MultiMAE: Multi-modal Multi-task Masked Autoencoders

Roman Bachmann*, David Mizrahi*, Andrei Atanov, and Amir Zamir

Swiss Federal Institute of Technology Lausanne (EPFL)
{firstname.lastname}@epfl.ch
https://multimae.epfl.ch/

Abstract. We propose a pre-training strategy called Multi-modal Multi-task Masked Autoencoders (MultiMAE). It differs from standard Masked Autoencoding in two key aspects: **I)** it can **optionally** accept additional modalities of information in the input besides the RGB image (hence "multi-modal"), and **II)** its training objective accordingly includes predicting multiple outputs besides the RGB image (hence "multi-task").
We make use of masking (across image patches and input modalities) to make training MultiMAE **tractable** as well as to ensure **cross-modality predictive coding** is indeed learned by the network. We show this pre-training strategy leads to a flexible, simple, and efficient framework with improved transfer results to downstream tasks. In particular, the same exact pre-trained network can be flexibly used when additional information besides RGB images is available or when no information other than RGB is available – in all configurations yielding competitive to or significantly better results than the baselines. To avoid needing training datasets with multiple modalities and tasks, we train MultiMAE **entirely using pseudo labeling**, which makes the framework widely applicable to any RGB dataset.
The experiments are performed on multiple transfer tasks (image classification, semantic segmentation, depth estimation) and datasets (ImageNet, ADE20K, Taskonomy, Hypersim, NYUv2). The results show an intriguingly impressive capability by the model in cross-modal/task predictive coding and transfer. Code, pre-trained models, and interactive visualizations are available at https://multimae.epfl.ch.

Keywords: Masked Autoencoders, Multi-modal Learning, Multi-task Learning, Transfer Learning, Vision Transformers

1 Introduction

Masked Autoencoders (MAEs) [28] have recently been demonstrated to be a powerful, yet conceptually simple and efficient, self-supervised pre-training strategy for Vision Transformers [22] (ViTs). Their training objective is to mask-out a high number of patches in an input image and to predict the missing regions.

* Equal contribution

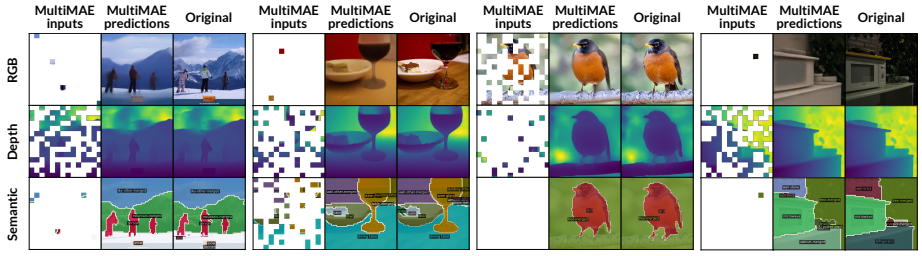

Fig. 1: MultiMAE pre-training objective. We randomly select 1/6 of all 16×16 image patches from multiple modalities and learn to reconstruct the remaining 5/6 masked patches from them. The figure shows validation examples from ImageNet, where masked inputs (left), predictions (middle), and non-masked images (right) for RGB (top), depth (middle), and semantic segmentation (bottom) are provided. Since we do not compute a loss on non-masked patches, we overlay the input patches on the predictions. More examples are shown in the supplementary and on our website.

To that end, only the small number of non-masked patches are first processed using a Transformer encoder [68], and then decoded with a light-weight Transformer that reconstructs the original image. To solve this task sufficiently well, it is assumed [28] that the network needs to learn representations that capture more than just low-level image statistics.

So far, however, the MAE pre-training objective has been limited to a single modality, namely RGB images, and does not make use of any other modalities that are optionally present. In practice, often more than only a single modality of information is available, either through sensing (e.g., a depth sensor) or pseudo labeling (e.g., a powerful pre-trained depth estimation network). Multi-modality is also argued to be employed by biological organisms to develop resilience and better representations [17,18,58]. As we demonstrate in our experiments, making use of such optionally present modalities has the potential to greatly improve the performance of downstream tasks, compared to using only RGB images.

Besides multi-modality (i.e., different inputs), multi-taskness (i.e., different outputs) is an important aspect, as it has been shown that there is usually no single pre-training objective that transfers best to all possible downstream tasks [43,54,79]. Instead, pre-training with a diverse set of tasks [8,66] has been observed to improve the performance on downstream tasks [26,63] and potentially learn a better representation. In general, modifying the training objectives is a powerful way to steer what representation the model will learn.

In this paper, we present Multi-modal Multi-task Masked Autoencoders (MultiMAE), a simple and effective method to make masked autoencoding include multiple modalities and tasks (see Fig. 2). In particular, in our current instantiation of this general method, we study adding dense scene depth to capture geometric information, as well as segmentation maps to include information about the semantic content of the scene. We created a multi-task dataset by pseudo labeling these tasks on ImageNet-1K [19,26]. This has the advantage that in order to train a MultiMAE, one only requires a large unstructured RGB

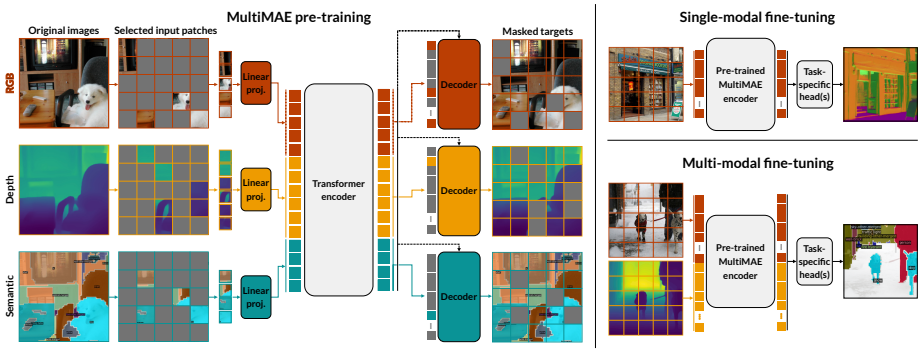

Fig. 2: (Left) MultiMAE pre-training: A small subset of randomly sampled patches from multiple modalities (e.g., RGB, depth, and semantic segmentation) is linearly projected to tokens with a fixed dimension and encoded using a Transformer. Task-specific decoders reconstruct the masked-out patches by first performing a cross-attention step from queries to the encoded tokens, followed by a shallow Transformer. The queries consist of mask tokens (in gray), with the task-specific encoded tokens added at their respective positions. **(Right) Fine-tuning**: By pre-training on multiple modalities, MultiMAE lends itself to fine-tuning on single-modal and multi-modal downstream tasks. No masking is performed at transfer time.

dataset without annotations and off-the-shelf neural networks to perform the pseudo labeling.

To train MultiMAE, we randomly sample a small set of patches from different input modalities, and encode them using a Transformer encoder. MultiMAE's objective is then to reconstruct the masked-out patches of all tasks using task-specific decoders. Figure 1 shows example predictions for the multi-task masked reconstruction that MultiMAE performs. MultiMAE has to learn not only the original MAE objective (within-RGB in-painting), but also to reconstruct any task from any input modality (cross-modal prediction) all from a very sparse set of input patches. The first objective leads to learning *spatial predictive coding* while the second one leads to *cross-modal predictive coding*.

2 Related Work

Masked image prediction consists of learning useful representations by learning to reconstruct images corrupted by masking. This approach was pioneered with denoising autoencoders [69] and context encoders [48]. With the introduction of Vision Transformers (ViT) [22] and motivated by the success of BERT [20] in NLP, many recent works propose a variety of masked image prediction methods for pre-training vision models in a self-supervised way, using reconstruction targets such as pixels [5, 13, 22, 25, 28, 74], discrete tokens [7, 81], and (deep) features [6, 70]. These methods scale very well and achieve strong results on various downstream tasks including motor control [72]. In particular, the masked autoencoder (MAE) [28] approach accelerates pre-training by using an asymmetric architecture consisting of a large encoder that operates *only* on unmasked patches

followed by a lightweight decoder that reconstructs the masked patches from the latent representation and mask tokens. Our approach leverages the efficiency of the MAE approach and extends it to multi-modal and multi-task settings.

Multi-modal learning involves building models capable of relating information from multiple sources. It can either involve training separate encoders or one unified architecture (e.g., a Transformer [68]) to operate on modalities such as images and text [3, 11, 15, 29, 31–34, 41, 42, 59, 62, 75], video and audio [4, 30, 44, 46], video, text and audio [2], and depth, images and video [27]. Our work proposes a simple approach to pre-train Transformers on multiple dense visual modalities and produce strong cross-modal interaction. Unlike most prior work which assumes that all modalities are available during inference, our approach is designed to perform well on any subset of the pre-training modalities.

Related to MultiMAE are several works that perform multi-modal autoencoding [45, 56, 60, 61, 71]. Our approach differs from them in that we use a more flexible architecture and perform masked autoencoding to learn cross-modal predictive coding among optional inputs (as demonstrated in Fig. 1).

Multi-task learning consists of training models to predict multiple output domains from a single input [10, 24, 35]. In computer vision, the input is usually an RGB image. A common approach for multi-task learning is to use a single encoder to learn a shared representation followed by multiple task-specific decoders [26, 67]. These methods differ from our approach as we use multiple tasks in both the input and the output along with masking.

In addition, many works study the importance of task diversity to improve transfer performance [26, 43, 54, 65, 79]. These works argue that learning from one task alone is insufficient and that a set of tasks can more effectively cover the many possible downstream tasks in vision. Our pre-training method operates on multiple tasks to learn more general representations capable of covering multiple downstream tasks.

Self-training is a technique to incorporate unlabeled data into a supervised learning setting [36, 53, 55, 77]. It is one of the earliest approaches to semi-supervised learning. Self-training methods use a supervised model to generate pseudo labels on unlabeled data and then train a student model on the pseudo labeled data. These approaches have been applied to a variety of vision tasks such as image classification [49, 73, 76], object detection [82], and segmentation [12, 82]. Most recently, multi-task self-training (MuST) [26] uses specialized teachers to create a multi-task pseudo labeled dataset and then trains a multi-task student model on this dataset to learn general feature representations. Our method also relies on pseudo labeling to produce a large-scale multi-task dataset. However, unlike prior work, pseudo labels are not only used as output targets but also as *masked* input modalities.

3 Method Description

In this Section, we describe the Multi-modal Multi-task Masked Autoencoder (MultiMAE) architecture (illustrated in Fig. 2), as well as the pre-training strat-

egy in more detail. We first give an architectural overview of both the multi-modal encoder (Sec. 3.1) and multi-task decoders (Sec. 3.2). We then describe our multi-modal token sampling strategy (Sec. 3.3) and introduce the pseudo labeled tasks we use for pre-training (Sec. 3.4). Finally, we display the most important pre-training details (Sec. 3.5).

3.1 Multi-modal Encoder

Our multi-modal Transformer encoder is a ViT [22], but with patch projection layers for each additional input modality. Specifically, 16×16 patches of each modality are projected to tokens with the correct Transformer dimension using a different linear projection for each modality. Projected patches are concatenated into a sequence of tokens and given as input to the same Transformer encoder. We also add an additional *global* token with a learned embedding, similar to the class-token used in ViT. Due to the architectural similarities to ViT, MultiMAE pre-trained weights can directly be used in a standard single-modal ViT by loading only the desired input projection and ignoring the others.

Positional, Modality and Class Embeddings. Since all our modalities have a 2D structure, we add 2D sine-cosine positional embeddings [14, 28] after the linear projection. We do not explicitly add any modality-specific embeddings, since the bias term in each linear projection can act as such. In order to perform the semantic segmentation patch projection, we first replace each class index with learned 64-dimensional class embeddings.

Low Computational Complexity. Just as in the RGB-only MAE [28], we only pass the small randomly sampled subset of all tokens to the Transformer encoder as part of the masked autoencoding objective. This is in contrast to the masked autoencoding approaches of SiT [1], BeiT [7] and SimMIM [74], that encode both the masked and visible tokens. Due to the quadratic complexity of standard self-attention as a function of the number of tokens, encoding only the random subset of visible tokens becomes increasingly important as the number of input modalities grows. Indeed, the speedup and reduction in memory are significant and crucial in enabling MultiMAE's multi-modal pre-training with three dense input modalities. A comparison of the pre-training time with and without masked tokens is given in the supplementary.

3.2 Decoders

To reconstruct the masked-out tokens from the visible tokens, we use a separate decoder for each task. The input to each decoder is the full set of visible tokens from the respective task it is reconstructing. As in MAE [28], these visible tokens are decoded jointly with a set of mask tokens, which serve as *placeholders* for the decoders to write the reconstructed patches (as shown in Fig. 2). To integrate information from the encoded tokens of other modalities, we add a single cross-attention layer in each decoder using these tokens as queries and all the encoded tokens as keys / values. Sine-cosine positional embeddings and learned modality

embeddings are added to the tokens before this step. This is then followed by a small MLP and Transformer blocks. Following MAE, we compute the losses only on the masked tokens.

As each task requires its own decoder, the computational cost of decoders scales linearly with the number of tasks. To keep pre-training efficient, we use shallow decoders (a single cross-attention layer and MLP, followed by two Transformer blocks) with a low dimensionality (256 dimensional). Compared to the encoder, these decoders add little to the overall computational cost, and as He et al. [28] show, they perform similarly to deeper decoders on ImageNet-1K fine-tuning.

3.3 Multi-modal Masking Strategies

For masked autoencoding to work well, a large percentage of tokens needs to be masked-out. He et al. [28] showed that the choice of mask sampling strategy can have a large impact on transfer performance. More specifically for MultiMAE and generally learning multi-task representations, masking across different modalities ensures the model develops predictive coding across different modalities besides different spatial patches. For efficiency and simplicity, we choose a constant number of visible tokens for all our experiments, which we fix at 98. This corresponds to 1/6 of all tokens when using three modalities of dimensions 224×224 pixels and a patch size of 16×16. Adapting the MAE mask sampling strategy by selecting the visible tokens uniformly from all tokens would result in most modalities being represented to similar degrees. Cases where one or more modalities have very few or no samples would be very rare. We propose a multi-modal token sampling strategy that allows for a more diverse sampling approach. It can be broken down into two steps: First, selecting the number of tokens per modality, and second, randomly sampling the set of tokens for each modality.

Number of Tokens per Modality. We select the proportion of tokens per modality λ by sampling from a symmetric Dirichlet distribution $(\lambda_{\text{RGB}}, \lambda_{\text{D}}, \lambda_{\text{S}}) \sim \text{Dir}(\alpha)$, where $\lambda_{\text{RGB}} + \lambda_{\text{D}} + \lambda_{\text{S}} = 1, \lambda \geq 0$. The sampling is controlled by the concentration parameter $\alpha > 0$. When $\alpha = 1$, the symmetric Dirichlet distribution is equivalent to a uniform distribution over the simplex (i.e., it is uniform over all points in its support). Smaller values ($\alpha << 1$) result in a sampling behavior where most of the tokens will be sampled from a single modality, while larger values ($\alpha >> 1$) result in an increasingly similar number of tokens to be sampled from each modality. As a design decision, we do not bias the sampling towards certain modalities (as we use a symmetric Dirichlet), since we want to be agnostic to the choice of downstream input modalities and tasks that users might want to consider. For simplicity and better representation of any possible sampled mask, we use a concentration parameter $\alpha = 1$ for all of our experiments. Random masks sampled using $\alpha = 1$ are shown in Figure 1, and an ablation on the choice of concentration parameter is given in the supplementary.

Sampling Tokens. From each modality, we sample the number of tokens, as specified by the above Dirichlet sampling step, uniformly at random without

replacement. Uniform sampling has been shown to work well for masked autoencoders, compared to less random alternatives [28].

3.4 Pseudo Labeled Multi-task Training Dataset

We pre-train MultiMAE with three tasks that we pseudo label on ImageNet-1K [19]. Pseudo labeling has the advantage that we do not need a large multi-task dataset with aligned task images. Instead, having access to a good set of pre-trained neural networks for the tasks we want to train on can be effective. Pseudo labeling scales to RGB datasets of arbitrary size and is a one-time pre-processing step. Compared to the cost of training, this step is computationally cheap and fast if parallelized.

Taskonomy [79] demonstrated computationally that common vision tasks cluster into three main categories, namely low-level, geometric, and semantic tasks. To have a coverage over such a space of vision tasks, we choose one representative task from each of these three clusters. We note that except for object detection and classification, these are the same pseudo labeled tasks that are used in MuST [26]. In the following, we will describe them in more detail.

RGB and Per-Patch Standardized RGB. We use RGB images due to their abundance and since RGB-only masked autoencoding is shown to be a powerful pre-training task. He et al. [28] study both predicting standard RGB patches, as well as per-patch standardized RGB patches. They find that predicting standardized patches slightly improves transfer performance. Since MultiMAE is naturally a multi-task model, we add both versions as separate decoder heads to get the representational benefits of predicting standardized patches, and to get a version that we can visualize better. Note that we only add the per-patch standardized version as an output task, and not as an input modality. For both RGB versions, we follow MAE and compute the MSE loss between the ground truth and predicted pixels. In the rest of the paper, we will refer to the RGB and per-patch standardized RGB output tasks simply as RGB.

Scene Depth. Depth is a key task informative about scene geometry. As with RGB, but unlike semantic segmentation, sensors exist to capture this modality, making it possible to use depth as an optional extra input for downstream tasks. To pseudo label depth, we use a DPT-Hybrid [50] that was trained on Omnidata [23]. Since monocular depth estimation is an inherently ill-posed task due to scale and shift ambiguity, we standardize the depth values in a robust way by ignoring the top and bottom 10% of values [78]. In addition, using standardized depth values as inputs allows us to use other depth images that might have different depth ranges and scales, without needing to match them to the Omnidata depth parameterization. We use the L1 loss for depth.

Semantic Segmentation. Lastly, we use a Mask2Former [16] with a Swin-S [38] backbone trained on COCO [37] to pseudo label semantic segmentation maps on ImageNet. For that, we extract 133 semantic classes by taking the argmax of the network predictions. Unlike RGB and depth, the main purpose of this task is to improve performance on downstream tasks, rather than using it as an

input modality (though we show results using pseudo labeled semantic inputs in Table 3). Since we use a network that was pre-trained on COCO, we do not evaluate semantic segmentation transfers on that dataset. For this task, we use the cross-entropy loss.

3.5 Pre-training Details

All our MultiMAE experiments use a *ViT-B* [22] with a patch size of 16×16 pixels. We pre-train the models for either 400 epochs (only for transfer ablation study in Sec. 4.4) or 1600 epochs (for best results and to be comparable to the MAE baseline) on 1.28M ImageNet images. We use the AdamW [40] optimizer with base learning rate 1e-4 and weight decay 0.05. We warm up training for 40 epochs, starting from learning rate 1e-6, and decay it to 0 over the course of training using cosine decay [39]. We set the batch size to a total of 2048 and train the models using 8 A100 GPUs with automatic mixed precision enabled. Our data augmentations are straightforward. We randomly crop the images, setting the random scale between 0.2 and 1.0 and the random aspect ratio between 0.75 and 1.33, after which we resize the crops to 224×224 pixels and apply a random horizontal flip with probability 0.5. Additional pre-training details can be found in the supplementary.

4 Experiments

Optimizing the pre-training objective of MultiMAE is successful as apparent in the various results shown in the main paper, the supplementary, and the interactive visualizations shown on our website. In this section we provide a transfer study to measure the effectiveness of MultiMAE pre-training compared to relevant baselines. This section is organized in the following manner: After introducing the downstream tasks and datasets (Sec. 4.1), we show transfer results for the case where the only available input modality is RGB (Sec. 4.2). Then, we show that MultiMAE can significantly improve downstream performance if other modalities like depth are either available as ground truth (sensor), or can be cheaply pseudo labeled (Sec. 4.3). We follow up with an ablation on the influence of pre-training tasks on the downstream performance (Sec. 4.4), and finally we visually demonstrate that MultiMAE integrates and exchanges information across modalities (Sec. 4.5).

4.1 Transfer Tasks and Datasets

We perform downstream transfers on a variety of semantic and dense regression tasks. For all transfers, we replace the pre-trained decoders by randomly initialized task-specific heads, and train them along with the pre-trained encoder. In the following, we give an overview over all tasks and datasets used in our transfer experiments. Exact training details are presented in the supplementary.

Classification. We evaluate our models and baselines by fine-tuning them on the supervised ImageNet-1K [19] 1000-way object classification task. We fine-tune our models for 100 epochs on the entire ImageNet-1K train split (1.28M images) and report the top-1 validation accuracy.

Semantic Segmentation. We further evaluate our models on semantic segmentation tasks on the ADE20K [80] (20'210 training images and 150 classes), NYUv2 [57] (795 training images and 40 classes), and Hypersim [52] (51'674 training images and 40 classes) datasets. NYUv2 and Hypersim contain ground-truth depth maps that allow us to evaluate semantic segmentation with RGB and depth as input modalities. For all datasets, we report the mean intersection over union (mIoU) metric. On ADE20K and Hypersim, we report it on the validation split, while on NYUv2, we show the test set mIoU.

Dense Regression Tasks. Finally, we study how our models transfer to geometric tasks, such as surface normals, depth and reshading, as well as tasks extracted from RGB images, such as keypoint or edge detection. For depth estimation, we use NYUv2 (795 training and 655 test images), while for all other tasks we train transfers on a subset of the Taskonomy dataset [79] (800 training images). As performance metrics, we report δ_1 on the NYUv2 test set, showing the percentage of pixels p with error $\max\{\frac{\hat{y}_p}{y_p}, \frac{y_p}{\hat{y}_p}\}$ less than 1.25 [21], while on Taskonomy we report L1 losses on the tiny-split test set.

In the tables, classification, semantic segmentation, and depth estimation are denoted by (C), (S), and (D), respectively.

4.2 Transfers with RGB-Only

In this section, we show our transfer results when fine-tuning using only the RGB modality as input.

Baselines. For this setting, we compare MultiMAE with various *ViT-B* models, namely *DeiT* [64] (without distillation) representing an ImageNet-supervised baseline, *MoCo-v3* [14], *DINO* [9], and *MAE* [28]. All these models are pre-trained on ImageNet-1K. We use the official weights for DeiT, MoCo-v3, and DINO, and reproduce MAE using the official PyTorch [47] codebase following the setting specified in [28] (i.e., decoder of depth 8 and width 512, per-patch standardized pixel loss, 1600 pre-training epochs, 75% mask ratio). In the supplementary, we compare the transfer performance of this MAE model to one with a shallower and narrower decoder (depth 2 and width 256), closer to the one used for MultiMAE.

We report the results in Table 1. We find that MultiMAE performs best on all tasks, matching MAE's performance on ImageNet-1K classification and ADE20K semantic segmentation, and outperforming it on all other tasks and datasets. These results show the effectiveness of MultiMAE as a pre-training strategy: it retains the benefits of MAE when RGB is the only fine-tuning modality but can also accept other modalities, as shown next.

Table 1: Fine-tuning with RGB-only. We report the top-1 accuracy (↑) on ImageNet-1K (IN-1K) [19] classification (C), mIoU (↑) on ADE20K [80] , Hypersim [52] , and NYUv2 [57] semantic segmentation (S), as well as δ_1 accuracy (↑) on NYUv2 depth (D). Text in **bold** and underline indicates the first and second-best results, respectively. All methods are pre-trained on ImageNet-1K (with pseudo labels for MultiMAE)

Method	IN-1K (C)	ADE20K (S)	Hypersim (S)	NYUv2 (S)	NYUv2 (D)
Supervised [64]	81.8	45.8	33.9	50.1	80.7
DINO [9]	83.1	44.6	32.5	47.9	81.3
MoCo-v3 [14]	82.8	43.7	31.7	46.6	80.9
MAE [28]	**83.3**	**46.2**	36.5	50.8	85.1
MultiMAE	**83.3**	**46.2**	**37.0**	**52.0**	**86.4**

Table 2: Fine-tuning with RGB and ground truth depth. We report semantic segmentation transfer results from combinations of RGB and depth, measured in mIoU (↑). MultiMAE can effectively leverage additional modalities such as depth, while MAE cannot. Text in gray indicates a modality that the model was not pre-trained on

	Hypersim (S)			NYUv2 (S)		
Method	RGB	D	RGB-D	RGB	D	RGB-D
MAE	36.5	32.5	36.9	50.8	23.4	49.3
MultiMAE	**37.0**	**38.5**	**47.6**	**52.0**	**41.4**	**56.0**

4.3 Transfers with Multiple Modalities

Since MultiMAE was pre-trained on RGB, depth, and semantic segmentation, it can optionally accept any of those modalities as input during transfer learning should they be available. In this set of experiments, we study on three semantic segmentation downstream tasks how much MultiMAE can benefit from using additional modalities during transfer. Often, ground truth depth maps are not available for a given downstream dataset and for that reason, we perform additional transfers using pseudo labeled depth. As there are several datasets that do in fact contain aligned RGB and depth images (e.g., Hypersim, NYUv2, Taskonomy, etc.) and since sensors exist that can measure depth, we consider it as a more realistic input modality compared to semantic segmentation. Since our model was trained with semantic segmentation as an input modality, we perform additional experiments using pseudo labeled semantic segmentation maps as inputs.

All multi-modal transfers are performed by concatenating the projected patches of all modalities into a single sequence (i.e., no masking is performed here). Using more than two modalities during transfer quickly becomes computationally expensive, since without masking, our method now scales with the full number of modalities and tokens. For performing multi-modal transfers with the standard MAE, we train a new input projection for the additional modalities while fine-tuning. Further training details can be found in the supplementary.

Transfers Using Sensory Depth. First, we consider that we have access to an aligned RGB-D dataset, like NYUv2 or Hypersim. We treat depth in the

Table 3: Fine-tuning with RGB and pseudo labels. Semantic segmentation transfer results using *pseudo labeled* depth and semantic segmentation maps, measured in mIoU (↑). MultiMAE benefits much more than MAE from pseudo labeled modalities as input. Text in gray indicates a modality that the model was not pre-trained on

	ADE20K (S)					Hypersim (S)					NYUv2 (S)				
Method	RGB	pD	RGB-pD	RGB-pS	RGB-pD-pS	RGB	pD	RGB-pD	RGB-pS	RGB-pD-pS	RGB	pD	RGB-pD	RGB-pS	RGB-pD-pS
MAE	**46.2**	20.0	46.3	46.2	46.3	36.5	21.0	36.9	37.7	37.3	50.8	23.8	49.1	50.1	49.3
MultiMAE	**46.2**	**34.4**	**46.8**	45.7	**47.1**	**37.0**	**30.6**	**37.9**	**38.4**	**40.1**	**52.0**	**39.9**	**53.6**	**53.5**	**54.0**

exact same way as during pre-training, i.e., pre-process it by standardizing it in a robust manner [78]. Because ground-truth depth maps might contain invalid measurements, we further set all these masked-out values to 0.

Table 2 shows RGB-D transfer results on Hypersim and NYUv2. Compared to the RGB-only results in Table 1, we see a substantial increase in performance when ground truth depth is available for MultiMAE. The standard MAE on the other hand is not able to sufficiently make use of the additional depth, since it was only trained on RGB images. We observe a similar story when evaluating transfers from depth-only, in that MultiMAE works well, even when no RGB information is available, while MAE does not. On Hypersim, MultiMAE depth-only transfer is even able to surpass MultiMAE RGB-only transfer, and, as expected, RGB-D works better than either RGB or depth alone.

Transfers with Pseudo Labels. In case ground truth modalities are not available, we can pseudo label them in the same way we did for pre-training. To pseudo label depth, we use the same Omnidata DPT-Hybrid model that we used for pre-training on both ADE20K and NYUv2. On Hypersim, we use a MiDaS [51] DPT-Hybrid, since the Omnidata depth model was partially trained on this dataset. For semantic segmentation pseudo labels, we use the same COCO Mask2Former model as in pre-training.

As shown in Table 3, MultiMAE can use pseudo labeled depth or semantic segmentation to boost performance beyond the RGB-only setting, although the gain is smaller than using real depth. Moreover, performance can further be improved by adding both of these pseudo labeled modalities to the input. This setting performs the best out of all settings involving pseudo labels.

4.4 Influence of Pre-training Task Choices and Masking on Transfer Performance

How does the choice of MultiMAE pre-training tasks affect downstream transfer performance? In this subsection, we aim to address this question by performing transfers from MultiMAE models that were pre-trained with RGB-D, RGB-S, or RGB-D-S. We further compare MultiMAE against MAE, single-task, and multi-task baselines.

All experiments are performed on ViT-B models that were pre-trained for 400 epochs. We transfer the pre-trained models to ImageNet, NYUv2 segmentation, as well as nine dense regression tasks on Taskonomy. On Taskonomy, we report

Table 4: Ablation experiments. We study the impact of additional modalities in Tab. 4a, and compare MultiMAE to non-masked pre-training in Tab. 4b. All models are pre-trained for 400 epochs. We report the top-1 accuracy (↑) on ImageNet-1K (IN-1K) [19] classification (C), mIoU (↑) on NYUv2 [57] semantic segmentation (S), δ_1 accuracy (↑) on NYUv2 depth (D) and avg. rank (↓) on Taskonomy [79]. While some specialized pre-trained models perform better at certain downstream tasks, they perform poorly at others. MultiMAE pre-trained with RGB, depth and semantic segmentation is a more generalist model that does well at transferring to a range of downstream tasks

(a) Impact of additional modalities. Transfer results of several MultiMAE models pre-trained on different input modalities / target tasks, compared against MAE (single-modal baseline). D2 = MAE pre-trained with a decoder of depth 2 and width 256, comparable in size to the decoders of MultiMAE

Method	IN-1K (C)	NYUv2 (S)	NYUv2 (D)	Taskonomy (D)
MAE (D2)	83.0	44.0	81.3	3.8
RGB-D	82.8	45.8	83.3	2.1
RGB-S	**83.2**	**51.6**	**85.5**	2.6
RGB-D-S	83.0	50.6	85.4	**1.5**

(b) Comparison to non-masked pre-training. We compare standard single-task and multi-task baselines pre-trained using *non-masked* RGB inputs against the RGB-D-S MultiMAE. The RGB→D-S model is conceptually similar to MuST using depth and semantic segmentation as target tasks

Method	IN-1K (C)	NYUv2 (S)	NYUv2 (D)	Taskonomy (D)
RGB→D	82.7	44.0	**87.1**	1.6
RGB→S	82.5	46.8	82.9	4.0
RGB→D-S	82.8	48.6	84.6	2.9
MultiMAE	**83.0**	**50.6**	85.4	**1.5**

the ranking of different pre-trained models, averaged over all nine tasks. Detailed per-task results on Taskonomy can be found in the supplementary.

Masked Multi-modal Pre-training. This experiment studies the influence that the choice of pre-training modalities has, when the input and output modalities are the same in MultiMAE pre-training. The transfer results are displayed in Table 4a. The RGB-S model performs best on ImageNet classification and NYUv2 semantic segmentation, whereas the RGB-D-S model has the best average rank on Taskonomy. The slight increase in performance of RGB-S on ImageNet and semantic segmentation compared to RGB-D-S comes at the cost of reduced flexibility, as models that were not pre-trained on depth can not as easily and effectively use it to boost performance (see Sec. 4.3).

Comparison to Non-masked Pre-training. We further compare MultiMAE against standard single-task and multi-task baselines, that were pre-trained with RGB as the only input modality and without applying any masking. Since we train on pseudo labels, the RGB→D-S multi-task model is conceptually similar to a MuST [26] model using depth and semantic segmentation targets. The transfer results are detailed in Table 4b. On nearly all categories, MultiMAE outperforms the supervised baselines.

To summarize, the results in this section show that using all modalities to pre-train a MultiMAE results in a more generalist model that does well at transferring to a range of downstream tasks. We find that there are some *specialized* pre-trained models that perform better at certain downstream tasks (e.g., models pre-trained with depth perform better at transferring to geometric tasks),

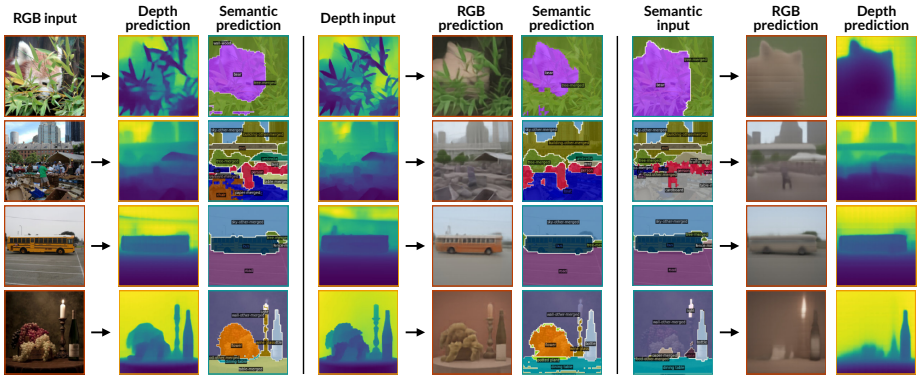

Fig. 3: Single-modal predictions. We visualize MultiMAE cross-modal predictions on ImageNet-1K validation images. Only a single, full modality is used as input. The predictions remain plausible despite the absence of input patches from other modalities.

but they will perform poorly at others. This is supported by previous findings [43, 54, 79] showing that there is usually no single visual pre-training task that transfers well to any arbitrary other task, and instead, a set is required.

4.5 Cross-modal Exchange of Information

In this section, we explore visually how MultiMAE predicts the three pre-training tasks by changing the inputs it receives. Figure 1 already showcased how MultiMAE is able to reconstruct images from various randomly sampled input patches. Here, we will further show non-masked cross-modal predictions, and will also give examples on how MultiMAE predictions change when we change certain details about the inputs.

Single-modal Predictions. Figure 3 displays several examples of cross-modal prediction without any masking. We show examples where, from one single modality, the two remaining ones are predicted. We note here that even though the number of patches we input to the model is $2\times$ higher than what was seen during training, the model still predicts very reasonable results despite the distribution shift.

Demonstration of Cross-modal Interaction. We demonstrate in Figure 4 how MultiMAE predicts completely different but plausible RGB images when given a full depth image and three edited versions of the same two RGB input patches (no semantic segmentation maps are given as inputs). We keep one RGB patch the same, while changing the hue of another patch (part of a lizard for the first image). We can see how MultiMAE recovers all the details in the image from the full depth input, but paints the entire lizard in the colors given in the modified patch. All the while, the background does not change. This suggests an intriguingly good representation is learned by the model as it extends the colors to the right segments without any segmentation provided in the input. More interactive examples can be seen on our website.

5 Discussion

We presented Multi-modal Multi-task Masked Autoencoders (MultiMAE), an effective and simple pre-training strategy for Vision Transformers. MultiMAE encodes a small random subset of visible tokens from multiple modalities and is trained to reconstruct the missing ones. By encoding only a fixed number of non-masked tokens, we can keep the bulk of the computation in the Transformer encoder constant, while only the shallow task-specific decoders scale with the number of tasks. Masking (across image patches and input modalities) ensures the network learns to perform predictive coding across different modalities, besides across different spatial patches. The experiments showed intriguing capabilities of MultiMAE at cross-modal coding and demonstrated this pre-training strategy can result in notable gains in transfer performance when additional input modalities are optionally available, either as ground truth or pseudo labels.

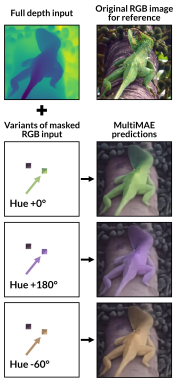

Fig. 4: Cross-modal interaction. By editing the hue of a single input token, the entire lizard's color can be changed, while keeping the background constant.

In the following, we briefly discuss some limitations to our approach and present exciting future directions:

Scaling Pre-training Modalities. We pre-trained MultiMAE on a set of three visual modalities, chosen to cover a large fraction of common vision problems based on prior studies [79]. It is, however, conceivable that our method can benefit from a rather straightforward inclusion of a more diverse set of modalities and tasks, such as videos, text, bounding boxes, sparse depth, feature maps, and more. In addition to providing more ways to use optional modalities as inputs, scaling up the number of pre-training modalities could have further transfer benefits by covering a larger space of useful vision problems and enabling more complex cross-modal predictive coding.

Scaling Pre-training Datasets. For pragmatic reasons and enabling comparison with prior works, we trained all of our models on pseudo labeled ImageNet-1K, but there is no reason to limit ourselves to a (classification) dataset of this size. Since we use pseudo labels, any dataset that is used for RGB-only self-supervised learning can be considered for training MultiMAE. Our method further benefits from any future improvements in model architectures, training strategy and supervised datasets that can be used to improve the quality of pseudo labels.

Masking Strategies. Lastly, we used a simple approach of sampling random tokens from each modality in an unbiased way. While this worked well for MultiMAE training, it does not have to be the optimal choice for learning a transferable representation. It will be an interesting direction to explore biasing the masking towards certain modalities and/or spatial locations.

Acknowledgments. We thank Stefan Stepanovic and Alexander Sax for their help and insightful discussions.

References

1. Ahmed, S.A.A., Awais, M., Kittler, J.: Sit: Self-supervised vision transformer. ArXiv **abs/2104.03602** (2021) 5
2. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems **34** (2021) 4
3. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems **33**, 25–37 (2020) 4
4. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017) 4
5. Atito, S., Awais, M., Kittler, J.: Sit: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602 (2021) 3
6. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022) 3
7. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. ArXiv **abs/2106.08254** (2021) 3, 5
8. Baxter, J.: A model of inductive bias learning. J. Artif. Intell. Res. **12**, 149–198 (2000) 2
9. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9630–9640 (2021) 9, 10
10. Caruana, R.: Multitask Learning. Machine Learning **28**(1), 41–75 (Jul 1997). https://doi.org/10.1023/A:1007379606734 4
11. Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2940–2949 (2016) 4
12. Chen, L.C., Lopes, R.G., Cheng, B., Collins, M.D., Cubuk, E.D., Zoph, B., Adam, H., Shlens, J.: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: European Conference on Computer Vision. pp. 695–714. Springer (2020) 4
13. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative Pretraining From Pixels. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1691–1703. PMLR (Nov 2020), iSSN: 2640-3498 3
14. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9620–9629 (2021) 5, 9, 10
15. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020) 4
16. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. ArXiv **abs/2112.01527** (2021) 7
17. De Sa, V.R.: Sensory modality segregation. In: NIPS. pp. 913–920. Citeseer (2003) 2

18. De Sa, V.R., Ballard, D.H.: Category learning through multimodality sensing. Neural Computation **10**(5), 1097–1117 (1998) 2
19. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 2, 7, 9, 10, 12
20. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 3
21. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2051–2060 (2017) 9
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2021) 1, 3, 5, 8
23. Eftekhar, A., Sax, A., Bachmann, R., Malik, J., Zamir, A.R.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10766–10776 (2021) 7
24. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015) 4
25. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint arXiv:2112.10740 (2021) 3
26. Ghiasi, G., Zoph, B., Cubuk, E.D., Le, Q.V., Lin, T.Y.: Multi-task self-training for learning general representations. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 8836–8845 (2021) 2, 4, 7, 12
27. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A single model for many visual modalities. arXiv preprint arXiv:2201.08377 (2022) 4
28. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. ArXiv **abs/2111.06377** (2021) 1, 2, 3, 5, 6, 7, 9, 10
29. Hu, R., Singh, A.: Unit: Multimodal multitask learning with a unified transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1439–1449 (2021) 4
30. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021) 4
31. Kaiser, L., Gomez, A.N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., Uszkoreit, J.: One model to learn them all. arXiv preprint arXiv:1706.05137 (2017) 4
32. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 4
33. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015) 4
34. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021) 4

35. Kokkinos, I.: Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6129–6138 (2017) 4
36. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML (2013) 4
37. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 7
38. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021) 7
39. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 8
40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 8
41. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019) 4
42. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10437–10446 (2020) 4
43. Mensink, T., Uijlings, J.R.R., Kuznetsova, A., Gygli, M., Ferrari, V.: Factors of influence for transfer learning across diverse appearance domains and task types. IEEE transactions on pattern analysis and machine intelligence **PP** (2021) 2, 4, 13
44. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems **34** (2021) 4
45. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.: Multimodal deep learning. In: ICML (2011) 4
46. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018) 4
47. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019) 9
48. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016) 3
49. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11557–11568 (2021) 4
50. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12159–12168 (2021) 7

51. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**, 1623–1637 (2022) 11
52. Roberts, M., Paczan, N.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10892–10902 (2021) 9, 10
53. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) (2005) 4
54. Sax, A., Emi, B., Zamir, A.R., Guibas, L.J., Savarese, S., Malik, J.: Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. (2018) 2, 4, 13
55. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3), 363–371 (1965) 4
56. Shi, Y., Siddharth, N., Paige, B., Torr, P.H.S.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. ArXiv **abs/1911.03393** (2019) 4
57. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 9, 10, 12
58. Smith, L., Gasser, M.: The development of embodied cognition: Six lessons from babies. Artificial life **11**(1-2), 13–29 (2005) 2
59. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019) 4
60. Sutter, T.M., Daunhawer, I., Vogt, J.E.: Multimodal generative learning utilizing jensen-shannon-divergence. ArXiv **abs/2006.08242** (2019) 4
61. Sutter, T.M., Daunhawer, I., Vogt, J.E.: Generalized multimodal ELBO. CoRR **abs/2105.02470** (2021), https://arxiv.org/abs/2105.02470 4
62. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019) 4
63. Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P.: Rethinking few-shot image classification: a good embedding is all you need? ArXiv **abs/2003.11539** (2020) 2
64. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., J'egou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021) 9, 10
65. Tripuraneni, N., Jordan, M., Jin, C.: On the theory of transfer learning: The importance of task diversity. Advances in Neural Information Processing Systems **33**, 7852–7862 (2020) 4
66. Tripuraneni, N., Jordan, M.I., Jin, C.: On the theory of transfer learning: The importance of task diversity. ArXiv **abs/2006.11650** (2020) 2
67. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE transactions on pattern analysis and machine intelligence (2021) 4
68. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. ArXiv **abs/1706.03762** (2017) 2, 4
69. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research **11**(110), 3371–3408 (2010), http://jmlr.org/papers/v11/vincent10a.html 3

70. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. arXiv preprint arXiv:2112.09133 (2021) 3
71. Wu, M., Goodman, N.D.: Multimodal generative models for scalable weakly-supervised learning. In: NeurIPS (2018) 4
72. Xiao, T., Radosavovic, I., Darrell, T., Malik, J.: Masked visual pre-training for motor control. arXiv preprint arXiv:2203.06173 (2022) 3
73. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020) 4
74. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. ArXiv **abs/2111.09886** (2021) 3, 5
75. Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., Huang, F.: E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. arXiv preprint arXiv:2106.01804 (2021) 4
76. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019) 4
77. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: ACL (1995) 4
78. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 204–213 (2021) 7, 11
79. Zamir, A.R., Sax, A., Shen, W.B., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2018) 2, 4, 7, 9, 12, 13, 14
80. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5122–5130 (2017) 9, 10
81. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021) 3
82. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020) 4