M^33D: Learning 3D priors using Multi-Modal Masked Autoencoders for 2D image and video understanding

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

Abstract

We present a new pre-training strategy called M^33D (Multi-Modal Masked 3D) built based on Multi-modal masked autoencoders that can leverage 3D priors and learned cross-modal representations in RGB-D data. We integrate two major self-supervised learning frameworks: Masked Image Modeling (MIM) and contrastive learning; aiming to effectively embed masked 3D priors and modality complementary features to enhance the correspondence between modalities. In contrast to recent approaches which are either focusing on specific downstream tasks or require multi-view correspondence, we show that our pre-training strategy is ubiquitous, enabling improved representation learning that can transfer into improved performance on various downstream tasks such as video action recognition, video action detection, 2D semantic segmentation and depth estimation. Experiments show that M^33D outperforms the existing state-of-the-art approaches on ScanNet, NYUv2, UCF-101 and OR-AR, particularly with an improvement of +1.3% mIoU against Mask3D on ScanNet semantic segmentation. We further evaluate our method on low-data regime and demonstrate its superior data efficiency compared to current state-of-the-art approaches.

1. Introduction

Pre-training Vision Transformers [24] (ViTs) using Masked Autoencoders (MAEs) followed by fine-tuning gives rise to the state-of-the-art results for various computer vision tasks such as image classification [9, 33, 66], video activity recognition [26, 60, 62], semantic segmentation [27, 36, 45] and 3D scene understanding [14, 52, 70]. Inspired by BERT [21], MAE masks high number of patches in an input image or video clip and predicts such missing regions. It usually consists of an asymmetric encoder-decoder architecture, where the encoder objective is to encode the unmasked patches into latent representations while the masked patches are predicted by the decoder using these representations. However, MAEs are mostly focused on learning from single-modality data (images or video clips) and can’t leverage other data modalities that are present in commonly used RGB-D and Time-of-Flight (ToF) cameras. The depth and point cloud data available in such cameras can provide an opportunity to learn geometric priors and avoid view-dependent effects for efficient representation learning. Particularly, various 3D understanding approaches [18, 37, 65, 73] have been leveraging RGB-D datasets for high-level scene understanding and low-level point matching [7, 72] tasks using contrastive learning. However, very little work has been done in exploring 3D priors for 2D image understanding with limited focus on tasks like semantic segmentation and depth estimation.

Our goal is to embed 3D priors into ViTs based backbones for various downstream tasks that include but not limited to video activity recognition, video action detection, semantic segmentation and depth estimation, to effectively learn geometric and structural priors as well as cross-modal features. Recently, Mask3D [36] effectively learns masked 3D priors from a single-view RGB-D frame by only reconstructing the depth through masking out patches in RGB and depth modalities. However, the approach is limited to 2D image understanding tasks and relies on MAE pre-training for any cross-modal representation learning. Pri3D [38] explores to embed the 3D prior by leveraging multi-view RGB-D data to introduce geometric constraints in a contrastive learning scheme for ResNet backbones. However, it relies on camera registration between multiple views for each frame. Instead, we consider to empower ViTs with such 3D priors using MAEs and cross-modal training strategies such as contrastive learning and matching loss that transfer well to various video and image based downstream tasks. Cross-modal learning and Masked Image Modeling are complementary to each other. While the former would leverage the very useful visual-depth pair information to encode modality-invariant features, the latter forces its representation to encode the majority of the input information. This motivates us to integrate both the pre-training strategies to extract different discriminative features.

In this paper, we present M^33D, a multi-modal masked autoencoder based approach which learns to embed 3D priors in a self-supervised manner by pre-training on RGB-D data.
image pair or video clips and also integrates cross-modal learning. To train M^3D, we randomly mask patches in both RGB and depth modalities and encode the unmasked patches by passing them through either modality-specific transformer encoders or modality-agnostic transformer encoder which is shared between both the modalities. The representations from the encoder are then appended with learned masked tokens to reconstruct masked regions in depth and RGB input. We equipped our approach with contrastive learning and RGB-D matching which are used to enhance cross-modal correspondence during pre-training. The objective of contrastive learning is to align the visual scene and the corresponding depth data by pulling the feature associated with the RGB patch and its corresponding depth while pushing away the other depth patches in the corresponding depth map. Finally, to further enhance the correspondence, the matching loss predicts whether the RGB-depth pair is matched (positive) or not matched (negative) by applying a linear layer followed by softmax on the top of the encoder features to predict a two-class probability. Our experiments demonstrate the effectiveness of M^3D on a variety of datasets for video and image understanding. We pre-train on UCF-101 [57], OR-AR [35] and fine-tune for video activity recognition, and video action detection respectively. We also pre-train the model with ScanNet [20] and fine-tune it for semantic segmentation. We also show the generalizibility of the model on NYUv2 [49] data for depth estimation besides semantic segmentation. In summary, the main contributions of our paper are:

- We propose a new self-supervised pre-training approach called M^3D which is equipped with masked autoencoders and cross-modal training for RGB-D representation learning. Our method can be applied to various video and image understanding tasks based on single view RGB-D image pair or video clips.
- Our approach learns to embed 3D priors without any camera pose registration, while enhancing the cross-modal correspondence between RGB and depth modalities using contrastive learning and matching loss.
- We demonstrate the efficacy of M^3D for current ViT backbones on variety of video (UCF-101, OR-AR surgical dataset) and image understanding (ScanNet, NYUv2) datasets.
- We further evaluate our method on low-data regime and demonstrate its superior data efficiency compared to current state-of-the-art approaches. Several ablation studies have been conducted and presented here to prove the effectiveness of various aspects of M^3D.

2. Related Work

Our work proposes a self-supervised learning algorithm for general computer vision tasks. We will review some of the approaches that are closely related to our approach from different aspects - Masked Autoencoder based pre-training for vision transformers, Multi-modal learning and pre-training by learning 3D priors.

**Masked Autoencoder based pre-training for transformers.** Self-supervised learning paradigm learns visual features based on some pretext task using large-scale unlabeled dataset, before fine-tuning for multiple downstream tasks. Earlier work on pretext task includes colorization [71], jigsaw puzzle [50], and predicting image rotation [28]. Instance-discrimination based learning uses two different views or augmentations of an image to learn view-invariant features by pulling them closer in the embedding space while pushing away the negative pair. SimCLR [16], MoCo [69] uses contrastive learning, BYOL [31] uses online and target network while SwAV [11], DeepCluster [10] uses clustering to discriminate between the clusters. Recently, the momentum in self-supervised learning has been shifted towards the masked image modeling which learns useful representations by learning to reconstruct the masked image patches. Inspired by the success in NLP domain such as bidirectional encoder (BERT) [21] and Generative Pre-Training (GPT) [54], several masked image prediction methods for pre-training vision transformers have been proposed with various reconstruction targets such as pixels [4, 15, 24, 25, 33, 66], features [6, 63], discrete tokens via dVAE [9, 74]. They have shown to outperform contrastive learning based approaches on various downstream tasks. One particular approach is MAE [33] which first masks large portion of the input (e.g., 75%) and then passes the unmasked patches to the large encoder followed by a lightweight decoder that reconstructs the masked patches. By masking out large portion of input patches, it accelerates the pre-training of vision transformers. Moreover, these approaches have been extended to the video domain where most of focus is to design a masking strategy that includes random masking [26], tube masking [60] and adaptive masking based on motion [41, 58] or a separate network [8]. On the other-hand, our work proposes multi-modal masked auto-encoder pre-training by learning 3D priors for several downstream tasks such as video action detection, video activity recognition, 2D semantic segmentation and depth estimation.

**Multi-Modal Learning.** It involves training the models by leveraging data from multiple modalities. The learning may involve training modality-agnostic unified encoder or modality-specific encoders using modalities such as im-
ages and text [2, 13, 19, 39, 42, 59, 68], video and audio [3, 40, 48, 51], video, text and audio [1], and images and depth [29]. One particular approach is MultiMAE [5] which extends MAE to multi-modal multi-task settings and includes depth modality. However, it requires depth during the fine-tuning stage as well as semantic segmentation labels during the pre-training. Our approach doesn’t rely on semantic segmentation, it only uses depth during the pre-training and it is not limited to the image domain only. Please refer to experimental section 4 for more results on MultiMAE without semantic segmentation modality.

Pre-training by learning 3D priors. There is a recent surge of learning cross-modal features particularly between languages and images. CLIP [53] learns visual concepts from natural language supervision, showing impressive zero-shot performance for image classification. Pri3D [38] learns view-invariant and geometry-aware representations using multi-view consistency during pre-training for 2D image understanding. It leverages contrastive learning for 2D-3D correspondence and embed 3D priors to pre-train ResNet [34] architectures. However, our approach is more ubiquitous, it is not limited to 2D image understanding tasks and doesn’t require camera pose registration across views. Mask3D [36] proposes a masked auto-encoder based self-supervised learning for ViT backbones to learn masked 3D priors for 2D image understanding without the reconstruction of RGB masked patches. In contrast, we formulate a self-supervised pre-training that can operate on both single-view images and videos and leverages masked 3D priors for several downstream tasks. Our approach also enhances the cross-modal learning and 2D-3D correspondence with contrastive learning and matching loss during pre-training which consequently boost the transfer learning performance.

3. Method

We propose M$^3$D to embed 3D priors, capture modality complementary feature and learn cross-modal representation by self-supervised pre-training using a single-view RGB-D image or a video clip. To learn such priors and representations, we formulate the problem from Masked autoencoder perspective and cross-modal training strategies.

Given the RGB-D input, we first the masked the modalities, and pass them through the modality-specific or modality-agnostic encoder. The latent representations from the encoder(s) are then used to reconstruct the masked RGB and depth input. To learn the cross-modal representations, we use contrastive learning and matching loss which are the two most common cross-modal training strategies. The objective of depth reconstruction task is to embed geometric awareness into the encoder while cross-modal training enhances cross-modal representation learning. After pre-training, the encoder can be fine-tuned for downstream tasks such video action recognition, 2D semantic segmentation, depth estimation etc. An overview of our approach is show in Figure 1. Below, we first explain the masking image modeling strategy from multi-modal input perspective. Then, we explain two types of cross-modal training strategies: contrastive, and matching loss. Finally, we describe the overall pre-training pipeline.

3.1. Multi-modal Masked Autoencoder

Our approach is not limited to the images and can be extended to the videos as well. Please see section 4 for performance on the video downstream tasks. For ease of notations, we describe our approach with a single-view RGB-D frame as an input. Given a RGB frame $I_R$ of size $3 \times H \times W$ and its corresponding depth map $I_D$ of size $1 \times H \times W$, we first divide them into $P \times P$ patches from which we randomly keep a percentage of patches by masking out others for both RGB and depth input. $P \times P$ patches are projected to correct dimension using a different linear projection layer for each modality. Please refer to supplementary material on masking strategies. We use Conv2d(Conv3d in case of videos) as a linear projection layer. We then add positional embeddings and pass the tokens correspond to the unmasked patches to the modality-specific encoders or shared encoder between the modalities to get the latent representations. We use ViT-B as an encoder. To reconstruct the masked-out tokens, we use a lightweight shared decoder based on transformers with two prediction head; one for each modality. The input to the decoder is the latent representations of the unmasked patches and the set of mask tokens which are actually placeholders for the decoder to reconstruct the masked patches at the original image resolution. Following [5], we add sine-cosine positional embeddings before passing them to the decoder. From the reconstruction output, we compute the losses only on the masked tokens.

Reconstruction Losses. Our model has two reconstruction loss; one for each modality. Following [35], we first normalize the output patches as well as target patches, and then used compute the MSE loss between the ground truth and the predicted pixels.

$$L_{rgb} = \frac{1}{\omega} \sum_{p \in \omega} ||I_R(p) - \hat{I}_R(p)||_2$$

where $p$ is the token index, $\omega$ is the set of masked tokens. $I_R$ corresponds to the reconstruction or the prediction of the model.

Similarly, for depth reconstruction task, we follow MultiMAE [5] to use L1 loss for image based pre-training. For
Figure 1. Overview of M$^3$3D pre-training. We introduce two self-supervised learning frameworks; Masked Image Modeling and contrastive learning. For MIM pre-training, we first mask out patches from RGB and depth input and then linearly project the unmasked patches to tokens with a fixed dimension. The tokens are then encoded using a Transformer encoder. The decoder takes the latent representations from the encoder and reconstruct the mask-out patches. Finally, we apply contrastive loss and matching loss on the encoded representations to enable cross-modal correspondence.

(Left) M$^3$3D pre-training setup for video understanding. (Right) M$^3$3D two-stage pre-training setup for image understanding.

For video based pre-training, we use MSE Loss.

$$L_{\text{depth}} = \frac{1}{\lambda \lambda} \sum_{m \in \Lambda} \| I_D(m) - \hat{I}_D(m) \|$$

where $m$ is the token index, $\lambda$ is the set of masked tokens. $I_D$ corresponds to the reconstruction or the prediction of the model.

3.2. Cross-Modal Representation Learning

Besides pre-training our model using masked image modeling, we also propose to use two cross-modal training strategies i.e. RGB-Depth contrastive learning and RGB-Depth matching. The objective of these strategies is to enhance the cross-modal representation learning.

**Contrastive Loss.** Inspired by the success of contrastive learning [47, 53], we present our cross-modal contrastive pre-training. For each RGB-Depth image pair, we use the latent representations from the transformer encoder. Specifically, we use InfoNCE loss [61] with temperature $\tau$ to pre-train the model. The goal of RGB-D contrastive learning is to align the RGB patch and the corresponding depth patch by pulling them closer and repelling unpaired ones apart by minimizing the similarity with the other patches in the corresponding map. Indeed, one can use instance-level contrastive learning, but we hypothesize it will fail to capture high-level information as it is relatively an easy pre-text task.

**Matching Loss.** Inspired by video-text matching, we also propose to use RGB-Depth matching to further enhance the cross-modal correspondence. It predicts whether the RGB-Depth pair is matched (positive) or not (negative). Specifically, we reuse the features from the encoder, and then passed them to the linear layer followed by a softmax classifier to solve two-class classification problem.

The final pre-training loss of M$^3$3D for video based understanding is,

$$L_\text{video} = \alpha L_{\text{rgb}} + \beta L_{\text{depth}} + \gamma L_c + \eta L_m$$
where \( \alpha, \beta, \gamma, \eta \) are the hyper-parameters tuned during pre-training. However, for image based understanding tasks, we pre-train our model in two-stages as shown in Figure 1. In the first stage, we pre-train the encoder using contrastive loss for few epochs and then in the second stage, we initialize the weights of the encoder with stage-1 weights, and then pre-train the encoder and decoder using masked image modeling.

\[
L_{\text{stage1}} = L_c \quad (6)
\]
\[
L_{\text{stage2}} = \alpha L_{\text{rgb}} + \beta L_{\text{depth}} \quad (7)
\]

4. Experiments

M\(^3\)3D aims to learn masked 3D priors to embed to ViT backbones and enhance the fine-tuning performance with further cross-modal learning using contrastive learning and matching loss. In this section, we first briefly describe the experimental setup for pre-training stage. Then, we provide a transfer study to measure the effectiveness of our approach across various downstream tasks. Moreover, we show the data-efficient nature of our approach by evaluating it under low-data regime setting. Finally, we present some ablation studies to study the effectiveness of each component in our approach.

4.1. Experimental Setup

**Pre-train:** We use vision-transformer base (ViT-B) as a encoder for all the experiments. For video activity recognition, we pre-train the model on UCF-101 [57] dataset, one of the benchmark datasets for video action recognition. We use monodepth2 [30] to extract the depths from the videos as original dataset contains RGB videos only. Please refer to [30] for more details on the approach. We use a lightweight decoder which consists of 4 blocks following VideoMAE [60]. We initialize the model weights with network weights trained on Kinetics-400 [43]. To maintain the self-supervised paradigm, we initialize the model with weights obtained by self-supervised Kinetics-400 pre-training [60]. We also show the efficacy of our approach by pre-training it from scratch. For surgical video action detection, we pre-train the model on OR-AR [35] dataset which consists of 820 full videos each having 9 temporal workflow phase labels. This dataset is collected using ToF cameras placed in two operating rooms in a single hospital. For 2D image understanding task, we pre-train the model on the ScanNet [20] dataset. ScanNet contains 2.5M RGB-D frames from 1513 video sequences. We regularly sample every 25\(^{th}\) frame without any filtering during pre-training. We initialize the encoder weights with the self-supervised ImageNet pre-training [33]. Please refer to the supplementary material for details on datasets, pre-training and hyper-parameters.

4.2. Results

4.2.1 Fine-tuning for video action recognition

We fine-tune M\(^3\)3D based ViT-B encoder on UCF-101. Table 1 shows the top-1 accuracy comparison with the recent state-of-the-art video based pre-training approaches. It can be seen from the table that our approach achieves 96.3% top-1 accuracy outperforming all the baselines including VideoMAE [60] which shows the effectiveness of leveraging RGB-D data to embed 3D priors for video understanding. Moreover, we also report the results when we pre-train the model from scratch. Although, it achieves the same performance as VideoMAE, M\(^3\)3D is computationally less expensive than VideoMAE as it requires less epochs and higher masking ratio during pre-training. For fair comparison, we also report the performance of VideoMAE when it is pre-trained for less epochs with higher masking ratio (e.g. 0.90). It is clearly seen from the table that our approach outperforms VideoMAE under the same pre-training epochs and masking ratio. To the best of our knowledge, M\(^3\)3D provides the first ViT-B backbone that is pre-trained from scratch under self-supervision using RGB-D for video action recognition. It is worth mentioning that the depth maps are extracted from an off-the-shelf model and we hypothesize that the performance will further enhance with an improved depth estimation model or readily available depth data.

4.2.2 Fine-tuning for surgical video action detection

We fine-tune M\(^3\)3D based ViT-B encoder on OR-AR, a benchmark dataset for surgical video action detection. Table 2 shows the results compared to the recent state-of-the-art video based masked autoencoder pre-training approaches. Following common practice [35, 56], we report the mean average precision (mAP). Our approach consistently outperforms the existing video based MAE models on low-data regime. More specifically, M\(^3\)3D achieves 80.90 mAP when fine-tuned using 5% labeled data which also shows that it is a more data-efficient learner than VideoMAE. Moreover, we report the results of VideoMAE when it is initialized using Kinetics-400 pre-trained weights. The weights can be found from VideoMAE’s github repository. Similar to video action recognition, it is the first multi-modal based transformer for surgical action detection which shows the effectiveness of the M\(^3\)3D on datasets in the medical domain. Finally, we report the results of Spatio-temporal MAE [26] with different masking strategies such as random, frame, etc. and they are termed as MAE-Random, MAE-Frame in the table.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Masking ratio</th>
<th>Backbone</th>
<th>Pre-train Epochs</th>
<th>Dataset</th>
<th>Frames</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>-</td>
<td>ViT-B</td>
<td></td>
<td>UCF-101</td>
<td>16</td>
<td>51.4</td>
</tr>
<tr>
<td>OPN [44]</td>
<td>-</td>
<td>VGG</td>
<td></td>
<td>UCF-101</td>
<td>N/A</td>
<td>59.6</td>
</tr>
<tr>
<td>VCOP [67]</td>
<td>-</td>
<td>R(2+1)D</td>
<td></td>
<td>UCF-101</td>
<td>N/A</td>
<td>72.4</td>
</tr>
<tr>
<td>CoCLR [32]</td>
<td>-</td>
<td>S3D-G</td>
<td></td>
<td>UCF-101</td>
<td>32</td>
<td>81.4</td>
</tr>
<tr>
<td>ViCLR [22]</td>
<td>-</td>
<td>S3D</td>
<td></td>
<td>UCF-101</td>
<td>32</td>
<td>82.8</td>
</tr>
<tr>
<td>CoCLR [32]</td>
<td>-</td>
<td>S3D-G</td>
<td></td>
<td>Kinetics-400</td>
<td>32</td>
<td>87.9</td>
</tr>
<tr>
<td>ViCLR [22]</td>
<td>-</td>
<td>S3D</td>
<td></td>
<td>Kinetics-400</td>
<td>32</td>
<td>89.1</td>
</tr>
<tr>
<td>VideoMAE [60]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>100</td>
<td>Kinetics-400 + UCF-101</td>
<td>16</td>
<td>96.3</td>
</tr>
<tr>
<td>M^3D [36]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>800</td>
<td>UCF-101</td>
<td>16</td>
<td>90.1</td>
</tr>
<tr>
<td>VideoMAE [60]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>3200</td>
<td>UCF-101</td>
<td>16</td>
<td>90.7</td>
</tr>
<tr>
<td>VideoMAE [60]</td>
<td>0.75</td>
<td>ViT-B</td>
<td>3200</td>
<td>UCF-101</td>
<td>16</td>
<td>91.2</td>
</tr>
</tbody>
</table>

Table 1. Video action recognition accuracy on UCF-101 [57]. † means we re-run the code using the original repository provided by VideoMAE.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Masking ratio</th>
<th>Backbone</th>
<th>Pre-train</th>
<th>Fine-tune Modality</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskFeat [63]</td>
<td>0.90</td>
<td>MViT-S</td>
<td>Kinetics-400</td>
<td>RGB</td>
<td>62.35</td>
</tr>
<tr>
<td>MAE-Random [26]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>Kinetics-400</td>
<td>OR-AR</td>
<td>64.66</td>
</tr>
<tr>
<td>MAE-Random [26]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>OR-AR</td>
<td>66.38</td>
<td>81.87</td>
</tr>
<tr>
<td>MAE-Frame [26]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>OR-AR</td>
<td>63.44</td>
<td>78.89</td>
</tr>
<tr>
<td>VideoMAE [60]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>OR-AR</td>
<td>65.57</td>
<td>81.74</td>
</tr>
<tr>
<td>VideoMAE [60]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>Kinetics-400 + OR-AR</td>
<td>70.93</td>
<td>83.73</td>
</tr>
<tr>
<td>M^3D [36]</td>
<td>0.90</td>
<td>ViT-B</td>
<td>800</td>
<td>RGB</td>
<td>80.90</td>
</tr>
</tbody>
</table>

Table 2. Comparison of M^3D with the other state-of-the-art methods under different data-regime setting on the OR-AR dataset [35].

<table>
<thead>
<tr>
<th>Methods</th>
<th>Reconstruction task</th>
<th>Backbone</th>
<th>Pre-train</th>
<th>Fine-tune Modality</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td></td>
<td>ViT-B</td>
<td>None</td>
<td>RGB</td>
<td>32.6</td>
</tr>
<tr>
<td>MultiMAE* [5]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>64.0</td>
</tr>
<tr>
<td>Pri3D [38]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>59.3</td>
</tr>
<tr>
<td>DINO [12]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>60.2</td>
</tr>
<tr>
<td>MAE [33]</td>
<td>RGB</td>
<td>ViT-B</td>
<td>ImageNet</td>
<td>RGB</td>
<td>64.8</td>
</tr>
<tr>
<td>Mask3D [36]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>66.0</td>
</tr>
<tr>
<td>M^3D [36]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>67.3</td>
</tr>
<tr>
<td>MultiMAE [5]</td>
<td>RGB + Depth + Segmentation</td>
<td>ViT-B</td>
<td>ImageNet</td>
<td>RGB</td>
<td>66.4</td>
</tr>
</tbody>
</table>

Table 3. ScanNet Semantic Segmentation. M^3D outperforms Mask3D and other state-of-the-art approaches that leverage RGB-D data during pre-training.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Reconstruction task</th>
<th>Backbone</th>
<th>Pre-train</th>
<th>Fine-tune Modality</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiMAE* [5]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>49.1</td>
</tr>
<tr>
<td>MAE [33]</td>
<td>RGB</td>
<td>ViT-B</td>
<td>ImageNet</td>
<td>RGB</td>
<td>46.9</td>
</tr>
<tr>
<td>Mask3D [36]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>50.5</td>
</tr>
<tr>
<td>M^3D [36]</td>
<td>RGB + Depth</td>
<td>ViT-B</td>
<td>ImageNet+ScanNet</td>
<td>RGB</td>
<td>51.2</td>
</tr>
</tbody>
</table>

Table 4. NYUv2 Semantic Segmentation. M^3D outperforms state-of-the-art approaches that leverage RGB-D data during pre-training. It demonstrates the effectiveness in transferring to out-domain dataset. † means we re-run the code using the original repository provided by MultiMAE.
4.2.3 Fine-tuning for 2D semantic segmentation

In this section, we show that our pre-training approach is capable of transferring the learned representations to the 2D image understanding tasks by performing the 2D semantic segmentation task. Following MultiMAE [5], we use segmentation head based on ConvNext architecture [46] on top of the ViT-B encoder. We use mean Intersection over Union (mIoU) as an evaluation metric. For fine-tuning, we sample every 100th frame, resulting in 20,000 training images and 5000 validation images which is common protocol of the ScanNet benchmark [20]. Table 3 shows the results compared to the recent state-of-the-art pre-training approaches for ScanNet semantic segmentation. It can be observed from the table that M3D significantly outperforms Mask3D [36] (+1.3 mIoU), a state-of-the-art approach. More notably, our approach improves over MAE [33] which is also pre-trained with ImageNet and ScanNet (+2.8 mIoU). Moreover, compared to Pri3D [38], a 3D-based pre-training method, our approach outperforms by a significant margin of 7.1 mIoU which shows that multi-view based 3D pre-training does not effectively embed 3D priors to ViT backbones, rather degrades the performance compared to ResNet backbone. Furthermore, we also report the results of MultiMAE [5] when it is pre-trained on ScanNet and it is denoted as MultiMAE* in the table. For fair comparison, we only use RGB and depth as the reconstruction task. We observe that M3D outperforms MultiMAE by a margin of 3.3 mIoU as well as original MultiMAE. Finally, we demonstrate the generalizability of M3D across datasets by fine-tuning the pre-trained model on NYUv2 [49] following the same setup. We draw the same observation that our pre-trained approach outperforms the competing baselines showing how well it transfers across datasets.

4.2.4 Fine-tuning for depth estimation

In this section, we study how M3D transfers the representation to the dense regression tasks. We use NYUv2 for the depth estimation and report δ1 on the NYUv2 test set. δ1 is the percentage of pixels that have an error ratio \( \max\{\frac{\hat{y}_p - y_p}{y_p}, \frac{y_p - \hat{y}_p}{y_p}\} \) below 1.25 [23]. Following MultiMAE, we use DPT [55] as dense prediction head on the top of the ViT encoder. Table 5 shows that M3D outperforms recent state-of-the-art approaches including MultiMAE, Mask3D. Notably, it outperforms CroCo [64] which is a cross view completion pre-training strategy based on MIM specifically designed for 3D vision tasks with 86.7% vs 85.6%. Although our approach is marginally outperforming original MultiMAE, but we want to reiterate that it is pre-trained with three different tasks. Moreover, when multiMAE is pre-trained on ScanNet with RGB and depth only, the performance drops to 85.3%.

4.3. Data-Efficient Learner

We fine-tune M3D based ViT-B encoder on ScanNet for 2D semantic segmentation mainly under low-data regime setting to study how data-efficient learner our approach is. Figure 2 shows that our approach consistently outperforms the competing approaches by a considerable margin across different percentage of available labeled training data. It is worth mentioning that M3D recovers more than 80% performance of the full labeled training set performance when fine-tune with only 20% of training data.

4.4. Ablation studies

In this section, we perform in-depth ablation studies on ScanNet 2D semantic segmentation.

Effect of masking ratio during pre-training. We study the influence of masking ratio and report the results in Table 6 on ScanNet semantic segmentation. It is clearly shown from the table that by masking more patches in RGB and depth modality, M3D achieves the best performance.

Without ImageNet Initialization. We observe a performance drop when the model is not initialized using ImageNet pre-trained weights in the pre-training stage as shown in Table 7. Because ScanNet has a relatively small amount of indoor data, so it’s harder to pre-train the ViT backbones from scratch. Since, ImageNet weights are readily available, so we initialize our models with it following [36, 38].
Methods | Reconstruction task | Backbone | Pre-train | Fine-tune Modality | $\delta_1$
--- | --- | --- | --- | --- | ---
MultiMAE [5] | RGB + Depth | ViT-B | ImageNet+ScanNet | RGB | 85.3
MAE [33] | RGB | ViT-B | ImageNet | RGB | 85.1
Mask3D [36] | Depth | ViT-B | ImageNet+ScanNet | RGB | 85.4
CroCo [64] | RGB + Depth | ViT-B | Habitat | RGB | 85.6
MultiMAE* [5] | RGB + Depth + Segmentation | ViT-B | ImageNet | RGB | 83.0
M$^3$3D | RGB + Depth | ViT-B | ImageNet+ScanNet | RGB | 86.7
MultiMAE* [5] | RGB + Depth + Segmentation | ViT-B | ImageNet | RGB | 86.4

Table 5. NYUv2 Depth Estimation. M$^3$3D outperforms Mask3D and other state-of-the-art approaches that leverage RGB-D data during pre-training. It demonstrates the effectiveness in transferring to dense prediction tasks for out-domain datasets. * means the number are reported from CroCo [64] paper.

<table>
<thead>
<tr>
<th>RGB ratio</th>
<th>Depth ratio</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.0%</td>
<td>20.0%</td>
<td>65.5</td>
</tr>
<tr>
<td>20.0%</td>
<td>50.0%</td>
<td>65.4</td>
</tr>
<tr>
<td>20.0%</td>
<td>80.0%</td>
<td>65.4</td>
</tr>
<tr>
<td>50.0%</td>
<td>20.0%</td>
<td>66.0</td>
</tr>
<tr>
<td>50.0%</td>
<td>50.0%</td>
<td>65.3</td>
</tr>
<tr>
<td>80.0%</td>
<td>20.0%</td>
<td>66.1</td>
</tr>
<tr>
<td>80.0%</td>
<td>80.0%</td>
<td>67.3</td>
</tr>
</tbody>
</table>

Table 6. We study the effect of different masking ratio for RGB and depth input on ScanNet 2D semantic segmentation where each ratio indicates the percentage of masked patches.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train from Scratch</td>
<td>ViT-B</td>
<td>32.6</td>
</tr>
<tr>
<td>Mask3D [36]</td>
<td>ViT-B</td>
<td>41.3</td>
</tr>
<tr>
<td>M$^3$3D</td>
<td>ViT-B</td>
<td>42.5</td>
</tr>
</tbody>
</table>

Table 7. Results on ScanNet 2D semantic segmentation without ImageNet initialization during pre-training.

<table>
<thead>
<tr>
<th>Loss</th>
<th>Backbone</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/out contrastive loss</td>
<td>ViT-B</td>
<td>65.5</td>
</tr>
<tr>
<td>w/out RGB task</td>
<td>ViT-B</td>
<td>66.8</td>
</tr>
<tr>
<td>M$^3$3D</td>
<td>ViT-B</td>
<td>67.3</td>
</tr>
</tbody>
</table>

Table 8. Results on ScanNet 2D semantic segmentation by pre-training without different loss functions.

**Effect of each loss function.** We also study the effect of loss functions during pre-training and report the performance on ScanNet 2D semantic segmentation. We compare the performance of models without contrastive loss and without RGB reconstruction tasks. The results are reported in Table 8. From the results, we observe that contrastive loss overall improves the mIoU compared to Mask3D which suggests that cross-modal learning is an important component besides Masked Image Modeling.

**5. Conclusion**

In this paper, we present M$^3$3D, a new self-supervised pre-training technique with two main functions: (1) learning to embed geometric priors into 2D representations using Masked Image Modeling, (2) learning cross-modal representations in RGB-D data using contrastive learning. M$^3$3D is a general pre-training method applicable to a variety of image/video understanding tasks, doesn’t require camera registration between multi-view input as found in recent self-supervised approaches such as Pri3D, and works well in low-data regime. Our extensive experiments on downstream tasks such as video action recognition, video action detection, 2D semantic segmentation and depth estimation show the superiority of M$^3$3D compared to current state-of-the-art 2D representation learning approaches. Future work includes extending the approach to applications with datasets that have more than two modalities.
References


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


[26] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. arXiv:2205.09113, 2022. 1, 2, 5, 6


Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019. 5


Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In Neurips, 2020. 6

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


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, 6, 7, 8

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

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning?, 2021. 1, 3, 6, 7

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


Muhammad Abdullah Jamal and Omid Mohareri. Surgmcae: Masked autoencoders for long surgical video analysis, 2023. 2

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


Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 7

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos, 2020. 4

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


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


 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. 3, 4

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


Hao Tan and Mohit Bansal. Lxmer: Learning cross-modality encoder representations from transformers, 2019. 3

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural


[68] 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.


