Autonomous Obstacle Avoidance Decision Method and System Based on Multi-Source Sensor Fusion

Technical Field

The present invention relates to the technical field of intelligent navigation for unmanned aerial vehicles (UAVs), in particular to an autonomous obstacle avoidance decision method and system based on multi-source sensor fusion.

Background Art

Traditional unmanned aerial vehicle (UAV) navigation relies primarily on geometric maps for path planning and obstacle avoidance, with its core lying in the perception of the physical boundaries of obstacles, yet it lacks in-depth semantic understanding of the environment.

This results in the inability of UAVs to make intelligent decisions when confronted with targets with specific semantics such as "dilapidated buildings", "ground fissures" or "survivors".

Existing semantic mapping technologies suffer from three major bottlenecks:

Weak adaptability to complex outdoor aerial scenarios (e.g., building occlusion, smoke interference, low illumination), where perception accuracy is susceptible to environmental factors;

A closed cognitive system that relies heavily on predefined label libraries, failing to identify and understand unknown objects or emergent scenarios;

Low data processing efficiency, which makes it difficult to meet the requirements of real-time mapping and dynamic obstacle avoidance.

The aforementioned drawbacks have severely restricted the intelligent application upgrading of UAVs in key scenarios of the low-altitude economy.

Therefore, there is an urgent need for a novel autonomous obstacle avoidance decision-making system that is capable of fusing multi-source heterogeneous perception data, equipped with open-vocabulary comprehension capability, and able to efficiently generate safe navigation strategies.

Summary of the Invention

The technical problem to be solved by the present invention is to overcome the above technical defects and provide an autonomous obstacle avoidance decision-making method and system based on multi-source sensor fusion, which is capable of accurately understanding complex environments and performing intelligent and safe path planning according to natural language instructions.

To solve the aforementioned technical problem, the technical solution provided by the present invention is as follows: an autonomous obstacle avoidance decision-making system based on multi-source sensor fusion, comprising:

a multi-source perception module configured to synchronously collect RGB image data, infrared image data, and laser point cloud data in the flight environment of an unmanned aerial vehicle (UAV);

a feature extraction and fusion module configured to perform spatiotemporal alignment on the collected data, separately extract RGB image semantic features, infrared image penetration features, and laser point cloud geometric features, and generate a fused feature map through cross-modal fusion;

an open-vocabulary semantic understanding module integrated with a vision-language large model, which is configured to receive the fused feature map, perform zero-shot recognition and semantic relationship reasoning on objects in the fused feature map based on natural language instructions, and generate an open-vocabulary semantic map containing object categories, attributes, and spatial relationships;

an autonomous obstacle avoidance decision-making module configured to construct a risk cost map according to the open-vocabulary semantic map and the received natural language task instructions, generate a semantically annotated safe navigation path via a multi-constraint path planning algorithm, and control the unmanned aerial vehicle (UAV) to perform autonomous obstacle avoidance flight;

Preferably, the processing of laser point cloud data in the said feature extraction and fusion module includes a dynamic adaptive sampling strategy based on local point cloud density:

reducing redundant sampling points in dense point cloud regions;

retaining key detailed sampling points in sparse regions;

introducing a self-attention mechanism during the local feature extraction phase;

dynamically focusing on and aggregating key information within local point cloud regions.

Preferably, the semantic processing of RGB images in the said feature extraction and fusion module includes the adoption of a cascaded structure of the YOLO object detection model and the SAM segmentation model.

Preferably, the spatiotemporal alignment in the said feature extraction and fusion module includes performing three-dimensional-to-two-dimensional reprojection of RGB images and infrared images by using laser point clouds, converting point cloud coordinates to the camera coordinate system via sensor calibration parameters, and projecting the coordinates onto the image plane.

which includes achieving precise temporal and spatial synchronization between image frames and point cloud data.

Preferably, the vision-language large model refines and distills the preliminary semantic descriptions, extracts key concepts related to navigation tasks, and generates a structured representation of key navigation features.

Preferably, the said open-vocabulary semantic map is a hierarchical semantic topological network, which sequentially comprises the following layers from the bottom to the top:

a point cloud layer configured to provide the original geometric structure of the scene;

a lane layer configured to characterize the network topology of passable areas;

an instance layer configured to identify and aggregate individual objects of the same semantic category into semantic nodes;

a road layer configured to characterize semantic regions of a larger scale and their attributes;

an environment layer configured to integrate information from the lower layers, construct global spatial relationships and topological connections, and provide a basis for high-level task decision-making.

In another aspect, the present invention discloses a decision-making method, characterized by comprising the following steps:

S1: Synchronously collecting RGB image data, infrared image data, and laser point cloud data via the multi-source perception module;

S2: Performing spatiotemporal alignment on the collected multi-source data, and separately extracting RGB semantic features, infrared penetration features, and point cloud geometric features;

S3: Performing cross-modal fusion on the three extracted features to generate a fused feature map;

S4: Inputting the fused feature map into the open-vocabulary semantic understanding module integrated with a vision-language large model, performing zero-shot object recognition and semantic relationship reasoning in combination with natural language task instructions, and constructing an open-vocabulary semantic map;

S5: Constructing a risk cost map based on the open-vocabulary semantic map, and generating a semantically annotated safe navigation path via a multi-constraint path planning algorithm, so as to realize autonomous obstacle avoidance decision-making for the unmanned aerial vehicle (UAV).

Preferably, said step S4 comprises the following sub-steps:

S4.1: Generating a detailed description containing a caption and spatial position for each object in the map;

S4.2: Calculating the 3D bounding box intersection over union (IoU) between each pair of object nodes based on the 3D point cloud reconstruction result to construct a similarity matrix;

S4.3: Pruning the similarity matrix by using the minimum spanning tree algorithm to generate fine-grained potential spatial relationship edges between objects, and finally forming a map rich in semantic information.

Preferably, the task-aware cost function of the multi-constraint path planning algorithm in said step S5 takes into account the following factors: path length, obstacle risk coefficient, and task target reachability.

The advantages of the present invention compared with the prior art are as follows: by fusing RGB image data, infrared image data and laser point cloud data, the perception limitations of a single sensor in complex environments such as smoke, haze and low

illumination are effectively overcome, and the integrity and robustness of environmental modeling are significantly improved.

By leveraging the vision-language large model, the limitations of traditional closed label libraries are broken through, zero-shot recognition and semantic relationship reasoning of unknown objects and complex scenarios are realized, and the unmanned aerial vehicle (UAV) is endowed with genuine environmental perception capability;

By improving the network and optimizing the fusion method, the data processing efficiency is greatly improved, which meets the stringent requirements of real-time semantic mapping and dynamic obstacle avoidance.

Brief Description of Accompany Drawings

Figure 1 is a schematic diagram of RGB image processing.

Figure 2 is a schematic diagram of point cloud data processing.

Figure 3 is a schematic diagram of map construction via multi-modal fusion.

Figure 4 is a comparison diagram of the processing effects with other image segmentation models for the same image.

Specific Embodiment of the Invention

The present invention is described in further detail below with reference to Figures 1–4.

The present invention is mainly directed to the spatiotemporal alignment and fusion of RGB image data, infrared image data, and laser point cloud data. By implementing 3D–2D reprojection, the temporal synchronization and spatial alignment between image frames and lidar point clouds are ensured, thereby achieving accurate mapping and integration of features.

By fusing RGB image data, infrared image data and laser point cloud data, the precise correspondence of multi-modal perception information and the collaborative extraction of features are realized, which lays a foundation for subsequent tasks such as semantic segmentation and target recognition.

The extrinsic parameters describe the transformation relationship from the lidar coordinate system to the camera coordinate system, which are usually represented by a rotation matrix R and a translation vector T. In the NuScenes dataset, the parameters can be obtained through a calibration process. The intrinsic parameters describe the optical

characteristics of the camera, including the focal length f_x, f_y and principal point c_x, c_y, and are usually represented by a 3×3 matrix:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix};$$

Point Cloud Coordinate Transformation:From the lidar coordinate system to the camera coordinate system: For a point $P_{lidar} = [X_{lidar}, Y_{lidar}, Z_{lidar}]^T$ in the lidar, it is transformed into the camera coordinate system. The transformation formula is as follows: $P_{cam} = R \cdot P_{lidar} + T$;

where $P_{cam} = [X_{cam}, Y_{cam}, Z_{cam}]^T$ is the coordinate of the transformed point in the camera coordinate system.

Project the point P_{cam} in the camera coordinate system onto the image plane. The projection is performed using the intrinsic parameter matrix K of the camera:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \cdot \begin{bmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \end{bmatrix};$$

where u and v are the pixel coordinates of the projected point on the image, and the specific calculation formulas are as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \cdot \begin{bmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \end{bmatrix};$$

Round the calculated values of u and v, and check whether the projected point falls within the valid range of the image (i.e., between $0 \le u < W$ and $0 \le v < H$, where W and H are the width and height of the image, respectively).

Thus, the projection of the point cloud onto the RGB image is achieved, facilitating further segmentation and fusion.

RGB Image and Infrared Image Processing

By analogy with indoor and outdoor ground segmentation, this invention realizes target instance segmentation by combining object detection (using the YOLO model) and segmentation (using the SAM model). The specific process is as follows:

When the light is normal and there is no obvious line-of-sight obstruction, the input is an RGB image, whose dimensions are generally H×W×3,where H and Wrepresent the height and width of the image respectively, and 3 represents the three color channels of RGB. Considering that most disaster rescue scenarios are accompanied by line-of-sight obstruction caused by heavy smoke and fog as well as search and rescue operations under nighttime conditions, the input will be converted to an infrared image under special circumstances.

Feature ExtractionThe encoder part of the Segment Anything Model (SAM) is based on the Vision Transformer (ViT).

1. PatchEmbedding The image is divided into patches of size P×P (16×16) by the PatchEmbedding module, and each patch is embedded into a feature vector of fixed dimension to form a serialized feature map. This process can be regarded as converting a two-dimensional image into a one-dimensional sequence for processing by the Transformer.

Transformer Encoder

The embedded feature sequence is input into the Transformer encoder. The encoder is composed of multiple Transformer layers, and each layer includes a multi-head self-attention mechanism and a feed-forward neural network. Through the multi-head self-attention mechanism, the model can capture the long-range dependencies between different positions in the image, thereby better understanding the global semantic information of the image. The feed-forward neural network performs non-linear transformation on the features to further extract the semantic information of the features.

Feature Map Output

After processing by the Transformer encoder, a feature map F_{img} is output, whose dimension is (C, H', W'), where C denotes the number of feature channels, and , H' and W' denote the height and width of the feature map respectively.

YOLO Bounding Box and Category Recognition

The YOLO model is used to perform object detection on the input RGB image, and the output detection results include the bounding boxes and category confidence levels of the objects. Feature maps are extracted from the feature layers of YOLO for subsequent fusion, whose dimension is $\left(C_{yolo}, H_{yolo}, W_{yolo}\right)$.

Overall Implementation Effect

The overall implementation effect is shown in Figure 1.

Point Cloud Data Processing

Improvements are made to the widely adopted PointNet++ model, based on two core aspects as follows:

Improved Sampling StrategyAiming at the limitation of the original fixed farthest point sampling in scenarios with uneven point cloud density, a dynamic adaptive sampling strategy is proposed, which adjusts the number and positions of sampling points according to the local density.

Enhanced Local Feature ExtractionTo address the problem of detail loss caused by the original model's reliance on max-pooling, a self-attention mechanism is introduced in the local feature extraction stage, endowing the model with the ability to dynamically focus on key points within local point cloud regions.

Input Stage

The input is point cloud data, whose dimension is generally $\left(N, D\right)$, where N denotes the number of points in the point cloud, and D denotes the feature dimension of each point (typically, $D = 4$, including three coordinate values of x, y, z and one intensity value I).

Improved Sampling Strategy

To address the problem of uneven point cloud density in different scenarios, the improvements made to the sampling strategy of PointNet++ in this project are as follows:

Calculation of Local Density

First, calculate the local density of each point. This is achieved by calculating the distances between each point and its neighboring points. For example, for each point p_i search for k nearest neighboring points within its neighborhood and calculate the average

distance d_i between these points and p_i. The local density of point ρ_i is defined as $\rho_i = \frac{1}{d_i}$.

The number and positions of sampling points are dynamically adjusted according to the local density. In regions with high density, the number of sampling points is reduced to avoid redundant information; in regions with low density, the number of sampling points is increased to retain more detailed information. Specifically, a density threshold ρ_{th} is first set. For regions where the local density is higher than ρ_{th}, sampling is performed at a lower sampling rate; for regions where the local density is lower than ρ_{th},sampling is performed at a higher sampling rate.

Multi-Scale Sampling

To capture structural information at different scales, the project introduces a multi-scale sampling strategy. At each level, different numbers of sampling points are selected according to the local density and global distribution of the point cloud. For example, N_1 sampling points are selected at the first level, N_2 sampling points at the second level, and so on. The number of sampling points at each level needs to be adaptively adjusted based on the overall density of the point cloud and the target application to achieve a better feature extraction effect.

Improved Local Feature Extraction

To better capture the detailed information in local point cloud patches, the project introduces a self-attention mechanism in the local feature extraction stage.

For each sampling point, search for neighboring points p_i within the spatial range around it to form a local point cloud block P_i. The search for neighboring points can be implemented by means of a spherical query method, i.e., searching for all points within a spherical region that takes the sampling point as the center and has a radius of r.

The feature of each point in the local point cloud block P_i is embedded into a high-dimensional space, which is implemented by a shared Multi-Layer Perceptron (MLP);

for example:

$f_{embed}(x) = MLP(x)$;

Self-Attention CalculationFor the embedded feature sequence $\left\{f_{embed}\left(p_{i,j}\right)\right\}_{j-1}^{k}$, the self-attention weights are calculated. Specifically, the correlation between each point $p_{i,j}$ and the other points is computed to derive the attention weights $\alpha_{i,j}$;

$$\alpha_{i,j} = \frac{\exp\left(Attention\left(f_{emped}\left(p_{i,j}\right), f_{embled}\left(P_i\right)\right)\right)}{\sum_{k-1}^{K}\exp\left(Attention\left(f_{mbbed}\left(p_{i,k}, f_{embed}\left(P_i\right)\right)\right)\right)};$$

Herein, Attention is a learnable attention function, such as algorithms including dot-product attention or additive attention;

Weighted Feature Aggregation:

Based on the attention weights $\alpha_{i,j}$, weighted aggregation is performed on the point features in the local point cloud block to obtain the feature representation f_i of the local region: $f_i = \sum_{j-1}^{k}\alpha_{i,j} f_{embed}\left(p_{i,j}\right)$

After processing with the improved PointNet++, a point cloud feature $\left(N, C_{point}\right)$ is output, where the dimension N represents the number of points in the point cloud, and the dimension C_{point} represents the number of channels of the point cloud feature. The processing of the point cloud data is illustrated in Figure 2.

In the present invention, the strictly aligned image feature F_{img} and point cloud feature F'_{point} are deeply fused, and the features of the two modalities are concatenated along the channel dimension.pecifically, for each pixel point with spatially aligned positions, the corresponding image feature vector $F_{img}(u,v)$ and point cloud feature vector $F'_{point}(u,v)$ are concatenated to form a fused feature vector. The dimension of the fused feature map is $\left(C_{img} + C_{point}, H, W\right)$, where C_{img} and C_{point} represent the number of channels of the image feature and the point cloud feature, respectively, and H and W represent the height and width

of the feature map, respectively.

The concatenated features integrate the dual-modality information F_{fused}, but further integration is required to extract high-level discriminative features. The present invention designs a feature fusion module, which adopts a Multi-Layer Perceptron (MLP) to implement the fusion instead of a Transformer encoder with higher computational complexity. The specific layer structure is as follows:

Input Layer: The concatenated feature F_{fused} is flattened into a two-dimensional tensor with a dimension of $\left(N, C_{img} + C_{point}\right)$, where $N = H \times W$ denotes the total number of pixels.

Hidden Layers: The input features are subjected to nonlinear transformation via one or more fully connected layers. In the present invention, two hidden layers are configured, where the number of output channels of each layer is $2\left(C_{img} + C_{point}\right)$ and C_{fused}, respectively.Activation Function: A ReLU activation function is applied after each hidden layer to enhance the nonlinear capability of the model.

Output Layer: The dimension of the finally output feature is $\left(N, C_{fused}\right)$, which is then reshaped into a three-dimensional tensor $\left(C_{fused}, H, W\right)$.

By means of the module, the present invention effectively integrates the RGB image features and point cloud features to establish a semantic segmentation map, as illustrated in Figure 3.

Analysis demonstrates that the original RGB images, infrared images and point cloud data all suffer from information deficiency. By fusing the RGB features and point cloud features, the model proposed herein successfully generates an environmental map with a more complete geometric structure. However, the map at this stage does not yet contain semantic information. To endow the map with high-level semantic understanding capability, a Large Language Model (LLM) will be introduced in the subsequent processing stage for semantic reasoning and annotation.

The present invention adopts the NuScenes dataset and DroneVehicle dataset for evaluation, which mainly involves the point cloud level and the instance level.

The model performs 3D point cloud semantic segmentation using 8 predefined classes in the NuScenes dataset, i.e., selecting the class with the highest similarity to the object features as the label of the object. In view of the lack of direct comparison baselines for zero-shot outdoor semantic segmentation, the present invention employs two types of fully supervised comparison methods: (1) the point cloud-based 3D segmentation network (PointNet++); (2) the 2D→3D segmentation networks that are trained on outdoor images and then projected onto point clouds (DeepLabv3, DecoupleSegNets).

The generation of the 3D semantic map is realized by projecting the 2D segmentation results onto the point cloud space. In terms of method comparison, the present invention evaluates alternative schemes that directly utilize the capability of Vision-Language Model (VLM), with a focus on comparing two methods based on CLIP and BLIP respectively. Figure 4 shows a comparative display of the segmentation effects of the aforementioned segmentation methods (including the CLIP-based method and the BLIP-based method) and the method proposed in the present invention.

Qualitative analysis shows that, compared with the baseline models, the segmentation results generated by the scheme proposed in the present invention have significant advantages in terms of visual fidelity. Especially in the near-field region, the segmentation masks exhibit lower noise levels and higher boundary precision.

In terms of quantitative evaluation, the present invention systematically compares the segmentation accuracy of each model for all 8 target classes on the target datasets, as shown in the table below:

Table 1. Comparison of Segmentation Accuracy

Method	Car	Road	Fence	Pole	Building	walkers	Sidewalk	Veg	Ave	Fl
PointNet++	0.38	0.78	0.57	0.77	0.57	0.58	0.75	0.12	0.49	0.62
DeeplabV3	0.42	0.69	0.72	0.52	0.72	0.64	0.78	0.23	0.57	0.58

Blip	0.46	0.76	0.62	0.85	0.62	0.59	0.84	0.18	0.6	0.53
CLIP	0.53	0.48	0.82	0.62	0.82	0.72	0.76	0.14	0.6	0.71
Blip+CLIP	0.39	0.65	0.76	0.76	0.76	0.76	0.82	0.32	0.64	0.65
Ours	0.35	0.82	0.78	0.78	0.83	0.78	0.86	0.21	0.66	0.73

The results of quantitative analysis show that although the method proposed in the present invention has limitations in the segmentation performance of plant categories, it exhibits significant advantages in the remaining seven target categories and outperforms the baseline models in multiple key evaluation metrics. Such advantages are particularly prominent in key categories including roads, buildings and sidewalks. In addition, the method achieves a higher mean Intersection over Union (IoU) index, which indicates that it can segment different target objects and regions in images more accurately on the whole. Compared with the separate application of the PointNet++ or CLIP model for semantic segmentation, the multi-modality fusion method proposed herein realizes a significant improvement in segmentation accuracy, which effectively verifies the effectiveness of the multi-modality feature fusion strategy in improving the performance of semantic segmentation tasks.

The present invention introduces a Large Language Model (LLM) to construct a navigation semantic map, so as to realize the automatic conversion from natural language task instructions to three-dimensional spatial semantic expressions, thereby supporting the understanding and execution of navigation tasks.

In one embodiment:Taking earthquake rescue as an example, the system receives a natural language task instruction: "Please locate the trapped personnel in the ruins and plan a safe rescue path." The system then invokes the Large Language Model (LLM) developed by Meta, namely LLaMA, to understand the task semantics and decompose the task into two subtasks: "trapped personnel positioning" and "path planning".

Based on the Information Bottleneck principle, the intelligent compression of the three-dimensional scene is realized by balancing the maximization of task-related information and the minimization of scene redundancy.

$$\min_{p(\tilde{x}|x)} \underbrace{I(X;\tilde{X})}_{\text{Compress Redundant Information}} - \beta \underbrace{I(\tilde{X};Y)}_{\text{Preserve Task-Associated Information}};$$

Herein, X denotes the task-irrelevant primitive, $\tilde{X}$denotes the compressed task-related concept, Y denotes the task target parsed from the natural language instruction, and β denotes the task correlation intensity coefficient.

Combined with the map and perception data, the model generates a spatial relationship reasoning engine. By extracting key object nodes (building debris, roads, survivors) and calculating spatial topological relationships (distance, orientation, inclusion relationship), the model generates an initial semantic description, e.g., "Trapped personnel may be located under the ruins of collapsed buildings. The area is situated at the intersection of the main road and the commercial street, with a large number of damaged buildings and scattered debris around."

On this basis, the system further outputs spatial relationship labels to support subsequent scene reconstruction and path planning.

After completing the preliminary semantic understanding and spatial position annotation, the present invention further utilizes the iFLYTEK Spark Large Model to perform target abstraction and semantic compression on large-scale descriptive texts, and extract keyword information related to navigation tasks.

This stage focuses on refining concepts such as "traversable area" and "hazardous obstacle", so as to enhance the structural expression of the semantic map and its adaptability to task execution.

The present invention constructs captions and spatial descriptions for each object. Combined with the reconstruction results of the 3D point cloud, it calculates the 3D bounding box IoU between each pair of object nodes to obtain a similarity matrix, estimates the Minimum Spanning Tree (MST) for pruning, generates a set of refined potential edges between objects, and finally yields a semantic map with rich semantic information.

The visualization results show that:

(1) The detection boxes and semantic segmentation masks of key elements such as traffic signs, motor vehicles and road infrastructure exhibit significant spatial consistency;

(2) The annotation scheme based on the multi-color system achieves the synergistic representation of the target category ("car") and the semantic description ("a white car on the road");

(3) From the panoramic aerial perspective of the unmanned aerial vehicle (UAV), the recognition rate of the system for multi-scale targets in complex urban scenes verifies its robust segmentation and recognition capability, and the output provides a high-precision spatio-temporal perception basis for traffic flow analysis, autonomous driving decision-making and urban digital twins.

The experimental test results conducted on the AerialVLN dataset show that the method proposed in the present invention achieves significant performance advantages in zero-shot transfer tasks and is significantly superior to other comparison models in multiple key metrics, as shown in the table below:

Table 2

Model	COCO (AP box)	LVIS -minival (AP all)	LVIS-Minival (AP rare)	LVIS-Val (AP all)	LVIS-Val (AP rare)	ODinW35 (AP avg)	ODinW13 (AP avg)
Other Best Open-Set Model	53.4	47.6	45.4	45.3	43.8	30.1	59.8
DetCLIPv3	-	48.8	49.9	41.4	41.4	-	-
Grounding DINO	52.5	27.4	18.1	-	-	26.1	56.9

T-Rex2 (text)	52.2	54.9	49.2	45.8	42.7	22	-
Ours	54.3	55.7	56.1	47.6	44.6	30.2	58.7

In the present invention:

Layer 1: Point Cloud Layer: Located at the bottommost layer of the architecture, it is based on the raw point cloud/depth data generated by the unmanned aerial vehicle (UAV) flight trajectory lines proposed in Innovation Points 1 and 2, provides the geometric structure and spatial reference of the scene, and further serves as the preliminary geometric structure and spatial reference for the upper layers.

Layer 2: Lane Graph Layer: Constructed on the basis of the underlying trajectory and perception data, it characterizes the network topological structure and connection relationship of traversable areas (e.g., air corridors, preset paths or ground lanes), and serves the local path planning of unmanned aerial vehicles (UAVs).

Layer 3: Instance Layer: This layer identifies and distinguishes objects of different semantic categories in the scene (e.g., pedestrians, plants, vehicles, buildings, etc.). The key innovation lies in the following: individual objects belonging to the same semantic category are assigned the same color identifier, and are aggregated and abstracted into a single, more representative "semantic node" by an algorithm based on spatial proximity and category consistency. The iFLYTEK Spark Large Model further optimizes description generation and introduces navigation key predicates such as "traversable area" and "hazardous obstacle". For key objects such as ruins and trapped personnel, detailed semantic descriptions and spatial relationship descriptions are constructed to generate an accurate semantic map.

Layer 4: Road Layer: It characterizes semantic regions and their attributes at a larger scale, and explicitly labels key objects such as ruins, trapped personnel and surrounding damaged buildings, thereby providing comprehensive and intuitive on-site information for rescuers and assisting them in formulating scientific and efficient rescue strategies.

Layer 5: Environment Layer: As the topmost layer of the architecture, it integrates information from the underlying layers and applies graph optimization technology to construct an overall high-level semantic topological network. This layer provides a

comprehensive understanding of the global spatial relationships and topological connections among environmental elements, and serves as the core basis for global path planning and advanced task decision-making.

The architecture adopts a hierarchical design, which performs step-by-step abstraction from raw perception data to high-level semantic topology, and finally constructs a large-scale semantic map with clear relationships that facilitates downstream navigation tasks, thereby providing comprehensive and intuitive on-site information for rescuers and assisting them in formulating scientific and efficient rescue strategies.

It should be noted that like reference numerals and letters denote analogous elements in the accompanying drawings hereinafter; thus, once an element is defined in one drawing, it is not required to be further defined and explained in subsequent drawings.

The present invention and its embodiments have been described above. Such description is not restrictive, and what is shown in the accompanying drawings is merely one of the embodiments of the present invention, and the actual structure is not limited thereto. In summary, if a person of ordinary skill in the art, inspired hereby, designs structural modes and embodiments similar to the technical solution without departing from the purpose of the present invention and without resorting to creative efforts, all such structural modes and embodiments shall fall within the scope of protection of the present invention.