

Huawei Cloud AI Video Service

Technical White Paper



CONTENTS

01

Introduction / 01

02

AI Meets Video / 05

2.1 Trends	05
2.2 Typical Use Cases	07

03

Introduction to Huawei's AI Video Service and Solutions / 09

3.1 Reference Architecture	09
3.2 Industry Video Management Service (IVM)	11
3.3 Video Intelligent Analysis Service (VIAS)	13
3.4 Pangu CV Model	15
3.5 Pangu Video Interpretation Model	21

04

Success Stories / 23

4.1 Huawei Stores	23
4.2 Logistics	25
4.3 Railways	26
4.4 Coal Mines	28
4.5 Electric Power	30

05

Looking Ahead: From Perception to Generation / 31



01 Introduction

Using cameras to photograph, record, and playback events can be dated back to the 19th century, when the French film "Workers Leaving the Lumière Factory in Lyon", often referred to as the first real motion picture ever made, was produced by Louis Lumière in 1895. Since then, photography has come a long way: from analog to digital, from cumbersome movie cameras to compact portable ones used by people all over the world, and now to built-in cameras in almost every mobile phone and portable computer. Cameras have become indispensable tools for recording and sharing information, and doing so has become part of many people's daily lives.

The use of cameras has also expanded from just movie-making to many parts of society, including city

governance, security, and industrial quality inspection. Most cities and enterprises today have numerous cameras that continuously record everything that happens. Massive amounts of video data, while enabling a wide range of use cases, can be difficult to store and manage. Extensive research has been conducted to address these challenges, seeking solutions for storing vast video datasets, utilizing video for event sensing and recording, efficiently extracting insights from massive video data, and alleviating the burden of manual video analysis.

Although AI is considered a relatively recent development, it has made steady progress and achieved significant milestones in recent years.



In 1950, Alan Turing introduced the renowned Turing Test as a benchmark for determining machine intelligence, in a paper titled "Computing Machinery and Intelligence". The landmark Dartmouth Conference in 1956, attended by luminaries such as John McCarthy and Marvin Minsky, laid the groundwork for exploring machine simulation of human intelligence and coined the term "artificial intelligence", marking the birth of AI as a field of study.

Since its inception, AI's mission has been to understand the world and free humans from mundane, repetitive tasks. From symbolism and expert systems to neural networks, deep learning, reinforcement learning, and pre-trained models, AI has witnessed remarkable

advancements in its relatively brief history. Envisioning the eventual realization of artificial general intelligence (AGI) – a form of strong AI capable of emulating human-like cognitive abilities – within a few decades or less, the application of AI for video analysis and generation emerges as a natural progression.

Leveraging extensive expertise in both video and AI domains, Huawei has developed the Huawei Cloud AI Video Service, a testament to the convergence of these technologies. This Huawei Cloud Video AI White Paper is a summary of our team's extensive research and practical insights, aiming to illuminate pathways for collective progress in driving the industry forward.



02 AI Meets Video

Technologies, far from existing in isolation, thrive when they converge. Although video and AI technologies set their sail at different coordinates in the vast tapestry of evolution, they intertwine seamlessly to create a symphony of progress in waves of ebb and flow. As shown in the following figure, both AI and video codec had navigated an extensive journey of exploration and charted new horizons in the 21st century. Deep learning introduces AI into production in various industries, while video services are introduced into the mobile Internet by H.264 encoding and integrated into every customer's daily routine. The spike of video data and the birth of pre-trained models turn the distinct trajectories of video and AI technologies into convergence, where all individuals and industries are empowered to steer their way towards potentials as witnesses to explosive business value.

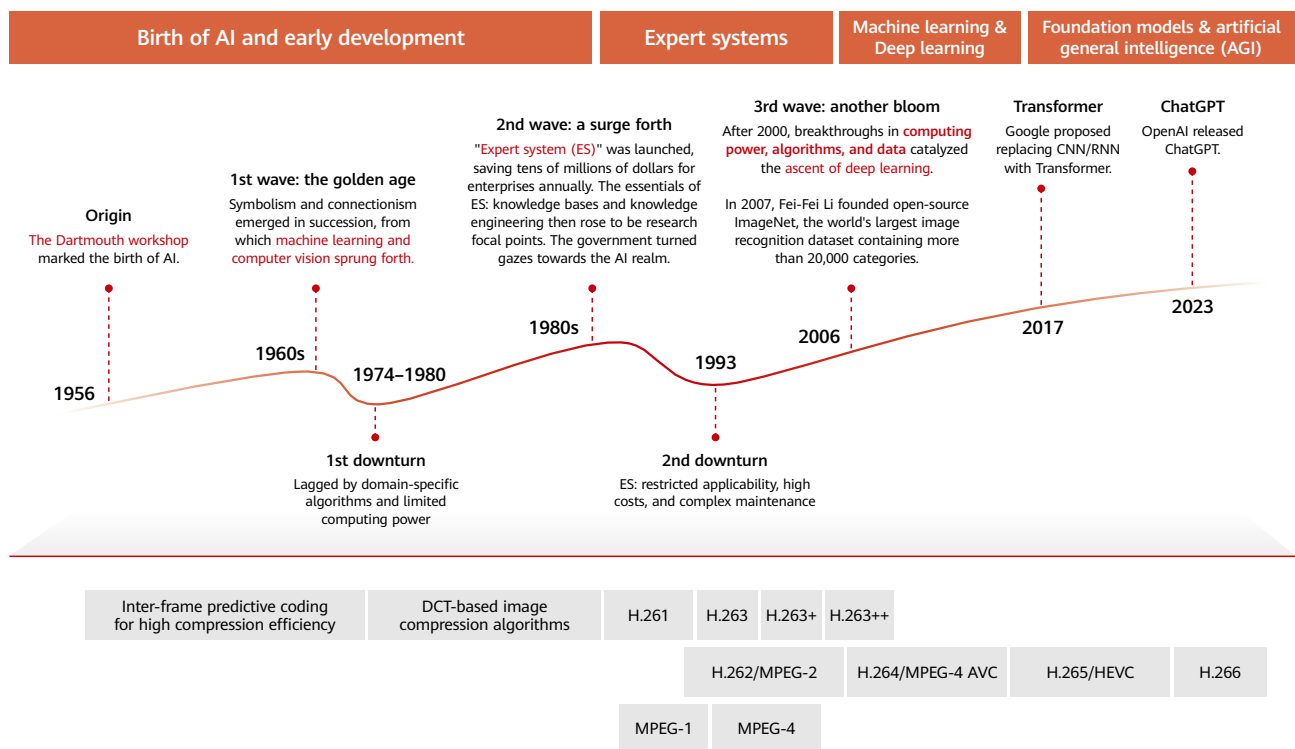


Figure 1 Development history of AI and video codec

2.1 Trends



Trend 1: Centrally Managing Video Streams on the Cloud

Deploying a large number of cameras may be easy, but managing them is not. A most wanted platform would manage cameras everywhere with rights- and domain-based permissions to ensure privacy and security. Such a platform is also expected to centrally store video streams on the cloud to avoid stream fragments and loss. A consulting report highlights over 27% compound annual growth rate (CAGR) of migrating video streams to cloud and storing them on cloud from 2023 to 2027, a number that indicates a wider future adoption of this strategy.

Trend 2: Expediting the Development of Task-specific Models with Pre-trained Models

AI technologies entailing video processing refer to computer vision (CV). CV deals with methods and algorithms to teach computers and systems to derive meaningful information from digital images, videos and other visual inputs, just like humans do. Examples include object detection, scene understanding, motion tracking, and 3D reconstruction. CV already has a wide range of use cases, such as autonomous driving, medical imaging analysis, and robot vision.

CV models are used to analyze video streams or images. A CV model refers to a deep neural network model trained to solve a range of problems in the field of computer vision. A CV model typically has millions to more than hundreds of millions of parameters, and is capable of



high-level understanding and analysis of visual data such as images and videos. Typical CV tasks include image classification, object detection, semantic segmentation, and facial recognition.

Big data and AI compute grow and model parameters expands, heralding the birth of foundation models. A foundation model refers to a deep learning model that has a very large number of parameters, usually in tens of billions. Researches show that a model's performance (or accuracy) is closely related to how many parameters it has. The more parameters, the more quickly a model learns, and the higher accuracy and the better generalization performance it will achieve in the end.

Foundation models improve the accuracy and adaptability of task-specific models, which are pre-trained on mass data with few or even no samples. By doing so, long-tail video algorithms can be improved.

Trend 3: Video Interpretation Model

Video models can analyze video streams based on specified rules, identify key events, assist in making informed decisions. However, the real world is akin to a vast mosaic, where every tile represents a unique facet of existence, forming an intricate gallery of diversity. Faced with countless rules, discriminant algorithms need to address new requirements and challenges continuously. Industries seek to employ models with better generalization to understand videos and interact in natural languages, freeing human beings from cumbersome and repetitive affairs.

Video interpretation models integrate multiple models, including the CV, multi-modal, and natural language processing (NLP) models, to analyze videos, images, voices, texts, and their combinations. By detecting the events in video streams, video interpretation models analyze, interact, and decide smart.

2.2 Typical Use Cases



City Management

A city usually has hundreds of event categories, requiring a large and complex event category system that may cover littering, road damage, and fence damage. Additionally, different cities may have different standards and areas of focus. For example, different cities may pay more attention to specific event categories. This is sometimes called fragmented AI requirements. For the large number of categories of events, it is difficult to collect sufficient samples to train task-specific models. This is called the long-tail problem of AI.

It has always been challenging to address the long-tail problem for AI development. Traditionally, for each category or each single event, a dedicated model needs to be developed, with separate data collection, model training, tuning, and deployment that heavily rely on AI and domain experts. Each model takes weeks or even months to develop. Clearly, traditional AI development cannot keep up with the fast pace of smart cities. With AI development workflows in the Huawei Cloud AI Video Cloud Solution, users can turn complex development

processes, such as data labeling, model training, and deployment into fixed steps in workflows. Zero coding is needed. With the proper data, anyone can use these workflows to quickly develop and roll out the needed AI applications. Each AI model can be developed in just days. The Pangu CV model, pre-trained on massive amounts of data, can achieve good generalization performance and robustness through few-shot learning.

To sum up, the pre-trained Pangu CV model and AI development workflows, when used together, can support the long tail of smart city applications, enabling cities to better utilize AI to improve city governance and services for both businesses and residents.

Emergency Handling

In terms of city governance, there are always emergencies to respond to, in addition to preset event categories. One example is rainstorms. When a rainstorm occurs, the government needs to quickly find out whether waterlogging has occurred in the city, so that prompt



measures can be taken to mitigate the impact or rescue missions carried. Another example is traffic accidents. When a traffic accident occurs, the traffic police department will need to know the scope of impact of this accident, so they can dispatch police forces to handle the accident or reroute traffic if necessary. The ability to properly respond to such emergencies is crucially important for city management, and the task is challenging, as a lot of the elements may vary, such as weather, location, and time.

This type of temporary, unplanned requirements, however, is catastrophic for traditional AI. With traditional AI development, for each type of event, dedicated data needs to be collected, a dedicated model needs to be trained on this data, and the trained model will be able to solve this specific task only. When a new, temporary AI requirement occurs, a new model needs to be developed, because the previously trained model cannot adapt to the new task, so the entire process is repeated all over again, which may take days even with the support of AI development workflows. In the case of an emergency, for example, waterlogging, time is crucial to saving lives. There is no

time for developing new models or algorithms. Traditional AI development may fail for city emergency responses.

Based on the latest technology in multimodal models, Huawei Cloud has developed an open-set object detection and segmentation model in the AI Cloud Video Solution. Pre-trained on massive amounts of data, this model is capable of general feature extraction. Additionally, it embeds a large language model, and this allows it to understand the semantic meaning of text entered by users. As such, users can enter text to instruct this model to start object detection tasks, and this model can recognize objects that are never in its training data. This capability is of great help to city emergency responses. For example, in the case of waterlogging, you do not need to train a dedicated model. Instead, you ask the model a simple question: "Is there waterlogging in the image?" By understanding your question and detecting objects in images, the model can give accurate answers. This shows you that the model is not task-specific. Instead, it can be easily adapted to a wide range of tasks in city management.



03

Introduction to Huawei's AI Video Service and Solutions



3.1 Reference Architecture

Combining domain knowledge with next-generation ICT technologies (connectivity, cloud, AI, and computing), AI Video Service facilitates profound synergy in sensing, perception, decision-making, and action with AI-empowered video monitoring and analytics. Leveraging the capabilities of the cloud, the service offers reliable, stable, and comprehensive features. A modular, cluster- and services-based design provides carrier-grade reliability, scalability, and maintainability and meets the need for seamless integration and interoperability between different systems, as well as application compatibility and sustainable development.

AI Video Service is a comprehensive solution that consists of three Huawei products: Industry Video Management (IVM), Video Intelligent Analysis Service (VIAS), and Pangu CV Model. AI Video Service offers a one-stop service: video ingestion from cameras, retrieval and storage management, analysis, identification of key events and abnormal events, and reporting events to upper-layer application systems. It empowers users with video AI and intelligence, facilitating enhanced oversight and control.

As shown in the figure below, the typical network for AI Video Service consists of the access layer, network



layer, platform layer, and application layer. The core services reside at the platform layer, interacting with the access, network, and application layers to deliver a comprehensive solution.

Cameras and NVRs are deployed at the access layer to collect video data. They are registered with the platform layer through standard protocols. NVRs can aggregate and manage data from multiple cameras and store video data locally.

The network layer consists of networking devices, which transmit the data captured by cameras and NVRs to the platform layer via the IP protocol. The network layer guarantees the quality of video data transmission, e.g., packet loss, jitter, and latency.

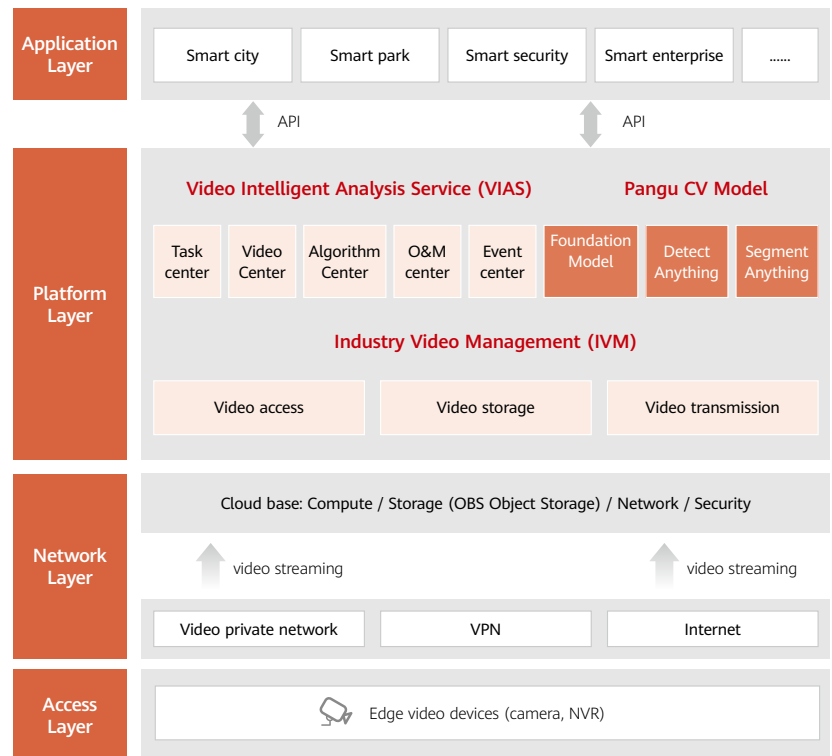


Figure 2 AI Video Service logical architecture

AI Video Service, situated at the platform layer, employs a cloud-based architecture with support for various cloud types, including public and hybrid clouds. Positioned as a SaaS offering, the service relies on cloud-based virtual machines, OBS storage, and network capabilities to manage cameras and the videos they generate, and it leverages AI to analyze and understand video streams, capture key events, and push the analysis results to upper-layer application systems.

The application layer provides graphical user interfaces (GUIs) and management functions tailored to users' requirements. Application systems vary depending on the industry to which the customer belongs. Examples of such systems include retail store customer flow management, smart security systems, and smart campus management. A loose coupling exists between AI Video Service and the application layer as the two connect through messaging interfaces.



3.2 Industry Video Management Service (IVM)

3.2.1 Requirements

Devices like software-defined cameras (SDCs) and IP cameras (IPCs) need centralized management for tasks such as device information registration, remote configuration, and rights- and domain-based management. It is imperative that their status and video content remain accessible over the Internet, from any location and at any time. Moreover, the system must ensure the recording and secure storage of video streams for future use.

IVM functions include device access, video retrieval, and recording management.

1. Device access

IVM supports access international standard protocols, China's GB/T28181, and proprietary protocols. Video streams can be decoded and displayed through proprietary protocols or SDKs.

2. Video retrieval

IVM provides media stream playback capabilities for the public and other service systems. Streams can be converted into RTMP, HTTP-FLV, HLS, and other formats that can be directly played on PCs for adaptation between the Internet web/HTML5 and video systems. This is a technical solution for service integration, openness, and video-based media applications.

Users can remotely view live video shot by cameras to observe onsite conditions in real time. Specific information can be displayed on live video images, which helps monitor live video streams and detect faults.

3.2.2 Solution

Leveraging Huawei Cloud infrastructure and technical advantages in the audio and video fields, IVM provides cloud-based video ingestion, transmission, and storage for Huawei and third-party devices such as cameras (SDCs/ IPCs), network video recorders (NVRs), and intelligent video storage (IVS1800). This service applies to scenarios such as security protection, production management, and intelligent operations. IVM helps enterprises quickly migrate video devices to the cloud and incorporate intelligence into them, facilitating enterprise digital transformation.



Live video feeds from multiple cameras can be viewed simultaneously. Users can click a camera to play its video in an idle pane that is selected from left to right and from top to down. If no idle pane is available, a new multi-pane layout can be added.

Live video information, including the current bit rate, average bit rate, encoding format, and resolution, can be displayed. Videos can be muted and unmuted, played or stopped by dragging cameras, and stopped separately or in batches.

3. Recording management

IVM comes with long-term, large-capacity, and secure cloud storage. Camera data is transferred to the cloud in

real time and backed up in reliable OBS buckets.

Users can play back recordings on clients and download recording files to their local computers for playback using a common player. They can also search for recordings for investigation purposes. Server-based and PU-based recordings in a user-defined time range can be searched. Search results are displayed in a progress bar that can be dragged. Videos can be zoomed in or out.

3.2.3 Summary

IVM provides camera management and video access, retrieval, and storage on Huawei Cloud. The following figure summarizes IVM proposal.

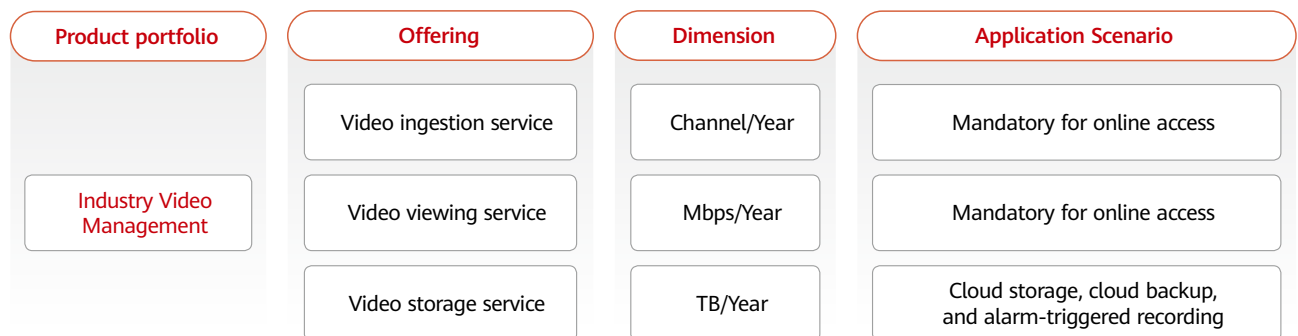


Figure 3 Proposal of Industry Video Management Service

3.3 Video Intelligent Analysis Service (VIAS)

3.3.1 Requirements

Manually examining videos can be labor-intensive, often relying heavily on individual skills for accuracy. This is where AI comes in. AI-powered video analytics automate the analysis of video streams, and key events can be precisely identified and reported. Related capabilities include but are not limited to:

- » Diverse video analysis algorithms tailored for complex scenarios;
- » A unified video analysis platform for centralized management and optimal utilization of video resources;
- » Unified algorithm management and compute power scheduling, decoupling of compute power from algorithms, and shared repository of algorithms from multiple vendors.

3.3.2 Solution

VIAS is an AI-driven, out-of-the box solution for video analytics, event identification, and decision-making support in a wide range of use cases, from smart cities, campuses, and logistics to workplace safety monitoring, public security, and chain store management. VIAS functions include analytics, algorithm center, video center, task center, and event center.



1. Analytics

The video analytics service provides an elastic computing engine that carries AI video algorithms. It is capable of video data ingestion, analysis, and alarm output. It supports service application development through APIs, helps AI developers improve AI video integration efficiency, and facilitates core service value development.

Video analysis involves the following technologies and techniques:

1) Object detection

Object detection is the first step of visual perception and an important branch of computer vision. Its goal is to use a box to mark the location of an object and predict the category of the object. In the current capabilities of the video analytics service, people or vehicle detection is the first and most critical step. The accuracy of people and vehicle detection also directly affects the effect of subsequent algorithms. However, due to the diversity and complexity of target environments, object detection is usually subject to environmental interference. So, to improve the accuracy of algorithms, custom training is performed based on actual application scenarios to eliminate interference caused by complex environments.

2) Image classification

Image classification is the task of assigning a label or class to an entire image. Generally, an image classification algorithm does this by using manually configured features or a feature learning method. Research on image classification typically derives the capability of detecting specific objects, for example, recognizing trucks and buses.

3) Object positioning

Object positioning aims to position an object in an image. It uses CV to position a target object in an image. Localization of an object can usually derive a wide range of application scenarios. For example, in the security field, algorithms such as intrusion detection, loitering detection,

and head counting can be used to determine the location of a target object.

Based on the techniques above, VIAS provides AI video analytics capabilities for smart campus, water conservancy, transportation, and emergency management. It not only works with Huawei-developed AI algorithms, but also supports third-party algorithms and industry-shared algorithms.

Various types of Huawei-developed video analytics algorithm capabilities are continuously accumulated and optimized based on more than 100 projects.

2. Algorithm Center

AI algorithms with different frameworks and functions from different vendors can be centrally managed. Users can centrally manage imported algorithm images. They can also manage the lifecycle of algorithm versions for subsequent algorithm deployment and monitoring. Users can view released algorithms. An account system is provided for third-party developers, allowing them to release algorithms and update algorithm versions. Users can be redirected from the algorithm center to the algorithm store, which displays the list of algorithms available. Users can subscribe to them as needed.

3. Video Center

The video center allows you to add and manage video

sources. You can select video sources when you create a job in the task center, facilitating intelligent video analysis. Video center functions include video source management, video quality inspection, and camera group management.

4. Task Center

The task center is the core configuration module for algorithms. Task center functions include job configuration, job management, batch configuration, public templates, and scheduled tasks.

5. Event Center

The event center provides unified event management. In the event center, government agency users can view the results of video analysis jobs and report the results to other application systems. If necessary, service tickets can be generated and allocated based on the discovered events, improving handling efficiency. Event center functions include event management, deduplication aggregation, event review, event subscription, and operations report generation.

3.3.3 Summary

VIAS provides video analytics services, algorithm management, computing power management, task management, and event management on Huawei Cloud. The following figure summarizes VIAS.

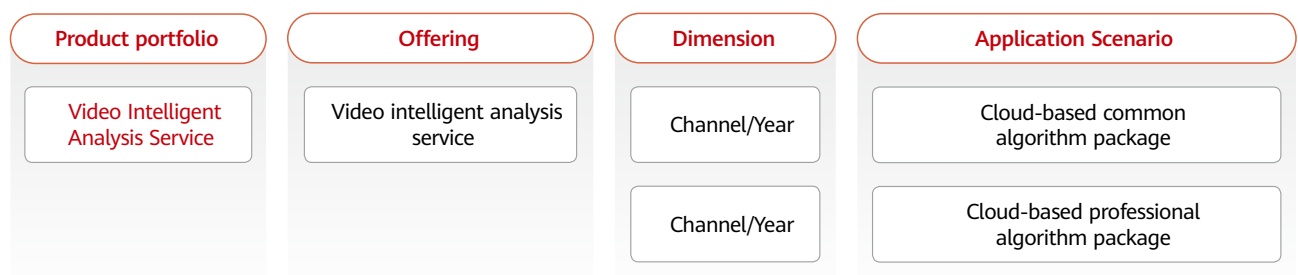


Figure 4 Proposal of Video Intelligent Analysis Service

3.4 Pangu CV Model

3.4.1 Requirements

Today, traditional industries are seeking ways to help them free workers from repetitive, labor-intensive tasks and improve productivity. This means AI algorithms will need to accommodate hugely diversified needs in vastly different application scenarios. This is why generalization is so important to AI models. Generalization refers to a model's ability to adapt to a wide range of downstream tasks. Today, many AI models in use are developed in isolated workshops. For each use scenario, a new model needs to be developed, trained, tuned, and iterated independently. This method is usually quite inefficient. Many developers also do not have the skills needed to develop and tune high-quality models in terms of accuracy, performance, and scalability. These are important reasons why it is currently difficult to scale up AI, especially in traditional industrial sectors.

Many people in the industry are working hard trying to

build pipelines that can quickly create task-specific models and train them using few-shot learning.

3.4.2 Solution

Aiming at scaling up computer vision in industrial applications, Huawei's Pangu CV Model, a computer vision model pre-trained on massive amounts of image/video data, serves as a training pipeline that allows you to quickly develop task-specific CV models. The Pangu CV Model is pre-trained on large datasets of images and image-text pairs using unsupervised or self-supervised learning methods to extract knowledge from the data, with the knowledge then being stored in the model's huge number of parameters. For any new task, the pre-trained model can be deployed to release the knowledge embedded in it. That knowledge can then be combined with industry knowledge and know-how to solve related problems.

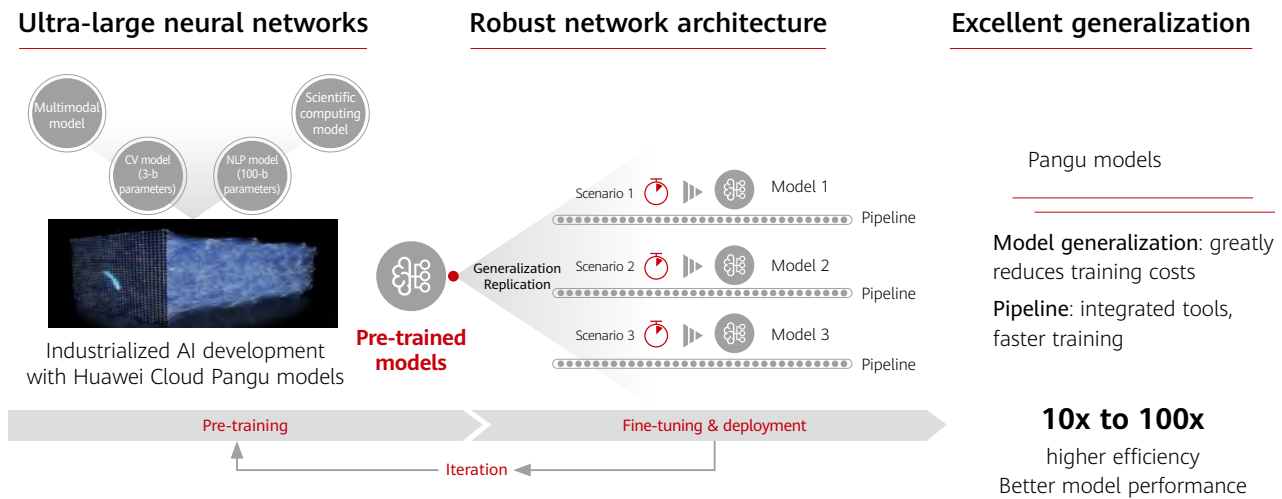


Figure 5 Pangu CV model training pipeline

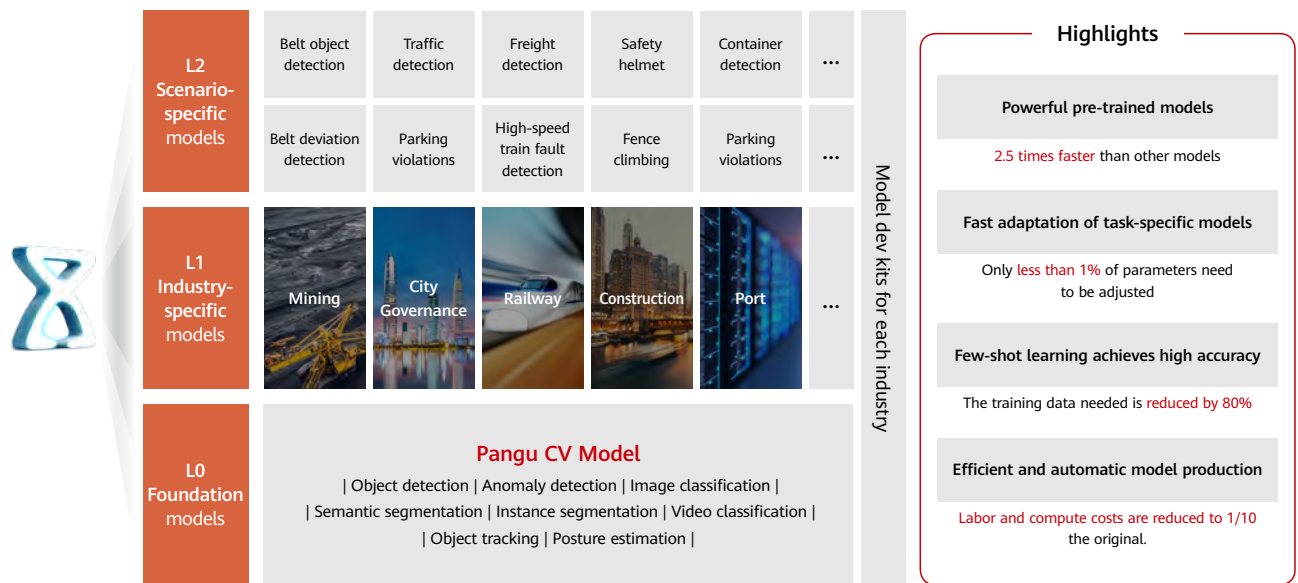


Figure 6 Use cases of Pangu CV Model

For common computer vision tasks, the Pangu CV Model can be used to quickly develop and train task-specific models through automatic model extraction and automatic parameter tuning. The Pangu CV Model offers several pre-training workflows for task-specific models, covering object detection, posture estimation, video classification, image classification, anomaly detection, object tracking, semantic segmentation, and instance segmentation. These workflows have been deployed and verified across a range of industries, such as coal mining, steel-making, railway, and transportation.

Small training samples required: The use of data retrieval and augmentation techniques reduces the demand for training data by 80% compared with conventional training methods.

High accuracy: Benefiting from better semantic alignment, the Pangu CV Model demonstrates excellent performance in few shot learning, significantly surpassing the compared method.

High efficiency: The efficient representation and data filtering capabilities of industry-tailored models are leveraged,

improving data processing efficiency by over 5 times.

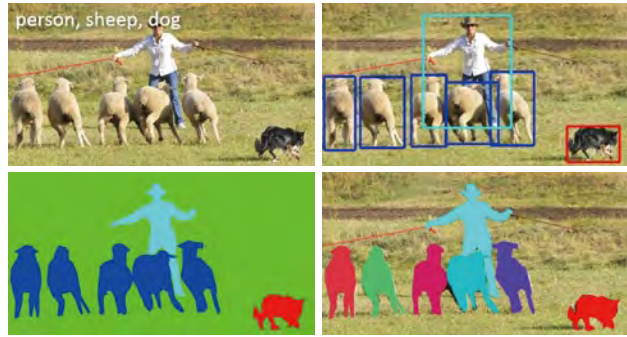
The Pangu CV Model also provides comprehensive engineering kits. With a graphical user interface (GUI), business users can perform data labeling, model development, and inference task deployment with zero coding. The Pangu CV Model is a truly viable option for enterprises looking for ways to accelerate video AI deployment in a sustainable way.

About the technology

Computer vision is about designing programs to automatically acquire, extract, process, and analyze visual signals, and to derive a high level of understanding from them. Put simply, computer vision is a discipline that studies how to teach computers to "see". Typical computer vision tasks include image classification, object detection, object segmentation, object tracking, and pose estimation. The following figure shows the famous ImageNet dataset (with over 20,000 object categories) and MS-COCO dataset (supporting different types of tasks, such as object detection and segmentation) for image classification.



The ImageNet dataset
~15M images, ~21K categories, ~1.5TB



The MS-COCO dataset
detection, segmentation, pose estimation, etc.

Figure 7 Datasets of CV model

In computer systems, visual signals are usually stored as "densely sampled intensities": the intensities of light rays coming in different directions on each channel (for example, red, green, and blue) are recorded and are used to generate a high-level representation of an image. Each basic unit in an image is called a pixel. Obviously, these pixels by themselves cannot represent any semantic information. Hence, there is a big gap between the way

images are stored digitally and the way semantics are understood by humans. In academia, this gap is called the "semantic gap", which is a core problem that almost all computer vision tasks must deal with.

Further exploring the storage formats of images, we can find several characteristics of digital image signals:



Complex content

Digital images are represented by pixels, but each pixel alone cannot express semantics. The task of image recognition is to build specific functions to output semantics from input pixels. Such functions are often very complex and difficult to define manually.



Low information density

Digital image signals can faithfully represent an object. The problem is that a large proportion of the data is used to represent low-frequency areas (such as the sky) or high-frequency areas without clear semantics (such as random noises). This means image signals have a low information density, especially when compared to text signals.



Volatile domains

The semantics derived from image signals are affected by their domains. For example, the same semantics can have distinctively different representations under different intensities of light. Furthermore, the same object can appear in different sizes, angles of view, and poses, leading to hugely different pixels. This creates challenges for image recognition algorithms.



In light of these characteristics, we believe large pre-trained models based on deep neural networks are one of the best ways to develop and deploy computer vision algorithms. The pre-training process is in a way the process of compressing visual signals. A deep neural network can extract visual features hierarchically, and pre-training combined with fine tuning can help the network adapt to different domains.

Data collection

Images are complex unstructured data that contains rich semantic information. Presently there are no good ways to accurately describe the mathematical patterns of image data, so all we can do is to collect large amounts of data to approximate real-world images. The ImageNet dataset first published in 2009 is an important milestone in the field of computer vision. It makes it possible to train and evaluate large-scale image processing methods. With

the advances made in computer vision technology and the emergence of new applications, the limitations of ImageNet datasets in terms of the scale and complexity began to show.

To solve this problem, we need image datasets that are larger and more complex than ImageNet.

We collect image data using many different methods, including downloading public datasets, expanding in-house developed datasets, search engine crawling, reverse image search, and image extraction from video. We also filter out low-quality image data, such as those with low resolution, underexposure or overexposure, or simple background, and then we use the pre-trained CV model to identify and delete duplicate images. Finally, we have developed a dataset with over 1 billion high-quality images and a total size of approximately 40 TB.



1+ billion
images



~40 TB
storage space



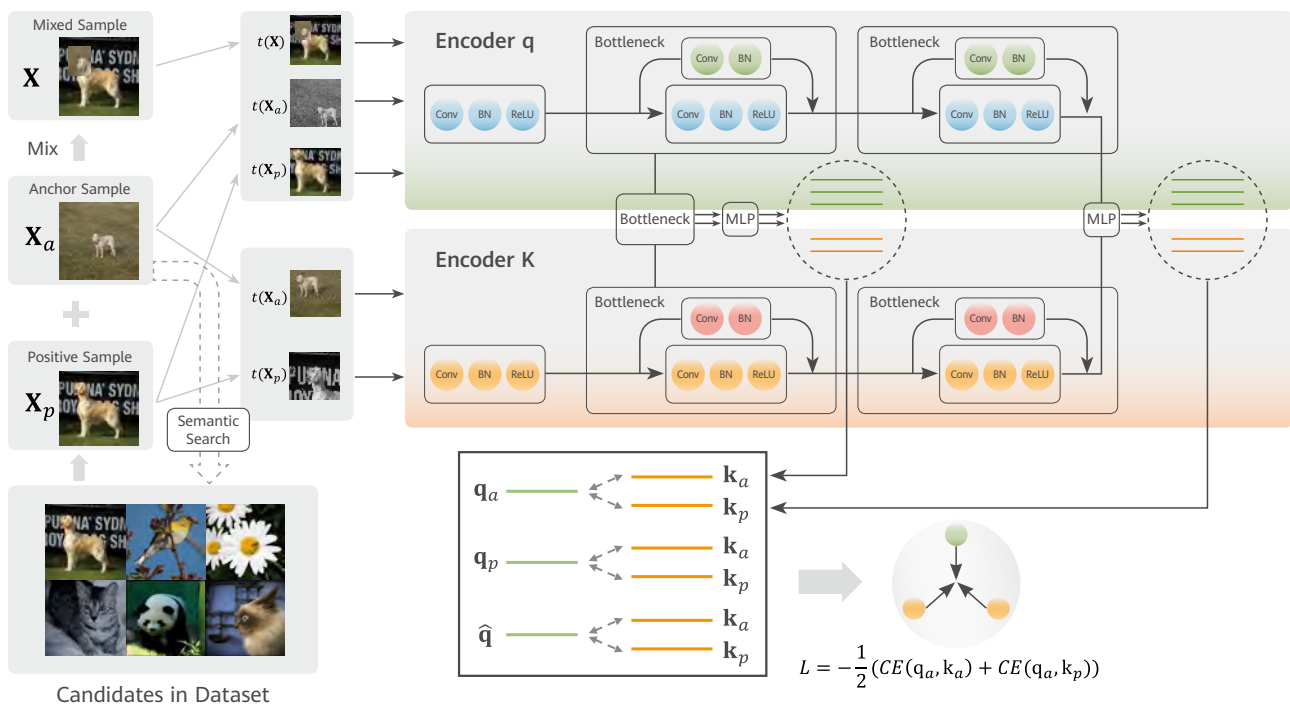
Wide application
Wide application
Autonomous driving,
electric power, railway,
remote sensing

Pre-training method

The neural network models we use include the most commonly used convolutional networks and Transformer architectures in the field of CV. They can be used separately or together to achieve optimal results. With automated machine learning algorithms, we can support and invoke neural networks of different sizes — nearly 3 billion parameters for the largest model or only hundreds of thousands of parameters for the smallest. This allows us to quickly adapt the models to different downstream CV tasks.

Most of the training data we have collected comes from the Internet and may have a high level of noise and inaccurate or no semantic labels. To fully utilize the training data, we used self-supervised learning methods.

That is, we use one or several types of proxy tasks to teach models how to understand visual data so that they can fit to complex data without any semantic labels. In particular, we optimized some proxy algorithms based on contrastive learning. We were the first to use hierarchical semantic similarity in contrastive self-supervised learning. That is, we select the neighbors nearest to the clustering centers as positive samples, and we use hybrid sample enhancement when gathering semantically similar samples. This way we reduce the impact of noise during sample selection. On top of this, we expand the number of positive samples for self-supervised learning algorithms, so that positive samples can be aggregated more efficiently and the impact of negative samples can be mitigated. The following is a simple illustration of the pre-training algorithm we use (published in TPAMI).



(Note: Contrastive self-supervised learning based on hierarchical semantic aggregation)

Performance

The Pangu CV model achieved results comparable to those of fully supervised learning models in linear classification tasks based on ImageNet datasets.

Thanks to better semantic alignment, our method also performs well in few shot learning. Trained on ImageNet - 1% and 10% labeled data, our method achieved 66.7% and 75.1% accuracy in image classification tasks, respectively, both surpassing the results of other models

by large margins. Based on this method, we designed a large model with 1 billion parameters and pre-trained it on a dataset consisting of over 1 billion unlabeled images. This model achieved 88.7% classification accuracy on ImageNet, and the accuracy of semi-supervised classification on 1% labeled data also reached 83.0%. Furthermore, the Pangu CV model achieved good generalization performance in over 20 downstream tasks, as shown in the tables below.

	Dataset	Benchmark model	Pangu pre-trained model
1	Aircraft (aircrafts)	90.43	89.32
2	CUB-200-2011 (birds)	86.90	91.80
3	DTD (texture)	80.05	85.00
4	EuroSAT (satellite images)	98.85	98.98
5	Flowers102 (flowers)	97.07	99.69
6	Food101 (food)	92.21	94.58
7	Pets	95.29	95.91
8	SUN397 (scenes)	71.51	78.92
9	Stanford Cars (cars)	92.48	94.09
10	Stanford Dogs (dogs)	87.41	91.28
11	Average	89.22	91.96

Figure 8 Pangu CV model: classification performance

	Dataset	Benchmark model	Pangu pre-trained model
1	VOC (natural scenes)	72.2	76.6
2	Comic (style transfer)	35.6	38.0
3	Clipart (style transfer)	57.5	61.0
4	Watercolor (style transfer)	34.4	36.9
5	DeepLesion (healthcare)	36.7	38.1
6	Dota 2.0 (remote sensing)	21.2	21.0
7	Kitti (autonomous driving)	29.6	32.9
8	Wider Face (human faces)	35.3	36.3
9	LISA (traffic lights)	43.5	42.7
10	Kitchen (kitchen scenes)	53.6	55.0
	average	41.96	43.85

Figure 9 Pangu CV model: object detection performance



3.5 Pangu Video Interpretation Model

3.5.1 Requirements

There is a growing demand for open-vocabulary object detection. This is especially true for emergence responses in cities. The requirements include but are not limited to the following:

Intelligent video search: Users can search for videos on cameras or in storage using natural language. For example, they can search for events that occurred at a specific time or location, or search for events using a combination of clues.

Visual tag library: Visual tags can be managed for all video streams in a refined manner. Accurate, practical

tags can be maintained and updated dynamically for all cameras.

Key frame positioning: With video data vectorization, you can pinpoint key frames for queried events, and easily access neighboring frames as well, allowing you to easily get to the bottom of the truth when investigating past events.

Intelligent video summarization: An LLM summarizes the key tags of cameras and generates a one-sentence abstract or analysis report. It also summarizes camera-captured videos and reports only key content to the supervisor.



3.5.2 Solution

Based on Video Intelligent Analysis Service (VIAS) and Pangu CV Model, the Pangu Video Interpretation Model integrates the capabilities of a multimodal foundation model to further extend the reach of AI in the field of computer vision. The bedrock of this solution is the CV model's ability to understand anything, supporting video retrieval, video tagging, and video summarization. It allows for accurate textual description of camera-captured videos and images.

As shown in the figure below, this solution consists of four parts: At the core are two foundation models: CV and multimodal. Together, they support open-vocabulary visual analysis, covering thousands of task scenarios. They

also support expert models for highly accurate recognition in certain special scenarios. On top of the foundation models, agents drive visual perception and serve as the human-machine interface. Visual capabilities can be flexibly orchestrated and assembled with plug-and-play support, allowing easy access and use of visual perception capabilities via LLM agents.

The top layer shows scenario-specific visual applications that support easy human interactions through visual-text coordination. A hierarchical visual tag system supports dynamic visual tagging, enabling refined management of video resources and helping unlock the value of video data.

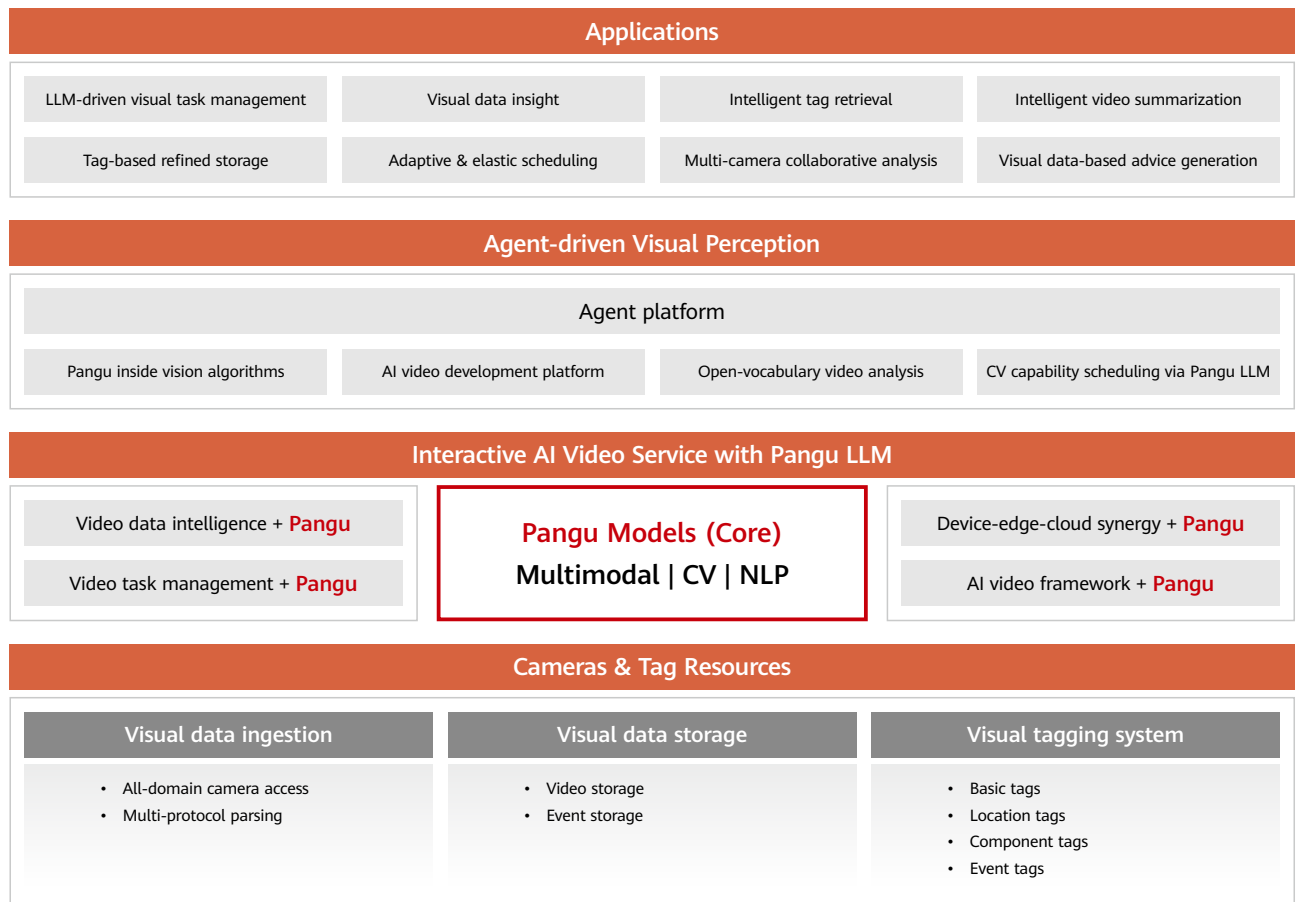


Figure 10 Architecture of Pangu Video Interpretation Model



04

Success Stories

4.1 Huawei Stores

One must be confident enough if they try their own innovations – that is what Huawei did. Huawei has deployed the Industry Video Management (IVM) system in their own operations. Huawei Consumer BG operates over 10,000 stores worldwide, with 100,000 cameras scattered in these stores for monitoring. To enhance efficiency, a centralized management system is essential for handling these cameras.



Key requirements	
 <ul style="list-style-type: none"> • Unified management of cameras • Video data security • Unified supervision of all stores • Exterior wall advertising • Clean walls 	 <ul style="list-style-type: none"> • Employee dress and behavior • In-store advertising • Merchandise and sample placement • Clean floor • In-store window branding

Figure 11 Key requirements of Huawei stores



The IVM system is a perfect fit to meet all the needs shown in the figure. It provides a cloud-based platform for simple video access, retrieval, and storage. One system can cover all requirements across stores.

The IVM system connects with cameras from different brands, so store operators do not need to rebuild their camera network. They can just update the software and get the benefits of unified management. Data is stored on the cloud for reliability and traceability, and videos are protected through watermarking and encryption during transmission and storage. With its role-based mechanism, the IVM system provides privacy for managers and operators across various levels and departments, ensuring data security. This system is crucial for Huawei store operations and has benefited Huawei in conducting business.

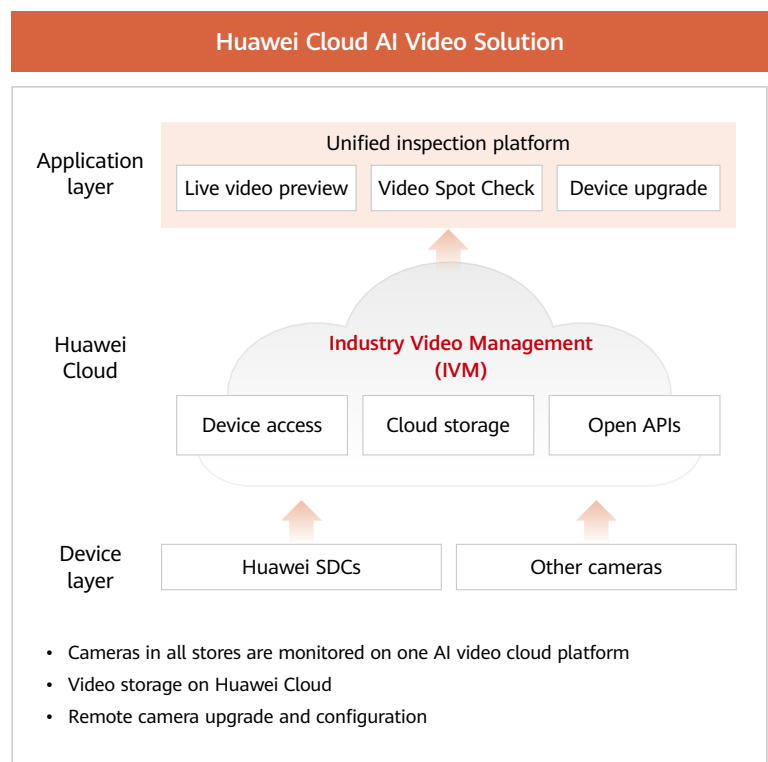


Figure 12 IVM solution for Huawei stores

4.2 Logistics

With the e-commerce booming, the logistics sector has seen rapid rise in turnover and coverage areas. This labor-intensive industry has numerous branches and warehouses, so remote management through cameras is commonly used to guarantee smooth and secure operations, ultimately enhancing efficiency and customer satisfaction. These features have brought about the following needs for the industry:

- » Unified management on tens of thousands of cameras across branches
- » Secure storage of video data
- » AI monitoring to prevent errors in handling and shipping

In response to these requirements, Huawei Cloud introduces Industry Video Management (IVM) and Video Intelligent Analysis Service (VIAS). IVM allows managers to remotely control and view videos stored on Huawei public cloud. VIAS uses AI to detect inappropriate actions at work, like rough handling of goods and smoking on the premises. This improves the quality and standardization of logistics operations.

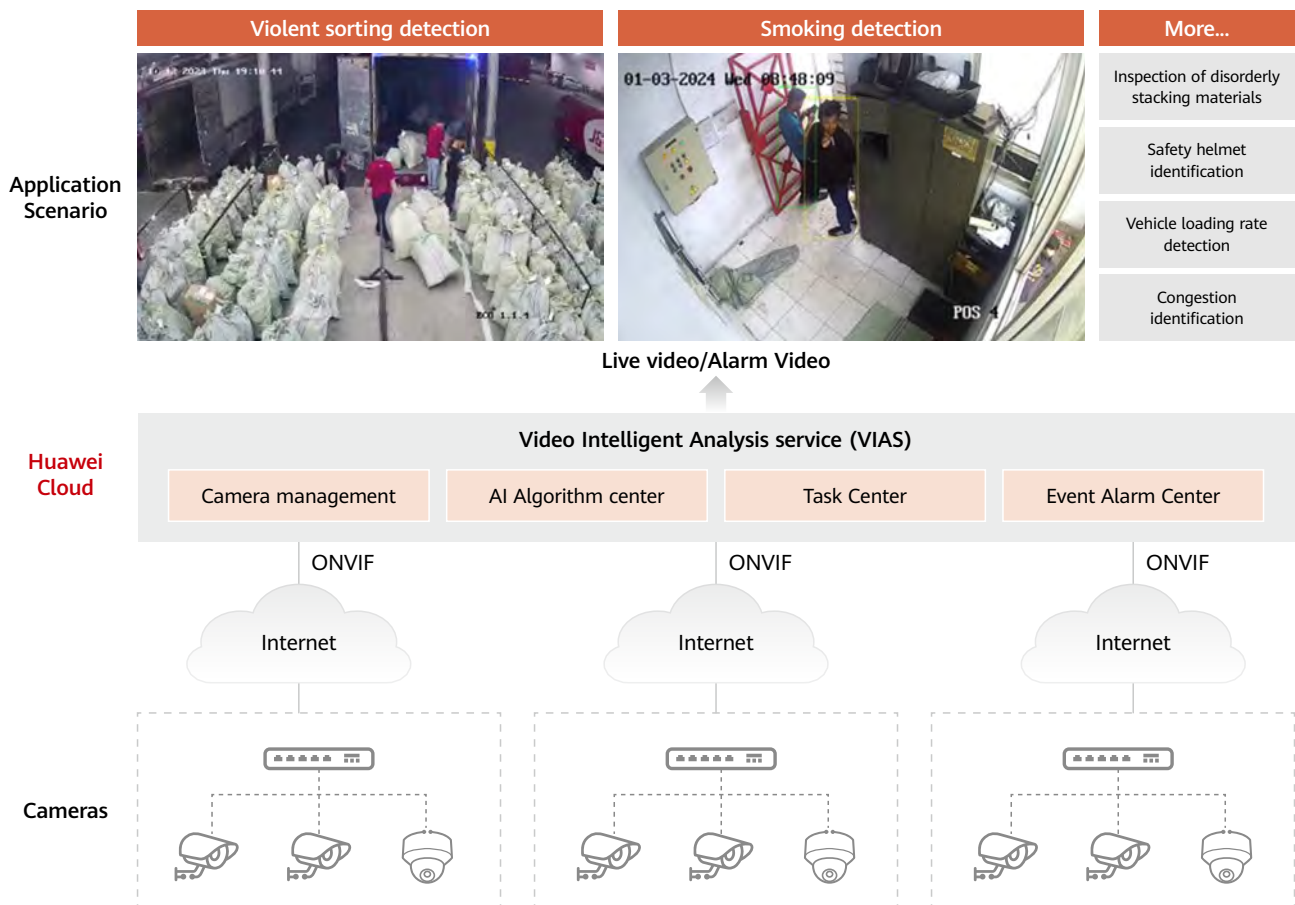


Figure 13 VIAS for the logistics industry

4.3 Railways

China's railway network spans 155,000 km with over 1 million freight trains. The Trouble of moving Freight car Detection System (TFDS) is used to detect faults and defects on moving freight trains. This system uses high-speed camera arrays installed at both sides of the railway track to capture images of the bottom and lower parts of the trains, which human inspectors then examine.

However, processing the sheer volume of the images is a daunting task. Take the 5T inspection center of Zhengzhou Railway Group as an example. This center processes over 2.8 million images from 40,000 cars of over 800 freight trains per day. An inspector has only 5 seconds to process each image and find anything suspicious. This is both intense and prone to human error due to fatigue or boredom, leading to potential misjudgments.

In 2021, China State Railway Group identified the intelligent TFDS solution as a key project. The 5T inspection center of Zhengzhou Railway Group, Huawei, and Huitie Technology were commissioned to work together on this project.

In this project, the Pangu CV model acted as the "AI trainer" of the TFDS system. Pre-trained on a huge number of unlabeled images, the Pangu model is capable of continuous learning and few-shot learning. For example, there is only sample available for a disconnected bolster center plate in all of China. By training on this single sample, the Pangu model can now accurately recognize this type of fault. The Pangu model can also generate new samples at scale based on existing samples and use the new samples to train the TFDS system.

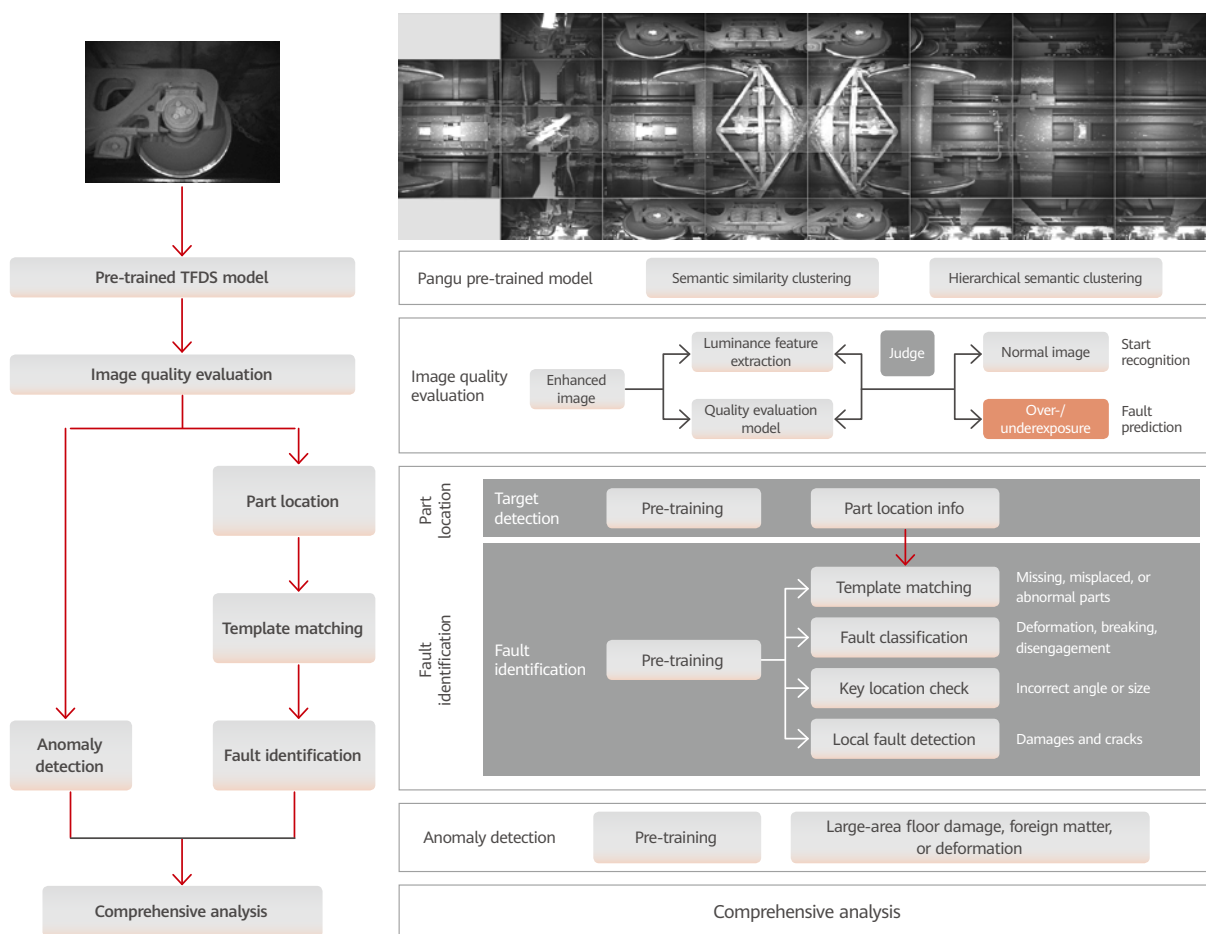
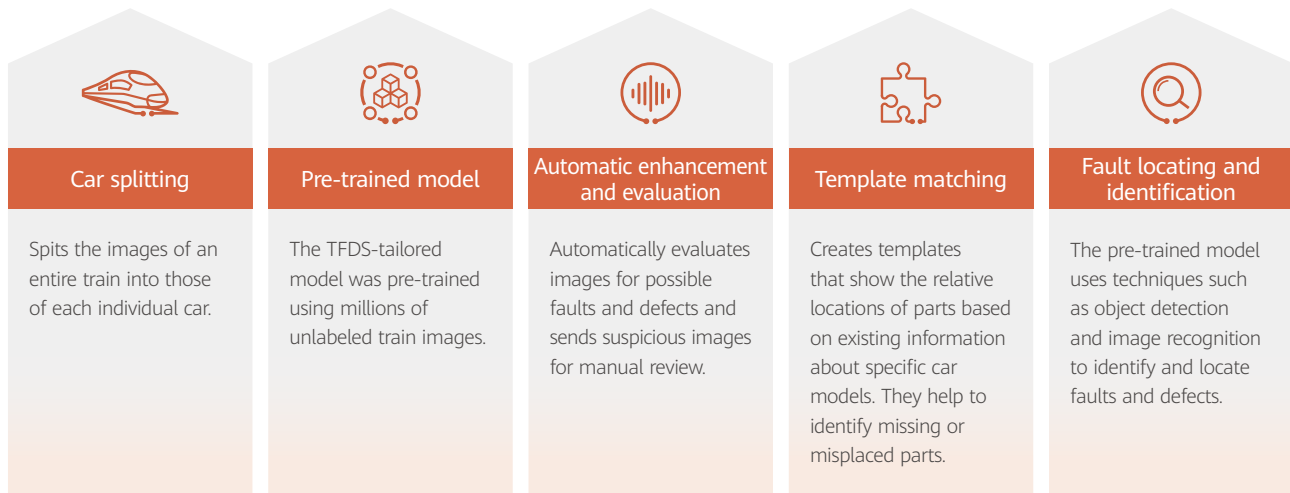
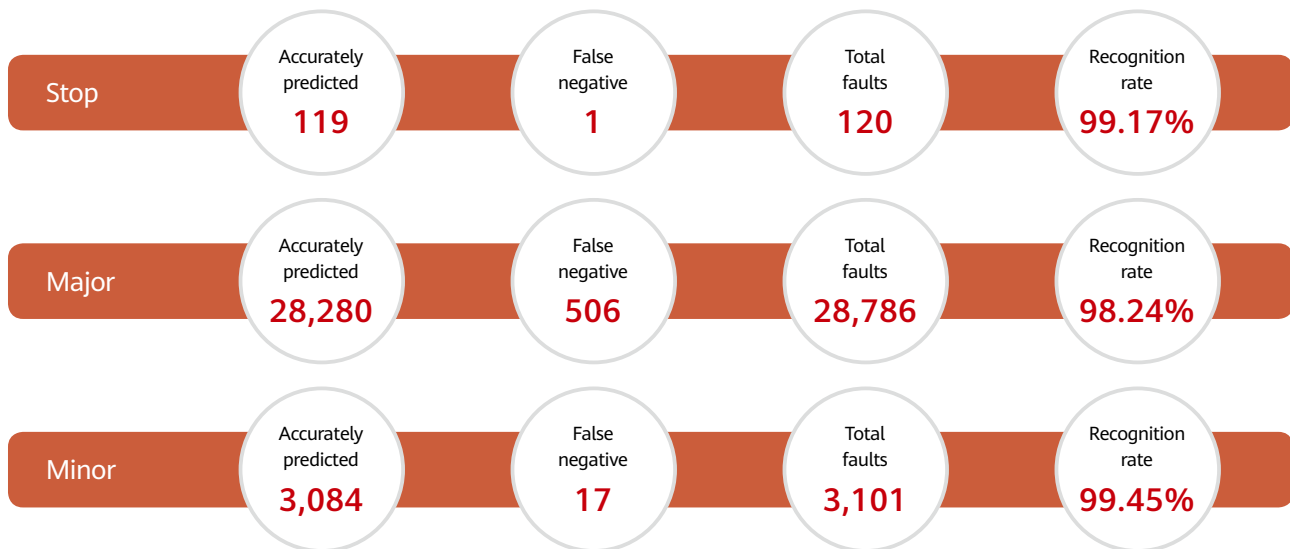


Figure 14 TFDS solution enabled by Pangu CV model

The figure above shows the tailored TFDS solution supported by the Pangu CV model. The solution includes several modules: car model screening, location classification, part screening, image quality evaluation, matching with existing templates, and multi-car cascaded analysis. Key techniques include the following:



The solution was verified on 14 railroads. In a test environment, a total of 32,007 fault samples (images that have been identified by inspectors as containing faults or defects and confirmed by their team leaders) were mixed with a large number of fault-free images and fed to the Pangu CV model. As shown in the table below, the test results show that the Pangu CV model has surpassed humans in image recognition accuracy.



In December 2022, the TFDS system, powered by Huawei Cloud's Pangu CV model, was deployed for trial use. By learning from collected samples, the new TFDS system can automatically extract component features and find anomaly patterns, and continuously improve its performance over time.

This new, AI-powered TFDS system can now accurately recognize more than 430 types of faults and defects on 67 types of freight cars. The detection rate of major faults has reached 100%, and the overall detection rate has reached 99.8%. The system filters out 95.75% of all camera-captured freight car images, leaving only 4.25% for human inspection. This significantly enhances fault detection accuracy and efficiency for freight trains, making them safer.

4.4 Coal Mines

Shandong Energy Group is a prominent energy conglomerate with diverse interests spanning mining, high-end chemical products, electric power, new energy materials, high-end equipment manufacturing, and modern logistics and trade. It is China's third largest coal producer and a global leader in smart coal mining, with nine of its mines serving as national-level exemplars for smart coal mine initiatives.

Utilizing Huawei Cloud Pangu models, Shandong Energy Group has established a corporate AI training center and applied the Pangu Mining Model across nine major coal mine processes, including mining, drivage, equipment control, transportation, ventilation, and separation. Integrating Pangu CV, prediction, and NLP models, the Pangu Mining Model facilitates large-scale use of AI applications, enabling Shandong Energy to streamline operations, enhance efficiency, and ensure safety across its coal mines.

1. Better AI models

1) High training efficiency

By harnessing cloud-edge synergy, data flows seamlessly between the corporate central cloud and edge locations, such as coal mines. The model excels in continuous learning and few-shot learning, allowing it to achieve high accuracy with limited training data. The model is trained at the Xinglongzhuang Coal Mine, Shandong Energy's smart coal mine showcase, and can be quickly replicated across its other 70+ mines.

2) Massive data processing capacity

The model, trained on over 1 billion images and 100 TB of video data using unsupervised learning, delivers extraordinary performance in visual representation and recognition.

3) Good generalization

Compared with small models, large models offer superior generalization performance. The pre-trained Pangu model



can be quickly adapted to new tasks at different coal mines, and achieve passable accuracy by generalizing over new data. There is no need to train new models from scratch.

4) High data screening efficiency

The Pangu model can efficiently obtain defect samples for brand new task scenarios, reducing the workload of data labeling by 85% compared to traditional methods.

5) High accuracy

Task-specific models can be quickly trained for a wide range of downstream tasks, covering production, safety monitoring, and decision-making, based on the principle of "anything unexpected is abnormal". With few-shot learning, these models achieve 10% higher accuracy than conventional models.

2. Higher productivity

In the past, human experts set the input parameters for processes like coal separation and coking coal blending based on their own experience. Now, by modeling real-world production data, the Pangu model accurately predicts outputs based on input parameters, optimizing these parameters to maximize productivity and benefits.

In the case of coal separation, the Pangu prediction model was used to build models that predict the outputs of heavy-media separation and ash content. Parameters are optimized for relevant processes, such as the cyclone. The density of the coal separation media and the inlet pressure can be automatically adjusted based on the ash content of the cleaned coal. The ash content of cleaned coal is stabilized and the yield of cleaned coal is increased by 0.1% to 0.2%, allowing a single coal mine to increase its production of cleaned coal by 8,000 tons a year. Replicating this solution, coal mines in China could increase their annual production of cleaned coal by an average of 2,000 tons per mine.

In the case of coking coal blending, graph neural network techniques are used to train a coal blending optimization model, which helps accelerate coal blending time from 1 to 2 days to mere minutes for the coking process.



3. Safer coal mines

The Pangu Mining Model powers intelligent video surveillance and inspection in challenging underground coal mine environments. Human inspection is now required only weekly, cutting labor costs and improving safety.

According to results from the Xinglongzhuang Coal Mine Project Phase I, the model achieved over 90% accuracy

in detecting human presence in hazardous areas, 10% more accurate than conventional small models. This helps prevent safety hazards from progressing into accidents by issuing prompt warnings. Prompt and accurate warnings on non-compliance worker behavior also help to improve workers' safety awareness.

Drilling depth is a key parameter in anti-burst and pressure relief projects. A coal mine in west Shandong used the Pangu Mining Model to monitor this parameter. Dedicated cameras are used to monitor construction in real-time, generating audio-visual alarms when insufficient drilling depth is detected. This real-time oversight, coupled with functions such as construction plan management, video recognition result query, drilling depth verification, and drilling hole counting, ensures prompt and accurate oversight and inspection while reducing human labor costs by 80%.



4.5 Electric Power



XX Company is responsible for the planning, construction, and operations management of the power grid.

The company uses drones powered by Huawei Cloud's pre-trained Pangu CV Model to inspect high-voltage power transmission lines in mountainous areas. One large model has replaced more than 20 small models, with 18.4% higher accuracy. There is no more tower climbing for field workers. All data is collected by drones operated by workers on the ground. The inspection time has been reduced from 16 days to just two days, and the fault rate has been reduced by 60%.

A field worker said they used to spend more than two weeks in the mountains inspecting power lines tower by tower. Drones are now their best assistant.



05

Looking Ahead: From Perception to Generation

VIAS is an out-of-the-box solution for video analytics, event detection, and decision-making support in a wide range of use cases, from smart cities and smart campuses to workplace safety monitoring. Huawei's Pangu CV Model serves as a training pipeline that allows you to quickly develop scenario-specific models. It is a truly viable option for enterprises looking for ways to accelerate AI transformation in a sustainable way. Pangu Video Interpretation Model is a multimodal model that can convert between videos, images, text, and voice. Many industries can benefit from using them.

For example, in the field of city governance, cameras monitor and promptly report fire hazards or extreme weather conditions. Likewise, cameras mounted on cars capture environmental details to assist self-driving capabilities. Enterprises utilize video-to-text to offer customer support or suggest relevant content. Additionally, students can easily search for and understand course materials.

Cameras are just like extensions of our eyes. Different types of cameras, such as infrared, ultraviolet, high-speed, and polarization cameras, can capture images that are not visible to the naked eye. However, the massive amount of video data generated by these cameras leads

to a data explosion and high storage costs. Consequently, a significant portion of video stream data is ignored, and images from many cameras go unused. To address this issue, we must combine AI and video technologies. By efficiently recording, storing, sensing, and analyzing data, and distinguishing between useful and useless images, AI can enhance safety in cities, improve industrial efficiency, and enhance overall quality of life. This is the beauty of "Tech for Good".

Richard Feynman, an American theoretical physicist and one of the greatest minds in human history, once said, "What I cannot create, I do not understand." This idea applies to video technology too. By embracing this concept, we can filter out irrelevant data in videos and produce videos of exceptional quality. Considering the applications of videos in gaming, industrial production, and everyday life, it becomes evident that they can simulate production processes, unlock new possibilities, and enhance communication efficiency.



It is crucial to further merge AI with video technology in the future. Despite the challenges along the journey, this fusion is poised to unleash human creativity and lead us towards greater innovations.



HUAWEI TECHNOLOGIES CO., LTD.

Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P. R. China
Tel: +86-755-28780808
www.huawei.com

Trademark Notice

 HUAWEI,  are trademarks or registered trademarks of Huawei Technologies Co., Ltd.
Other Trademarks, product, service and company names mentioned are the property of their respective owners.

General Disclaimer

The information in this document may contain predictive statement including, without limitation, statements regarding the future financial and operating results, future product portfolios, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

Copyright © 2024 HUAWEI TECHNOLOGIES CO., LTD. All Rights Reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.