Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29677

From Convolutional Neural Networks to Vision Foundation Models: A Comprehensive Review of Intelligent Evolution in Computer Vision

Xiwei Luo

School of Resources and Environmental Engineering, Hefei University of Technology, Hefei, China
270833795@qq.com

Abstract. Artificial intelligence (AI) technology is developing rapidly, and computer vision has become a cornerstone for enabling machines to perceive and interpret the visual world. This paper systematically reviews the evolution of computer vision from manually designed features to convolutional neural networks (CNNs), and then to large-scale visual models centered on the Transformer. This paper focuses on the global and hierarchical attention modeling of ViT and Swin, as well as the cross-modal alignment of CLIP and the zero-shot segmentation capability of SAM, demonstrating the generalization and transfer advantages brought by large-scale pre-training. In terms of core technologies, this paper analyzes CNN optimization, attention mechanisms, self-supervised learning, generative models, and multimodal fusion technology systems, clarifying the innovation and synergy of each module. Regarding applications, it outlines representative solutions and key implementation points in five typical scenarios. Moreover, it discusses the major challenges currently faced. It also looks forward to the development direction of lightweight, edge AI federated learning and general multimodal basic models, hoping to provide a systematic reference for scientific research and industrial practice.

Keywords: Computer vision, convolutional neural network (CNN), vision foundation model, deep learning, industrial applications

1. Introduction

In the wave of rapid development of artificial intelligence technology, computer vision, has become a key bridge connecting the digital space and the physical world. Its technological breakthroughs have directly promoted artificial intelligence from theory to practical application. From early simple image recognition achieved through manually designed features to today's automatic extraction of complex semantic features based on deep learning, computer vision has completed a paradigm shift from "feature engineering driven" to "data and model driven".

In the industrial sector, the value of computer vision has penetrated into many key scenarios. For example, in autonomous driving, perception models can identify vehicles, pedestrians, and traffic signs in real time; in medical imaging, it can achieve precise segmentation of organs and lesions,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29677

helping doctors make early disease diagnoses. These applications have changed the production and service models of traditional industries and created enormous economic and social value.

This article reviews the technological evolution from traditional methods to CNNs and then to large-scale visual models; delves into the key technologies supporting the field's development; elaborates on specific solutions and achievements for technology implementation in five typical application scenarios; and finally summarizes the current core challenges to future technological breakthroughs. It is hoped that this article will provide researchers with a comprehensive and valuable reference for grasping the forefront of the field and industry practitioners for promoting technology implementation.

2. Core methods and technological evolution

2.1. Convolutional Neural Networks (CNNs) and variants

Convolutional Neural Networks (CNNs) are the core technology in deep learning for processing visual data. Their emergence solved two major problems faced by traditional multilayer perceptrons when processing images: parameter explosion and loss of spatial structure information. Their basic logic achieving feature extraction and dimensionality compression through a three-step process of "convolution-pooling-activation." Figure 1 illustrates the architecture of a typical convolutional neural network, visually depicting the processes of feature extraction and classification.

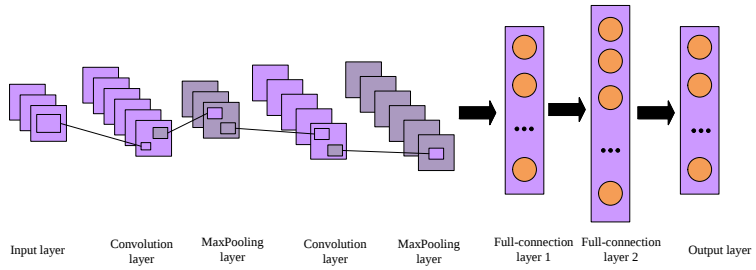

Figure 1. Architecture of a convolutional neural network (CNN)

Convolution is the fundamental mathematical operation underpinning CNNs. Feature maps are generated through the cross-correlation between convolutional kernels and input feature maps. The core computation can be expressed using the two-dimensional discrete convolution formula:

$$y_{i,j} = b + \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} \omega_{m,n} \bullet x_{i+m,j+n} \tag{1}$$

Here, $y_{i,j}$ denotes the pixel value at position (i, j) of the output feature map, $x_{i+m,j+n}$ represents the pixel values of the corresponding local region in the input feature map, $\omega_{m,n}$ denotes the weight parameters of the convolutional kernel of size $k \times k$, and b is the bias term used to adjust the overall offset of the feature map.

The size of the output feature map can be calculated using the following formula:

$$H_{out} = \left\lfloor \frac{H_{in} + 2P - K}{S} \right\rfloor + 1 \tag{2}$$

$$W_{out} = \left\lfloor \frac{W_{in} + 2P - K}{S} \right\rfloor + 1 \tag{3}$$

DOI: 10.54254/2755-2721/2025.LD29677

Here, H_{in}/W_{in} denote the height and width of the input feature map respectively, while H_{out}/W_{out} denote the height and width of the output feature map. P represents the number of padding pixels, S is the convolutional stride, $\lfloor \bullet \rfloor$ indicates the floor operation.

LeNet-5 was the first successfully implemented CNN model, laying the foundation for subsequent network designs. Figure 2 illustrates the complete architecture of LeNet-5.

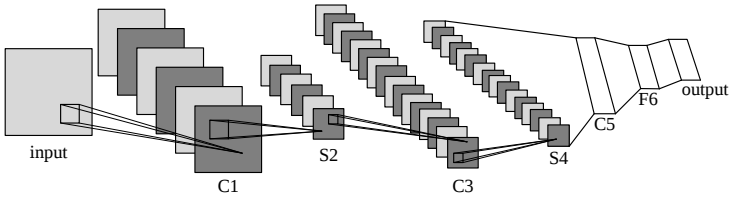

Figure 2. Complete architecture of LeNet-5

Regarding the multi-channel mechanism, color images contain three channels: RGB, allows each input channel to have an independent convolution kernel, and finally, the convolution results of all channels are summed to generate the output feature map. This mechanism enables CNNs to process multi-dimensional information simultaneously, improving their ability to process complex images. Figure 3 visually illustrates the process of the multi-channel feature fusion mechanism.

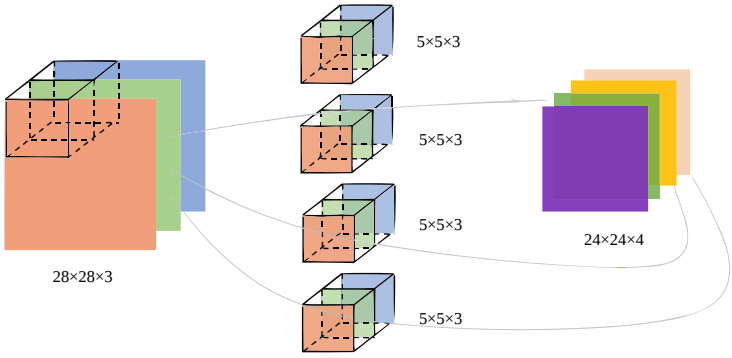

Figure 3. Multi-channel image convolution

Pooling is another essential component of CNNs. It is commonly expressed in two forms mathematically, as shown below for max pooling and average pooling, respectively:

$$y_{i,j} = max_{m,n \in poolarea} x_{i+m,j+n} \tag{4}$$

$$y_{i,j} = \frac{1}{k^2} \sum_{m,n \in poolarea} x_{i+m,j+n} \tag{5}$$

Pooling layers not only reduce computational complexity but also provide the model with a degree of translation, rotation, and scale invariance, which is crucial for image recognition tasks.

Given the computing power and data scale at the time, LeNet-5 had shortcomings such as only being able to process low-resolution images, using average pooling to cause information loss, using tanh as the activation function, and lacking modern regularization techniques. Before LeNet-5, AlexNet had a training architecture. It introduced the ReLU activation function to replace sigmoid; applied dropout in fully connected layers to prevent over-processing; used max pooling to replace

DOI: 10.54254/2755-2721/2025.LD29677

average pooling; and performed data augmentation through image transformation and other methods to expand the dataset [1].

In recent years, many variant architectures have emerged. The VGG network constructs an extremely deep network by using repeated convolutional blocks. Studies have found that stacking multiple small convolutional kernels is more advantageous than a single large convolutional kernel. For example, two 3×3 convolutional kernels are equivalent to a 5×5 receptive field, but with fewer parameters and stronger nonlinearity [2].

2.2. Attention mechanisms and transformers

The introduction of attention mechanisms has brought a transformative change to deep learning. The core principle is to emulate the human cognitive process of "selectively focusing on important information" by computing the association weights between queries, keys, and values, assigning high weights to important information and low weights to less relevant information. The essence of attention is that the model can dynamically compute the association weights between each position and all other positions, thereby selectively focusing on key information. The attention weight computation can be simplified as follows [3]:

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Here, Q denotes the query matrix, K the key matrix, and V the value matrix; $\sqrt{d_k}$ is a scaling factor to prevent excessively large values; and the softmax function ensures that the attention weights sum to 1.

The Transformer architecture is entirely built on self-attention mechanisms, abandoning the sequential or local processing methods of traditional recurrent and convolutional neural networks. The core innovation of multi-head attention can be expressed mathematically as follows:

$$MultiHead\left(Q,K,V\right)=Concat\left(head1,head2,\ldots,head_h\right)W^O \tag{7}$$

Here:

$$head_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{8}$$

Here, *h* denotes the number of attention heads, and W^{O} represents the output projection matrix.

The Transformer architecture has been widely applied in the field of computer vision. For instance, the Vision Transformer (ViT) splits an image into multiple patches, treats them as a sequence, and models the relationships between patches using self-attention mechanisms. Experimental results show that ViT achieves performance comparable to or even surpassing traditional CNNs on large-scale image datasets [4]. The ViT architecture is shown in Figure 4.

DOI: 10.54254/2755-2721/2025.LD29677

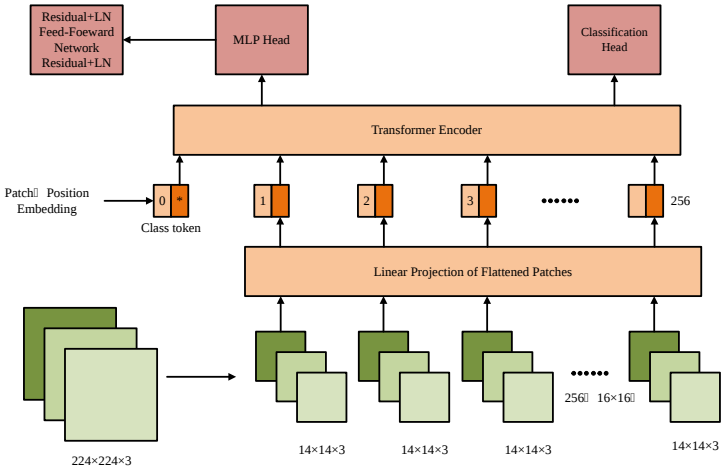

Figure 4. Schematic of the ViT architecture

2.3. Self-supervised learning and pretraining

Self-supervised learning aims to leverage the inherent structure and features of data to perform unsupervised learning. SimCLR, based on contrastive learning, generates multiple views of the same image through different data augmentations and maximizes the agreement between these views, thereby learning discriminative image representations [5]. BYOL constructs a predictor and a target network to perform self-supervised learning without requiring negative samples, thereby avoiding the reliance of traditional contrastive learning on large numbers of negatives and simplifying the training process [6]. MAE performs self-supervised learning via masked image modeling: it randomly masks parts of an image and tasks the model with predicting the pixel values of the masked regions, encouraging the model to learn the global structure and semantic information of the image [7].

The development of the pre-training paradigm has had a profound impact on deep learning. OpenAI's GPT series of models pioneered the pre-training-fine-tuning approach, which significantly improved the performance of the model in various natural language processing tasks by performing unsupervised pre-training on large-scale text data and then fine-tuning it on specific downstream tasks [8].

2.4. Generative models and data augmentation

Generative Adversarial Networks (GANs) consist of two adversarial networks: a generator and a discriminator. The generator is responsible for generating realistic data samples, while the discriminator is used to distinguish between generated samples and real samples. Through adversarial training, the generator gradually learns to generate high-quality data that is difficult to distinguish from real data [9]. Diffusion models achieve high-quality image generation by progressively adding noise to the data and learning the reverse denoising process. Compared with traditional generative models, it can generate more diverse and higher quality images [10]. In the intersection of vision and generative AI, CycleGAN introduces cycle consistency loss, which enables the generator to perform bidirectional transformation between images in different domains. This greatly expands the application scope of generative AI in vision tasks [11].

DOI: 10.54254/2755-2721/2025.LD29677

2.5. Multimodal and cross-modal fusion

Multimodal large models are dedicated to integrating data from multiple modalities. CLIP model is pre-trained on large-scale image-text pairs through contrastive learning. This process can enabling tasks such as image retrieval based on text descriptions [12]. Multimodal fusion methods can be divided into three categories, the specific differences are shown in Table 1.

Table 1. Comparison of multimodal fusion methods

Fusion Level	Operation Stage	Core Logic	Applicable Scenarios	Typical Case
Data-level Fusion	Data preprocessing stage	Directly merge raw modal data (e.g., image + depth map)	3D object detection, remote sensing analysis	Camera+LiDAR data stitching
Feature-level Fusion	Post feature extraction	Concatenate/weighted fusion of feature vectors from different modalities	Image caption generation, emotion computing	Image feature+text feature semantic alignment
Decision-level Fusion	Post single-modal model decision	Fusion of prediction results from each modality (e.g., voting, weighting)	Multi-sensor fault diagnosis	Video behavior recognition+ audio anomaly detection

Cross-modal alignment techniques are key to achieving multimodal fusion, including explicit alignment and implicit alignment [13]. Multimodal large models demonstrate great potential in understanding and reasoning tasks. In visual question answering, the model can accurately answer questions related to image content [14].

3. Representative application scenarios

3.1. Autonomous driving and intelligent transportation

Deep learning plays a critical role in environmental perception and object detection for autonomous driving. The CNN-based YOLO series, specifically YOLOv3, can detect vehicles, pedestrians, and traffic lights in real time, achieving a mean Average Precision (mAP) of 57.9% and 32 FPS on the COCO dataset, satisfying real-time requirements [15]. The Transformer-based BEVFormer model maps multi-camera image features to a bird's-eye view, achieving a NuScenes Detection Score (NDS) of 60.0% and a 15% improvement in occluded object detection [16].

In intelligent transportation management, LSTM-based time series prediction models can be used for traffic flow analysis; on the METR-LA dataset, vehicle count prediction achieves a mean absolute error (MAE) of 2.3 vehicles per minute, providing data support for traffic signal control [17]. Multimodal sensor fusion techniques integrate data from cameras, LiDAR, and millimeter-wave radar, leveraging the advantages of each sensor [4], thereby improving perception reliability of autonomous driving systems in complex environments. A comparison of key performance metrics of different core models is provided in Table 2.

DOI: 10.54254/2755-2721/2025.LD29677

Table 2. Performance comparison of core models in autonomo driving

Model Name	Core Architecture	Task Type	Public Dataset	Key Metrics
YOLOv3	CNN	Real-time object detection	COCO	mAP=57.9%, Frame Rate=32 FPS
BEVFormer	Transformer	Bird's-eye view environment modeling	nuScenes	NDS=60.0%, Occlusion Recognition Rate +15%
LSTM	Recurrent Neural Network	Traffic flow prediction	METR-LA	MAE=2.3 (vehicles per minute)

3.2. Intelligent medical imaging analysis

In the field of medical image diagnosis, the U-Net model with CNN architecture can be used for the segmentation of organs and lesions in CT and MRI images. In typical segmentation tasks such as liver, the Dice coefficient reaches 0.89 on the BraTS brain tumor dataset [18]. The TransUNet medical image analysis model based on Transformer achieves a Dice coefficient of 0.87 on the CBIS-DDSM mammogram dataset, which improves the ability to identify early lesions [19].

Generative models also have important applications in medical image processing. CycleGAN technology can achieve style transfer of different modal medical images [19]. On the IXI brain image dataset, SSIM reaches 0.92, which can maintaining the diagnostic quality of the images. The self-supervised learning model MAE learns general image features by pre-training on large-scale unlabeled medical images [4], providing feature support for small sample medical image tasks.

In 2024, CVPR Highlight proposed a multi-level visual feature adaptation and comparison framework, which uses a lightweight residual adapter and a cross-modal alignment strategy. In zero-shot anomaly detection tasks of five medical modalities, the average AUC of this method was improved by 6.24% compared with the state-of-the-art method; the average AUC was improved by 7.33% in the few-shot scenario, and it maintained a generalization accuracy of more than 85% for imaging modalities not seen during the training phase [20].

3.3. Remote sensing and geospatial intelligence

The DeepLabv3+ model with CNN architecture has performed well in semantic segmentation of remote sensing images thanks to its dilated convolution and multi-scale feature fusion technology. It can be used for land use classification and other scenarios. Its classification accuracy on the public remote sensing dataset ISPRS Potsdam has become the performance benchmark in this field [21]. The Siamese network structure change detection model can identify areas of change in land cover and is suitable for tasks such as disaster monitoring and urban expansion analysis [22].

The Swin Transformer has made a breakthrough in the task of small target recognition in remote sensing images due to the advantages of the global attention mechanism, and can effectively identify small ground objects such as street lights and manhole covers in the images [4].

3.4. Industry and manufacturing

In the field of industrial quality inspection, ResNet series CNN models can be used for surface defect detection of industrial products [23]. Super-resolution reconstruction techniques such as SRGAN can improve the clarity of low-resolution industrial images [20].

DOI: 10.54254/2755-2721/2025.LD29677

In terms of equipment operation and maintenance, the LSTM model can predict the remaining life of equipment and provide early warning of faults [24]. Reinforcement learning technology is used for industrial production process scheduling. There are already application cases in complex production scenarios such as automobile assembly.

Spiking neural networks (SNNs) in the field of neuromorphic computing have become an ideal choice for industrial edge devices. However, traditional SNNs perform far worse than ANNs (artificial neural networks) in complex target detection tasks. SpikeYOLO, proposed by the 2024 ECCV Award Candidate, is the first to achieve high-precision and low-energy application of SNNs in industrial defect detection. On the industrial defect dataset, the mAP reaches 48.9%, which is 18.7% higher than the previous SOTA results in the SNN field; compared with the same architecture ANN (YOLOv5n), the inference energy efficiency is improved by 5.7 times [25]. Table 3 provides a detailed comparison of the performance of the models introduced above.

Table 3. Performance comparison of industrial defect detection models

Model Type	Dataset	mAP (%)	Energy Efficiency (GOPS/W)	Frame Rate (FPS)
YOLOv5n (ANN)	NEU-DET	51.2	2.3	35
SpikeYOLO (SNN)	NEU-DET	48.9	13.1	28
Traditional SNN	NEU-DET	30.2	8.5	15
SpikeYOLO (SNN)	PCB Defect Dataset	45.6	12.8	25

3.5. Security and human–computer interaction

In the security field, the abnormal behavior detection model based on the AutoEncoder architecture of the CNN-based pedestrian re-identification model can identify abnormal behaviors in monitored scenarios by learning the feature patterns of normal behavior. It has application value in security early warning in public places such as subways and shopping malls [26].

In the field of human-computer interaction, the speech recognition model that combines CNN and LSTM can improve the speech recognition accuracy in noisy environments and is suitable for scenarios such as intelligent vehicles. Multimodal emotion computing technology can realize the perception of user emotions by analyzing facial expressions and speech tone features [4,27].

The Segmentation Model (SAM) has performance fluctuation issues in multi-resolution image segmentation in security scenarios. The BA-SAM proposed in the 2024 CVPR paper dynamically adjusts the dot product scaling factor of the attention layer $\sqrt{d_k \times N_{ref}/N}$ (N_{ref} =1024×1024/16×16=4096) according to the number of patches (N) of the input image. In the zero-shot segmentation task of 4K resolution image, the mask AP is improved by 18.3% compared with the original SAM; the boundary IoU of 480P low-resolution image segmentation is improved by 15.7%, and the inference speed is on par with the original SAM. Currently, BA-SAM has been applied to the pedestrian abnormal behavior segmentation task in shopping mall monitoring [28].

4. Current challenges and future development trends

Deep learning currently faces challenges across multiple dimensions. At the algorithmic level, its robustness, interpretability, and generalization ability are all insufficient. Insufficient robustness makes it difficult to cope with interference in complex scenarios; poor interpretability makes it difficult to gain trust in key domains; and weak generalization ability limits its transfer and application across different scenarios. At the data level, high-quality labeled data increases the

DOI: 10.54254/2755-2721/2025.LD29677

difficulty of model training; there is a conflict between privacy protection and data sharing. In terms of computing resources, training large models requires massive amounts of computing power and storage. This makes it difficult for many small and medium-sized enterprises and research teams to participate in related research. In engineering deployment, balancing model compression, inference acceleration, and real-time requirements is challenging. Unified modeling and knowledge transfer are very difficult in cross-modal and multi-task domains. This restricts the model's ability to work collaboratively across multiple scenarios.

However, the future development direction is relatively clear. Algorithmically, robustness and interpretability will be further strengthened; at the data level, semi-supervised and self-supervised learning will be promoted to reduce reliance on manually labeled data, while privacy protection will be enhanced through technologies such as federated learning; on the computing resource side, more efficient model architectures will be developed, and dedicated AI chips will be researched to improve computing efficiency; in terms of engineering deployment, the focus will be on lightweight models and edge AI technologies, enabling models to run efficiently on more devices; in cross-modal and multi-task domains, a unified basic model will be built to achieve effective knowledge transfer, while continuous learning technologies will be explored to allow models to continuously learn new knowledge.

5. Conclusion

This paper combs through the technological evolution of computer vision, spanning from manually designed features and Convolutional Neural Networks to large vision models. It analyzes core technologies such as CNN optimization and attention mechanisms, and expounds on the application of computer vision in scenarios including autonomous driving and medical imaging. At present, the field still faces challenges such as insufficient robustness, high data costs, and strong dependence on computing power. In the future, lightweight models, edge AI, and general multi-modal foundation models will become important development directions, providing references for scientific research and industrial practice.

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (Vol. 25, pp. 1097-1105). Curran Associates, Inc.

[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556. https: //arxiv.org/abs/1409.1556

[3] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998-6008). Curran Associates, Inc.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929. https: //arxiv.org/abs/2010.11929

[5] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020) (Vol. 119, pp. 1597-1607). PMLR.

[6] Grill, J. B., Strub, F., Altché, F., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. In Advances in Neural Information Processing Systems (Vol. 33, pp. 21271-21284). Curran Associates, Inc.

[7] He, K., Chen, X., Xie, S., et al. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022) (pp. 16000-16009). IEEE. https: //doi.org/10.1109/CVPR.2022.01597

[8] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report. https: //cdn.openai.com/research-covers/language-

DOI: 10.54254/2755-2721/2025.LD29677

unsupervised/language_understanding_paper.pdf
[9] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672-2680). Curran Associates, Inc.
[10] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840-6851). Curran Associates, Inc.
[11] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017) (pp. 2223-2232). IEEE. https: //doi.org/10.1109/ICCV.2017.244
[12] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021) (Vol. 139, pp. 8748-8763). PMLR.
[13] He, J., Zhang, C. Q., Li, X. Z., et al. (2020). Research review on multi-modal fusion technology for deep learning. Computer Engineering, 46(5), 1-11. https: //doi.org/10.19678/j.issn.1000-3428.0057370
[14] Xiang, Y., & Gong, Y. (2015). Aggregating deep convolutional and LSTM networks for multi-modal gesture recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops (ICCV Workshops 2015) (pp. 1-8). IEEE. https: //doi.org/10.1109/ICCVW.2015.128
[15] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767. https: //arxiv.org/abs/1804.02767
[16] Wang, T., Zhu, X., Pang, J., et al. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022) (pp. 11674-11683). IEEE. https: //doi.org/10.1109/CVPR.2022.01134
[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016) (pp. 770-778). IEEE. https: //doi.org/10.1109/CVPR.2016.90
[18] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015) (pp. 234-241). Springer. https: //doi.org/10.1007/978-3-319-24574-4_28
[19] Chen, H., Qi, X., Yu, L., & Dou, Q. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv: 2102.04306. https: //arxiv.org/abs/2102.04306
[20] Huang, C. Q., Jiang, A. F., Wang, Y. F., et al. (2024). Adapting visual-language models for generalizable anomaly detection in medical images. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (pp. 12345-12355). IEEE.
[21] Chen, L. C., Zhu, Y., Papandreou, G., et al. (2018). DeepLabv3+: Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018) (pp. 8392-8401). IEEE. https: //doi.org/10.1109/CVPR.2018.00873
[22] Wang, M., Chen, M., Zheng, H. T., et al. (2022). Siamese network with pyramid pooling and edge attention for remote sensing image change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 3318-3333. https: //doi.org/10.1109/JSTARS.2022.3172164
[23] Zhang, X., Yao, J., Sun, Y., et al. (2022). A survey of deep learning-based object detection in industrial scenes. IEEE Transactions on Industrial Informatics, 18(11), 8098-8111. https: //doi.org/10.1109/TII.2022.3176366
[24] Zhao, R., Yan, R., Chen, Z., et al. (2019). Deep learning based remaining useful life prediction of bearings. IEEE Transactions on Industrial Electronics, 66(11), 8505-8513. https: //doi.org/10.1109/TIE.2019.2901741
[25] Luo, X. H., Yao, M., Chou, Y. H., et al. (2024). Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In Proceedings of the 2024 European Conference on Computer Vision (ECCV 2024). Springer.
[26] Sun, Y., Liang, D., Zheng, N., et al. (2017). Person re-identification with part-aware convolutional features. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017) (pp. 716-724). IEEE. https: //doi.org/10.1109/CVPR.2017.79
[27] Zhang, J., Yao, S., Sun, A., et al. (2021). A survey on deep learning based emotional speech recognition. IEEE Transactions on Affective Computing, 13(3), 1110-1131. https: //doi.org/10.1109/TAFFC.2020.2979670
[28] Song, Y. R., Zhou, Q. Y., Li, X. T., et al. (2024). BA-SAM: Scalable bias-mode attention mask for segment anything model. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (pp. 8765-8774). IEEE.