CoMAE: Single Model Hybrid Pre-training on Small-Scale RGB-D Datasets

Jiange Yang[1], **Sheng Guo**[2], **Gangshan Wu**[1], **Limin Wang**[1*]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] MYbank, Ant Group, China
{jiangeyang.jgy, guosheng1001}@gmail.com, {gswu, lmwang}@nju.edu.cn

Abstract

Current RGB-D scene recognition approaches often train two standalone backbones for RGB and depth modalities with the same Places or ImageNet pre-training. However, the pre-trained depth network is still biased by RGB-based models which may result in a suboptimal solution. In this paper, we present a single-model self-supervised hybrid pre-training framework for RGB and depth modalities, termed as **CoMAE**. Our CoMAE presents a curriculum learning strategy to unify the two popular self-supervised representation learning algorithms: contrastive learning and masked image modeling. Specifically, we first build a patch-level alignment task to pre-train a single encoder shared by two modalities via cross-modal contrastive learning. Then, the pre-trained contrastive encoder is passed to a multi-modal masked autoencoder to capture the finer context features from a generative perspective. In addition, our single-model design without requirement of fusion module is very flexible and robust to generalize to unimodal scenario in both training and testing phases. Extensive experiments on SUN RGB-D and NYUDv2 datasets demonstrate the effectiveness of our CoMAE for RGB and depth representation learning. In addition, our experiment results reveal that CoMAE is a data-efficient representation learner. Although we only use the small-scale and unlabeled training set for pre-training, our CoMAE pre-trained models are still competitive to the state-of-the-art methods with extra large-scale and supervised RGB dataset pre-training. Code will be released at https://github.com/MCG-NJU/CoMAE.

Introduction

Thanks to the flourishing development of deep learning, a series of vision tasks, including image classification (He et al. 2016), object detection (Ren et al. 2017), and semantic segmentation (Chen et al. 2018), have achieved impressive success by learning discriminative features in an end-to-end manner. Scene recognition (Zhou et al. 2014; Wang et al. 2017) is a more challenging task that could benefit from these object-centric understanding. However, it also tends to depend on more representative features than object-centric tasks due to illumination changes, complex spatial layout, and the diversity of objects and their correlations. Therefore, with the widespread use of cost-affordable depth sensors, e.g., Kinect, the research community has witnessed the advancement of RGB-D scene recognition (Silberman et al. 2012; Song, Lichtenberg, and Xiao 2015). The depth modality can provide the cues on the shape of a local geometry and is less sensitiveness to illumination changes and texture variation compared with RGB in this task.

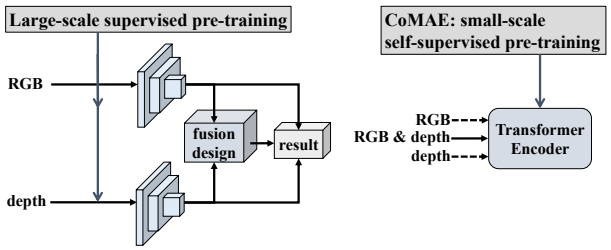

Figure 1: Comparison of RGB-D scene recognition pipeline. (a) Traditional methods often use supervised pre-training to initialize CNNs and train two separate networks for RGB and depth modalities. Then, these two networks are combined with a fusion module to yield final recognition results. (b) Our CoMAE presents single-model and self-supervised pre-training for RGB and depth two modalities with a Transformer encoder. Our pre-trained singel-model encoder could be deployed for unimodal or multimodal input due its flexibility of self-attention operation.

As shown in Fig. 1, the most straightforward method for RGB-D scene recognition (Zhu, Weibel, and Lu 2016; Xiong, Yuan, and Wang 2021; Du et al. 2021) is to train two separate networks (CNNs) for two modalities, and combine their results with some fusion modules to yield the final recognition results. In this paradigm, depth data is expected to learn fine and complementary geometrical features to enhance the performance of RGB CNN. However, data scarcity is an important issue for depth modality and training from scratch is prone to over-fitting. Therefore, some methods (Zhu, Weibel, and Lu 2016) try to use RGB pre-trained models on ImageNet (Deng et al. 2009) or Places (Zhou et al. 2014) to initialize the depth CNNs to mitigate this issue. Yet, the modality gap between RGB and depth makes

*Corresponding author.

this cross-modal pre-training strategy to be biased and less effective (Song, Herranz, and Jiang 2017; Du et al. 2021). Meanwhile, this recognition paradigm of fusing two-stream CNNs makes the whole scene recognition pipeline less efficient and incapable of handling missing modalities in practice. Therefore, it still remains as a challenge that *how to effectively learn RGB and depth representations on the relatively small-scale RGB-D datasets without using extra pre-training or data, that could be easily and efficiently generalized to downstream tasks with multimodal or unimodal data.*

Recently Transformer (Vaswani et al. 2017) has shown the unique advantages in multi-modal understanding tasks, such as image-text retrieval (Radford et al. 2021), image-text caption (Yu et al. 2022), and video-audio understanding (Akbari et al. 2021). The attention operation exhibits high flexibility in handling the multi-modality inputs, and can yield a unified architecture to process different modalities with a single encoder (Likhosherstov et al. 2021; Girdhar et al. 2022b). However, these multi-modal transformers are all pre-trained on large-scale datasets of millions of samples with human labels or noisy texts. It still remains unclear how to develop an effective self-supervised multi-modal transformer on the relatively small-scale RGB-D datasets. Inspired by the results in VideoMAE (Tong et al. 2022), we aim to adapt the data-efficient masked image autoencoder (MAE) (He et al. 2022) to multi-modal and small-scale RGB-D datasets for better representation learning.

Based on the above analysis, in this paper, we present a single-model hybrid pre-training method on the relatively small-scale RGB-D datasets. Our method is based on a self-supervised multi-modal transformer, termed as **CoMAE**. To better learn the multi-modal information from limited training samples, we devise a self-supervised curriculum learning framework. Our CoMAE unifies the two types of self-supervised learning paradigms (contrastive learning and masked image modeling), and trains a single encoder for both RGB and depth modalities. Specifically, we first build a cross-modal patch-level alignment task to guide the encoder training in a self-supervised manner. This contrastive learning objective would encourage the encoder to capture the coarse structure information to build correspondence between modalities. Then, the learned encoder will be passed into a multi-modal masked autoencoder to perform another round of pre-training from a generative perspective. The multi-modal generative objective would pose a more challenging pre-training task that requires for more detailed granularity to predict the exact pixel values of masked tokens. Unlike the original MAE (He et al. 2022), our CoMAE presents a shared encoder and decoder among RGB and depth modalities and acts as a kind of regularizer to guide pre-training. In addition, the pre-trained transformer encoder pre-trained by our CoMAE breaks the conventional two-stream fusion paradigm and can handle both multi-modal and unimodal inputs in the subsequent downstream deployment. In summary, our main contribution is threefold:

- We present a single-model self-supervised hybrid pre-training framework (CoMAE) for RGB-D representation learning with application on scene recognition. Our CoMAE presents a minimalist design in curriculum learning manner to unify two types of popular self-supervised learning methods for RGB-D pre-training on a relatively small-scale dataset.
- Our proposed single-model encoder is modality-agnostic, which simultaneously deal with RGB and depth with a parameter-sharing encoder. Hence it is flexible to generalize to various settings of multi-modal or unimodal data in both training and testing phases.
- For first time, our CoMAE demonstrates that a multi-modal transformer could be successfully trained on limited RGB-D samples, without using extra data or pre-trained models. Our CoMAE pre-trained models are competitive to those methods with strong and large-scale data pre-training on the two challenging RGB-D scene recognition datasets of NYUDv2 and SUN RGB-D.

Related Work

RGB-D Scene Recognition. Several previous works explicitly modeled the object correlations to obtain a more comprehensive understanding of the scene, such as (Yuan, Xiong, and Wang 2019) and (Song et al. 2019). Recently some works focused on excavating the inter-modality and intra-modality relationships. For example, MSN (Xiong, Yuan, and Wang 2020) simultaneously extracted the local modal-consistent and global modal-specific features. (Du et al. 2021) proposed to enhance the modality-specific discriminative ability by a cross-modal translation fashion, which both effectively transfers complementary cues to the backbone network. However, these methods all use RGB pre-trained models on ImageNet or Places to initialize the depth CNNs, which is biased and less effective. In addition, the two-stream fusion paradigm makes the recognition less efficient and incapable of handling missing modalities in practice. Instead, our CoMAE does not rely any extra data or pre-trained models and can handle both inputs of multi-modal or unimodal data.

Self-Supervised Learning. Self-supervised representation learning automatically constructs pretext tasks from data itself to learn rich representations that can be transferred to downstream tasks. Currently two types of popular pretext tasks have been designed and applied for self-supervised learning: masked image modeling (Bao et al. 2022; He et al. 2022) and contrastive learning (Chen et al. 2020; He et al. 2020; Grill et al. 2020). These two types of self-supervised learning methods focus on modeling data distribution from generative and discriminative perspectives. Some methods focuses on improving downstream tasks for dense prediction (Wang et al. 2021; Gao et al. 2022). Recently some works (Tao et al. 2022; Yi et al. 2022) also attempted to unify these two types self-supervised representation learning algorithms through the patch-level contrastive loss between masked prediction tokens and the augmented view from a momentum encoder. Different from these methods, we propose a curriculum learning strategy to combine these two types of self-supervised learning methods in a simpler way. Besides, we extend this setting to multi-modal data and only perform pre-training on very small-scale dataset.

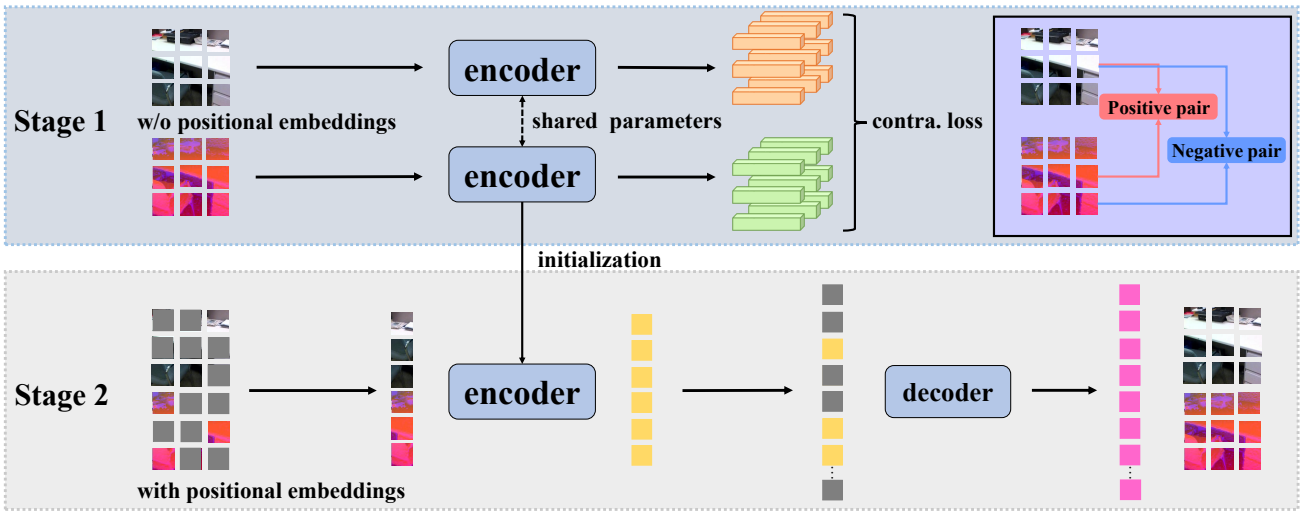

Figure 2: The CoMAE framework. Our CoMAE performs cross-modal patch-level contrastive pre-training in the first stage (Upon) which provides weights initialization for following reconstruction task and multi-modal masked pre-training in the second stage (Down) which provides weights initialization for downstream recognition task. In the figure, all encoders have the same structure, and the cuboids denote vision tokens of RGB and depth.

Multi-Modal Transformer. Transformer architectures are very expert in handling multiple input modalities data due its flexibility of self-attention operation. We note that transformers have been used to process multiple modalities in a single encoder such as vision and language (Kim, Son, and Kim 2021), vision and point cloud (Wang et al. 2022), as well as video and audio (Nagrani et al. 2021). Recently pre-training multi-modal transformers by masked modeling on large-scale data has become a popular approach towards solving kinds of downstream tasks. (Girdhar et al. 2022a) jointly pre-trains a single unified model on images and videos on unpaired data with masked modeling. (Geng et al. 2022) and (Bachmann et al. 2022) build the multi-modal masked autoencoders for paired image-text data and RGB-depth data, respectively. Significantly different from (Bachmann et al. 2022), our CoMAE proposes to use a single decoder for different modalities, while (Bachmann et al. 2022) used different decoders for different modalities. Meanwhile, our CoMAE conducts multi-modal masked reconstruction on small-scale RGB-D datasets instead of large-scale ImageNet with pseudo depth from depth estimation network.

Methodology

In this section, we describe our CoMAE in detail. First, we give an overview of the CoMAE framework, which is composed of the cross-modal patch-level contrastive pre-training and the multi-modal masked pre-training. Then, we give a technical description on these two components. Finally, we describe how to fine-tune on downstream scene recognition task.

Overiew of CoMAE

As analyzed above, RGB-D scene recognition often uses a two-stream CNN with extra pre-training on large-scale datasets. To handle the issue of representation learning from limited RGB-D samples, we present the CoMAE framework as shown in Fig. 2. Our CoMAE aims to train a single unified model for RGB and depth modalities on small-scale datasets with a simple Vision Transformer (ViT) (Dosovitskiy et al. 2021). We employ a hybrid self-supervised pre-training strategy to progressively tune the parameters of an encoder in curriculum learning manner. First, we propose a relatively easier cross-modal contrastive learning framework to capture the correspondence between two modalities in patch level. This correspondence task will guide transformer encoder to capture structure information to discriminate different patch instances. Then, we present the multi-modal masked autoencoder with a simple parameter sharing scheme in both encoder and decoder among two modalities. This masking and reconstruction task will encourage the transformer encoder to extract detailed granularity to regress the exact pixel values of missing tokens. Meanwhile, the shared encoder-decoder approach would greatly reduce the parameter numbers and also act as a kind of regularizer for pre-training. This customized hybrid training strategy generally follows the principle of curriculum learning (Bengio et al. 2009) and is able to improve the generalization ability of learned representation to the downstream tasks. Thanks to this single-model pre-training scheme, our ConMAE pre-trained encoder could be deployed for fine-tuning with either multi-modal or unimodal inputs, which is efficient to handle the issue of lacking modality in practice.

Cross-modal Patch-level Contrastive Pre-training

Inspired by the success of contrastive pre-training in vision and language (Li and Wang 2020; Miech et al. 2020; Radford et al. 2021; Jia et al. 2021), we present our cross-modal patch-level contrastive (CPC) pre-training. To the best of our

knowledge, the joint learning objective of enforcing RGB-depth alignment has not been well explored. Specifically, for each paired RGB-depth image in a training iteration, the input image r and corresponding depth d are separately fed into respective patch projection layers and the shared encoder to generate the individual patch feature representations $\{f(r)_i|_{i=1}^N\}$ and $\{f(d)_i|_{i=1}^N\}$, where i is the patch index, N is the total number of patches within an RGB or depth (e.g., $N = 196$ for a 224×224 image with a patch size of 16×16). Note that we remove the sine-cosine positional embeddings of ViT to avoid information leakage in contrastive learning. Moreover, different from conventional contrastive learning methods like (He et al. 2020; Chen et al. 2020), our query and dictionary keys are cross-modal patch within a paired RGB-depth.

Specifically, we use the InfoNCE loss (van den Oord, Li, and Vinyals 2018) to train our transformer encoder, with a fixed temperature of 0.07. This loss maximizes the similarity of an RGB patch and its corresponding depth patch in the same location, while minimizes similarity to all other patches in corresponding depth map. Our cross-modal contrastive loss function $Loss_{cpc}$ is as follows:

$$\ell_{rgb}(i) = -\log\left(\frac{\exp(s(f(r)_i, f(d)_i)/\tau)}{\sum_{k=1}^{N}\left(\exp(s(f(r)_i, f(d)_k)/\tau)\right)}\right), \quad (1)$$

$$\ell_{depth}(i) = -\log\left(\frac{\exp(s(f(d)_i, f(r)_i)/\tau)}{\sum_{k=1}^{N}\left(\exp(s(f(d)_i, f(r)_k)/\tau)\right)}\right), \quad (2)$$

$$Loss_{cpc} = \frac{1}{2N}\left(\sum_{i=0}^{N}\left(\ell_{rgb}(i) + \ell_{depth}(i)\right)\right). \quad (3)$$

where τ and $s(.)$ are temperature coefficient and similarity functions respectively.

This patch-level contrastive pre-training explicitly utilizes the inter-modal alignment relations as the self-supervisory signal. It expects the cross-modal alignment may be implicitly learned from paired small-scale RGB-depth datasets. With the above optimized objectives, the encoder network is encouraged to better capture the discriminative local context representations across RGB and depth modalities, which could also facilitate their the following multi-modal masking and reconstruction task.

Multi-modal Masked Pre-training

After the cross-modal contrastive learning, our transformer encoder can learn some structure information to discriminate the corresponding RGB and depth patch pairs. To further enhance its representation power, we propose to perform another round training of transformer encoder in a generative view. Driven by the good self-supervised learning performance of MAE (He et al. 2022) in images, we design a multi-modal masked autoencoder (MM-MAE) in RGB-D data. RGB and depth are homogeneous modalities and therefore our MM-MAE could easily deal with them with a single encoder and decoder with parameter sharing for both modalities. Our MM-MAE just requires the modality-specific patch projection layers and prediction heads to distinguish the different modalities and reconstruction tasks. This *parameter sharing scheme* would greatly reduce the pre-training parameter numbers and also act as a kind of regularizer to improve the pre-training performance.

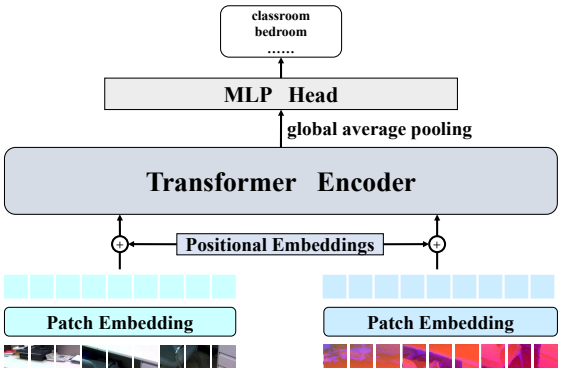

Figure 3: Illustration of the fine-tuning on downstream scene recognition task.

Specifically, we use the corresponding patch projection layers to obtain the RGB and depth tokens, and then add the fixed and shared 2D sine-cosine positional embeddings for two modalities. All patch tokens of two modalities are concatenated into a sequence. Afterward we randomly mask a subset of RGB and depth tokens both with a 75% masking ratio, and the encoder takes all visible (unmasked) tokens as input. After encoder, our lightweight decoder takes all encoded visible tokens and learnable mask tokens with 2D positional embeddings as input. The mask tokens are shared among two modalities. Finally, we apply two modality-specific prediction heads to reconstruct RGB and depth patches, respectively. Our reconstruction target $\hat{P}_r$ and $\hat{P}_d$ are per-patch normalization pixels. The whole MM-MAE pipeline can be formally described as follows:

$$F = \text{Decoder}(\text{Encoder}(T_{vis\_r}, T_{vis\_d}), T_{mask}), \quad (4)$$

$$\begin{cases} \hat{P}_r = \text{Linear}_{RGB}(F_r) \\ \hat{P}_d = \text{Linear}_{depth}(F_d) \end{cases}, \quad (5)$$

where T_{vis_r} and T_{vis_r} are visual tokens, T_{mask} is mask token, $\hat{P}$ is the reconstructed patches for RGB or depth modality. We compute the MSE loss between the reconstructed and original pixels of all masked patches. So our MM-MAE loss function $Loss_{mm-mae}$ is as follows:

$$Loss_{mm-mae} = \text{MSE}(\hat{P}_{m\_r}, P_{m\_r}) + \text{MSE}(\hat{P}_{m\_d}, P_{m\_d}). \quad (6)$$

By reconstructing all the masked RGB and depth tokens, our MM-MAE could enforce the encoder to capture modality-complementary features in a joint space manner while simultaneously preserving modality-specific features.

Fine-tuning on Scene Recognition

After our CoMAE pre-training, the pre-trained encoder can be directly fine-tuned on the downstream scene recognition task by appending light linear classifier on top. Specifically, we also concatenate all RGB and depth tokens in the

same process with the multi-modal masked reconstruction as shown in Fig 3. We obtain the global representations via global average-pooling (GAP) instead of a class token. It is worth noting that during fine-tuning process, we randomly drop a modality with a probability of 0.5 as data augmentation to obtain better results.

Experiments

In this section, we present the experiment results of our CoMAE framework on the task of RGB-D scene recognition. First, we introduce the evaluation datasets and the implementation details of our proposed approach. Then we elaborate the ablation studies on our CoMAE design. We also show some visualization results to further analyze our the effectiveness of our CoMAE. Finally, we evaluate the performance of our CoMAE and compare with the previous state-of-the-art methods. We quantitatively report the average accuracy over all samples (acc_s) and the average accuracy over all scene classes (acc_c) according to the previous evaluation scheme.

Datasets

SUN RGB-D Dataset is the most popular RGB-D scene recognition dataset. It contains RGB-D images from NYU depth v2, Berkeley B3DO (Janoch et al. 2013), and SUN3D (Xiao, Owens, and Torralba 2013), which is composed of 3,784 Microsoft Kinect v2 images, 3,389 Asus Xtion images, 2,003 Microsoft Kinect v1 images and 1,159 Intel RealSense images. Following the official setting in (Song, Lichtenberg, and Xiao 2015), we only use the images from 19 major scene categories, containing 4,845 / 4,659 train / test images.

NYU Depth Dataset V2 (NYUDv2) contains 1,449 well labeled RGB-D images for scene recognition. 795 images are for training and 654 are for testing. Following the standard split in (Silberman et al. 2012), the 27 scene categories are grouped into 9 major categories and an other category.

Implementation Details

We employ the three-channel HHA (Gupta et al. 2014) to encode depth. HHA offers a representation of geometric properties at each pixel, including the horizontal disparity, the height above ground, and the angle with gravity, which has been proven to have better performance to capture the scenes structural and geometrical information than raw depth. Our proposed CoMAE is implemented on the Pytorch toolbox, and we pre-train and fine-tune our models on eight TITAN Xp GPUs using AdamW (Loshchilov and Hutter 2019) optimizer with a weight decay 0.05. In hybrid pre-training stage, we separately train cross-modal patch-level contrastive learning for 75 epochs and multi-modal masked autoencoder for 1200 epochs. The base learning rate is set to 1.0×10^{-3} for pre-training and 5.0×10^{-4} for fine-tuning, as well as the batch size is set to 256 for pre-training and 768 for fine-tuning, the effective learning rate follows the linear scaling rule in (Goyal et al. 2017): $lr = base\_lr \times batchsize \div 256$. The warm-up and layer decay strategies are used to adjust the learning rate. Finally, it is worth noting that our pre-training uses cropping and flipping as data augmentation following MAE, but our fine-tuning only adds mixup, cutmix and erasure, which is much easier to implement than MAE.

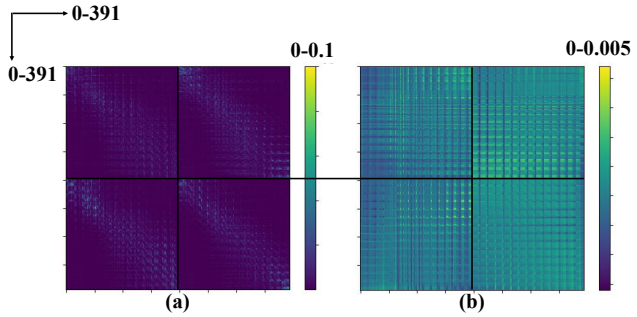

Figure 4: Visualization of the attention maps of a representative attention head in encoder at the beginning of fine-tuning. (a) Transformer encoder with cross-modal patch-level contrastive pre-training. (b) Transformer encoder with random initialization. The four sub-squares from top to bottom and left to right in attention maps represent the attention between RGB and RGB, RGB and depth, depth and RGB, as well as depth and depth respectively. Due to the ability of inter-modal alignment and intra-modal local features extraction, the former focuses its attention around the main diagonal of the four sub-squares, while the latter is relatively uniform. Please refer to the supplementary material for more attention heads visualization.

Method	acc_c	acc_s
From scratch	40.5	47.5
Patch-level Contrast Pre-training	**50.2**	**57.9**
Instance-level Contrast Pre-training	32.3	40.8

Table 1: Ablation experiments of the cross-modal patch-level contrast pre-training.

Results on SUN RGB-D Dataset

We begin our experiments by studying the effectiveness of CoMAE for RGB-D scene recognition task on the SUN RGB-D dataset. In this subsection, unless otherwise specified, all pre-training and fine-tuning experiments are performed with ViT-B on SUN RGB-D training set.

Study on cross-modal patch-level contrastive pre-training. In Tab. 1, we verify the effectiveness of the proposed cross-modal patch-level contrastive pre-training. We train the same RGB-D transformer classification model from scratch as a baseline. Moreover, we implement a conventional cross-modal contrast framework which predicts which RGB goes with which depth in instance level. It can be clearly seen that the performance of our cross-modal patch-level contrastive pre-training significantly outperforms the other two methods. Conventional instance-level cross-modal contrastive learning fails since naive RGB-depth matching is a relatively easy pretext task and fails to capture high-level

RGB	depth	Method	acc_c	acc_s
✓		RGB-MAE	42.9	52.7
	✓	depth-MAE	41.2	47.8
✓	✓	AVG Fusion	46.8	57.1
✓	✓	MM-MAE	**54.2**	**62.6**

Table 2: Ablation experiments of the fusion paradigm of MM-MAE.

	RGB	Depth	mask	acc_c	acc_s
scratch	✓	✓	-	40.5	47.5
	✓		-	27.5	34.8
		✓	-	38.6	45.3
MAE	✓	✓	mi	**54.2**	**62.6**
	✓	✓	mc	52.7	61.6
	✓		-	42.9	52.7
		✓	-	41.2	47.8

Table 3: Ablation experiments of the multi-modal masked reconstruct pre-training. 'mi', 'mc' denotes mutual independency and mutual consistency respectively.

semantic information. In contrast, our cross-modal patch-level contrast can not only facilitate inter-modal alignment between RGB and depth, but also capture intra-modal local discriminative context features. We also visualize the attention maps of contrastive learning model in Fig. 4. Our patch-level contrastive objective can encourage the encoder to find the correspondence between RGB and depth patches.

Study on multi-modal masked autoencoder. In Tab. 2, we evaluate the effectiveness of the proposed multi-modal masked autoencoder. We compare the baseline method of training two separate MAEs for RGB and depth modalities, and then fine-tuning the corresponding encoder on the each modality. We see that the RGB MAE obtains the accuracy of 42.9% and the depth modality performance is 41.2%. We also report the fusion results of RGB and depth modality by taking an average of two MAE recognition results, and the performance is 46.8%, which is worse than our CoMAE result. The superior performance of CoMAE demonstrates the effectiveness of jointly training a single encoder on two modalities and it can capture finer cross-modal structural information. We also provide the reconstruction results on SUN RGB-D and NYUDv2 in Fig. 5.

In addition, we explore two simple random masking strategies: mutual independency and mutual consistency of masked spatial location. As shown in Tab. 3, the former is better although there exists the risk of indirect cross-modal information leakage in the masking and reconstruction pipeline. We suspect that the cross-modal information leakage could promote the multi-modal fusion, because the important and complementary information from cross-modal aligned patches is encouraged to capture in reconstruction task.

Pre-training approach	acc_c	acc_s
CPC	50.2	57.9
MM-MAE	54.2	62.6
CPC + MM-MAE	50.6	58.2
MM-MAE → CPC	51.4	60.1
CPC → MM-MAE	**55.2**	**64.3**

Table 4: Ablation experiments of the hybrid pre-training in a curriculum learning manner.

Pre-train		Fine-tune		Test		acc_c	acc_s
R	D	R	D	R	D		
✓	✓	✓	✓	✓	✓	55.2	64.3
✓	✓	✓		✓		44.6	55.1
✓	✓		✓		✓	49.6	58.3
✓	✓	✓	✓	✓		38.1	43.8
✓	✓	✓	✓		✓	39.2	47.0
✓		✓		✓		42.9	52.7
	✓		✓		✓	41.2	47.8

Table 5: Ablation experiments of the robustness of unimodal data. 'R', 'D' denotes RGB and depth respectively.

Study on the hybrid pre-training strategy. After the specific ablations on the cross-modal patch-level contrastive pre-training and multi-modal masked autoencoder, we are ready to investigate the effectiveness of our proposed hybrid pre-training strategy of CoMAE. We compare with other two training alternatives: (1) joint training of transformer encoder with two objectives (CPC+MM-MAE), (2) the cascade training of transformer encoder in a reverse order (MM-MAE→CPC). The results are reported in Tab. 4. From results, we see that joint training could lead to worse performance than the single MM-MAE pre-training (50.6% vs. 54.2%). This might be because that the patch-level cross-modal contrast discards positional embedding, which degrades its optimization consistency with multi-modal masked reconstruction and thus leads to inferior performance in downstream task. We also notice that the MM-MAE→CPC baseline obtains a worse performance than our CoMAE, which demonstrates the effectiveness of our curriculum learning design in our CoMAE, by learning easier task first and then training a harder task.

Study on the random modality dropping. We propose a random modality dropping strategy with a probability of 0.5 as a very strong data augmentation in fine-tuning process. It is likely to degrade performance due to inconsistency between training and testing. However, our model benefits from this data augmentation in fine-tuning (55.2% vs. 52.7%). This result demonstrates that our model has great flexibility in handling modality inputs, and it can benefit from more unpaired unimodal data in fine-tuning phase.

Study on the robustness of unimodal data. It is difficult to collect paired multimodal data from the real world, due to the unavailability of a sensor. Therefore multimodal mod-

Method	Backbone	Additional labeled training data	acc_c	acc_s
Multimodal fusion (Zhu, Weibel, and Lu 2016)	AlexNet	Places	41.5	-
Modality and component fusion (Wang et al. 2016)	AlexNet	Places	48.1	-
RGB-D-CNN (wSVM) (Song, Herranz, and Jiang 2017)	AlexNet	Places	52.4	-
RGB-D-OB (wSVM)* (Song et al. 2018)	AlexNet	Places	53.8	-
G-L-SOOR* (Song et al. 2019)	AlexNet	Places	55.5	-
DF^{2}Net (Li et al. 2018)	AlexNet	Places	54.6	-
ACM (Yuan, Xiong, and Wang 2019)	AlexNet	Places	55.1	-
MSN (Xiong, Yuan, and Wang 2020)	AlexNet	Places	56.2	-
TRecgNet (Du et al. 2019)	ResNet18	Places, 5K unlabeled, gen	56.7	-
ASK (Xiong, Yuan, and Wang 2021)	ResNet18	Places	**57.3**	-
CNN-randRNN (Caglayan et al. 2022)	ResNet101	ImageNet	-	60.7
Omnivore (Girdhar et al. 2022b)	Swin-B	ImageNet-22K, Kinetics-400	-	**67.2**
Omnivore (Girdhar et al. 2022b)	Swin-L	ImageNet-22K, Kinetics-400	-	67.1
Ours (w/o pre-training)	ViT-B	None	40.5	47.5
Ours (CoMAE pre-training)	ViT-S	None	52.6	62.1
Ours (CoMAE pre-training)	ViT-B	None	**55.2**	**64.3**
Ours (Image MAE pre-training)	ViT-B	ImageNet-1K (unlabeled)	55.2	63.6

Table 6: Comparison with state-of-the-art methods on SUN RGB-D. * denotes using object detection to obtain object annotations. 'gen' denotes using generated data for the training data sampling enhancement.

Method	Backbone	Additional labeled training data	acc_c	acc_s
Modality and component fusion (Wang et al. 2016)	AlexNet	Places, SUN RGB-D	63.9	-
RGB-D-CNN (Song, Herranz, and Jiang 2017)	AlexNet	Places	65.8	-
RGB-D-OB (wSVM)* (Song et al. 2018)	AlexNet	Places	67.5	-
G-L-SOOR* (Song et al. 2019)	AlexNet	Places, SUN RGB-D	67.4	-
DF^{2}Net (Li et al. 2018)	AlexNet	Places	65.4	-
ACM (Yuan, Xiong, and Wang 2019)	AlexNet	Places	67.4	-
MSN (Xiong, Yuan, and Wang 2020)	AlexNet	Places	68.1	-
TRecgNet (Du et al. 2019)	ResNet18	Places, SUN, 5K unlabeled, gen	69.2	-
ASK (Xiong, Yuan, and Wang 2021)	ResNet18	Places, SUN	**69.3**	-
Omnivore (Girdhar et al. 2022b)	Swin-B	ImageNet-22K, Kinetics-400, SUN	-	79.4
Omnivore (Girdhar et al. 2022b)	Swin-L	ImageNet-22K, Kinetics-400, SUN	-	**79.8**
Ours (w/o pre-training)	ViT-B	None	40.7	49.8
Ours (CoMAE pre-training)	ViT-S	None	61.4	66.8
Ours (CoMAE pre-training)	ViT-B	None	62.7	66.5
Ours (SUN RGB-D pre-training)	ViT-S	SUN RGB-D	68.6	71.4
Ours (SUN RGB-D pre-training)	ViT-B	SUN RGB-D	**70.7**	**76.3**

Table 7: Comparison with state-of-the-art methods on NYUDv2. * denotes using object detection to obtain object annotations. 'gen' denotes using generated data for the training data sampling enhancement.

els are expected to show robustness against unimodal data. As shown in Tab. 5, we evaluate our CoMAE's robustness of unimodal data in fine-tuning and test phases. Different from some existing methods, our CoMAE is very flexible with its training and testing setting with multimodal or unimodal data. Surprisingly, our pre-trained multi-modal model significantly outperforms RGB or depth single-modal model pre-trained by image masked autoencoder under the fine-tuning and testing with only RGB or depth modality. These results demonstrate that our hybrid multimodal pre-training can significantly promote unimodal learning, and we suspect that our CoMAE pre-training can enforce model implicitly obtain some complementary modal information when only using unimodal data to inference.

Comparison with the state-of-the-art methods. We report the performance of our CoMAE on SUN RGB-D test set and compare with state-of-the-art methods in Tab. 6. Although these methods utilize supervised pre-trained models or many additional labeled data, our CoMAE can still achieve competitive results without utilizing any extra data or pre-trained models, demonstrating the effectiveness of our hybrid pre-training approach on relatively small-scale datasets. This property is quite important in real applications as sometimes it is really hard to collect lots of samples. In addition, our CoMAE instantiated with ViT-B also achieves

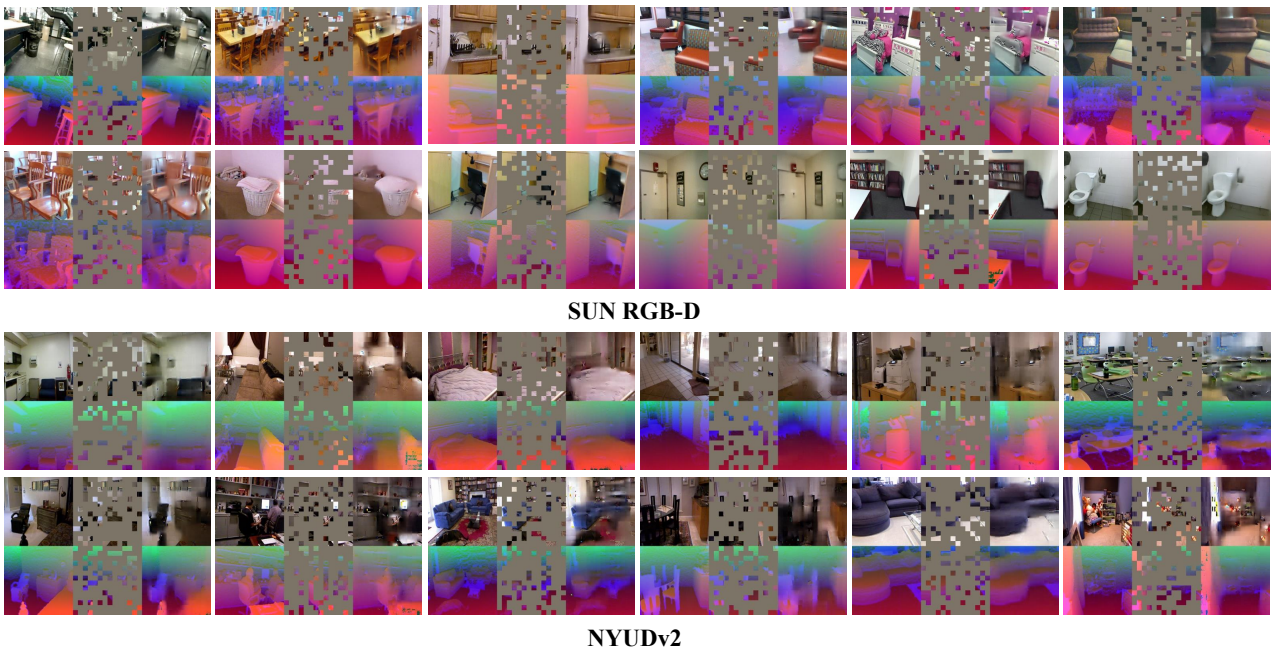

Figure 5: Some examples of reconstruction result on SUN RGB-D test set(Upon) and NYUDv2 test set(Down). For each triplet, we show the ground-truth (left), masked RGB-D (middle), and our MM-MAE reconstruction (right). We simply overlay the output with the visible patches to improve visual quality. The MM-MAE models are pre-trained using only their respective training set.

competitive and even better results of the same ViT-B model pre-trained by ImageMAE (He et al. 2022), this reveals the data efficiency advantage of our CoMAE pre-training.

Results on NYUDv2 Dataset

We also evaluate the performance of our CoMAE on the NYUDv2 test set and compare with other methods in Tab. 7. NYUDv2 is very small and the training data is seriously imbalanced, and consequently some works load the weights pre-trained on SUN RGB-D dataset for model initialization on NYUDv2 dataset. Our CoMAE achieves much better result than training from scratch on NYUDv2 when only using its training set for pre-training, but it is not excellent enough. We conjecture that NYUDv2 train set is too small to fully unleash the pre-training power of CoMAE. In particular, it lacks enough labeled data for explicit semantic supervision during fine-tuning, which is very critical for masked image modeling self-supervised learning paradigms. Therefore, following the previous methods, we transfer the pre-trained model from SUN RGB-D training set and fine-tune on NYUDv2 training set, achieving excellent results, which confirms our conjecture and also verifies the generalization ability of our CoMAE.

Conclusions and Future Works

In this work, we have proposed CoMAE, a data-efficient and single-model self-supervised hybrid pre-training framework for RGB and depth representation learning. Our CoMAE presents a curriculum learning strategy to unify two types of self-supervised learning methods, which is composed of two critical pretext tasks of cross-modal patch-level contrast and multi-modal masked reconstruction. In addition, our single model design without requirement of fusion module is very flexible and robust to generalize to unimodal data, and has significant practical value for scenarios with limited modality available. Extensive experiments on SUN RGB-D and NYUDv2 datasets demonstrate the effectiveness of our CoMAE on multi-modal representation learning, in particular for small-scale datasets. Our CoMAE simply pre-trained on small-scale RGB-D datasets obtains very competitive performance to those models with large-scale pre-training.

As the first method for self-supervised pre-training on small-scale RGB-D datasets, we hope it can be taken as a strong baseline and facilitate further research along this direction, in particular for research groups with limited computational resources. Finally, we believe that our CoMAE has a large space for exploration in negative examples mining and masking strategies of spatial location relationships between modalities. In addition, transferring our pre-trained ViT for downstream dense prediction tasks such as RGB-D semantic segmentation is also possible future work.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62076119, No. 61921006), the Fundamental Research Funds for the Central Universities (No. 020214380091), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.; Chang, S.; Cui, Y.; and Gong, B. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS*, 24206–24221.

Bachmann, R.; Mizrahi, D.; Atanov, A.; and Zamir, A. 2022. MultiMAE: Multi-modal Multi-task Masked Autoencoders. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697 of *Lecture Notes in Computer Science*, 348–367. Springer.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *ICML*, 41–48.

Caglayan, A.; Imamoglu, N.; Can, A. B.; and Nakamura, R. 2022. When CNNs meet random RNNs: Towards multi-level analysis for RGB-D object and scene recognition. *Computer Vision and Image Understanding*, 217: 103373.

Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, 833–851. Springer.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 248–255. IEEE Computer Society.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Du, D.; Wang, L.; Li, Z.; and Wu, G. 2021. Cross-modal pyramid translation for RGB-D scene recognition. *International Journal of Computer Vision*, 129(8): 2309–2327.

Du, D.; Wang, L.; Wang, H.; Zhao, K.; and Wu, G. 2019. Translate-to-recognize networks for RGB-D scene recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11836–11845.

Gao, P.; Ma, T.; Li, H.; Lin, Z.; Dai, J.; and Qiao, Y. 2022. ConvMAE: Masked Convolution Meets Masked Autoencoders. arXiv:2205.03892.

Geng, X.; Liu, H.; Lee, L.; Schuurams, D.; Levine, S.; and Abbeel, P. 2022. Multimodal Masked Autoencoders Learn Transferable Representations. arXiv:2205.14204.

Girdhar, R.; El-Nouby, A.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2022a. OmniMAE: Single Model Masked Pretraining on Images and Videos. arXiv:2206.08356.

Girdhar, R.; Singh, M.; Ravi, N.; van der Maaten, L.; Joulin, A.; and Misra, I. 2022b. Omnivore: A single model for many visual modalities. In *CVPR*, 16102–16112.

Goyal, P.; Dollár, P.; Girshick, R. B.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. arXiv:1706.02677.

Grill, J.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. Á.; Guo, Z.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *NeurIPS*, 21271–21284.

Gupta, S.; Girshick, R.; Arbeláez, P.; and Malik, J. 2014. Learning rich features from RGB-D images for object detection and segmentation. In *European conference on computer vision*, 345–360. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 15979–15988. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Janoch, A.; Karayev, S.; Jia, Y.; Barron, J. T.; Fritz, M.; Saenko, K.; and Darrell, T. 2013. A category-level 3d object dataset: Putting the kinect to work. In *Consumer depth cameras for computer vision*, 141–165. Springer.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In Meila, M.; and Zhang, T., eds., *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 4904–4916. PMLR.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Li, T.; and Wang, L. 2020. Learning Spatiotemporal Features via Video and Text Pair Discrimination. arXiv:2001.05691.

Li, Y.; Zhang, J.; Cheng, Y.; Huang, K.; and Tan, T. 2018. Df 2 net: Discriminative feature learning and fusion network for rgb-d indoor scene classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Likhosherstov, V.; Arnab, A.; Choromanski, K.; Lucic, M.; Tay, Y.; Weller, A.; and Dehghani, M. 2021. PolyViT: Co-training Vision Transformers on Images, Videos and Audio. arXiv:2111.12993.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Miech, A.; Alayrac, J.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *CVPR*, 9876–9886. Computer Vision Foundation / IEEE.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention Bottlenecks for Multimodal Fusion. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS*, 14200–14213.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*, 746–760. Springer.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 567–576. IEEE Computer Society.

Song, X.; Herranz, L.; and Jiang, S. 2017. Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs. In *Thirty-first AAAI conference on artificial intelligence*.

Song, X.; Jiang, S.; Herranz, L.; and Chen, C. 2018. Learning effective RGB-D representations for scene recognition. *IEEE Transactions on Image Processing*, 28(2): 980–993.

Song, X.; Jiang, S.; Wang, B.; Chen, C.; and Chen, G. 2019. Image representations with spatial object-to-object relations for RGB-D scene recognition. *IEEE Transactions on Image Processing*, 29: 525–537.

Tao, C.; Zhu, X.; Huang, G.; Qiao, Y.; Wang, X.; and Dai, J. 2022. Siamese Image Modeling for Self-Supervised Vision Representation Learning. arXiv:2206.01204.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *NeurIPS*, 5998–6008.

Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2016. Modality and component aware feature fusion for rgb-d scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5995–6004.

Wang, L.; Guo, S.; Huang, W.; Xiong, Y.; and Qiao, Y. 2017. Knowledge Guided Disambiguation for Large-Scale Scene Classification With Multi-Resolution CNNs. *IEEE Trans. Image Process.*, 26(4): 2055–2068.

Wang, X.; Zhang, R.; Shen, C.; Kong, T.; and Li, L. 2021. Dense Contrastive Learning for Self-Supervised Visual Pretraining. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 3024–3033. Computer Vision Foundation / IEEE.

Wang, Y.; Ye, T.; Cao, L.; Huang, W.; Sun, F.; He, F.; and Tao, D. 2022. Bridged Transformer for Vision and Point Cloud 3D Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 12104–12113. IEEE.

Xiao, J.; Owens, A.; and Torralba, A. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, 1625–1632.

Xiong, Z.; Yuan, Y.; and Wang, Q. 2020. MSN: Modality separation networks for RGB-D scene recognition. *Neurocomputing*, 373: 81–89.

Xiong, Z.; Yuan, Y.; and Wang, Q. 2021. ASK: Adaptively selecting key local features for RGB-D scene recognition. *IEEE Transactions on Image Processing*, 30: 2722–2733.

Yi, K.; Ge, Y.; Li, X.; Yang, S.; Li, D.; Wu, J.; Shan, Y.; and Qie, X. 2022. Masked Image Modeling with Denoising Contrast. arXiv:2205.09616.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917.

Yuan, Y.; Xiong, Z.; and Wang, Q. 2019. Acm: Adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9176–9184.

Zhou, B.; Lapedriza, À.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning Deep Features for Scene Recognition using Places Database. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *NeurIPS*, 487–495.

Zhu, H.; Weibel, J.-B.; and Lu, S. 2016. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2969–2976.