hards.	Park.	Buil.
Mask2Former	45.8	34.3	62.0	65.2	46.1	75.5	71.5	55.9	53.4	75.0	60.8	74.9	31.8	39.6	43.1	27.7	73.6	45.5	54.3	59.1	82.0	36.4	81.7	34.0	50.8
Ours w/ Mask2Former	46.0	38.3	63.7	67.9	49.5	75.7	69.4	55.9	54.6	75.1	59.9	75.1	39.3	39.0	49.9	30.9	73.3	45.3	52.6	64.2	81.7	35.9	82.5	35.4	51.2
gain	+0.2	+4.0	+1.7	+2.7	+3.4	+0.2	-2.1	0.0	+1.2	+0.1	-0.9	+0.2	+7.5	-0.6	+6.8	+3.2	-0.3	-0.2	-1.7	+6.1	-0.3	-0.5	+0.8	+1.4	+0.4

[21] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2017.

[22] X. Sun, Y. Lv, Z. Wang, and K. Fu, "Scan: Scattering characteristics analysis network for few-shot aircraft classification in high-resolution sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[23] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *International conference on pattern recognition applications and methods*, vol. 2. SciTePress, 2017, pp. 324–331.

[24] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 28–37.

[25] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[26] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE transactions on geoscience and remote sensing*, vol. 56, no. 12, pp. 7147–7161, 2018.

[27] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[28] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[30] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4872–4881.

[31] Z. Li, G. He, H. Fu, Q. Chen, B. Shangguan, P. Feng, and S. Jin, "Rs dino: A novel panoptic segmentation algorithm for high resolution remote sensing images," in *2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 2023, pp. 1–5.

[32] T. Fernando, C. Fookes, H. Gammulle, S. Denman, and S. Sridharan, "Toward on-board panoptic segmentation of multispectral satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[33] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, pp. 157–173, 2008.

[34] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[35] A.-z. yatengLG and horffmanwang, "ISAT with segment anything: Image segmentation annotation tool with segment anything," 2023. [Online]. Available: https://github.com/yatengLG/ISATwithsegmentanything

[36] J. Chen, Z. Yang, and L. Zhang, "Semantic segment anything," https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023.

[37] 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.

[38] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, "Samrs: Scaling-up remote sensing segmentation dataset with segment anything model," *arXiv preprint arXiv:2305.02034*, 2023.

[39] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[40] Y. Wenqi, C. Gong, W. Meijun, Y. Yanqing, X. Xingxing, Y. Xiwen, and H. Junwei, "Mar20: A benchmark for military aircraft recognition in remote sensing images," *National Remote Sensing Bulletin*, vol. 27, no. 12, pp. 2688–2696, 2023.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[43] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 864–17 875, 2021.

[44] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[45] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[47] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[48] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[49] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1290–1299.

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[51] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.