# Contrastive Masked Video Modeling for Coronary Angiography Diagnosis

Zhiming Shao[1,8][0009−0008−0210−8513], Yingqian Zhang[2], Zechen Wei[1], Yong Ge[3], Chen Wang[4], Guodong Ding[5], Lei Gao[2], Liwei Zhang[2], Yundai Chen[2(✉)], Jie Tian[6,7(✉)][0000−0003−0498−0432], and Hui Hui[1,7,8(✉)][0000−0002−6732−4232]

[1] Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, China
[2] Senior Department of Cardiology, Chinese PLA General Hospital, Beijing, China
[3] Bo Ai Hospital of Huanghua City, Hebei Province, China
[4] Qingyun County People's Hospital of Shandong Province, China
[5] Qixia People's Hospital of Yantai City, Shandong Province, China
[6] School of Engineering Medicine and School of Biological Science and Medical Engineering, Beihang University, Beijing, China
[7] National Key Laboratory of Kidney Diseases, Beijing, China
[8] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
cyundai@vip.163.com, tian@ieee.org, hui.hui@ia.ac.cn

**Abstract.** Patients with valvular heart disease often exhibit motion characteristics such as artery movements and anatomic characteristics, thus extracting dynamic features from coronary angiography (CAG) is of great significance for diagnosing. Given the challenge of limited annotated medical imaging data, we propose a novel self-supervised learning framework that integrates masked video modeling (MVM) and video contrastive learning, enabling the model to learn representations with both strong instance discriminability between video segments and local perceptibility between neighboring frames. Specifically, our framework consists of three key components: an off-the-shelf frozen encoder, an online encoder-decoder following the MVM pipeline and a momentum encoder composed of an exponential moving average of previous students. We enhance the integration of contrastive learning and MVM in mainly two ways: the frozen encoder converts the supervision of masked reconstruction from low-level pixels to high-level features; an augmentation strategy called *frame shifting*, is introduced specifically for video contrastive learning. To validate the effectiveness of our proposed method, we first conducted self-supervised pre-training on over 50,000 self-collected, unlabeled CAG sequences. Subsequently, we performed supervised fine-tuning using two small-scale labeled CAG diagnostic datasets, achieving state-of-the-art performance (98.1% and 75.0% F1-Score, respectively) in both supervised and self-supervised video recognition domains. Our code is publicly available at: https://github.com/ZmingShao/ConMVM.

**Keywords:** Masked Video Modeling · Contrastive Learning · Self-Supervised Learning · Coronary Angiography.

## 1 Introduction

Coronary artery disease (CAD) continues to be the leading cause of death in the developed world, presenting a significant challenge for global health policies. Significant advancements in deep learning have led to notable improvements in the diagnosis and prognosis of CAD patients with invasive coronary angiography (CAG) over the past few decades. However, current deep learning methods for CAG primarily focus on supervised image recognition, with major challenges: 1) the high cost of annotating CAG imaging data, which is a common issue in medical imaging; 2) some conditions are difficult to diagnose confidently by analyzing features of single-frame image, e.g. patients with arrhythmia, which affects cardiac remodeling and abnormal blood, causes irregularities in their CAG sequences; 3) the difficulty to identify subtle differences in coronary artery blood flow patterns, ventricular and atrium anatomy, and epicardial vessels motion that are altered in patients with valvular heart disease [22, 23]. In response, we attempt to introduce a video self-supervised learning (SSL) method to build a video foundation model for multiple CAG-related downstream diagnostic tasks.

Currently, there are two mainstream paradigms in visual SSL: contrastive learning (CL) and masked learning. Contrastive learning works by minimizing the distance between positive samples in the feature space while pushing away negative samples, which promotes the model's ability to learn discriminative features. Masked learning, on the other hand, involves predicting the information of masked patches from a small number of visible patches, encouraging the model to learn semantic relationships between neighboring patches in an image. Recently, many works in the field of image SSL have attempted to combine both approaches [16, 17, 32], aiming to leverage both advantages. Such combination is effective for video SSL as well, allowing the model to capture representations with both strong instance discriminability across video segments and local coherence among adjacent frames, while there has been little progress in research on it currently. This paper, therefore, conducts an in-depth study on this subject.

Firstly, we identify several key challenges in combining masked learning and contrastive learning: (1) In contrastive learning, the encoder encodes the complete input into the feature space, while in masked learning, the output of the encoder contains incomplete information, and the decoder typically predicts pixel values, neither of which are suitable for directly constructing contrastive learning. (2) Contrastive learning requires two encoders that input different augmented views of the original image. However, the high proportion of masking in masked learning is not compatible with too strong augmentations, since the disparity between views caused by them will be excessively amplified, resulting in false positive views [16]. To address the challenge (1), this paper introduces an off-the-shelf frozen encoder, which allows the supervision signal in masked reconstruction to be encoded from low-level pixels into high-level features, making the output of decoder applicable for both reconstruction loss and contrastive loss. To address the challenge (2), the paper proposes a weaker augmentation tailored to the characteristics of video data, called *frame shifting*, which enhances video

contrastive learning while avoiding conflicts with masked learning. In summary, the main contributions of this paper are as follows:

1. We combine masked learning and contrastive learning in the video SSL domain, and apply this to pretraining on over 50,000 unlabeled CAG sequences.
2. We change the routine masked pixel modeling to masked feature modeling and propose an novel augmentation specifically for video data, *frame shifting*, both of which promote combining masked learning with contrastive learning.
3. Our foundation model achieves 98.1% and 75.0% F1-Score on two CAG downstream tasks, respectively, surpassing the state-of-the-art (SOTA) algorithms in both supervised and self-supervised video recognition domains.

## 2 Related Work

**Contrastive Learning** focuses on learning instance discriminative representations by ensuring that multiple views of the same image are mapped closer in the feature space while pushing apart representations of different images. SimCLR [4] extensively explores data augmentations in contrastive learning to create different view for the same image. To efficiently manage and utilize negative samples, MoCo [15] introduces a memory queue to maintain a diverse set of negative examples. BYOL [13], on the other hand, eliminates the need for explicit negative samples by incorporating an online encoder that predicts the output of a momentum-updated encoder, effectively preventing training collapse. To further simplify BYOL, SimSiam [5] replaces the momentum updates with a stop-gradient mechanism, reducing computational complexity while maintaining stability. Recent advancements in contrastive learning have integrated Vision Transformers (ViT) [8] as the backbone architecture, leading to further improvements. For instance, MoCov3 [6] extends MoCo with transformer-based feature extraction, while DINO [3] builds upon BYOL, employing a self-distillation strategy to enhance representation learning.

**Masked Visual Modeling.** Inspired by the success of masked language modeling in natural language processing [7], masked visual modeling seeks to learn effective visual representations by reconstructing the original input from partially observed data. SimMIM [30] and MAE [14] reconstruct raw pixel values, albeit with different masking strategies. SimMIM reconstructs entire image patches, whereas MAE masks a significantly larger portion of the input and reconstructs only the visible patches, making it more effective for pretraining. To capture richer semantic features, MaskFeat [29] introduces low-level local features (HOG) as the reconstruction target. BEiT [1] leverages discretized tokens produced by an offline tokenizer to guide the encoder's training. In recent years, the framework of masked image modeling have been extended to video self-supervised learning, leading to the development of several notable methods, including BEVT [27], MaskedFeat [29], MAE-ST [11] and VideoMAE [25, 26].

# 3 Method

## 3.1 Framework

The overall framework is shown in pre-training stage of Figure 1, which consists of three parts: an off-the-shelf frozen encoder, an online encoder-decoder, and a momentum encoder. Contrastive learning is built between the features reconstructed by the online encoder-decoder and those obtained from the momentum encoder, which encodes an augmented view of the complete input. Specifically, given a video clip $\boldsymbol{V}_s \in \mathbb{R}^{T\times H\times W\times C}$, we obtain a sequence of tokens $\boldsymbol{x}_s = \{x_s^i\}_{i=1}^N$ according to the cube embedding strategy [25]. On the other hand, we create an augmented view $\boldsymbol{V}_t$ from $\boldsymbol{V}_s$, and similarly process it to get a sequence of tokens $\boldsymbol{x}_t$.

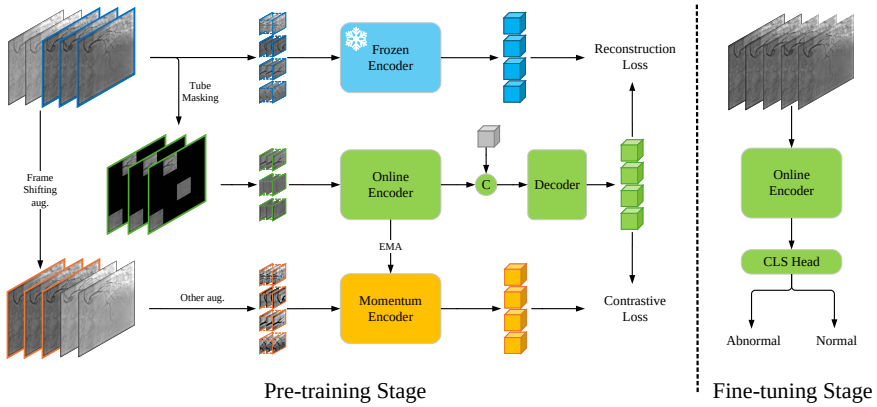


**Fig. 1.** An illustration of the overall pipeline.

**Frozen encoder $\mathcal{F}$** is a powerful encoder pre-trained in a MVM manner, which encodes all tokens of $\boldsymbol{x}_s$ into high-level features $\hat{\boldsymbol{z}}_s = \mathcal{F}(\boldsymbol{x}_s)$, serving a role similar to that of a teacher model in feature distillation. We select a larger scale ViT [8] architecture (e.g. ViT-B/L/H) and perform pre-training in advance. During pre-training of the entire framework, we freeze the weights of this encoder.

**Online encoder-decoder $\mathcal{F}_s$-$\mathcal{G}$** is a typical MVM pipeline. First, the tube masking strategy $\mathcal{M}$ [25] is applied to the input token sequence $\boldsymbol{x}_s$ to obtain visible tokens $\boldsymbol{x}_s^v = \mathcal{M}(\boldsymbol{x}_s)$. The encoder $\mathcal{F}_s$ encodes only the visible tokens $\boldsymbol{x}_s^v$ to obtain their representations $\boldsymbol{z}_s^v = \mathcal{F}_s(\boldsymbol{x}_s^v)$, with the corresponding positional embeddings incorporated. Then, the masked portion $\boldsymbol{m}$, in the form of learnable tokens with positional embeddings added, is concatenated back with the visible portion $\boldsymbol{z}_s^v$. The concatenated full token sequence is fed into the decoder $\mathcal{G}$

for reconstruction $\boldsymbol{z}_s = \mathcal{G}(\text{concat}(\boldsymbol{z}_s^v, \boldsymbol{m}))$, where loss is calculated only for the masked portion of predictions from decoder by comparing them with the corresponding features $\hat{\boldsymbol{z}}_s$ encoded by the frozen encoder $\mathcal{F}$ mentioned above. Due to distillation-like design of the frozen encoder, the online encoder still performs well even using a smaller scale ViT (e.g., ViT-S).

**Momentum encoder $\mathcal{F}_t$** processes the augmented view $\boldsymbol{V}_t$ of the original video to implement contrastive learning. Its architecture is identical to that of the online encoder, with parameters updated through the exponential moving average (EMA). The momentum encoder $\mathcal{F}_t$ takes all augmented tokens $\boldsymbol{x}_t$ as input, and the feature sequence it encodes $\boldsymbol{z}_t = \mathcal{F}_t(\boldsymbol{x}_t)$ is compared with that reconstructed by the online encoder-decoder $\boldsymbol{z}_s$. After simple mean pooling and projection, a contrastive loss is calculated between the two.

### 3.2 Combining MVM with CL

**Masked Feature Modeling.** To combine MVM with contrastive learning, a natural idea is to use the features output by the masked encoder for two purposes: feeding them to the decoder for reconstruction loss and using them to compute contrastive loss with the output from another encoder. However, since the masked encoder receives incomplete input information, CMAE [16] designed an extra decoder called *feature decoder* to reconstruct feature vectors that represent the complete information, which are then used to calculate contrastive loss. In fact, we can unify them into one single decoder to reconstruct merely features instead of pixels. Specifically, we adopts a new masked learning paradigms called *masked feature modeling* [28], introducing an additional off-the-shelf frozen encoder to supervise the masked reconstruction.

**Frame Shifting.** In contrastive learning, siamese encoders receive different augmented views of the original data. The augmentations typically used include random resized cropping, flipping, color jittering, random gray-scaling, etc. However, a large proportion of masking severely damages the input information to the online encoder, which may significantly amplify the disparity between inputs of the two encoders caused by too strong spatial transformations (e.g. cropping, flipping) [16]. Based on the above, we propose a weaker augmentation strategy called *frame shifting* according to video characteristics. Specifically, let us denote $\boldsymbol{V} \in \mathbb{R}^{L\times H\times W\times C}$ as a raw video with $L$ frames. We need to sample two video views, each having $T$ frames, as the aforementioned $\boldsymbol{V}_s$ and $\boldsymbol{V}_t$. First, we define $d = \lfloor \frac{L-T+1}{2} \rfloor$, and then we randomly sample the starting frame indexes based on the following *uniform distribution* $t_s \sim U(0, d), t_t \sim U(d, 2d)$ to obtain two views $\boldsymbol{V}_s = \boldsymbol{V}[t_s : t_s + T], \boldsymbol{V}_t = \boldsymbol{V}[t_t : t_t + T]$ that are equally transformed by a shift along the frame dimension. In addition to *frame shifting* and some inappropriate augmentations mentioned above, we also applied two extra augmentations commonly used in contrastive learning, random Gaussian blurring and color jittering.

### 3.3 Training Objective

**Contrastive Loss.** We adopt the commonly used loss function in contrastive learning, InfoNCE loss [24]. Following the common practice in contrastive learning [4, 15, 13, 5], we add *projection-prediction and projection* heads after the online encoder-decoder and the momentum encoder respectively. After mapping $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$ through the heads, we calculate the cosine similarity $\rho$ between them, then the InfoNCE loss is computed using the following formula:

$$L_{ct} = -\log \frac{\exp(\rho^+/\tau)}{exp(\rho^+/\tau) + \sum_{j=0}^{B} \exp(\rho_j^-/\tau)} \tag{1}$$

where $\rho^+$ and $\rho_j^-$ represents the cosine similarity of positive pair, i.e. two views from the same input, and of the j-th negative pair, which comes from other data within the same batch $B$ as the current input, respectively.

**Reconstruction Loss.** Traditional masked learning methods based on pixel reconstruction typically use the L2 loss function, while we adopts the smooth L1 loss function for *Masked Feature Modeling*, as it is more robust to outliers. The mathematical formula for the reconstruction loss function is:

$$L_{rc} = \begin{cases} 0.5\|\hat{\boldsymbol{z}}_s - \boldsymbol{z}_s\|_2^2 & \text{if} \quad \|\hat{\boldsymbol{z}}_s - \boldsymbol{z}_s\|_2 < 1 \\ \|\hat{\boldsymbol{z}}_s - \boldsymbol{z}_s\|_1 - 0.5 & \text{otherwise} \end{cases} \tag{2}$$

The final loss function is computed by weighted summing the reconstruction loss $L_{rc}$ and contrastive loss $L_{ct}$:

$$L = \lambda_{rc} L_{rc} + \lambda_{ct} L_{ct} \tag{3}$$

## 4 Experiments

### 4.1 Implementation and Datasets

**Pre-training.** We generally follow the settings of VideoMAEv2 [26] for basic hyperparameters during pre-training. Besides, the frozen encoder is scaled to ViT-Large model by default, while the online encoder is scaled to ViT-Small model for lightweight and efficient. Regarding the optimization objective, we set equal weights of 1.0 by default for both the reconstruction loss and contrastive loss. All pre-training experiments are conducted on 1 NVIDIA A800 GPUs.

A total of 57,857 unlabeled CAG sequences collected from Chinese PLA General Hospital, Bo Ai Hospital of Huanghua City, Qingyun County People's Hospital of Shandong Province and Qixia People's Hospital of Yantai City served as pre-training dataset, first used for preparing the frozen encoder, followed by the entire model. These imaging sequences have a frame rate of 15 FPS, a resolution of $512 \times 512$, and an average length of 52 frames. We filtered out sequences with fewer than 21 frames, leaving a final dataset of 47,794 sequences. Each sequence was uniformly sampled at 3-frame intervals to extract 16 frames.

**Fine-tuning.** Given the scarcity and limited feature diversity of CAG imaging data, we discarded most commonly used data augmentation strategies (e.g. random resized cropping, mixup, label smoothing), and fine-tuned the model for only 10 epochs with a linearly scaled learning rate $lr = base\_lr \times batch\_size/64$, while keeping all other settings consistent with VideoMAEv2 [26].

We selected two CAG diagnostic tasks as downstream benchmarks: severe mitral regurgitation (MR) and severe aortic stenosis (AS), both of which have altered ventricular, atrium anatomy and epicardial vessels motion. This was a retrospective study enrolling patients from 2016.12 to 2025.1, from Chinese PLA General Hospital. The MR dataset consists of 605 CAG sequences, including 316 patient cases and 289 controls, while the AS dataset consists of 361 CAG sequences, including 121 patient cases and 240 controls. During fine-tuning, the imaging parameters of all CAG sequences remained consistent with those in pre-training stage. The dataset was split into a 3:1 ratio for training and testing. Performance was evaluated on the test set using *precision*, *recall*, and *F1-score* as the final assessment metrics.

### 4.2 Comparison with the SOTA Methods

We selected several recent SOTA methods in both supervised and self-supervised video recognition fields and conducted comparative experiments (See Table 1) on two benchmarks mentioned above: MR and AS. 1) For supervised algorithms, we selected both Convolution-based [31, 10, 9, 12] and ViT-based [2, 20, 19, 18, 21] methods. Our algorithm improves the F1-score by 3.4% on MR and 5.9% on AS compared to the best-performing supervised algorithm, Video Swin [21]. 2) For self-supervised algorithms, considering the limitation in available data, we only selected algorithms pre-trained with pure video data [11, 26], i.e. no mixture of video and other modalities (e.g. image [27, 28]) was used for pre-training. Our method outperforms the best-performing self-supervised approach, VideoMAEv2 [26], with a F1-score gain of 2.2% on MR and 2.9% on AS.

### 4.3 Ablation Study

**Frame Shifting.** As shown in Table 2, directly introducing video contrastive learning into the MVM pipeline, using only the commonly applied augmentations from image contrastive learning, resulted in almost no performance gain. However, with *frame shifting* introduced, there was a significant performance improvement (+1.5% F1), indicating that this augmentation is crucial for video contrastive learning and meanwhile is well-suited to the MVM paradigm.

**Contrastive Loss.** Since the batch size was not large enough due to the limited data scale, we also tested two extra contrastive loss functions, SimSiam [5] loss and BYOL [13] loss, which do not depend heavily on mass negative samples. However, the experimental results (Table 2) showed that the InfoNCE loss performed the best, probably owing to limited feature diversity of data, while

**Table 1.** Comparison with SOTA methods on MR and AS

| Method | CAG MR | | | CAG AS | | |
|---|---|---|---|---|---|---|
| | Precision(%) | Recall(%) | F1(%) | Precision(%) | Recall(%) | F1(%) |
| *supervised* | | | | | | |
| TPN [31] | 51.1 | 95.8 | 66.7 | 87.5 | 46.7 | 60.9 |
| CSN [12] | 98.0 | 81.0 | 88.7 | 59.4 | 63.3 | 61.3 |
| SlowFast [10] | 74.7 | 89.9 | 81.6 | 83.3 | 50.0 | 62.5 |
| SlowOnly [10] | 88.0 | 95.8 | 91.7 | 83.3 | 50.0 | 62.5 |
| X3D [9] | 93.1 | 73.8 | 82.4 | 70.8 | 56.7 | 63.0 |
| TimeSformer [2] | 91.9 | 95.3 | 93.6 | 57.1 | 77.4 | 65.7 |
| MViTv2 [20] | 98.6 | 89.5 | 93.8 | 81.0 | 56.7 | 66.7 |
| UniFormer [19] | 93.3 | 93.7 | 93.5 | 61.5 | 77.4 | 68.6 |
| UniFormerv2 [18] | 99.1 | 90.3 | 94.5 | 74.1 | 64.5 | 69.0 |
| Video Swin [21] | 94.9 | 94.5 | 94.7 | 79.2 | 61.3 | 69.1 |
| *self-supervised* | | | | | | |
| MAE-ST [11] | 98.1 | 88.6 | 93.1 | 66.7 | 60.0 | 63.2 |
| VideoMAEv2 [26] | 97.8 | 94.1 | 95.9 | 73.3 | 71.0 | 72.1 |
| **Ours** | **97.1** | **99.2** | **98.1** | **84.0** | **67.7** | **75.0** |

the SimSiam loss, due to requirement for $\mathcal{F}_t$ to share weights with $\mathcal{F}_s$ instead of EMA updating, even resulted in a negative performance gain (-0.2% F1).

We also tested different settings for $\lambda_{ct}$. The experimental results indicate that masked learning and contrastive learning are equally important, thus only by assigning $\lambda_{rc} = \lambda_{ct} = 1.0$ can optimal performance be achieved. This also reflects the effectiveness of our strategy in integrating both paradigms.

**Frozen encoder.** Building on the above, we introduced the frozen encoder $\mathcal{F}$ to change supervision of MVM from pixels to features and tested at different ViT scales. The experimental results (Table 2) showed that *Masked Feature Modeling* at larger scale further facilitated the integration of contrastive learning and masked learning (+0.3% F1 for ViT-B, +0.7% F1 for ViT-L).

## 5 Conclusion

We propose a novel video SSL method that combines MVM and video contrastive learning, applied to pre-training on large-scale unlabeled CAG sequences. We delve into an key question: how to integrate masked learning and contrastive learning so that the two complement each other effectively, and we propose two main solutions: 1) changing the reconstruction target of masked learning from pixel values to encoded features; 2) introducing *frame shifting* to generate augmented views for video contrastive learning. We validated the transferability of the foundation model on two small-scale manually labeled CAG diagnostic datasets and demonstrated the effectiveness of the two proposed solutions.

**Table 2.** Ablation study on MR. We integrate the proposed components step by step and conduct ablation experiments on relevant parameters at each step, where * indicates the optimal parameters we selected as the default settings for other experiments.

| | $L_{ct}$ | $\mathcal{F}$ | $\lambda_{ct}$ | F1(%) |
|---|---|---|---|---|
| Baseline [26] | - | - | - | 95.9 |
| + Momentum encoder | InfoNCE | - | 1.0 | 95.9 |
| + Frame shifting aug. | SimSiam | - | 1.0 | 95.7 |
| | BYOL | | | 96.6 |
| | InfoNCE* | | | 97.4 |
| + Frozen encoder | InfoNCE | ViT-B | 1.0 | 97.7 |
| | | ViT-L* | 0.1 | 97.0 |
| | | | 1.0* | 98.1 |
| | | | 2.0 | 97.2 |

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
6. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
7. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition (2020)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
11. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems **35**, 35946–35958 (2022)

12. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12046–12055 (2019)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
16. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
17. Lee, Y., Willette, J., Kim, J., Lee, J., Hwang, S.J.: Exploring the role of mean teachers in self-supervised masked auto-encoders. arXiv preprint arXiv:2210.02077 (2022)
18. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. arXiv preprint arXiv:2211.09552 (2022)
19. Li, K., Wang, Y., Peng, G., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatial-temporal representation learning. In: International Conference on Learning Representations (2022)
20. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4804–4814 (2022)
21. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
22. Lu, W., Zhang, X., Yan, G., Ma, G.: The differences of quantitative flow ratio in coronary artery stenosis with or without atrial fibrillation. Journal of Interventional Cardiology **2023**(1), 7278343 (2023)
23. Mehta, R.H., Harjai, K.J., Grines, L., Stone, G.W., Boura, J., Cox, D., O'Neill, W., Grines, C.L., in Myocardial Infarction (PAMI) Investigators, P.A.: Sustained ventricular tachycardia or fibrillation in the cardiac catheterization laboratory among patients receiving primary percutaneous coronary intervention: incidence, predictors, and outcomes. Journal of the American College of Cardiology **43**(10), 1765–1772 (2004)
24. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
25. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
26. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14549–14560 (2023)

27. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14733–14743 (2022)
28. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.G.: Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6312–6322 (2023)
29. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
30. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
31. Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal pyramid network for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
32. Yao, Y., Desai, N., Palaniswami, M.: Masked contrastive representation learning for self-supervised visual pre-training. In: 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2024)