IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 August 2023, accepted 7 October 2023, date of publication 11 October 2023, date of current version 18 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323383

SURVEY

Masked Autoencoders in Computer Vision: A Comprehensive Survey

ZEXIAN ZHOU AND XIAOJING LIU
Department of Computer Technology and Application, Qinghai University, Xining 810016, China

Corresponding author: Xiaojing Liu (645020710@qq.com)

This work was supported in part by the Qinghai Provincial Department of Science and Technology through the Applied Basic Research Project under Grant 2021-ZJ-717.

ABSTRACT Masked autoencoders (MAE) is a deep learning method based on Transformer. Originally used for images, it has now been extended to video, audio, and some other temporal prediction tasks. In the field of computer vision, MAE performs well in classification, prediction, and target detection tasks. In terms of specific application, MAE has made many achievements in medical treatment, geography, 3D point cloud and machine troubleshooting. Since its introduction at the end of 2021, there have been more than 300 related preprints, and MAE has been significantly performed in tier one computer vision conferences during 2022 and 2023. In view of the current popularity of MAE and its future development prospects, we conduct a relatively comprehensive survey of MAE mainly covering officially published articles so far. We comb through and classify the improvements in MAE, demonstrating relatively representative applications in computer vision. Finally, as a summary, we discuss the possible future research directions and development areas based on the characteristics of MAE, hoping our work could be a reference for the future work of MAE.

INDEX TERMS Computer vision survey, MAE, masked autoencoders, masked image modeling.

I. INTRODUCTION

Masked Autoencoders (MAE) is a new masked image modeling method proposed by He et al. [1] in November 2021 and published in CVPR 2022. Based on Transformer [2], MAE learns the features of a image by first masking partially and then reconstructing. At present, MAE can be well applied in image classification, image segmentation, target detection and other fields. During the past one and a half years, MAE not only appeared in the top computer vision conferences such as CVPR, ICLR, WACV, but also showed great potential in the medical field, geographical remote sensing and other aspects.

As a pre-training method, the contribution of MAE can be summarized as follows:

- MAE uses a simple NLP-like approach to perform self-supervised learning on images. It involves masking a large portion of the image (e.g., 75%) and then reconstructing the image based on the unmasked portion, thereby learning the image features. Experimental results [1] have shown that models pretrained with MAE achieve better performance in downstream tasks such as image classification.
- Self-supervised learning does not require annotated data, significantly reducing the workload and meeting the needs of training with large-scale datasets. It enables efficient training of large models.
- The effectiveness of MAE demonstrates the presence of significant redundancy in images. Even when a large portion of an image is masked, it can still be well reconstructed. The encoder of MAE only needs to process a small portion of the original image, which greatly reduces the time required for large model pretraining. Additionally, it improves accuracy while reducing memory consumption.

MAE has attracted a great deal of attention since its preprint, with the original article being cited nearly 2,000 times. At CVPR 2022 where MAE was published officially, 20 concurrent papers referenced MAE. There are also numerous research papers based on MAE at 2023 CVPR and ICLR conferences. Currently, there are approximately 300 articles related to MAE on the preprint website arXiv.

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

Given the current trend, we believe it is necessary to organize a survey focused on MAE. Here, we only examine the work of MAE in the field of computer vision. It should be noted that MAE also shows remarkable performance in time-series related predictions, such as mechanical anomalies detection [3], [4], [5], [6], [7], [133]. It can be said that MAE is a new research focus following methods such as the YOLO series and Deeplab series. Fu et al.'s works [134], [135] introduce information on related cutting-edge research.

The remaining parts of this survey introduce the following topics:

- Section II introduces the existing survey preprint of MAE (2.1), provides a detailed explanation of the MAE training process (2.2), introduces commonly used datasets (2.3), and discusses model evaluation metrics (2.4).
- Section III introduces the current improvements and variations of MAE. According to different modalities, we discuss fundamental MAE for images, multimodal MAE, and MAE for videos separately. For the improvement of fundamental MAE, we categorize it into three major classes: adding information, combining with contrastive learning, and integrating with Convolutional Neural Networks, according to specific methods. For the improvement of multimodal MAE, we provide detailed introductions from various aspects, including multimodal image, image-text multimodal, image-video, and image-audio multimodal tasks. As for MAE for videos, there are currently two overall tendencies: removing temporal information and utilizing temporal information.
- Section IV introduces the current applications of MAE for images in different domains, as well as its development in the field of videos and 3D. Regarding the former, we categorize them into three major classes: medical images, unmodified images, and geographic and remote sensing images.
- Sections V and VI summarize and discuss the future research directions, applications, and potential improvement approaches.

II. BACKGROUND

A. EXISTING MAE SURVEY

In August 2022, Zhang et al. published the first and currently only survey on mask autoencoders [8]. This survey mainly focuses on discussing related work on masked image modeling, with limited coverage on the research and improvements of MAE itself. Considering that the survey was written less than a year after the publication of MAE, its summary of recent achievements in top conferences and preprints is limited. Additionally, this survey has limited coverage of the applications of MAE and overlooks some achievements. Here, we focus on the work of MAE in the field of computer vision, analyze and summarize the recent improvements of MAE, and conduct a more comprehensive investigation of its specific applications in the image domain.

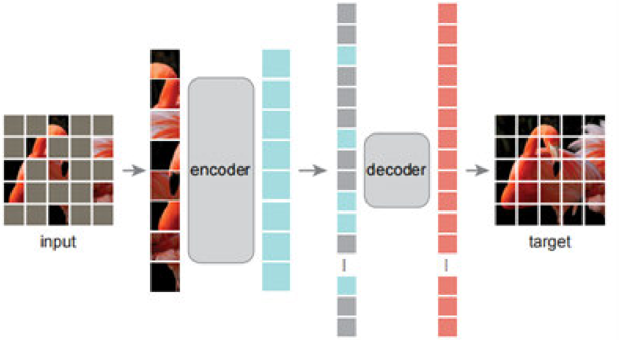

FIGURE 1. MAE training process [1]. In the MAE approach, the original image is first divided into patches, with a majority of them (e.g., 75%) masked out. The remaining portion is used as input. After going through encoding and decoding steps, the reconstructed image is obtained as the end result.

B. MASKED AUTOENCODERS (MAE) THEORY

Overall, MAE adopts an asymmetric encoder-decoder architecture. The encoder is essentially a ViT [9] model that takes as input only the visible image patches without any masking. Corresponding, the decoder is a Transformer model that operates only on the visible blocks and reconstructs the masked regions based on them. The loss function employed is MSE, as shown in (1). For an input image X, it is divided into non-overlapping image patches Xi, with a subset of blocks M being masked. Yi represents the reconstructed image block. The training process of MAE is shown in Figure 1.

$$Loss = \frac{1}{|M|} \sum_{i \in M} ||X_i - Y_i||_2^2 \quad (1)$$

On the details, MAE first performs patchification on the original image. This is actually the conventional approach of ViT [9]. Since the Transformer model itself is designed for NLP, in order to apply Transformer to image processing, it is necessary to partition the image into blocks. Patchification involves dividing the given 2D image into small squares and converting these squares into one-dimensional vectors. The dimensions of the given image are HxWxC (height x width x channels), and the side length of the small square blocks is denoted as P, i.e., they are PxPxC images. In total, N=HxW/PxP image blocks are obtained. After applying a linear transformation to each image block, they are projected into a D-dimensional space, resulting in N one-dimensional vectors of size PxPxC.

1) MASKING STRATEGY

MAE utilizes a random masking strategy, with 75% of patches masked. Additionally, it employs a uniform distribution to prevent excessive masking near the center of the image. The remaining visible image patches obtained after masking are used as input for the encoder.Recent research [10] has demonstrated that the effectiveness of MAE stems from its ability to extract high-level semantic representations from low-level features such as image pixels. The masking ratio and patch size determine which latent variables need to be recovered.

2) ENCODER

For the MAE encoder, which is ViT, Kong et al. [11] demonstrated that the reason MAE can achieve reasonable reconstruction results is due to the stable propagation of representations brought by ViT. Furthermore, comparing ViT and Convolutional Neural Network (CNN), Transformer-based models are simpler than CNN because Transformers do not rely on temporal sequences, while CNN operates in a layer-wise manner. This enables large-scale parallel computing. Cao further pointed out that because CNNs are locally supported and operate on small receptive fields, they need to be deeply stacked. ViT is globally supported, so it does not require a massive architecture like CNN. This may explain why MAE can reconstruct the entire image better using only a small number of image patches.

The original MAE uses vanilla ViT [9]. Depending on the task, there are also works that replaced it with plain ViT [9] or Swin Transformer [12], [13], [14], [15].

In addition, there are also negative evaluations of utilizing ViT. Li et al. [16], who designed CoTMAE, believe that ViT, as an encoder network for MAE, suffers from high computational cost and a large number of parameters, which presents significant obstacles in industrial detection applications. Therefore, they replaced the encoder with a convolutional-transformer hybrid structure. In the improvement of MAE, there are also studies [17], [18] that achieve multimodal training by changing the encoder.

3) DECODER

There are currently two different opinions regarding the role of the decoder in MAE. Cao et al. believe that the decoder is of great importance. Although the decoder is only used during the pre-training stage, it helps the encoder establish better representations. On the contrary, the authors of SimMIM, Xie et al., argue [19] that a more complex decoder is not necessarily better, and the decoder may not be as important as thought. In fact, SimMIM uses a simpler linear layer as the decoder. On the other hand, in the improvement of MAE, some studies [20], [21], [22] incorporate additional information through the decoder.

C. DATASETS

1) IMAGES

In order to better compare the performance, current evaluations of MAE for image processing generally include pre-training using the ImageNet-1K dataset and use ViT-B as the backbone, which means that image features are extracted by ViT-B.

The ImageNet-1K dataset [23] is derived from the large-scale visual recognition challenge ILSVRC. The "1K" represents the presence of 1,000 classes, with over one million images sourced from search engines such as Flickr. ImageNet-1K meets the requirements for large-scale datasets. Many experiments and studies currently use the ImageNet-1K dataset to assess the performance of their final models.

2) VIDEOS

There is not much practice of MAE in videos. However, most of them involve the use of the Kinetics-400 and Something-Something V2 datasets.

Kinetics-400 (K400) [24] was released in 2017 and covers 400 kinds of human actions extracted from YouTube videos. The video clips in this dataset have an average length of around 10 seconds, with a total quantity of over 200,000 videos, making it one of the largest video datasets available.

Something-Something V2 (SSv2) [25] was released at ICCV 2017. It covers 174 human actions and, like K400, is a large-scale video dataset with over 200,000 videos.

In addition, smaller datasets like AVA, UCF101, and HMDB51 are also involved. AVA [26] contains over 400 videos covering 80 basic and atomic human activities. UCF101 and HMDB51 are classic small-scale datasets that often appear in earlier studies. UCF101 [27] consists of over 10,000 videos representing 101 action classes, while HMDB51 [28] contains over 6,000 videos representing 51 action classes.

3) MULTIMODAL TASKS

For image-text multimodal tasks, the Conceptual 12M dataset is used. Conceptual 12M [29] comprises over 12 million paired images and textual descriptions. Since its release in 2018, Conceptual 12M has become an important dataset for multimodal tasks and has been continuously expanding. Conceptual 12M is not manually annotated; its image-text pairs are filtered from web content.

In the current research on image-video multimodal tasks, there is no dedicated multimodal dataset. Instead, separate image datasets like ImageNet-1K and video datasets like SSv2 are used individually.

Regarding image-audio multimodal tasks, both audio datasets and dedicated multimodal datasets are employed. AudioSet [30], released by Google, covers a wide range of sounds from various categories and events. Its content is derived from the audio portions of videos on YouTube. As the dataset includes original video URLs, it is commonly used in multimodal tasks.

VGGSound [31] consists of audio and video and, like AudioSet, is a large-scale dataset sourced from YouTube. However, VGGSound reduces the number of labels and covers sounds from more daily life scenarios.

D. EVALUATION METRICS

MAE adopts the pretraining-finetuning paradigm, which involves pretraining the model on a larger dataset to obtain certain model parameters, followed by fine-tuning on a downstream task-specific dataset. For example, after training

MAE on ImageNet-1K, fine-tuning on a dataset specific to a particular species can yield better image classification performance. In essence, this approach builds upon the large amount of data and labels available in ImageNet-1K and further refines the model. The benefit of this approach is that it allows for the utilization of a large amount of data while also saving training time. Additionally, the model benefits from the generalizable features learned during pretraining, enabling faster convergence and mitigating overfitting issues that may arise from small datasets.

There are currently two methods for evaluating the performance of pretrained models: fine-tuning and linear probing. Both methods involve adjusting the pretrained model and then assessing its accuracy.

Fine-tuning [32] refers to freezing most of the pretrained model and only training the output layers and fully connected layers. For MAE using ViT, fine-tuning involves dividing the image into patches and converting them into visual tokens based on the requirements of ViT. The adjusted model's performance is then evaluated to assess the effectiveness of the pretrained model.

Linear probing [33] involves making minimal changes to the model. This is based on the belief that the pretrained model itself should already possess certain image classification capabilities and can extract features properly. Therefore, in linear probing, only the linear layers are updated, while the other parameters of the model remain unchanged. Typically, linear probing is used as an evaluation method for the pretrained model, but in practice, fine-tuning the pretrained model is usually performed before use. It is important to note that the accuracy achieved through linear probing in MAE is significantly lower than that achieved through fine-tuning in general.

Accuracy [34] refers to the proportion of correctly classified samples in the total number of samples in a classification task. It is a traditional method for evaluating the performance of classification. The formula for accuracy is shown in (2).

$$Accuracy = \frac{correct\ classifications}{all\ classifications} \tag{2}$$

Lastly, the number of training epochs is also an aspect to consider. Training on large datasets can be time-consuming, so reducing the number of training epochs is important for the practical use and deployment of the model. The original MAE, for example, required training for 1600 epochs to achieve optimal performance when using ImageNet-1K. However, with the current advancements in MAE, comparable performance can be achieved with as few as 300 epochs or even fewer, as shown in [20] and [35].

E. RELATED THEORIES AND METHODS

1) DENOISING AUTOENCODERS: FROM NLP TO CV

The method used by MAE is fundamentally based on denoising autoencoders [36]. It aims to reconstruct the original appearance of a corrupted input and learn its features through this reconstruction process.

This approach has been widely used in natural language processing and has achieved significant results. BERT [37] and GPT [38] are two notable examples. Compared to BERT, GPT adopts an autoregressive method, which incorporates sequential elements during training. Although there have been attempts to apply similar methods in the field of image processing, the performance of convolution-based masked image modeling is not ideal. In the visual domain, contrastive learning methods have achieved more success and have various variations. Recent examples include SimCLR [39], Cao et al.'s Parametric Instance Discrimination [40], BYOL [41], and DeiT [42], while earlier examples include VGG [43].

These differences may be attributed to the distinctions between visual tasks and natural language. Some of these differences must be overcome, while others can be considered as directions for method improvements.

In terms of differences that must be overcome, natural language is inherently segmented, consisting of individual words, phrases, and sentence components, whereas images are continuous by nature. This raises the question of how to segment an image into separate objects that can be processed individually, similar to words in text.

Regarding directions for improvement, firstly, images exhibit stronger locality, with higher correlation between neighboring pixels. Unlike methods used for natural language tasks, both MAE and SimMIM perform image reconstruction tasks without semantic guidance. In fact, incorporating semantic information into the training process of MAE is one of the directions for improvement.

Secondly, Xie, the author of SimMIM, argues that [19] visual signals are relatively low-level compared to text. Current improvements in MAE also involve focusing more on low-level semantic information. Adapting the model to the characteristics of images can improve its performance to some extent.

Regarding the segmentation of images to enable processing similar to text, [9] demonstrated through experiments that when using ViT (applying Transformers to images), dividing the image into 16 × 16 patches yields the best results. This approach has also been adopted in MAE with ViT as backbone.

2) MASKED IMAGE MODELING: COMPARISONS OF NOTABLE MIM METHODS

The specific method of denoising autoencoders, when applied to the field of natural language processing (NLP), is referred to as masked language modeling (MLM). In the context of visual tasks, it is known as masked image modeling (MIM). The MAE method is a type of MIM. Here, we compare MAE with BEiT, SimMIM, and MaskFeat - that is, the MIM methods before, concurrent with, and after MAE. Table 1

TABLE 1. Comparisons of notable MIM methods. We compare and provide detailed explanations of the research conducted before, concurrently, and after MAE, namely BEiT, SimMIM, and MaskFeat. These four methods have some similarities in terms of overall architecture and training process, making them comparable. Additionally, in the subsequent improvements and developments of MAE, there are cases of mutual inspiration and influence among these methods.

Method	Masking strategy	Masking range	Loss function	Calculation object
MAE [1]	random masking	masking ratio of 75%	mean squared error	pixel values of the reconstructed and original images
SimMIM [19]	random masking	a large masked patch size of 32x32	mean absolute error	pixel values of the masked area
BEiT [44]	block-wise masking	masking ratio of 40%	mean absolute error	the normalized RGB values of visual tokens
MaskFeat [45]	block-wise masking	masking ratio of 40%	mean squared error	HOG Feature

illustrates the differences among these methods in terms of the masking process, loss function, and computational object.

Prior to MAE, BEiT [44] is the first to apply MLM from NLP to the visual domain. BEiT establishes a visual vocabulary based on the approach used in NLP, making the training process more complex. In comparison, MAE training is much simpler. Among the MIM studies conducted concurrently with MAE, SimMIM [19] gains considerable attention. SimMIM utilizes different masking strategies and loss functions compared to MAE, but still achieves impressive results. Another notable MIM method that emerged after MAE is MaskFeat [45]. MaskFeat calculates the histogram of gradient feature instead of directly computing pixels, as MAE does.

BEiT [44] introduces the concept of masked image modeling as a pretraining task. In the pretraining of BEiT, an image is first divided into image patches, and corresponding visual tokens are established. The input to the Transformer consists of masked image patches, and the output is the visual tokens. Finally, the visual tokens are decoded to reconstruct the original image.

In terms of masking strategies, BEiT uses the block-wise masking method. In simple terms, random block sizes and aspect ratios are obtained, and the process is repeated until the masking exceeds 40%. The mean squared error (MSE) loss function is used for image classification tasks.

Similar to MAE, SimMIM [19] has a simpler training process compared to BEiT. Both directly predict image patches instead of predicting visual tokens and reconstructing the image. Both the encoder and MAE utilize Transformers, but SimMIM uses a linear layer for the decoder, resulting in lower computational complexity compared to MAE. SimMIM suggests that a more complex decoder may not necessarily be better and may not play a significant role.

While SimMIM uses random masking like MAE, it uses larger masking blocks instead of increasing the masking ratio, as in MAE.

MaskFeat [45] was proposed after MAE, with a focus on video prediction. The main improvement of MaskFeat lies in the prediction target. MaskFeat compared pixel RGB values, Histogram of Oriented Gradients (HOG) features, and tokens encoded from image patches, ultimately proving that HOG as the prediction target yielded the best results. HOG is an image descriptor method proposed by Dalal in 2005 [46]. It calculates the gradient direction and magnitude of pixels in an image, divides the image into small regions, and calculates histograms of gradient directions within each region, resulting in a vector that describes the image features.

Furthermore, MaskFeat uses the same blocking-wise masking strategy as BEiT.

III. IMPROVEMENTS TO MAE

We discuss the current improvements separately by fundamental MAE, multimodal MAE, and MAE for video. Figure 2 shows the classification of improvements on MAE.

A. FUNDAMENTAL

We categorized the improvements on the fundamental MAE into four aspects: adding additional information, combining with contrastive learning, combining with CNN, and other approaches. This section provides detailed explanations and analysis for each of these aspects. Table 2 compares the differences in training process and performance among these fundamental MAE approaches.

1) ADDING INFORMATION

One direction of improvement for MAE during training is to add information through different methods. Specifically, this can be categorized into two directions: adding semantic information and adding noise.

By incorporating semantic information, the model training can be guided, reducing the memory requirements of the encoder and improving pretraining speed. MAE variants that have been improved by incorporating semantic information include BootMAE, SemMAE, and AdaMAE. The encoder of MAE is essentially a ViT that focuses on the unmasked parts of the image, as the proportion of unmasked parts is relatively small compared to the entire image. The original MAE already requires less memory compared to other ViT methods. By guiding the model with semantic information, further memory savings can be achieved, highlighting the lightweight advantage of MAE. In this regard, BootMAE transfers semantic information to the decoder, while SemMAE and AdaMAE focus on the masking strategy.

AdaMAE is specifically designed for a video-based MAE, known as VideoMAE. It incorporates semantic information to improve the performance. Further details about the improvements in MAE for videos will be discussed in the subsequent part.

The approach of BootMAE [20] (2022.7) involves incorporating low-level semantic information into the decoder, as the original MAE primarily focuses on high-level semantic

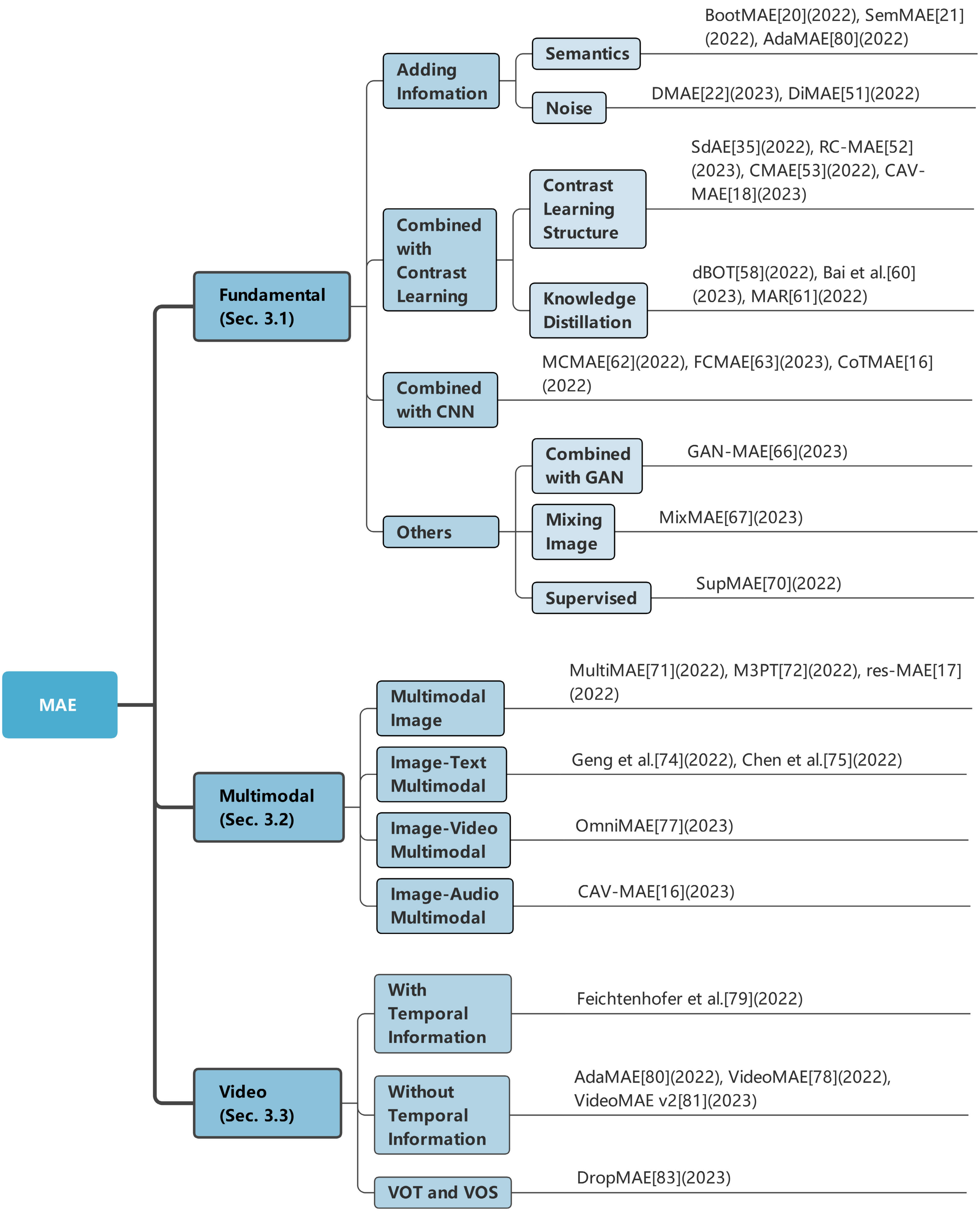
MAE
Fundamental (Sec. 3.1)
Adding Infomation
Semantics
BootMAE[20](2022), SemMAE[21](2022), AdaMAE[80](2022)
Noise
DMAE[22](2023), DiMAE[51](2022)
Combined with Contrast Learning
Contrast Learning Structure
SdAE[35](2022), RC-MAE[52](2023), CMAE[53](2022), CAV-MAE[18](2023)
Knowledge Distillation
dBOT[58](2022), Bai et al.[60](2023), MAR[61](2022)
Combined with CNN
MCMAE[62](2022), FCMAE[63](2023), CoTMAE[16](2022)
Others
Combined with GAN
GAN-MAE[66](2023)
Mixing Image
MixMAE[67](2023)
Supervised
SupMAE[70](2022)
Multimodal (Sec. 3.2)
Multimodal Image
MultiMAE[71](2022), M3PT[72](2022), res-MAE[17](2022)
Image-Text Multimodal
Geng et al.[74](2022), Chen et al.[75](2022)
Image-Video Multimodal
OmniMAE[77](2023)
Image-Audio Multimodal
CAV-MAE[16](2023)
Video (Sec. 3.3)
With Temporal Information
Feichtenhofer et al.[79](2022)
Without Temporal Information
AdaMAE[80](2022), VideoMAE[78](2022), VideoMAE v2[81](2023)
VOT and VOS
DropMAE[83](2023)

FIGURE 2. Classification of improvements on MAE. The selected studies are from high-level journals and conferences, and some research papers are highly cited preprints.

TABLE 2. Comparisons of fundamental MAE improvements. Best accuracy and corresponding pretraining epochs when using the ImageNet-1K dataset and ViT-B as the backbone are demonstrated. Masking strategies, encoders, and decoders are also shown as below.

Method	Pretrain epochs	Fine-tuning	Backbone	Masking strategy	Encoder	Decoder
SimMIM [19]	800	83.3	ViT-B	random masking	ViT	linear layer
MAE [1]	1600	83.6	ViT-B	random masking	ViT	Transformer
BootMAE [20]	800	84.2	ViT-B	block-wise masking	ViT	ViT, Transformer with MLP layers
SemMAE [21]	800	84.5	ViT-B	semantic-guided masking	ViT	StyleGAN
SdAE [35]	300	84.1	ViT-B	multi-fold masking	ViT	Transformer
RC-MAE [53]	1600	83.6	ViT-B	random masking	ViT	Transformer
CMAE [54]	1600	85.3	ViT-B	random masking	ViT	Transformer
dBOT [58]	1600	84.5	ViT-B	random masking	ViT	Transformer
Bai et al. [60]	100	84.0	ViT-B	random masking	ViT	Transformer
MCMAE [62]	1600	85.0	CViT-B	block-wise masking	hybrid convolution -transformer architecture	multi-scale decoder
FCMAE [63]	1600	84.9	ConvNeXt V2-B	random masking	ConvNeXt	ConvNeXt block
GAN-MAE [66]	800	84.3	ViT-B	random masking	ViT	Transformer
MixMAE [67]	600	85.1	Swin-B/W14	random masking	Swin Transformer	Transformer
SupMAE [70]	400	83.6	ViT-B	random masking	ViT	Transformer

information. Specifically, BootMAE introduces two decoders with identical structures. The difference lies in how they select elements from the embedding dimension of the encoder. One decoder selects elements from the modified encoder output, which changes the dimension, while the other decoder selects elements from the embedding dimension of the decoder. In contrast, the original MAE only selects elements from the decoder's embedding dimension without shuffling.

The authors believe that this design allows one decoder to handle low-level semantic information while the other decoder focuses on contextual and high-level semantic information. Experiments show that BootMAE outperforms MAE on the ImageNet-1K and COCO datasets, and it requires fewer pretraining epochs.

SemMAE [21] (2022.9) primarily focuses on proposing its own masking strategy called Semantic-Guided Masking. SemMAE first increases the number of patches to four times that of the original MAE and performs masking based on this. The masking strategy of SemMAE involves dividing the image into different semantic parts based on objects and then masking a specific part among them. When increasing the masking percentage, SemMAE partially masks a semantic part first, gradually increasing the percentage until the entire part is masked, and then moves on to the next semantic part. This way, semantic information is implicitly encoded in the masking.

In addition, SemMAE combines the pretraining results of iBOT and uses Style Gan [47] to reconstruct the spatial and texture information of the image. iBOT [48] itself is designed to capture semantic information from images. Finally, SemMAE utilizes the argmax function to process attention maps, calculating the set corresponding to the maximum value, which allows obtaining a semantic segmentation map of the original image and performing masking accordingly.

According to experiment results, SemMAE performs better than BootMAE on ImageNet-1K. Additionally, SemMAE outperforms MAE in downstream tasks of fine-grained dataset classification. The authors also evaluated SemMAE's performance on semantic segmentation of the ADE20K dataset [49]. Although the results were not entirely satisfactory, SemMAE showed better performance compared to similar methods such as BEiT and MAE.

There are currently two approaches to utilizing noise. One approach is to disrupt the image through noise. In essence, this has a similar effect to masking a portion of the image, as both involve initially damaging the image and learning its characteristics during the reconstruction process. The other approach is to utilize the information within the noise, essentially adding information through noise, which is similar to adding semantic information. DMAE adopts the former approach to improve robustness and enhance the classification performance of downstream tasks, while DiMAE adopts the latter approach.

The goal of the DMAE [22] is primarily to increase the model's robustness. DMAE first adds Gaussian noise to each pixel value and randomly masks several patches to disrupt each image. Then, a Transformer-based encoder-decoder model is trained to reconstruct the original image from the corrupted model. This approach is an extension of the Gaussian smoothing model, which takes the noise-corrupted image as input for classification. Similar to the work of Carlini et al. [50], DMAE consists of two stages. The first stage is to process the noise-corrupted input, and in the second stage a classifier is applied to predict the labels in the denoised image. The resulting model can better withstand adversarial attacks, maintaining correct classification even in the presence of small perturbations.

The overall idea of DiMAE [51] is to add noise from different domains to the image and then reconstruct it to learn domain-invariant features. On one hand, DiMAE preserves the content of the input while adding style information from other domains. On the other hand, DiMAE restores the original domain style through multiple domain-specific decoders. Regarding "content-preserved style blending," DiMAE's experiments found that content noise leads to a performance decrease in cross-domain reconstruction tasks. Therefore, DiMAE adopts a new non-parametric content-preserved style blending method to utilize cross-domain reconstruction while avoiding the performance degradation caused by content noise. As for ''restoring domain style,'' the encoder transfers domain information to the decoder, guiding the reconstruction of the input image's style.

2) COMBINED WITH CONTRASTIVE LEARNING

From an overall perspective, in the MAE improvements combined with contrastive learning, there are relatively few changes to the training process of MAE itself, and the research mainly focuses on how to achieve the integration.

Contrastive learning is a popular self-supervised learning method in recent years, mostly used for classification tasks. As the name suggests, contrastive learning focuses on the commonalities among similar instances and the differences between different classes. Therefore, contrastive learning often incorporates data augmentation techniques. Among them, BYOL [41] has achieved better performance than previous contrastive learning methods. BYOL adopts an asymmetric structure, using one neural network to predict the representations in another network, where the second network is referred to as the target or teacher network, and its parameters are periodically copied from the first network (online or student network). The student network is trained by minimizing the difference between itself and teacher networks. The success of BYOL has inspired many subsequent works, and SdAE and RC-MAE have adopted the BYOL paradigm to improve the performance of MAE.

In SdAE [35], the student component utilizes both a decoder and an encoder to reconstruct the masked portion, while the teacher component is used to explore the latent representation of the masked tokens. Additionally, SdAE employs multiple masked views to enhance performance.

Specifically, the output of the teacher component in SdAE utilizes the EMA, which is an important component of BYOL. EMA [52], standing for Exponential Moving Average, is used for parameter updates and optimization. EMA is a special type of average that assigns greater weight to recent data in a given time period. The EMA formula is shown in (3). Experiments have shown that EMA helps maintain the differences between the two networks in BYOL, thereby preventing model collapse and significantly improving model performance.

$$EMA_{cur.} = Weight \times (Value_{cur.} - EMA_{pre.}) + EMA_{pre.} \tag{3}$$

In the process of incorporating contrastive learning into MAE, SdAE proposes a multi-fold masking strategy to handle the images inputted into the teacher network. This strategy involves dividing the masked tokens into several groups on top of random masking and independently inputting them into a shared teacher network. The teacher network then calculates the tokens for each group separately.

In terms of results, although SdAE does not show a significant improvement in accuracy, it greatly reduces the number of pretraining epochs. The original MAE, when using the ViT-Base model for ImageNet-1k classification, requires 1600 epochs to achieve an accuracy of 83.6%. In contrast, SdAE only requires 300 epochs to achieve an accuracy of 84.1%.

Compared to SdAE, RC-MAE [53] is simpler as both the student and teacher networks use the same architecture. RC-MAE first randomly masks the original image and then feeds the remaining parts separately to the student and the teacher. The student component is responsible for reconstructing the image, while the teacher component predicts the missing parts. Similar to SdAE, RC-MAE also utilizes the EMA from BYOL to update the parameters of the teacher component. The main improvement of RC-MAE lies in modifying the loss function to be a combination of the reconstruction loss between the original image and the teacher network and the consistency loss.

$$Loss = \frac{1}{|M|}\sum_{i \in M}(||X_i - Y_i||^2 + ||Y_i - Y_i'||^2) \tag{4}$$

In (4), i represents the token index, Y and Y' represent the reconstructed parts from the decoder of the student and teacher networks, respectively. X denotes the segmented image patches, and M represents the mask token. Therefore, Xi represents the randomly masked image patches.

In terms of results, taking the classification task using the ImageNet-1k dataset as an example, RC-MAE does not show significant improvement when fine-tuning with ViT-B, but it shows some improvement when fine-tuning with ViT-L compared to MAE. Additionally, RC-MAE performs slightly better than MAE in downstream tasks such as image detection and segmentation, while requiring fewer memory resources during training.

In CMAE [54], the student component utilizes an encoder-decoder to learn latent representations and reconstruct images, while the teacher component utilizes a momentum encoder [55] to provide contrastive learning supervision. To better leverage the benefits of contrastive learning, CMAE also incorporates an auxiliary feature decoder in the student component. In order to ensure semantic integrity, the momentum encoder in the teacher component of CMAE uses complete image patches.

In terms of data augmentation, CMAE generates two different views to align with the characteristics of contrastive learning, and these views are inputted into the student and teacher branches, respectively. According to [55], color augmentation methods applied to views in contrastive learning can degrade the results. Therefore, spatial and color data augmentation is applied to the input of the teacher branch. Additionally, to prevent excessive differences between the two branches, CMAE employs a weak augmentation method to generate inputs for both the student and teacher networks. The core idea of this method is to first obtain the primary image through random cropping from the original image. Then, this primary image is used as a shared object for both branches, and each branch generates its own view by fine-tuning the cropping positions based on this shared object.

Regarding the training objective, the reconstruction loss in CMAE uses mean squared error (MSE) as the loss function, specifically calculating the loss between the masked patches and the predicted results. For the contrastive loss [56], the authors of CMAE experimentally demonstrate that the InfoNCE loss [57] performs better than the BYOL loss [41]. The former focuses on extracting similar positive samples from the same sample, while the latter expands the commonalities between positive samples.

CMAE achieves the best performance when trained for 1600 epochs, with a fine-tuned accuracy of 85.3%.

The mentioned works represent typical cases of improving MAE. In addition to those, CAV-MAE [18] applies contrastive learning and MAE to audio-visual multimodal tasks, while dBOT [58] incorporates MAE and knowledge distillation [59] into contrastive learning with a guided teacher network, without making significant changes to MAE itself. Bai et al. [60] also utilize MAE to enhance knowledge distillation. Due to MAE's ability to achieve good reconstruction even when most of the information is masked, incorporating MAE leads to higher efficiency and robustness in knowledge distillation. MAR [61] introduces MAE to improve the generalization ability of knowledge distillation.

3) COMBINED WITH CONVOLUTIONAL NEURAL NETWORKS (CNN)

Combining with Convolution Neural Networks (CNN) is also a direction of improvement for MAE. MAE itself is based on Transformer. Although CNN has dominated the field of computer vision for the past decades, replacing CNN with Transformer has become a rising trend.

On one hand, in the context of the masking-reconstruction task, the challenge with CNN is that finding ways to integrate masked tokens or position embeddings into convolutional networks is required. Although the earliest masked image modeling was actually implemented by CNN. On the other hand, while Transformer networks often have larger capacity, their generalization performance may be worse than CNN due to a lack of proper inductive bias.

Currently, methods that combine convolution with MAE can be concluded into two kinds. One approach is to use CNN to replace or improve a specific training process of MAE, while the other approach is to construct architectures that combine CNN with MAE.

MCMAE [62] utilizes masked convolution to prevent information leakage within convolutional blocks. To address the issue of heavy computational cost associated with the original MAE masking strategy, MCMAE employs a block-wise masking strategy to improve computational efficiency.

Specifically, the encoder of MCMAE progressively, on one hand, abstracts the input image to generate multi-scale token embeddings. The decoder, on the other hand, reconstructs pixels based on the masked tokens. In the early stages of high-resolution token embeddings, convolutional blocks are used to encode local content. In the later stages of low-resolution token embeddings, Transformer blocks aggregate global context. As a result, the encoder obtains both local and global perspectives at different stages, generating diverse multi-scale features.

In the current masked autoencoder frameworks, all tokens in the masking strategy need to be preserved for the later Transformer processing, resulting in high computational costs and losing the efficiency advantage of MAE. To address these issues, MCMAE adopts a block-wise masking strategy. It first obtains the mask for the later Transformer stage and gradually upsamples the mask to larger resolutions in the early convolutional stages. This way, the processed tokens can be fully divided into masked tokens and visible tokens, inheriting the computational efficiency of MAE. By equipping the early convolutional blocks with masked convolution, information leakage is prevented, avoiding the mixing of features from masked and visible regions in the later stages, thereby improving training performance.

Although MCMAE involves minor modifications, it yields significant results. According to its performance on ImageNet-1K, it achieves performance second only to CMAE, which utilizes contrastive learning.

FCMAE [63] adopts a masking strategy where 60% of the original input is randomly masked. It uses the ConvNeXt [64] model as the encoder and a lightweight ConvNeXt block as the decoder, forming an asymmetric structure with a heavy encoder and a light decoder.

CoTMAE [16] takes inspiration from CoAtNet and combines CNN and self-attention in the encoder using a hybrid convolution-transformer pyramid network. The pyramid structure progressively downsamples the input, with four stages that shrink the input by 1/4, 1/8, 1/16, and 1/32. The first two stages utilize convolution for local feature encoding, while the latter two stages employ convolution and self-attention fusion modules. CoAtNet uses convolution for downsampling and global relative attention operations, incorporating progressive pooling in multiple stages.

Specifically, CoTMAE first masks the original image by dividing it into equally sized parts and randomly masking each part with a fixed 75% masking ratio. Then, it extracts visible blocks from each part, reorders and reassembles them as the input image for the encoder. This input is then passed through the hybrid convolution-transformer pyramid network. Additionally, CoTMAE designs a Transformation-Convolution Fusion (TCF) module, which combines convolutional layers and self-attention layers. To facilitate the transition between attention and convolution blocks, DW-Conv [65] is added as implicit positional encoding in the Multi-Head Self-Attention (MHSA) module [2].

Finally, CoTMAE uses mean squared error (MSE) as the loss function, similar to BERT, only calculating the loss on the masked patches. In terms of performance, on an industrial segmentation dataset, CoTMAE demonstrates better fine-tuning performance compared to supervised models.

4) OTHERS

GAN-MAE [66] combines MAE with GAN by using a complete MAE as the generator. The discriminator is responsible for determining whether each image patch is generated or original. This approach reduces the number of iterations required and improves accuracy. The authors believe that this discriminator can provide guidance for the image generation process.

MixMAE [67] addresses the slow training speed of MAE by proposing a novel approach. It randomly masks two images and combines them to form a mixed image, which is then used as input to reconstruct both original images. For MAE itself, MixMAE makes minimal modifications by only replacing the decoder with Swin Transformer. Similar studies have attempted to use heterogeneous ViT for MIM methods, such as [14], [15], [16], [68], and [69]. However, there are not many research studies that have explored the application of this method for improving MAE, which could potentially become a future research direction. In terms of results, the model's performance has been significantly improved.

Finally, the above approaches maintain the self-supervised learning nature of MAE or combine it with other unsupervised methods. SupMAE [70], which stands for supervised MAE, introduces supervised methods into MAE improvements for the first time. It adds a supervised branch parallel to the original MAE training process. This branch processes the masked input using a two-layer MLP and utilizes cross-entropy as the classification loss function. Although SupMAE does not significantly improve precision, it greatly enhances training efficiency and robustness while requiring fewer computational resources.

B. MULTIMODAL

The comparison of different multimodal MAE methods is shown in Table 3. Considering that different multimodal tasks may involve multiple different datasets, in addition to the masking strategies, encoders, decoders, and loss functions involved in the training process, the datasets used for each task are also listed. On one hand, this is because comparing performance directly based on different datasets is challenging, and on the other hand, it aims to provide insights and references for research in this direction.

1) MULTIMODAL IMAGE

Image multimodality primarily refers to different visual modalities, such as depth maps, semantic segmentation graphs, in addition to the common RGB images for the same scene. In practical applications like the medical field, there are various image modalities like CT and MRI, and in geographic remote sensing, there are point clouds, GS images, and so on. Currently, there are two main approaches for applying MAE to multimodal image tasks.

The first approach is to process each visual modality separately. However, this method often leads to a more complex training process.

The second approach is to process the images themselves by fusing different modalities, allowing a image to convey multimodal information. This enables the transformation of multimodal tasks into regular image processing tasks. However, whether this method is suitable for downstream tasks needs to be further considered in specific work.

MultiMAE [71] addresses image multimodal tasks by masking the RGB mode, depth map, and semantic segmentation graph separately. It employs different loss functions for different modalities and conducts pretraining on different datasets. The encoder and decoder follow the MAE pattern. Specifically, the RGB mode uses MSE as the loss function, while the depth map and semantic segmentation graph use mean absolute error and cross-entropy loss, respectively. Additionally, for training of the RGB mode on the ImageNet-1K dataset, MultiMAE utilizes pseudo-labeling. Pseudo-labeling is a technique where a small portion of labeled samples is used to predict the labels of unlabeled samples, significantly reducing the workload.

M3PT [72] takes an opposite approach to MultiMAE in terms of masking. In M3PT, different modes of the image (depth map and RGB image) share the same mask. The authors of M3PT argue that if the masks for the two modes are different, it would lead to information leakage. The purpose of the MIM method is to disrupt a portion of the input and then predict it. If there is information leakage, it is not truly masking the image. Additionally, to better adapt the model to downstream tasks, M3PT follows a similar training process for both pretraining and fine-tuning stages, instead of discarding the decoder during the fine-tuning stage as done in MAE.

res-MAE [17] and Zekai Chen [73] both employ the MAE method and make improvements for multimodal tasks in the medical field, specifically for CT and MRI. Chen's approach is relatively simple, directly applying MAE to 3D images without significant modifications to the training process. While the training process of res-MAE consists of a pretraining stage and a multimodal fusion stage. The former adopts the idea of MAE but utilizes ResNet18 as the encoder and a combination of convolution and upsampling as the decoder. The masking strategy and loss function remain consistent with MAE. In the latter stage, a ViT model is used for multimodal fusion.

2) IMAGE-TEXT MULTIMODAL

Currently, the research approaches for image-text multimodal tasks are similar to those for image multimodal tasks. One approach is to convert both images and text into sequences that can be processed together, while the other approach is to separately encode and decode images and text.

Geng et al. [74] propose a method that transforms text and images into sequences of the same dimensionality for joint processing. However, they treat images and text differently

TABLE 3. Comparisons of multimodal MAE improvements. Due to the different nature of the targeted multimodal tasks, it is not possible to directly compare the performance of models. Here, we list the datasets used in different works. Additionally, for multimodal tasks, model improvements are focused on the input and output parts. Therefore, a comparison of the loss functions used is also provided here.

Method	Masking strategy	Encoder	Decoder	Loss function	Dataset
MultiMAE [71]	random masking for each individual input	ViT	a single cross -attention layer and MLP, followed by two Transformer blocks	MSE(RGB image) /L1 loss(depth map) / the cross-entropy loss (semantic segmentation graph)	ImageNet-1K, Omnidata, COCO
M3PT [72]	shared random masking for different inputs	ViT	Transformer	MSE	Matterport3D, Stanford2D3D, 3D60
res-MAE [17]	random masking (mixed images as inputs)	ResNet18	upsampling -convolution -upsampling	MSE	self-made dataset (CT, MRI)
Geng et al. [74]	random masking for each individual input	ViT	Transformer	MSE(image) /the cross entropy loss(text)	Conceptual 12M, CIFAR-100, CIFAR-10
Chen et al. [75]	random masking for each individual input	ViT(image) /BERT(text)	Transformer(image) /MLP(text)	MSE	medicalimage-text datasets, ROCO, MedICaT
OmniMAE [77]	tube masking	ViT	Transformer	l2 distance	ImageNet-1K, SSv2
CAV-MAE [16]	random masking	Transformer	Transformer	MSE	AudioSet-20K, AudioSet-2M, VGGSound

in terms of masking and loss functions. The advantage of this approach is that it requires minimal modifications to the encoder and decoder, which are the main components of the training process. For image masking and loss functions, they adopt the same approach as MAE, masking 75% of the image and using MSE as the loss function. For text masking, they follow the BERT approach, masking 15% of the text and using cross-entropy loss as the loss function.

Chen et al. [75] focus on medical images and text. Instead of simply improving MAE, their approach is more like a combination of MAE and BERT. They mask and process images and text separately, using different encoders and decoders. The fusion module, after the encoders, combines the processing tasks for images and text. Specifically, the encoders for images and text are similar to those in MAE and BERT, respectively, while the decoders adopt ViT and MLP (Multi-Layer Perceptron) architectures. There are also works [76] in recent years that have replaced ViT with MLP in image recognition tasks. The fusion module employs a collaborative attention mechanism composed of two Transformer layers. Each Transformer layer includes self-attention, cross-attention, and feed-forward layers.

3) IMAGE-VIDEO, AND IMAGE-AUDIO MULTIMODAL

OmniMAE and CAV-MAE are two methods that improve MAE for image-video and image-audio multimodal tasks, respectively. Both methods involve masking videos or audio along with images for processing. OmniMAE treats images as a special type of video, while CAV-MAE masks the spectrogram of the audio. In terms of the training process, OmniMAE has minimal modifications to the training process of MAE, while CAV-MAE utilizes three encoders.

OmniMAE [77] focuses on input processing and masking strategies. It represents images or videos as four-dimensional tensors with dimensions T×H×W×3, where T represents the time dimension, H and W represent the spatial dimensions, and 3 represents the color channels. An image can be seen as a single-frame video with T=1. In terms of masking strategy, OmniMAE defaults to using random masking for images (90%) and videos (95%). These high masking rates are closer to what MAE uses for videos [78]. Additionally, the authors compare different masking strategies and find that for videos, tube masking performs slightly better than random masking. Random masking for videos is random for each frame, while tube masking applies the same mask to each frame. Considering that the video clips in the SSv2 dataset used by OmniMAE are relatively short, tube masking for videos may be closer to random masking for images.

In contrast, CAV-MAE [16] introduces more significant modifications to the training process and incorporates contrastive learning. In terms of masking strategy, CAV-MAE masks the spectrogram of the audio, allowing it to leverage the masking method used by MAE for images. In practical experiments, the best performance is achieved when masking 50% of the audio, but the improvement compared to the original MAE masking rate of 75% is not substantial. In the training process, although CAV-MAE incorporates three encoders for audio encoding, image encoding, and shared encoding, these three encoders are actually Transformers with different numbers of layers. Afterward, CAV-MAE uses contrastive loss to calculate the results obtained from the audio and image encoders. Contrastive loss is commonly used in contrastive learning to handle paired data. Finally, the results are fed into the decoder and the mean squared error (MSE) loss function is computed.

C. VIDEO

At present, there is not much research on extending the MAE method to videos. The current methods for migrating the MAE method from images to videos start with masking strategies. In terms of results, the masking proportion for MAE on videos is higher, at 90% or above. This is consistent with the fact that videos generally contain more redundant information compared to images. However, it is important

TABLE 4. Comparisons of MAE for videos. Due to the lack of a large-scale dataset for video, similar to ImageNet-1K for images, there are variations in the datasets used. We list the datasets here. However, because of the differences in targeted tasks and training datasets, direct comparisons of model performance are not feasible.

Method	Masking strategy	Encoder	Decoder	Dataset
Feichtenhofer et al. [79]	spacetime-agnostic sampling	vanilla ViT	vanilla ViT	Kinetics-400, AVA, Something-Something v2(SSv2)
VideoMAE [78]	tube masking	vanilla ViT	vanilla ViT	Kinetics-400, AVA, Something-Something V2(SSv2), UCF101, HMDB51
AdaMAE [80]	an adaptive masking strategy	ViT	Transformers	Kinetics-400, Something-Something V2(SSv2)
VideoMAE v2 [81]	a dual masking scheme	vanilla ViT	vanilla ViT	Kinetics-400, Something-Something V2(SSv2), WebVid2Mt

to note that there are two different improvement approaches for MAE on videos. The FAIR team from Meta AI uses a masking strategy that is independent of temporal and spatial information [79], while VideoMAE [78] and its extended research use masking strategies that incorporate temporal and spatial information, even introducing more semantic information. Table 4 shows comparisons of MAE for videos.

Feichtenhofer et al. [79] argue that there is higher redundancy and stronger continuity of information in both spatial and temporal dimensions. Therefore, if temporal information is present in the masking process, it can lead to less preserved information and affect the subsequent reconstruction results. Thus, in the research by the FAIR team, particular attention is paid to using masking strategies that do not involve temporal and spatial information. The encoders and decoders of MAE remain largely unchanged in their research. In contrast, VideoMAE [78] uses tube masking, where the masked blocks have temporal correlations. VideoMAE believes that this approach enables learning of the spatiotemporal structure of videos. Essentially, although the strategies used in these two approaches are different, they both aim to preserve more temporal and spatial information. In terms of results, the model from the FAIR team generally outperforms VideoMAE.

AdaMAE [80] improves upon VideoMAE by incorporating semantic information. AdaMAE samples visible tokens based on semantic context to mask 95% of the tokens, aiming for lower memory requirements and faster pre-training. Specifically, AdaMAE adds an independent adaptive token sampler to the existing tokenizer, encoder, and decoder of MAE. The output of the tokenizer is inputted into a multi-head attention network (MHA) and activated by softmax. This assigns scores to different tokens. According to the scores, AdaMAE keeps tokens that contribute to the reconstruction of the image while masking the rest. As for the results, AdaMAE achieves a classification accuracy of 70.0% on the SSv2 dataset and 81.7% on the Kinetics-400 dataset, surpassing previous methods for video clip classification.

VideoMAE v2 [81] focuses on improving VideoMAE for large-scale tasks. The main improvements can be summarized in terms of masking strategies and the pre-training process. In terms of masking strategies, inspired by [61], VideoMAE v2 adds an additional masking step in the decoder to alleviate the training burden. Therefore, the masking strategy used here is referred to as dual masking, where tube masking is applied in the encoder and running cell masking is used in the decoder. In terms of the pre-training process, self-supervised pre-training is conducted on unlabeled datasets, followed by supervised post-pre-training fine-tuning on labeled datasets. This maximizes the utilization of both labeled and unlabeled data. In the experiments, the authors also include Instagram data and a video dataset called WebVid2M [82] obtained from website scraping.

In addition, there are also works that apply MAE to video object tracking and segmentation (VOT and VOS) data. DropMAE [83] suggests adaptively removing intra-frame clues during the decoding process to facilitate better learning of inter-frame clues.

IV. APPLICATIONS

The applications of MAE in computer vision mainly focuses on different types of images, including medical images, natural images, and geographical images, as well as video surveillance and the 3D domain. Related applications are shown in Figure 3.

A. MEDICAL

In the medical field, MAE is primarily applied in the domain of disease image classification. There are also studies that involve MAE in the areas of image segmentation and cross-modality tasks.

1) PATHOLOGICAL IMAGE CLASSIFICATION

In the case of image classification, the objects primarily include professional charts and scans such as electrocardiograms (ECG) and electroencephalograms (EEG), as well as medical images such as tissue slices. Medical images are characterized by a large workload for annotating data and various types of noise and individual variations. MAE, as a form of self-supervised learning, can significantly reduce the workload of annotating data. By masking parts of the image and then reconstructing it, MAE can effectively capture the main characteristics of the learning object, reduce the impact of noise, and thus increase the robustness and transferability of the model.

ECG and EEG both contain temporal and spatial information. On one hand, they represent the activity of organs over a period of time, indicating their temporal nature. On the other hand, ECG and EEG describe the state of organ health, indicating their spatial nature. Therefore, MaeFE [84] (2022.12) focuses on different mask patterns specifically designed for ECG, regarding their temporal and spatial characteristics. MV-SSTMA [85] designs a multi-view convolution-transformer hybrid structure for

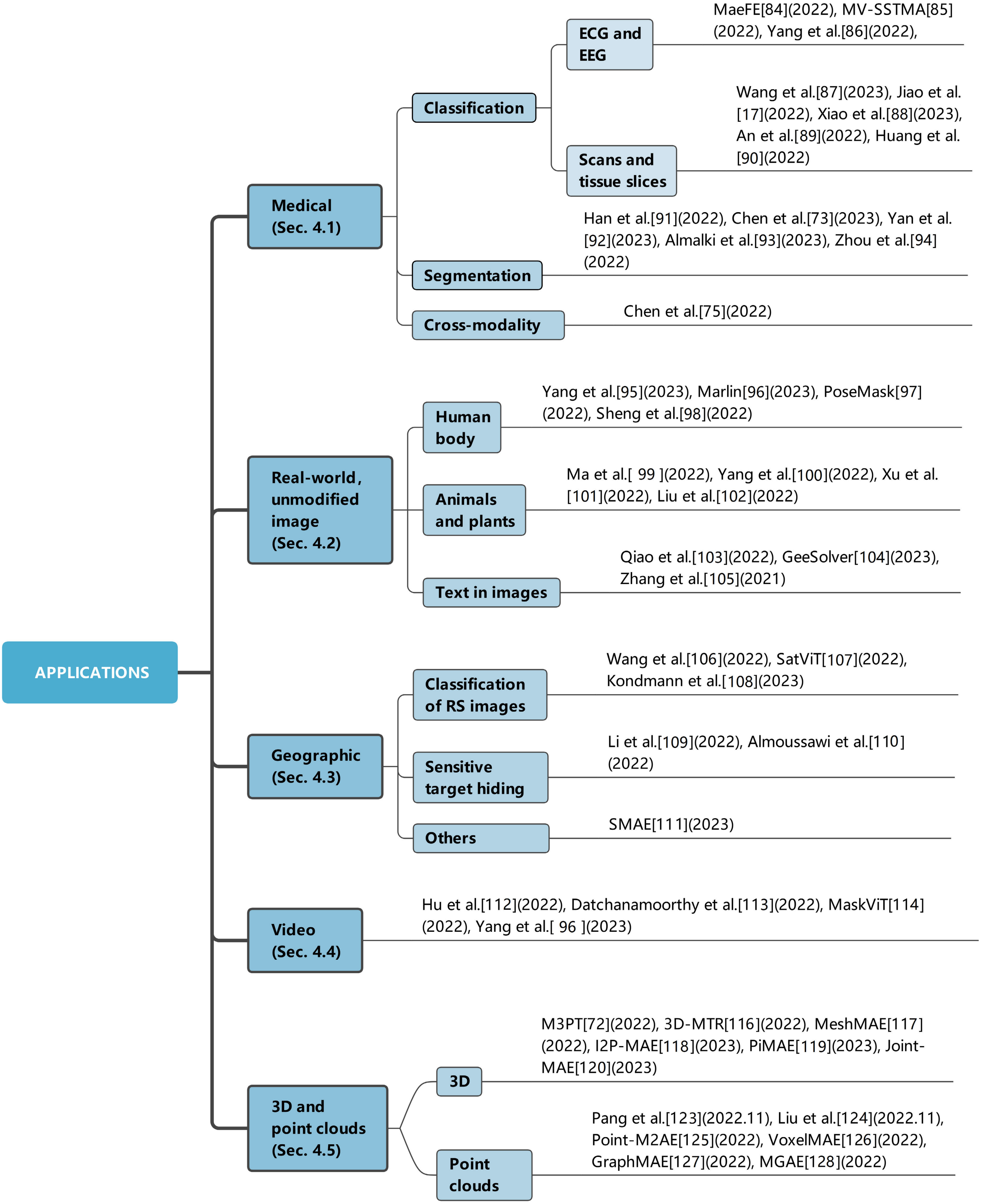
APPLICATIONS
Medical (Sec. 4.1)
Classification
ECG and EEG
MaeFE[84](2022), MV-SSTMA[85](2022), Yang et al.[86](2022),
Scans and tissue slices
Wang et al.[87](2023), Jiao et al.[17](2022), Xiao et al.[88](2023), An et al.[89](2022), Huang et al.[90](2022)
Segmentation
Han et al.[91](2022), Chen et al.[73](2023), Yan et al.[92](2023), Almalki et al.[93](2023), Zhou et al.[94](2022)
Cross-modality
Chen et al.[75](2022)
Real-world, unmodified image (Sec. 4.2)
Human body
Yang et al.[95](2023), Marlin[96](2023), PoseMask[97](2022), Sheng et al.[98](2022)
Animals and plants
Ma et al.[99](2022), Yang et al.[100](2022), Xu et al.[101](2022), Liu et al.[102](2022)
Text in images
Qiao et al.[103](2022), GeeSolver[104](2023), Zhang et al.[105](2021)
Geographic (Sec. 4.3)
Classification of RS images
Wang et al.[106](2022), SatViT[107](2022), Kondmann et al.[108](2023)
Sensitive target hiding
Li et al.[109](2022), Almoussawi et al.[110](2022)
Others
SMAE[111](2023)
Video (Sec. 4.4)
Hu et al.[112](2022), Datchanamoorthy et al.[113](2022), MaskViT[114](2022), Yang et al.[96](2023)
3D and point clouds (Sec. 4.5)
3D
M3PT[72](2022), 3D-MTR[116](2022), MeshMAE[117](2022), I2P-MAE[118](2023), PiMAE[119](2023), Joint-MAE[120](2023)
Point clouds
Pang et al.[123](2022.11), Liu et al.[124](2022.11), Point-M2AE[125](2022), VoxelMAE[126](2022), GraphMAE[127](2022), MGAE[128](2022)

FIGURE 3. Applications of MAE in Various Fields. The selected articles in this figure mainly focus on papers that have been published or included in journals and conferences.

EEG, interpreting the emotion-related information of the EEG signals from the perspectives of spectrum, space, and time. Additionally, Yang et al. [86] extract both local and global features from ECG to capture the key information, overcoming the scarcity of labeled ECG samples and achieving higher performance than existing state-of-the-art self-supervised models. Similarly, the training strategy of MV-SSTMA also adopts a phased processing approach. The overall samples are first randomly masked and learned, followed by learning specific samples. The resulting models also achieve the highest performance in their respective fields.

MAE achieves good results in the classification of medical images such as scans and tissue slices. Wang et al. [87] apply MAE to self-supervised classification of cervical cancer OCT images, and the proposed classification model demonstrated better transfer learning capability and comparable performance to medical experts. Jiao et al. [17] use MAE for self-supervised learning of spinal tumor CT and MRI scan images, significantly improving the classification accuracy for complex tumor subtypes. The team from Johns Hopkins University [88] compares the classification performance of CNN and MAE on different imaging scan datasets for chest diseases and conclude that MAE performs better when used for pre-training. An et al. [89] apply MAE to the classification of breast cancer tissue slice images, combining it with multiple instance learning (MIL), and achieve performance superior to the state-of-the-art method CLAM in this field. In addition, Huang et al. [90] use MAE for feature extraction in medical images and develop an evaluation mechanism called EXAMINE for extracting categories. SwinMAE [15], specifically designed for medical small datasets such as BTCV CT images, replaces the backbone with Swin Transformer, and achieves good results.

2) ORGAN IMAGE SEGMENTATION

In terms of segmentation, MAE's performance cannot be considered excellent, but it demonstrates considerable potential. Methods like MAE, which utilize masked image modeling (MIM), have opened up new possibilities for medical image segmentation. In terms of MAE's strengths, a team from Dongguan University of Technology [91] finds that MAE performs well in hybrid pre-training for medical segmentation tasks. Chen et al. [73] from pharmaceutical company Bristol Myers Squibb use MAE for image segmentation in abdominal CT, brain tumor MRI, and COVID-19 CT scans, resulting in robust models with accelerated training speed and reduced costs.

In terms of MAE's shortcomings, Yan et al. [92] point out that conventional MAE performs better in recovering coarse high-level semantic information but may struggle with detailed low-level information. Therefore, for tasks like multi-organ segmentation, using MAE directly may not yield ideal results. Similarly, Almalki and Latecki [93] compare the performance of UM-MAE and SimMIM in tooth numbering and X-ray image segmentation, and find that the image segmentation performance of the pre-trained model based on pyramid Vision Transformers using the uniform masking strategy in MAE is inferior to SimMIM, which also belongs to the masked image modeling approach. To address this issue, Yan proposes using convolutional encoders to extract low-level semantic information to complement MAE's ability to extract dense downstream information.

It is worth mentioning that Zhou et al. [94] once again demonstrates the outstanding performance of MAE in medical image classification and segmentation.

3) MEDICAL CROSS-MODALITY

Text and images are two fundamental forms of expression in medical knowledge. Chen et al. [75] propose a method to extract general knowledge from medical text and images by reconstructing missing pixels and labels from randomly masked images and text. And the results can then be applied to various medical visual and language tasks. Different decoders and masking ratios are designed for text and images respectively. The resulting model achieves better performance compared to other cross-modal studies in the medical domain.

B. REAL-WORLD AND UNMODIFIED IMAGES

Talking about real-world images and unaltered images, the applications can be classified into three categories: human body, animals/plants, and text.

For applications related to the human body, MAE is primarily used in graph-related techniques. Yang et al. [95] address the problem of face anti-spoofing (FAS), distinguishing between real and fake faces, by defining a facial region as a point. This transforms the FAS problem from a binary classification task into a graph classification problem. Yang simplifies the complex graph relationship information using MAE. Marlin et al. [96] also employ MAE for learning facial features, with a focus on facial details and facial features. PoseMask [97] applies MAE for pose estimation in classroom scenarios, using heatmaps as reference masks to estimate poses in crowded or occluded scenes. Furthermore, Sheng et al. [98] use MAE for feature extraction of gestures in Spatial-Temporal Motion Maps (STMM), improving gesture recognition accuracy. These applications utilize the ability of MAE to reconstruct images for graph-related tasks.

In the case of animals and plants, MAE is mainly used to handle small datasets consisting of real nature images. In chicken face detection, Ma et al. [99] employ MAE to generate more samples, thereby enhancing the dataset. Yang et al. [100] achieve second place in the SnakeCLEF 2022 fine-grained snake dataset classification competition using a pre-trained MAE model. These datasets are characterized by a scarcity of data for individual categories. Similarly, MAE is used in the PlantCLEF2022 classification competition [101]. The PlantCLEF2022 dataset contains millions of plant photos, but on average, each class has only 36 images, leading to a scarcity in samples. Additionally, MAE is applied for the classification of grape powdery mildew [102], effectively addressing the challenge of limited labeled data.

Regarding text in images, Qiao et al. use MAE for text recognition [103] to tackle the difficulty of extracting fine-grained features. In GeeSolver [104], MAE is employed to improve captcha solvers and reduce manual labeling efforts. Zhang et al. [105] use MAE to restore distorted backgrounds after modifying text in images.

C. GEOGRAPHIC AND REMOTE SENSING (RS)

In summary, MAE is primarily used for classification of remote sensing (RS) images and pre-training on a large amount of unlabeled data. Additionally, MAE's generation capability is also utilized for hiding sensitive targets. In the context of RS image classification, a team from Tsinghua University [106] performs pre-training on unlabeled data from the Hainan dataset acquired by the polarimetric synthetic aperture radar (PolSAR) remote sensing system and the flight dataset collected by NASA/JPL AIRSAR using MAE. This approach significantly reduces the workload while maintaining good model performance. SatViT [107] utilizes MAE for pre-training on unlabeled remote sensing images obtained from Sentinel-1 and 2. After fine-tuning, it outperforms existing state-of-the-art methods in downstream tasks such as peatland classification and land cover classification. In addition to specific terrain and landform classification tasks, Kondmann et al. [108] use MAE for change detection in remote sensing. They pretrain their model on the RapidAI4EO corpus dataset, and the MAE model performs the best in remote sensing change detection compared to conventional models.

Regarding the hiding of sensitive targets in RS images, Li and Bai [109] use MAE pretrained on the ImageNet-1K dataset to restore images with targets already masked, resulting in natural-looking images without sensitive targets. The advantage of MAE is that it has faster parameter tuning time and more stable models compared to GANs for image generation. Additionally, Almoussawi et al. [110] use MAE for correctness detection in the classification of real-world images of fires, achieving an F1-score of above 0.9.

Furthermore, [111] employs MAE for the reconstruction of non-saturated HDR images, optimizing the HDR effect.

D. VIDEO PREDICTION AND SURVEILLANCE

The application of MAE in videos mainly utilizes its scalability and focuses on detecting abnormal situations in surveillance. Since MAE generates logically consistent videos, its predictions for videos with anomalies can be particularly poor. This allows for better identification of videos with abnormal situations.

Reference [112] applies MAE to unsupervised video anomaly detection (UVAD), which aims to identify abnormal events from completely unlabeled videos. They use spatiotemporal cubes (STCs) to represent video events, which are constructed from temporally contiguous foreground patches of unlabeled videos. Then, half of the patches in the STC are masked along the temporal dimension, and a ViT is trained to predict the masked patches using the unmasked patches. [113] further uses MAE for recognizing anomalous human activities.

In addition, MaskViT [114] utilizes two types of window attention: spatial attention and spatiotemporal attention, and designs tokens with variable percentages to improve video prediction.

Reference [95] applies MAE to deepfake detection. The purpose of deepfake detection is to distinguish between forged faces and real faces. Deepfake detection [115] is defined as a graph classification problem, where each facial region corresponds to a vertex. However, the presence of redundant relational information hinders the expressiveness of the graph. Inspired by the success of mask modeling, mask relation learning is chosen to reduce the redundancy of learning information relationship features.

E. 3D AND POINT CLOUDS

1) 3D IMAGE

In addition to the multimodal scenario mentioned earlier [72], MAE has other applications in the 3D domain.

In 3D-MTR [116], MAE is used to process input 2D images to enhance 3D reconstruction. 3D-MTR consists of three parts: a 3D reconstruction network, MAE, and a CNN-based inpainting network. The inputs for the first two parts are 2D images, and the inpainting network combines the outputs of the previous two networks to obtain the reconstructed 3D image.

MeshMAE [117] focuses on processing 3D mesh data. The research here is mainly on how to handle meshes and utilize MAE. The mesh is divided into non-overlapping local patches, each containing the same number of faces, and the 3D positions of the center points of each patch are used to form position embeddings. Then, the MAE method is applied, randomly masking some patches in the mesh, and the damaged mesh is input into a Transformer to reconstruct the information of masked patches, allowing the network to learn discriminative representations of the mesh data.

I2P-MAE [118] addresses the lack of 3D data in datasets and obtains 3D representations from pre-trained 2D models. In terms of improvements to MAE, this approach deliberately ensures that important point labels are not masked in the masking strategy, rather than completely random masking.

Additionally, PiMAE [119] designs a dual-branch MAE to facilitate 3D and 2D interactions. Joint-MAE [120] is similar but uses joint encoders and joint decoders.

2) POINT CLOUDS

Point clouds are an important component of computer vision, representing datasets of objects or spaces. In addition to MAE, there have been other MIM methods applied to point clouds in recent years, such as [121] and [122].

Currently, the application of MAE in point clouds mainly focuses on the processing of inputs. Reference [123] divides the input point cloud into irregular patches and directly applies the MAE method. Reference [124] improves the discriminative capability of MAE for point clouds by representing them as discrete occupancy values (1 if a point belongs to the point cloud, 0 otherwise) and performing simple binary classification between masked object points and sampled noise points as a proxy task. This approach increases robustness and enriches the learned representation.

In terms of specific task improvements, Point-M2AE [125] improves the learning of irregular point clouds 3D representation by modifying the encoder and decoder into a pyramid architecture, capturing fine-grained and high-level semantic information of 3D shapes. VoxelMAE [126] addresses the sparse density of points and large variations in the same scene in autonomous driving point clouds. It specifically designs the discrimination between empty and non-empty points, similar to [124].

GraphMAE [127] combines MAE with graph learning, using scaled cosine error as the loss function instead of MSE.

Similar work includes MGAE [128], where [128] masks edges and reconstructs important edges instead of points.

V. DISCUSSION

A. IMPROVEMENT DIRECTIONS

Apart from improving accuracy, there are three noteworthy directions for performance improvement in MAE.

Currently, the training phase of MAE requires a large number of training epochs, around 1600 epochs. This not only prolongs the model training speed but also poses certain requirements for the computational resources. Some researches have shown limited improvement in model performance but significantly reduces the number of epochs required. According to Table 2 of Section III, it is particularly notable that SdAE [35] and SupMAE [70] have significantly reduced batch sizes.

Furthermore, enhancing the robustness of the model using the characteristics of MAE as an MIM method is another research direction. In the works adding noise to the training process [22], [51] has shown improvements in model robustness. Lastly, in terms of improving model performance, CMAE [53] has demonstrated the most significant enhancement, but it does not show significant improvement in reducing the number of epochs.

Lastly, expanding the application scope of MAE is indeed a direction worth exploring. Although MAE was initially used for image processing, particularly image classification [1], it has been extended to other tasks such as object detection, image segmentation (e.g., DropMAE [83]), as well as applications in video and multimodal tasks. One of the challenges in multimodal tasks is the lack of large-scale datasets that are applicable to different downstream tasks, similar to imagenet for image classification. This presents an opportunity for further research to focus on creating more diverse and comprehensive multimodal datasets that can effectively support various types of tasks.

It should be noted that in terms of model performance improvement, CMAE [53] has demonstrated the most significant enhancement. However, it does not show significant improvement in reducing the number of epochs.

B. APPLICATIONS

In terms of application, based on the summary of MAE in different fields in the previous section (Section IV), here we categorize the aspects of the role of MAE in practical applications into four categories.

Based on pre-training with larger datasets, MAE can achieve better fine-tuning results in 1) situations where downstream task datasets are smaller. MAE is also used in 2) situations to reduce data redundancy or extract features. Additionally, many works utilize MAE for: 3) its scalability and 4) as a self-supervised method to reduce the labeling workload.

In scenarios where the downstream task dataset is small, some works [84], [86], [98] leverage the good performance of MAE in small data scenarios, while others use the prediction process of MAE to augment the dataset. In the latter case, small datasets often come with imbalanced data, and some works [14] use MAE to simulate abnormal situations. Moreover, fine-grained classification tasks [100], [101] suffer from insufficient sample data. Reference [129] extensively discusses the use of MAE for data augmentation by generating input images.

As a denoising autoencoder, MAE reconstructs inputs based on partially corrupted data, thereby reducing dataset redundancy and extracting features. In addition to the works mentioned in Section IV [85], [95], this characteristic of MAE is also applied in industrial tasks [3].

In the original paper by He [1], the scalability of MAE is emphasized. This means that even if the reconstructed results differ from the original, they still possess a certain level of coherence and can connect with contextual information. Therefore, MAE has significant applications in image generation [97], [103], [109], and also performs well in tasks that require temporal coherence [4], [132].

Lastly, since MAE is a self-supervised method, its applications [84], [86], [87], [98] are often mentioned for greatly reducing the labeling workload. Overall, the contributions of MAE are particularly prominent in imperfect datasets with high workload. Compared to other pretraining-finetuning methods (mainly contrastive learning) that rely on a large number of negative sample pairs and data augmentation, MAE has significant advantages.

C. IMPROVEMENT STRATEGIES

In terms of improvement strategies, the improvements of MAE can be roughly categorized into: 1) modifying the training process, and 2) combining with other training methods. (Regarding modifying the training process, adjusting the masking strategy and loss function are relatively straightforward in terms of implementation, and many works [20], [21], [35] have focused on this aspect. In general, the selection of masking strategies and loss functions refers to other MIM methods such as BEiT and SimMIM. For example, the masking strategy of BEiT is used in [20] and [62]. Improvements to the encoder and decoder [62], [63] involve changes to the backbone architecture.

Regarding combining with other training methods, one more prominent approach is to improve the overall structure, while another approach is to integrate similar methods. As for the first approach, there are currently works that combine MAE with contrastive learning and convolution. Sections III-A2-III-A4 provide detailed introductions to the currently published research. Overall, the former is essentially the structure of contrastive learning with the masking process, while the latter focuses on improving the encoder. In fact, these combined methods are similar to MAE, targeting similar tasks but with different approaches to pre-training. Following this line of thinking, other self-supervised learning techniques such as knowledge distillation also have the potential to be combined and improved with MAE.

In terms of referencing similar methods, in addition to MIM methods, considering other autoencoder methods is also an idea. On the basis of other improvements, there is also room for adjustments specific to a particular downstream task. In fact, most of the MAE applications discussed in Section IV involve steps for adapting to specific downstream tasks.

VI. CONCLUSION

Since the publication of the original MAE paper, applications based on MAE have been widely seen in various journals and international conferences, along with research on improving MAE and combining it with other pre-training and self-supervised learning methods. In terms of applications, MAE has been extensively used in medical, natural image, and geographic remote sensing image domains. In fact, MAE has also been applied and extended to other fields such as audio [67], [130], [131], and dealing with machine malfunctions. Additionally, MAE has garnered more attention in the field of self-supervised training, contributing to the popularity of MIM methods in recent years.

MAE achieves good results while maintaining a relatively simple and non-redundant structure, which is significant in the context of increasingly large datasets. Through this survey, we have summarized the contributions and developments of MAE and discussed potential directions for future research, hoping to provide insights for practitioners in the field.

REFERENCES

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[2] A. Vaswani, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 5999–6009.

[3] W. Yu, M. Huang, S. Wu, and Y. Zhang, "Ensembled masked graph autoencoders for link anomaly detection in a road network considering spatiotemporal features," *Inf. Sci.*, vol. 622, pp. 456–475, Apr. 2023.

[4] M. Duan, W. Liu, R. Liu, L. Wang, L. Mao, Q. Qiu, and G. Ling, "Intercity railway risk space anomaly detection based on train predeparture key frame extraction and IADN network," *IEEE Sensors J.*, vol. 23, no. 3, pp. 1693–1706, Feb. 2023.

[5] H. Guo, H. Zhu, J. Wang, P. Vadakkepat, W. K. Ho, and T. H. Lee, "Masked self-supervision for remaining useful lifetime prediction in machine tools," in *Proc. IEEE 20th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2022, pp. 353–358.

[6] Y. Min and Y. Li, "Self-supervised railway surface defect detection with defect removal variational autoencoders," *Energies*, vol. 15, no. 10, pp. 35–92, May 2022.

[7] S. Xie, R. Liu, L. Du, and H. Tan, "Anomaly detection in rolling bearings based on the Mel-frequency cepstrum coefficient and masked autoencoder for distribution estimation," *Struct. Control Health Monitor.*, vol. 29, no. 11, Nov. 2022, Art. no. e3096.

[8] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," 2022, *arXiv:2208.00173*.

[9] A. Dosovitskiy, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[10] L. Kong, M. Q. Ma, G. Chen, E. P. Xing, Y. Chi, L.-P. Morency, and K. Zhang, "Understanding masked autoencoders via hierarchical latent variable models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7918–7928.

[11] S. Cao, P. Xu, and D. A. Clifton, "How to understand masked autoencoders," 2022, *arXiv:2202.03670*.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[13] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2022, pp. 280–296.

[14] J. Jiang, J. Zhu, M. Bilal, Y. Cui, N. Kumar, R. Dou, F. Su, and X. Xu, "Masked Swin transformer UNet for industrial anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2200–2209, Feb. 2023.

[15] Z. Xu, Y. Dai, F. Liu, W. Chen, Y. Liu, L. Shi, S. Liu, and Y. Zhou, "Swin MAE: Masked autoencoders for small datasets," *Comput. Biol. Med.*, vol. 161, Jul. 2023, Art. no. 107037.

[16] C. Li, "CoTMAE: Hybrid convolution-transformer pyramid network meets masked autoencoder," in *Proc. 9th Int. Conf. Asian Soc. Precis. Engg. Nanotechnol. (ASPEN)*, Nov. 2022, pp. 1–7. [Online]. Available: https://rpsonline.com.sg/proceedings/aspen2022/pdf/OR-08-0105.pdf

[17] M. Jiao, "Self-supervised learning based on a pre-trained method for the subtype classification of spinal tumors," in *Proc. Comput. Math. Modeling Cancer Anal. (CMMCA)*, Sep. 2022, pp. 58–67.

[18] Y. Gong, "Contrastive audio-visual masked autoencoder," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2023. [Online]. Available: https://arxiv.org/abs/2210.07839 and https://openreview.net/forum?id=QPtMRyk5rb

[19] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9643–9653.

[20] X. Dong, "Bootstrapped masked autoencoders for vision BERT pretraining," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2022, pp. 247–264.

[21] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng, "SemMAE: Semantic-guided masking for learning masked autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Nov. 2022, pp. 14290–14302.

[22] Q. Wu, H. Ye, Y. Gu, H. Zhang, L. Wang, and D. He, "Denoising masked autoencoders help robust classification," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2023. [Online]. Available: https://openreview.net/forum?id=zDjtZZBZtqK

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[25] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5843–5851.

[26] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[27] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[28] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[29] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3557–3567.

[30] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[31] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 721–725.

[32] J. Quinn, "Computer vision," in *Dive Into Deep Learning: Tools for Engagement*. Thousand Oaks, CA, USA: Corwin Press, 2020, p. 551.

[33] A. Kumar, "Fine-tuning can distort pretrained features and underperform out-of-distribution," presented at the ICLR, 2022.

[34] C. E. Metz, "Basic principles of ROC analysis," *Seminars Nucl. Med.*, vol. 8, no. 4, pp. 283–298, Oct. 1978.

[35] Y. Chen, "SDAE: Self-distillated masked autoencoder," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2022, pp. 108–124.

[36] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[38] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Nov. 2020.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. IEEE Int. Conf. (ICML)*, Jul. 2020, pp. 1597–1607.

[40] YH. Cao, H. Yu, and J. Wu, "Training vision transformers with only 2040 images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 220–237.

[41] J. B. Grill, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Nov. 2020, pp. 21271–21284.

[42] H. Touvron, "Training data-efficient image transformers and distillation through attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 10347–10357.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[44] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[45] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14648–14658.

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[47] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[48] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "IBOT: Image BERT pre-training with online tokenizer," 2021, *arXiv:2111.07832*.

[49] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," 2016, *arXiv:1608.05442*.

[50] N. Carlini, F. Tramer, K. Dj Dvijotham, L. Rice, M. Sun, and J. Zico Kolter, "(certified!!) adversarial robustness for free!" 2022, *arXiv:2206.10550*.

[51] H. Yang, "Domain invariant masked autoencoders for self-supervised learning from multi-domains," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 151–168.

[52] F. Klinker, "Exponential moving average versus moving exponential average," *Mathematische Semesterberichte*, vol. 58, no. 1, pp. 97–107, Apr. 2011.

[53] Y. Lee, "Exploring the role of mean teachers in self-supervised masked auto-encoders," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2023. [Online]. Available: https://openreview.net/pdf?id=7sn6Vxp92xV and https://arxiv.org/abs/2210.02077

[54] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," 2022, *arXiv:2207.13532*.

[55] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[56] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[57] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[58] X. Liu, J. Zhou, T. Kong, X. Lin, and R. Ji, "Exploring target representations for masked autoencoders," 2022, *arXiv:2209.03917*.

[59] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[60] Y. Bai, Z. Wang, J. Xiao, C. Wei, H. Wang, A. Yuille, Y. Zhou, and C. Xie, "Masked autoencoders enable efficient knowledge distillers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24256–24265.

[61] Z. Qing, S. Zhang, Z. Huang, X. Wang, Y. Wang, Y. Lv, C. Gao, and N. Sang, "MAR: Masked autoencoders for efficient action recognition," *IEEE Trans. Multimedia*, early access, Mar. 30, 2023, doi: 10.1109/TMM.2023.3263288.

[62] P. Gao, "MCMAE: Masked convolution meets masked autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Nov. 2022, pp. 35632–35644.

[63] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16133–16142.

[64] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[65] A. G. Howard, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[66] Z. Fei, M. Fan, L. Zhu, J. Huang, X. Wei, and X. Wei, "Masked autoencoders meet generative adversarial networks and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24449–24459.

[67] A. A. Attia and C. Y. Espy-Wilson, "Masked autoencoders are articulatory learners," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2023, pp. 1–5.

[68] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality," 2022, *arXiv:2205.10063*.

[69] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Green hierarchical vision transformer for masked image modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 19997–20010.

[70] F. Liang, Y. Li, and D. Marculescu, "Supmae: Supervised masked autoencoders are efficient vision learners," 2022, *arXiv:2205.14540*.

[71] R. Bachmann, "Multimae: Multi-modal multi-task masked autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 348–367.

[72] Z. Yan, "Multi-modal masked pre-training for monocular panoramic depth completion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 378–395.

[73] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M. M. Balan, and K. Brown, "Masked image modeling advances 3D medical image analysis," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1969–1979.

[74] X. Geng, "Multimodal masked autoencoders learn transferable representations," 2022, *arXiv:2205.14204*.

[75] Z. Chen, "Multi-modal masked autoencoders for Medical vision-and-language pretraining," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2022, pp. 679–689.

[76] I. Tolstikhin, "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[77] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "OmniMAE: Single model masked pretraining on images and videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10406–10417.

[78] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Nov. 2022, pp. 10078–10093.

[79] C. Feichtenhofer, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Nov. 2022, pp. 35946–35958.

[80] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel, "AdaMAE: Adaptive masking for efficient spatiotemporal learning with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14507–14517.

[81] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "VideoMAE v2: Scaling video masked autoencoders with dual masking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14549–14560.

[82] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1708–1718.

[83] Q. Wu, T. Yang, Z. Liu, B. Wu, Y. Shan, and A. B. Chan, "DropMAE: Masked autoencoders with spatial-attention dropout for tracking tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14561–14571.

[84] H. Zhang, W. Liu, J. Shi, S. Chang, H. Wang, J. He, and Q. Huang, "MaeFE: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.

[85] R. Li, Y. Wang, W.-L. Zheng, and B.-L. Lu, "A multi-view spectral–spatial–temporal masked autoencoder for decoding emotions with self-supervised learning," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6–14.

[86] S. Yang, C. Lian, and Z. Zeng, "Masked autoencoder for ECG representation learning," in *Proc. 12th Int. Conf. Inf. Sci. Technol. (ICIST)*, Oct. 2022, pp. 95–98.

[87] Q. Wang, K. Chen, W. Dou, and Y. Ma, "Cross-attention based multi-resolution feature fusion model for self-supervised cervical OCT image classification," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 4, pp. 2541–2554, Jul./Aug. 2023.

[88] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3577–3589.

[89] J. An, Y. Bai, H. Chen, Z. Gao, and G. Litjens, "Masked autoencoders pre-training in multiple instance learning for whole slide image classification," in *Proc. Med. Imag. With Deep Learn. (MIDL)*, Dec. 2022. [Online]. Available: https://openreview.net/forum?id=rV5gzFDn5PF

[90] CY. Huang, Q. Lei, and X. Li, "Efficient medical image assessment via self-supervised learning," in *Proc. MICCAI Workshop Data Augmentation, Labelling, Imperfections*, Sep. 2022, pp. 102–111.

[91] Y. Han, "Hybrid pre-training based on masked autoencoders for medical image segmentation," in *Proc. Natl. Conf. Theor. Comput. Sci. (NCTCS)*, Jul. 2022, pp. 175–182.

[92] X. Yan, J. Naushad, S. Sun, K. Han, H. Tang, D. Kong, H. Ma, C. You, and X. Xie, "Representation recovering for self-supervised pre-training on medical images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2684–2694.

[93] A. Almalki and L. J. Latecki, "Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5583–5592.

[94] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image analysis," 2022, *arXiv:2203.05573*.

[95] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for DeepFake detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1696–1708, 2023.

[96] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat, "MARLIN: Masked autoencoder for facial video representation LearnINg," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1493–1504.

[97] S. Liu, M. Ma, H. Li, H. Ning, and M. Wang, "Pose Mask: A model-based augmentation method for 2Dx pose estimation," *Sensors*, vol. 22, no. 21, p. 8331, Oct. 2022.

[98] Z. Sheng, H. Xu, Q. Zhang, and D. Wang, "Facilitating radar-based gesture recognition with self-supervised learning," in *Proc. 19th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Sep. 2022, pp. 154–162.

[99] X. Ma, X. Lu, Y. Huang, X. Yang, Z. Xu, G. Mo, Y. Ren, and L. Li, "An advanced chicken face detection network based on GAN and MAE," *Animals*, vol. 12, no. 21, p. 3055, Nov. 2022.

[100] L. Yang, X. Li, R. Song, K. Zhu, and G. Li, "Solution for Snake-CLEF 2022 by tackling long-tailed categorization," in *Proc. Working Notes (CLEF)*, Sep. 2022. [Online]. Available: https://ceur-ws.org/Vol-3180/paper-180.pdf

[101] M. Xu, S. Yoon, Y. Jeong, J. Lee, and DS. Park, "Transfer learning with self-supervised vision transformer for large-scale plant identification," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, Sep. 2022, pp. 2253–2261.

[102] E. Liu, K. M. Gold, D. Combs, L. Cadle-Davidson, and Y. Jiang, "Vision transformer with masked autoencoder pretraining for quantification of grape downy mildew," in *Proc. Houston, Texas July 17-20*, 2022, p. 1.

[103] Z. Qiao, Z. Ji, Y. Yuan, and J. Bai, "A vision transformer based scene text recognizer with multi-grained encoding and decoding," in *Proc. Int. Conf. Frontiers. Handwriting Recognit. (ICFHR)*, Nov. 2022, pp. 198–212.

[104] R. Zhao, X. Deng, Y. Wang, Z. Yan, Z. Han, L. Chen, Z. Xue, and Y. Wang, "GeeSolver: A generic, efficient, and effortless solver with self-supervised learning for breaking text captchas," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 1524–1541.

[105] Y. Zhang, "Natural image investigation using masked image encoding for text editing," in *Proc. 3rd Int. Academic Exchange Conf. Sci. Technol. Innov. (IAECST)*, Dec. 2021, pp. 758–762.

[106] H. Wang, C. Xing, J. Yin, and J. Yang, "Land cover classification for polarimetric SAR images based on vision transformer," *Remote Sens.*, vol. 14, no. 18, p. 4656, Sep. 2022.

[107] A. Fuller, K. Millard, and J. R. Green, "SatViT: Pretraining transformers for Earth observation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[108] L. Kondmann, "Geography-aware masked autoencoders for change detection in remote sensing," in *Proc. EGU Gen. Assem. Conf. Abstr.*, Apr. 2023, p. EGU-2843, doi: 10.5194/egusphere-egu23-2843.

[109] P. Li and W. Bai, "Automatic hiding method of sensitive targets in remote sensing images based on transformer structure," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 47, no. 8, pp. 1287–1297, Aug. 2022.

[110] Z. A. Almoussawi, "Fire detection and verification using convolutional neural networks, masked autoencoder and transfer learning," *Majlesi J. Electr. Eng.*, vol. 16, no. 4, pp. 159–166, Dec. 2022.

[111] Q. Yan, S. Zhang, W. Chen, H. Tang, Y. Zhu, J. Sun, L. Van Gool, and Y. Zhang, "SMAE: Few-shot learning for HDR deghosting with saturation-aware masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5775–5784.

[112] J. Hu, G. Yu, S. Wang, E. Zhu, Z. Cai, and X. Zhu, "Detecting anomalous events from unlabeled videos via temporal masked auto-encoding," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

[113] K. Datchanamoorthy, B. Padmavathi, V. V. Prakash, S. Yogesh, and R. S. Sahabudeen, "Anomaly and activity recognition in a video surveillance using masked autoencoder," in *Proc. Int. Conf. Innov. Comput., Intell. Commun. Smart Electr. Syst. (ICSES)*, Jul. 2022, pp. 1–7.

[114] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei, "MaskViT: Masked visual pre-training for video prediction," 2022, *arXiv:2206.11894*.

[115] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[116] T. Gao, J. Qin, H. Xue, L. Xue, and C. Wang, "3D-MTR: 3D reconstruction algorithms for deep understanding of images," in *Proc. 2nd Int. Conf. Comput. Sci., Electron. Inf. Eng. Intell. Control Technol. (CEI)*, Sep. 2022, pp. 621–626.

[117] Y. Liang, S. Zhao, B. Yu, J. Zhang, and F. He, "MeshMAE: Masked autoencoders for 3D mesh data analysis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 37–54.

[118] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3D representations from 2D pre-trained models via image-to-point masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21769–21780.

[119] A. Chen, K. Zhang, R. Zhang, Z. Wang, Y. Lu, Y. Guo, and S. Zhang, "PiMAE: Point cloud and image interactive masked autoencoders for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5291–5301.

[120] Z. Guo, X. Li, and P. A. Heng, "Joint-MAE: 2D-3D joint masked autoencoders for 3D point cloud pre-training," 2023, *arXiv:2302.14007*.

[121] K. Fu, P. Gao, SL. Liu, R. Zhang, Y. Qiao, and W. Manning, "POS-BERT: Point cloud one-stage BERT pre-training," 2022, *arXiv:2204.00989*.

[122] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19291–19300.

[123] Y. Pang, W. Wang, FEH. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 604–621.

[124] H. Liu, M. Cai, and Y. J. Lee, "Masked discrimination for self-supervised learning on point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 657–675.

[125] R. Zhang, "Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Nov. 2022, pp. 27061–27074.

[126] C. Min, X. Xu, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "Occupancy-MAE: Self-supervised pre-training large-scale LiDAR point clouds with masked occupancy autoencoders," 2022, *arXiv:2206.09900*.

[127] Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang, "GraphMAE: Self-supervised masked graph autoencoders," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 594–604.

[128] Q. Tan, N. Liu, X. Huang, R. Chen, S.-H. Choi, and X. Hu, "MGAE: Masked autoencoders for self-supervised learning on graphs," 2022, *arXiv:2201.02534*.

[129] H. Xu, S. Ding, X. Zhang, H. Xiong, and Q. Tian, "Masked autoencoders are robust data augmentors," 2022, *arXiv:2206.04846*.

[130] P. Y. Huang, "Masked autoencoders that listen," in *Proc. Adv. Neural Inf. Process. syst.*, vol. 35, Nov. 2022, pp. 28708–28720.

[131] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," in *Proc. HEAR, Holistic Eval. Audio Represent.*, Dec. 2022, pp. 1–24.

[132] C. Xia, Y. Zhan, Y. Tan, and W. Wu, "Research on forecasting aeroengine vibration signals based on the MAE model," *IEEE Access*, vol. 10, pp. 110676–110688, 2022.

[133] Y. Tang, Z. Huang, Z. Chen, M. Chen, H. Zhou, H. Zhang, and J. Sun, "Novel visual crack width measurement based on backbone double-scale features for improved detection automation," *Eng. Struct.*, vol. 274, pp. 115–158, Jan. 2023.

[134] F. Wu, J. Duan, P. Ai, Z. Chen, Z. Yang, and X. Zou, "Rachis detection and three-dimensional localization of cut off point for vision-based banana robot," *Comput. Electron. Agricult.*, vol. 198, Jul. 2022, Art. no. 107079.

[135] F. Wu, Z. Yang, X. Mo, Z. Wu, W. Tang, J. Duan, and X. Zou, "Detection and counting of banana bunches by integrating deep learning and classic image-processing algorithms," *Comput. Electron. Agricult.*, vol. 209, Jun. 2023, Art. no. 107827.

ZEXIAN ZHOU was born in Shaanxi, China, in 1997. She received the B.Phil. degree in philosophy from Jilin University, China, in 2019. She is currently pursuing the M.S. degree in artificial intelligence with Qinghai University, China.

From 2019 to 2022, she was a Staff Member with Guangzhou Xingzhong Advertising Company. She is the coauthor of two books and three articles. Her research interests include image processing and knowledge graph. She received the qualification certificate of computer and software technology proficiency, in 2022.

XIAOJING LIU was born in Anhui, China. She is currently a Professor with the Department of Computer Science, Qinghai University, the Director of the Institute of Information Visualization and Media Computing, the Director of the Chinese Society of Image and Graphics and the Qinghai Youth Joint Committee, and the Deputy Secretary-General of Qinghai Computer Society. She received 32 awards, including eight provincial awards, presided over two National Natural Science Foundation projects, and two provincial and ministerial projects, and published more than 30 teaching and research articles.