This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-Modal Contrastive Masked Autoencoders: A Two-Stage Progressive Pre-training Approach for RGBD Datasets

Muhammad Abdullah Jamal and Omid Mohareri
Intuitive Surgical Inc., Sunnyvale, CA

Abstract

In this paper, we propose a new progressive pre-training method for image understanding tasks which leverages RGB-D datasets. The method utilizes Multi-Modal Contrastive Masked Autoencoder and Denoising techniques. Our proposed approach consists of two stages. In the first stage, we pre-train the model using contrastive learning to learn cross-modal representations. In the second stage, we further pre-train the model using masked autoencoding and denoising/noise prediction used in diffusion models. Masked autoencoding focuses on reconstructing the missing patches in the input modality using local spatial correlations, while denoising learns high frequency components of the input data. Moreover, it incorporates global distillation in the second stage by leveraging the knowledge acquired in stage one. Our approach is scalable, robust and suitable for pre-training RGB-D datasets. Extensive experiments on multiple datasets such as ScanNet, NYUv2 and SUN RGB-D show the efficacy and superior performance of our approach. Specifically, we show an improvement of +1.3% mIoU against Mask3D on ScanNet semantic segmentation. We further demonstrate the effectiveness of our approach in low-data regime by evaluating it for semantic segmentation task against the state-of-the-art methods.

1. Introduction

Self-supervised learning (SSL) has paved the way for improving performance in various computer vision tasks such as object recognition [9, 38, 82], semantic segmentation [30, 45, 52] and depth estimation [36] by leveraging extensive unlabeled data and large models such as Vision Transformers (ViT) [26, 54]. There are two self-supervised frameworks commonly used to pre-train large vision models. The first family, contrastive learning [17, 18, 35, 86] pre-trains a model by maximizing the similarity between different augmentations of an image while minimizing the similarity to the other data samples. It essentially learns discriminative representations by enforcing invariance to data augmentations. Some prior works [19, 43, 81, 89] have explored contrastive learning for high-level scene understanding and low-level point matching [8, 88]. The second family of SSL is Masked Image Modeling (MIM) which is inspired by the recent success of masking in the NLP domain [23]. The seminal work in MIM is Masked Autoencoders (MAE) [38] which masks out high number of patches in the input before passing them into a Transformer encoder. Then, a lightweight decoder is used to reconstruct the masked patches in the input. Such reconstruction task provides the model with understanding of the data beyond low-level image statistics. It has achieved state-of-art performance in a variety of image recognition tasks. MAEs have also been extended to pre-train transformers using multi-modal data for tasks such as video-text retrieval [55, 78], image-text retrieval [68] and video-audio-text understanding [1].

However, utilizing MAEs for dense prediction tasks like semantic segmentation and depth estimation requires understanding beyond image level statistics. With the availability of RGB-D or Time-of-Flight (ToF) cameras, various approaches [6, 27, 32, 45] have been developed to improve the dense prediction tasks using multi-modal datasets [4, 14, 21, 62, 74]. A few methods [6, 45] also leverage RGB-D datasets to pre-train ViTs backbones under MAE paradigm. One such work is MultiMAE [6] which pre-trains the vision transformers using RGB, depth and segmentation masks in a multi-task MAE fashion. However, MultiMAE requires semantic segmentation labels during pre-training and access to multiple modalities during fine-tuning for downstream tasks. This method learns cross-modal relationships using a cross-attention module but under the paradigm of MAE to help reconstruct the masked patches in each modality. Another work in this line is Mask3D [45] which embeds 3D priors in ViT backbones to learn structural and geometric priors by reconstructing the masked patches in the depth input using RGB-D data. However, it doesn't learn cross-modal relationships to capture more local context beyond 3D priors.

Motivated by this, we believe contrastive learning and Masked autoencoders can be complementary to each other as both can provide unique data representations by extract-

ing different discriminative features from RGB-D datasets. We have seen recent works that combine instance-level contrastive learning and masked autoencoders and leverage multi-modal data in one single framework for image recognition [47], video action recognition [59], video-text [60] and audio-video [33]. However, such methods are not suitable for RGB-D datasets. For example, dense prediction tasks such as semantic segmentation and depth estimation require understanding beyond image-level representation. We hypothesize that the instance-level contrastive learning would not benefit the aforementioned downstream tasks when leveraging RGB-D datasets for pre-training. So, to capture useful discriminative local context features using these modalities, patch-level or pixel-level contrastive learning would be more beneficial. However, it is not easy to combine patch-level or pixel-level contrastive learning with Masked Autoencoders in a single framework because of the masking step. Recently, CoMAE [84] proposed a single model self-supervised hybrid pre-training framework for small-scale RGB-D datasets. It incorporates contrastive learning and multi-modal masked autoencoding to pre-train a single encoder to capture cross-modal representations. However, it requires both RGB and depth data during fine-tuning and evaluation, as omitting either results in a significant drop in downstream performance. It is also limited to pre-training with small-scale datasets and fails to distill features learned in stage-1, as stage-2 continues to pre-train the encoder using only the low-level pixel reconstruction task. This might not capture rich and diverse representations, and it is an easier task and can be learned quickly because of stage-1 initialization.

Additionally, recent approaches frequently struggle to capture high-frequency components of the data, which further motivates us to explore alternative learning mechanisms. Notably, recent advancements in diffusion models have achieved remarkable success in image and video generation [40–42] by leveraging denoising techniques, a strategy that has been absent in previous approaches, prompting us to investigate its potential in our context. Inspired by this, we introduce a noise prediction loss with the masked autoencoder reconstruction objective. To do so, we add a Gaussian noise in the input and pre-train the model to predict the masked patches as well as the noise. We hypothesize that the denoising would encourage the model to extract high-frequency features from the input.

Based on the above challenges, we propose a new strategy based on these self-supervised frameworks to pre-train ViTs using RGB-D datasets. Our pre-training strategy consists of two stages, 1) contrastive learning at patch-level to align RGB and depth modalities to capture local context for cross-modal relationships 2) MIM + Denoising would encourage the encoders to learn discriminative low-frequency and high-frequency features respectively. Besides, we capitalize on the feature distillation by distilling the global embeddings from the first stage, improving upon MIM + Denoising.

Lastly, we pre-train ViTs on ImageNet [22], ScanNet [21] and SUN RGB-D [74] with our progressive pre-training and evaluate it on multiple downstream tasks.

In summary, our main contributions are:

- We present a new strategy for pre-training ViTs using RGB-D dataset. Our approach provides different flavours of self-supervised frameworks i.e, contrastive learning, masked autoencoding and feature distillation.
- Inspired by the success of denoising in the recent diffusion models, we also propose to add a noise prediction objective to learn different discriminative features.
- We empirically demonstrate the effectivness of our approach on multiple datasets such as ScanNet [21], NYUv2 [62] and SUN RGB-D [74] for downstream tasks like semantic segmentation and depth estimation.
- Finally, we show the data-efficient nature of our pre-training approach by evaluating it under limited labeled data scenarios.

2. Related Work

Contrastive learning based pre-training. SSL consists of designing a pre-text task to learn discriminative representations from unlabeled data. Earlier work on constructing pre-text tasks include colorization [87], jigsaw puzzle [64], and predicting the rotation angle of image [31]. Currently, the two popular approaches for self-supervised learning are constrastive learning and masked image modeling. Contrastive learning groups similar data samples in the feature space while pushing away the dissimilar ones. SimCLR [17] uses contrastive learning to maximize the similarity between different augmentations of an image. MoCo [86] uses dictionary as a queue to facilitate contrastive learning. BYOL [35] uses two networks termed as online and target, to learn from each other by using augmented views of an image. On the other-hand, there are clustering based unsupervised methods that either use k-means assignments as pseudo-labels [10] or enforce similarity between the cluster assignments of different augmented views of the image [11].

Masked Autoencoder based pre-training. Masked Image Modeling (MIM) has received a lot of attention for pre-training vision models since the advent of Vision Transformers (ViT). The goal of MIM is to learn discriminative features for downstream tasks by reconstructing the missing patches in the input. Masked autoencoding for vision tasks is inspired by the success of methods such as BERT [23] and GPT [67] in NLP domain. Several MIM approaches have been proposed with various reconstruction objectives based on pixels [5, 16, 26, 28, 38, 49, 82], dVAE tokens [9, 91]

and features such as histogram of oriented gradients (HOG) or deep neural network embeddings [7, 79]. Specifically, MAE [38] uses an asymmetric encoder-decoder architecture based on ViT to pre-train models using masked autoencoding. It first masks large portions of an image and pass the unmasked patches into encoder followed by a decoder for reconstructing the missing patches. The masking accelerates the pre-training of ViT. Contrastive learning is different from MAE in the sense that MAE learns local spatial dependencies, while contrastive approaches learn representations such as invariance to different augmentations. This steers us to hypothesize that these two SSL frameworks learn complimentary features and therefore we combine them in a multi-stage manner to get the best out of each. Indeed, one can combine them in a single framework [33, 47], but our empirical experiments show poor performance of such methods on RGB-D datasets.

Multi-Modal Learning. It involves learning from multiple modalities such as images and text [2, 13, 20, 46, 51, 75, 83], video and audio [3, 48, 61, 65], video, text and audio [1], and RGB and depth images [32, 50]. Learning can be performed either using modality-specific encoders or shared ones. In terms of MIM, MultiMAE [6] pre-trains the ViT using RGB, depth and semantic segmentation while reconstructing the missing patches in these modalities. It also uses additional modality during the fine-tuning. Mask3D [45] uses RGB and depth modalities under MIM paradigm to encode the 3D priors by reconstructing the missing patches in the depth input. CoMAE [84] proposes a hybrid pre-training framework consists of contrastive and masked autoencoding for small-scale RGB-D datasets.

Denoising. It involves reconstructing clean data given a noisy input. Denoising has recently revolutionized generative models that produce high quality images using diffusion modeling [24, 40]. It consists of forward Gaussian diffusion process and backward generation process. In the forward process, it adds Gaussian noise to the sample while in reverse process, it learns to generate high quality image from the Gaussian noise. This has been used for generating text-conditioned images [63, 69, 72] and videos [41, 42, 73]. The success of denoising hasn't been utilized in representation learning and in this work, we leverage the denoising process to investigate the effectiveness of representations learned from RGB-D data.

Knowledge Distillation. Knowledge distillation [34, 39, 66] aims to transfer knowledge from one model mainly called teacher to the second model called student. It treats the output of the teacher model as the target prediction for the student model. Some approaches [15, 39, 90] use logits of the teacher model to distill knowledge while some methods [56, 71, 77, 85] leverage the intermediate features for distillation. In this paper, we present the first attempt to incorporate feature distillation for RGB-D pre-training.

3. Method

Our approach is a simple combination of constrastive learning, masked autoencoding, denoising used in diffusion models and feature distillation, designed in a two-stage pre-training for vision transformers using RGB-D data. Our goal is to learn representations from different learning mechanisms that can complement each other. We first give an overview of our approach and then explain the learning objectives i.e, contrastive, masked image modeling, noise prediction and distillation. Finally, we describe the overall progressive pre-training pipeline.

3.1. Overview

Given a batch of n RGB-D pair $\{\mathbf{x}^{rgb}, \mathbf{x}^{depth}\}_{i=1}^{n}$, we pre-train ViT [26] using the combination of self-supervised frameworks. Our pre-training consists of two stages. In the first stage, we pre-train our model to learn cross-modal representations to capture correspondence using constrastive learning at patch level. In the second stage, we first initialize the encoders using the weights learned in the previous stage. Then, we use multi-modal masked autoencoder to learn more fine-grained features beyond low-level image statistics to reconstruct the missing patches in the depth input. To pre-train using masked autoencoding, we first split the RGB-D input in $T = (h/p) \times (w/p)$ patches of size $p \times p$: $\mathbf{x}^{rgb}_{i,\text{patch}}, \mathbf{x}^{depth}_{i,\text{patch}} \in \mathbb{R}^{T \times p \times p \times 3}$ to pass them to the encoders. We use modality-specific encoders in the both stages. We then randomly mask patches in both inputs by keeping a percentage of patches and obtain two masks $\mathbf{M}^{rgb}_i, \mathbf{M}^{depth}_i$. Following MAE [38], we pass the unmasked patches to the encoders to obtain the latent representations. The representations are then concatenated with the learnable masked tokens which are essentially the placeholders for these tokens. A lightweight decoder is used to reconstruct the masked patches. Furthermore, we also add denoising to learn the high frequency information of data. Additionally, we utilize global distillation loss to efficiently leverage the knowledge acquired in the first stage. Figure 1 shows the overview of our approach.

3.2. Contrastive learning

To learn the correspondence between RGB and depth modalities, we use contrastive learning to discriminate the instances at patch-level. One can also apply contrastive learning on pixel-level by up-sampling the features to the input resolution. There has been a few works [53, 57] using constrastive learning to learn correspondence between RGB and pointclouds. The depth values can be indirectly

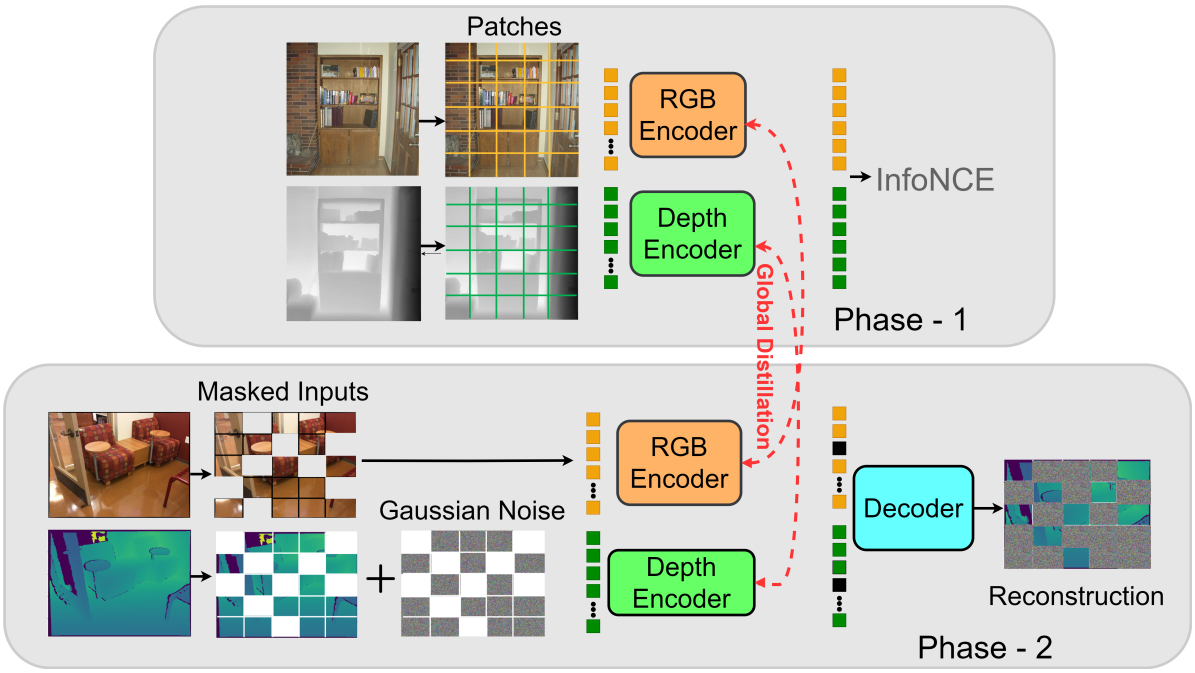


Figure 1. **Overview of our progressive pre-training.** In the first stage, we pre-train the encoders using **Contrastive learning** to align the RGB and the depth patches. In the second stage, we initialize the modality-specific encoders with the stage-1 weights and pre-train them using **Multi-modal Masked autoencoding** and **Denoising** to reconstruct the masked patches in depth input and the noise in the unmasked patches respectively. Moreover, we incorporate **Feature distillation** to leverage the knowledge acquired in the stage-1.

related to the pointclouds and projected to 3D pointclouds using camera parameters. Therefore, we can also leverage the same framework for pixel-level of pixel-to-point [57] to learn the correspondence between RGB and depth. This is left for the future work. Specifically, for a given RGB $\mathbf{x}_i^{rgb}$ and depth $\mathbf{x}_i^{depth}$, we pass them through the ViT encoders to get the latent representations $\mathbf{z}_i^{rgb}, \mathbf{z}_i^{depth}$. We use `Conv2d` as patch projection layer. Following [76], we use InfoNCE loss to pre-train the encoder with a fixed temperature τ:

$$\mathcal{L}_{\text{PNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau)+\exp(s_{i,i}/\tau)}\right] \tag{1}$$

where $s_{i,j} = \|\mathbf{z}_i^{rgb}\|^T\|\mathbf{z}_i^{depth}\|$. $\|\mathbf{z}_i^{rgb}\|$ and $\|\mathbf{z}_i^{depth}\|$ correspond to the encoder features for RGB and depth respectively for a patch or pixel i. In the experiments, we apply the InfoNCE loss between $\mathbf{z}_i^{rgb}$, $\mathbf{z}_i^{depth}$, $\mathbf{z}_i^{depth}$ and $\mathbf{z}_i^{rgb}$, and take the average of the losses. This loss aligns the RGB and depth patch by maximizing the similarity using the latent representations, while minimizing the similarity with other depth patches for a given RGB patch. Indeed, an instance-level contrastive learning can be used but it would fail to capture local discriminative features across modalities as the task becomes much easier by only learning high-level semantics.

3.3. Masked Autoencoding

The goal of masked autoencoding is to understand local spatial statistics of the input modality. To further improve the capability of our approach, we propose to pre-train the model using MAE [38] as a second learning step. We initialize our modality-specific encoders with the stage-1 weights. Unmasked patches of two modalities are passed to the encoders to get latent embeddings. Embeddings are then concatenated with learnable masked tokens which are then passed to the lightweight decoder. We then apply depth prediction MLP head to reconstruct the missing patches. We also experiment with RGB reconstruction task but we found that empirically it doesn't improve the downstream performance (c.f. Table 9). Following [38, 45], we normalize the target patches and compute the loss only on masked patches. Specifically, we use MSE loss which is given as:

$$\mathcal{L}_{\text{depth}} = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{M}_i^{depth}\circ(\mathbf{x}_i^{depth}-\hat{\mathbf{x}}_i^{depth})\|_2^2 \tag{2}$$

3.4. Denoising

Denoising approaches has been shown to be a catalyst for models generating high quality images and videos [40, 41]. Inspired by them, we investigate whether denoising can also be used in self-supervised learning. During pre-training,

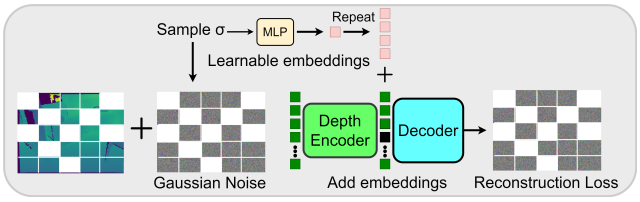


Figure 2. **Denoising:** We first add Gaussian noise to the unmasked patches. We then pass the noise level σ through MLP, and add it to encoded tokens before passing them through the decoder for reconstruction.

we add an independent isotropic Gaussian noise to the depth modality $\mathbf{x}_i^{depth} \leftarrow \mathbf{x}_i^{depth} + \sigma_i^{depth}\mathbf{e}_i^{depth}$ with $\mathbf{e}_i^{depth} \sim \mathcal{N}(\mathbf{0}, I)$ and σ_i^{depth} uniformly sampled from an interval $[0, \sigma_{\text{max}}]$. This noisy input is masked and then passed to the depth encoder as mentioned in section 3.3 which follows MAE [38]. Before passing the encoded representations to the decoder, we also add the information about noise level σ_i^{depth} to help separate the noise from the clean input, which is inspired from the denoising diffusion model [40]. To differentiate, we treat σ_i^{depth} as the positional encoding to the decoder. We first produce a 1D sinusoidal embedding of σ_i^{depth} which is then passed to the MLP with ReLU activation to produce a learnable embedding $\mathbf{p}_i^{depth} \in \mathbb{R}^d$. The dimension of this embedding matches the dimension of the representation from the encoder. We then add this embedding to the encoded embeddings as $(\mathbf{z}_i^{depth}) \leftarrow (\mathbf{z}_i^{depth}) + \mathbf{p}_i^{depth}$ before concatenating it with the RGB encoder representations and learnable masked tokens. The concatenated result is passed to the decoder to produce $\hat{\mathbf{x}}_i^{depth}$. We use the MSE loss and compute the denoising loss on the unmasked patches as:

$$\mathcal{L}_{\text{denoise}} = \frac{1}{n}\sum_{i=1}^{n} \|(1-\mathbf{M}_i^{depth}) \circ (\sigma_i^{depth}\mathbf{e}_i^{depth} - \hat{\mathbf{x}}_i^{depth})\|_2^2 \tag{3}$$

Denoising would encourage the encoder to extract high frequency components which can be complementary to the MAE pre-training in the second stage. The computation of the denoising doesn't bring any overhead to the pre-training compared to the vanilla MAE, since we use a lightweight MLP and compute the loss on the reconstructed unmasked patches. The decoder also reconstructs the unmasked patches which may go wasted in the vanilla MAE pre-training. We also provide an ablation study (c.f. Table 6) on the effect of each component in the denoising stage. Lastly, the figure 2 shows the overall setup of the denoising.

3.5. Feature Distillation

To distill the global modality-specific embeddings in the second stage, we use the output of the encoders as the visual embeddings for each modality. We then apply max pooling to output a global embeddings. Afterwards, we align this second-stage global embedding to the global visual embedding from the first-stage model using smooth ℓ_1 loss which is given as:

$$\mathcal{L}_{\text{distill}}(\mathbf{f_2}, \mathbf{f_1}) = \begin{cases} \frac{1}{2}(\mathbf{f_2} - \mathbf{f_1})^2/\beta, & |\mathbf{f_2} - \mathbf{f_1}| \leq \beta \\ (|\mathbf{f_2} - \mathbf{f_1}| - \frac{1}{2}\beta), & \text{otherwise} \end{cases}, \tag{4}$$

where β is set to 2.0; $\mathbf{f_2}$ and $\mathbf{f_1}$ are the max-pooled features from the stage-2 and stage-1 encoders respectively. In practice, we employ the feature distillation for both RGB and depth. So, the final distillation loss is the sum of modality-specific distillation losses.

3.6. Progressive Pre-training

We are now ready to present our full algorithm for pre-training ViT using RGB-D datasets. In the first stage, we pre-train the model using contrastive loss as mentioned in section 3.2.

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{PNCE}} \tag{5}$$

In the second stage, we pre-train the model using Masked autoencoding, denoising and feature distillation. Therefore, the overall stage-2 objective can be defined as:

$$\mathcal{L}_{\text{stage2}} = \alpha\mathcal{L}_{\text{depth}} + \beta\mathcal{L}_{\text{denoise}} + \gamma\mathcal{L}_{\text{distill}} \tag{6}$$

Where α, β and γ are the weights for each loss function during pre-training. Please see appendix on the details of these hyper-parameters.

4. Results

In this section, we show the efficacy of our approach for ViT-B backbone on semantic segmentation, instance segmentation and depth estimation. In the first stage, we pre-train the ViT-B encoders using ImageNet [22] dataset. We use off the shelf depth estimation model [29] to extract the depths. In the second stage, we continue pre-training the encoders on the ScanNet [21] dataset which contains 2.5M RGB-D frames from 1513 training video sequences. We regularly sample every 25^{th} frame without any filtering during pre-training following Mask3D [45]. To study the effectiveness of our pre-training approach for small scale datasets, we also pre-train the model using SUN RGB-D dataset [74] in the second stage. It contains 10,335 RGB-D images captured from 4 different sensors. We use the official split of train/test for pre-training and evaluation. We also initialize the second stage encoders with the weights learnt in the first stage. Please refer to the appendix for more details on implementations and hyper-parameters.

Methods	Reconstruction task	Backbone	Pre-train	Fine-tune Modality	mIoU
Scratch	-	ViT-B	None	RGB	32.6
Pri3D [44]		ViT-B	ImageNet+ScanNet	RGB	59.3
Pri3D [44]	-	ResNet-50	ImageNet+ScanNet	RGB	60.2
DINO [12]	-	ViT-B	ImageNet+ScanNet	RGB	58.1
MAE [38]	RGB	ViT-B	ImageNet	RGB	64.8
MAE [38]	RGB	ViT-B	ImageNet+ScanNet	RGB	64.5
MultiMAE* [6]	RGB + Depth	ViT-B	ImageNet+ScanNet	RGB	65.1
Mask3D** [45]	Depth	ViT-B	ImageNet+ScanNet	RGB	66.2
Mask3D [45]	RGB + Depth	ViT-B	ImageNet+ScanNet	RGB	65.5
Ours	Depth	ViT-B	ImageNet+ScanNet	RGB	**67.5**
MultiMAE [6]	RGB + Depth + Segmentation	ViT-B	ImageNet	RGB	66.4

Table 1. **ScanNet Semantic Segmentation**. Our approach outperforms Mask3D and other SOTA approaches that leverage RGB-D data during pre-training. * means MultiMAE was pre-trained using ScanNet by initializing the ViT-B encoder with ImageNet SSL weights. ** means ConvNext decoder is used instead of transformer decoder.

4.1. Semantic Segmentation

We first present the results for semantic segmentation. Following MultiMAE [6], we use ConvNext architecture [58] as a segmentation decoder for our approach and the baselines. All the baselines are initialized with ImageNet self-supervised weights [38]. For evaluation, we use mean Intersection over Union (mIoU). Table 1 shows the performance of our approach and other baselines on ScanNet 2D semantic segmentation task. For fine-tuning, we follow Mask3D [45] and sample every 100^{th} frame, resulting in 20,000 training images and 5000 validation images. Our approach achieves 67.5 mIoU compared to 66.2 mIoU for Mask3D and 65.1 mIoU for MultiMAE. For fair comparison, we pre-train MultiMAE with RGB and depth only and use ConvNext decoder for Mask3D. Moreover, compared to MAE pre-trained on ScanNet RGB modality, our approach improves the performance by **+3.0** mIoU. Furthermore, compared to Pri3D, a multi-view contrastive learning approach, we outperforms it by **+7.3** mIoU. Table 2 shows the results on SUN RGB-D 2D semantic segmentation. We observe that our method outperforms other state-of-the-art approaches. More notably, we achieve 48.7 mIoU compared to 47.4 mIoU for Mask3D and 47.1 mIoU for MultiMAE. Table 4 shows the comparison of our approach and CoMAE [84] on the SUN RGB-D dataset in terms of mean accuracy, as CoMAE did not report mIoU results for this dataset. We observe that our approach outperforms the CoMAE by 1.1%. These results demonstrate that our second-stage pre-training is effective even on relatively small-scale RGB-D datasets. We also present results on NYUv2 semantic segmentation, where similar observations can be made when comparing our method with the competing approaches. For further details, please refer to the supplementary materials.

4.2. Depth Estimation

In this section, we study the efficacy of our pre-trained model for depth estimation task. We use NYUv2 [62] for this downstream task and report δ_1 on the test set which shows the percentage of pixels that have an error ratio $(\max\{\frac{\hat{y}_p}{y_p}, \frac{y_p}{\hat{y}_p}\})$ below 1.25 [25]. We use DPT [70] as dense prediction head following MultiMAE for all the approaches and use the model pre-trained on ScanNet dataset for fine-tuning. The results on the depth prediction task are reported in Table 3. We can draw the same observation from the table that our approach outperforms state-of-the-art approaches. Specifically, it achieves 87.1 δ_1 compared to 85.4 δ_1 for Mask3D and 85.3 δ_1 for MultiMAE. We also compare our approach with CroCo [80], a cross-view completion pre-training approach specifically designed for 3D vision tasks. In particular, we achieve an improvement of **+1.5** δ_1 over CroCo. These results again verify the generalizibility of our two-stage pre-training strategy by transferring the representations to various downstream tasks and datasets.

4.3. Instance Segmentation

We also demonstrate the effectiveness of our approach on 2D instance segmentation. We use ScanNet [21] for comparison against other baselines and use MaskRCNN [37] as the instance segmentation head on top of the ViT-B encoder. Table 5 shows the results on 2D instance segmentation. We see the same observation that our approach transfers the learned representations better by outperforming the state-of-the-art approaches including Mask3D and MultiMAE.

4.4. Ablation Studies

Denoising. We consider three settings in this ablation study: 1) Without any noise 2) Input with noise only 3) full method with all the denoising components. Table 6 shows

Methods	Reconstruction task	Backbone	Pre-train	Fine-tune Modality	mIoU
MAE [38]	RGB	ViT-B	ImageNet	RGB	47.0
MAE [38]	RGB	ViT-B	ImageNet+SUN RGB-D	RGB	47.3
Mask3D [45]	Depth	ViT-B	ImageNet+SUN RGB-D	RGB	47.4
MultiMAE [6]	RGB + Depth	ViT-B	ImageNet+SUN RGB-D	RGB	47.1
Ours	Depth	ViT-B	ImageNet+SUN RGB-D	RGB	**48.7**

Table 2. Comparison with the state-of-the-art approaches on **SUN RGB-D Semantic Segmentation**.

Methods	Reconstruction task	Backbone	Pre-train	Fine-tune Modality	δ_1
MAE [38]	RGB	ViT-B	ImageNet	RGB	85.1
Mask3D [45]	Depth	ViT-B	ImageNet+ScanNet	RGB	85.4
CroCo [80]	RGB + Depth	ViT-B	Habitat	RGB	85.6
MultiMAE* [6]	RGB + Depth + Segmentation	ViT-B	ImageNet	RGB	83.0
MultiMAE [6]	RGB + Depth	ViT-B	ImageNet+ScanNet	RGB	85.3
Ours	Depth	ViT-B	ImageNet+ScanNet	RGB	**87.1**
MultiMAE [6]	RGB + Depth + Segmentation	ViT-B	ImageNet	RGB	**86.4**

Table 3. **NYUv2 Depth Estimation**. Ours approach outperforms Mask3D and MultiMAE which shows that learned representations can be transferred to the dense prediction tasks even for different dataset.* means the number are reported from CroCo [80] paper.

Methods	Pre-train	acc_s
CoMAE [84]	SUN RGB-D	64.3
Ours	ImageNet + SUN RGB-D	**65.4**

Table 4. Comparison with CoMAE [84] using average accuracy (acc_s) on **SUN RGB-D Semantic Segmentation**.

Methods	Pre-train	AP
Scratch	None	12.2
ImageNet baseline	Supervised ImageNet	17.6
Pri3D [44]	ImageNet + ScanNet	18.3
MoCov2 [86]	ImageNet + ScanNet	18.3
MAE [38]	ImageNet + ScanNet	20.7
MultiMAE [6]	ImageNet + ScanNet	22.4
Mask3D [45]	ImageNet + ScanNet	22.8
Ours	ImageNet + ScanNet	**23.7**

Table 5. Comparison with other state-of-the-art approaches using average precision (AP) on **ScanNet 2D Instance Segmentation**.

Method	Backbone	mIoU
w/out noise	ViT-B	66.5
noise only	ViT-B	66.9
Full	ViT-B	**67.5**

Table 6. Ablation study on the components of the denoising method. We report the performance on **ScanNet 2D semantic segmentation**.

Contrastive	**Reconstruction**	**Denoising**	**Distill**	**mIoU**
✓	✗	✗	✗	63.4
✓	✓	✗	✗	66.3
✓	✓	✗	✓	66.5
✓	✓	✓	✗	67.0
✓	✓	✓	✓	**67.5**

Table 7. Effect of different pre-training objectives on **ScanNet 2D semantic segmentation**.

the performance of the different components of denoising loss on ScanNet 2D segmentation. It can be seen from the table that by simply adding noise in the input, the performance improves 0.4 mIoU. It is further improved by 0.6 mIoU with full denoising method that consists of noise prediction and positional encoding in the decoder.

Effect of pre-training objectives. In Table 7, we highlight the importance of using different self-supervised frameworks. It shows that contrastive learning alone can perform reasonably well. By pre-training using only depth reconstruction loss in the second stage, the performance jumps from 63.4 to 66.3 which is on-par with Mask3D while integrating all three objectives in the stage-2 significantly enhances the final performance. Lastly, without our feature distillation, the performance drops from 67.5 to 67.0 mIoU.

RGB ratio	Depth ratio	mIoU
20.0%	20.0%	65.6
20.0%	50.0%	65.2
20.0%	80.0%	65.5
50.0%	20.0%	65.9
50.0%	50.0%	65.4
80.0%	20.0%	65.9
80.0%	80.0%	**67.5**

Table 8. Effect of different masking ratios for RGB and depth modality on **ScanNet 2D semantic segmentation**.

Effect of masking ratio during pre-training. Table 8 shows the performance on ScanNet 2D semantic segmentation by varying the masking ratio of RGB and depth modalities. It can be seen that our approach achieves the best performance with higher masking ratio for both of the modalities, which accelerates the pre-training.

Loss	Backbone	mIoU
Depth + RGB Reconstruction	ViT-B	66.9
Full	ViT-B	**67.5**

Table 9. Ablation study of RGB and Depth reconstruction in the second stage on **ScanNet 2D semantic segmentation**.

Effect of RGB + Depth Reconstruction. We also study the effect of reconstructing the RGB input besides the depth modality. It can be seen from the table 9 that it doesn't improve the performance. A plausible explanation for this is that reconstructing the RGB modality during pre-training is relatively simpler, as the 3D priors are encoded through the depth prediction process, providing additional reinforcement.

Method	Backbone	mIoU
MAE [38]	ViT-B	64.8
MultiMAE [6]	ViT-B	65.1
Ours	ViT-B	**67.5**
MAE [38]	ViT-L	68.2
MultiMAE [6]	ViT-L	69.3
Ours	ViT-L	**70.8**

Table 10. Performance on **ScanNet 2D semantic segmentation** using variants of ViT.

Results on different ViT variants. In the main manuscript, we showed the comparison of different approaches using ViT-B architecture. In Table 10, we show the performance on ScanNet 2D segmentation using ViT-L. We can infer the same observation that our approach outperforms the baselines that include MAE and MultiMAE.

4.5. Low-data Regime Scenarios

We show the data-efficient nature of our pre-training strategy by fine-tuning it in limited labeled data scenarios. Figure 3 shows the performance on ScanNet 2D semantic segmentation. Our approach consistently outperform Mask3D, MAE and Pri3D across different ranges of limited labeled data. More notably, it marginally outperforms MAE with only 60% of the original training set.

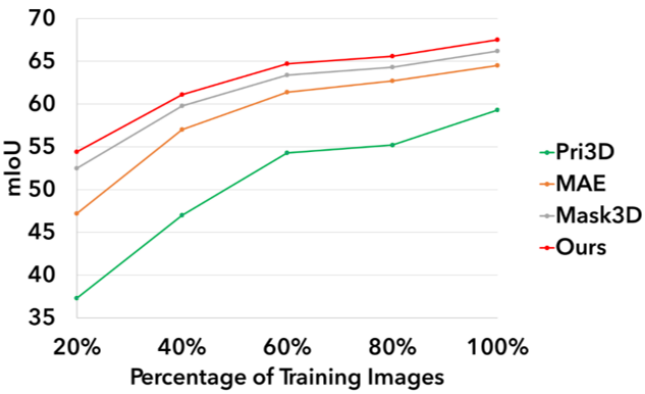


Figure 3. To show data-efficiency feature of our approach, we compare with recent state-of-the-art pre-training approaches on **ScanNet 2D semantic segmentation** under limited labeled data scenarios.

5. Conclusion

We propose a new progressive pre-training strategy that effectively combines powerful self-supervised learning methods in a single framework to pre-train ViTs using RGB-D datasets. Our pre-training consists of two phases of learning: (1) contrastive learning which aligns the RGB and depth patches to extract discriminative local context features, (2) Masked Image Modeling which reconstructs the masked patches in the depth modality using RGB-D input to learn spatial statistical correlations and embed 3D priors. To learn high frequency features, we introduce a denoising loss inspired by the recent success of diffusion models to predict the noise added in the input. To further leverage the knowledge learnt in the stage-1, we employ feature distillation based on the global embeddings. We demonstrate the effectiveness of our approach on various downstream tasks such as semantic segmentation, depth prediction and instance segmentation. Finally, we show scalability and data-efficiency by pre-training the second stage using small scale RGB-D datasets.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021. 1, 3

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks, 2020. 3

[3] Relja Arandjelović and Andrew Zisserman. Look, listen and learn, 2017. 3

[4] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[5] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer, 2022. 2

[6] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. 2022. 1, 3, 6, 7, 8

[7] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022. 3

[8] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences, 2021. 1

[9] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 1, 2

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features, 2019. 2

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924. Curran Associates, Inc., 2020. 2

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 6

[13] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data, 2016. 3

[14] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1

[15] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[16] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020. 2

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2

[18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 1

[19] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding, 2022. 1

[20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020. 3

[21] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2, 5, 6

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2, 5

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019. 1, 2

[24] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 3

[25] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning, 2017. 6

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[27] Dapeng Du, Limin Wang, Zhaoyang Li, and Gangshan Wu. Cross-modal pyramid translation for rgb-d scene recognition. *International Journal of Computer Vision*, 2021. 1

[28] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training?, 2021. 2

[29] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[30] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 1

[31] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. 2

[32] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities, 2022. 1, 3

[33] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[34] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021. 3

[35] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. 1, 2

[36] Vitor Guizilini, Rares Ambrus, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers, 2022. 1

[37] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 6

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[39] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3

[40] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2, 3, 4, 5

[41] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 3, 4

[42] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2, 3

[43] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts, 2021. 1

[44] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning?, 2021. 6, 7

[45] Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors, 2023. 1, 3, 4, 5, 6, 7

[46] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer, 2021. 3

[47] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners, 2022. 2, 3

[48] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. 3

[49] Muhammad Abdullah Jamal and Omid Mohareri. Surgmae: Masked autoencoders for long surgical video analysis, 2023. 2

[50] Muhammad Abdullah Jamal and Omid Mohareri. M33d: Learning 3d priors using multi-modal masked autoencoders for 2d image and video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[51] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all, 2017. 3

[52] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders, 2022. 1

[53] Lanxiao Li and Michael Heizmann. A closer look at invariances in self-supervised pre-training for 3d vision, 2022. 3

[54] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection, 2022. 1

[55] Yizhuo Li, Kunchang Li, Yinan He, Yi Wang, Yali Wang, Limin Wang, Yu Qiao, and Ping Luo. Harvest video foundation models via efficient post-pretraining, 2023. 1

[56] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring interchannel correlation for diversity-preserved knowledgedistillation, 2022. 3

[57] Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining, 2021. 3, 4

[58] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 6

[59] Cheng-Ze Lu, Xiaojie Jin, Zhicheng Huang, Qibin Hou, Ming-Ming Cheng, and Jiashi Feng. Cmae-v: Contrastive masked autoencoders for video action recognition, 2023. 2

[60] Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. Simvtp: Simple video text pre-training with masked autoencoders, 2022. 2

[61] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion, 2022. 3

[62] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2, 6

[63] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 3

[64] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2017. 2

[65] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features, 2018. 3

[66] Mary Phuong and Christoph H. Lampert. Towards understanding knowledge distillation, 2021. 3

[67] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 2

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[69] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3

[70] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 6

[71] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. 3

[72] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3

[73] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 3

[74] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 5

[75] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019. 3

[76] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 4

[77] Can Wang, Defang Chen, Jian-Ping Mei, Yuan Zhang, Yan Feng, and Chun Chen. Semckd: Semantic calibration for cross-layer knowledge distillation. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 3

[78] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. *arXiv preprint arXiv:2309.10091*, 2023. 1

[79] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 3

[80] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 6, 7

[81] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding, 2020. 1

[82] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[83] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning, 2021. 3

[84] Jiange Yang, Sheng Guo, Gangshan Wu, and Limin Wang. Comae: Single model hybrid pre-training on small-scale rgb-d datasets, 2023. 2, 3, 6, 7

[85] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[86] Chaoning Zhang, Kang Zhang, Trung X. Pham, Axi Niu, Zhinan Qiao, Chang D. Yoo, and In So Kweon. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco, 2022. 1, 2, 7

[87] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016. 2

[88] Yu Zhang, Junle Yu, Xiaolin Huang, Wenhui Zhou, and Ji Hou. Pcr-cg: Point cloud registration via deep color and geometry, 2023. 1

[89] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud, 2021. 1

[90] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation, 2022. 3

[91] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. 2