MIMIC: Masked Image Modeling with Image Correspondences

Kalyani Marathe[1,2*] **Mahtab Bigverdi**[1,2*] **Nishat Khan**[1] **Tuhin Kundu**
Patrick Howe **Sharan Ranjit S** [1] **Anand Bhattad** [3] **Aniruddha Kembhavi**[2]
Linda G. Shapiro[1] **Ranjay Krishna**[1,2]
[1]University of Washington, [2]Allen Institute for Artificial Intelligence,
[3]Toyota Technological Institute at Chicago
{kmarathe,mahtab,nkhan51,shapiro,ranjay}@cs.washington.edu,
anik@allenai.org, tuhinkundu@outlook.com, {pdh, sharanrs}@uw.edu

Abstract

Dense pixel-specific representation learning at scale has been bottlenecked due to the unavailability of large-scale multi-view datasets. Current methods for building effective pretraining datasets heavily rely on annotated 3D meshes, point clouds, and camera parameters from simulated environments, preventing them from building datasets from real-world data sources where such metadata is lacking. We introduce a pretraining dataset-curation approach that does not require any additional annotations. Our method allows us to generate multi-view datasets from both real-world videos and simulated environments at scale. Specifically, we experiment with two scales: MIMIC-1M *with* 1.3*M and* MIMIC-3M *with* 3.1*M multi-view image pairs and train models with different masked image modeling objectives. Through our comprehensive experimental analysis we show that: Representations trained on our automatically generated* MIMIC-3M *outperform those learned from expensive crowdsourced datasets (ImageNet-1K) and those learned from synthetic environments (*MULTIVIEW-HABITAT*) on three dense geometric tasks: depth estimation on NYUv2 (↑1.7%), and surface normal estimation on Taskonomy (↓2.05%), and depth estimation on Taskonomy (↓7.5%) and performs on-par with* MULTIVIEW-HABITAT *on Taskonomy edges and curvature tasks. Larger dataset (*MIMIC-3M*) improves performance, which is promising since our curation method can arbitrarily scale to produce even larger datasets. The code and instructions to download, access, and use* MIMIC-3M *can be found here.*

1. Introduction

Today, dense vision tasks—depth prediction, surface normal estimation, semantic segmentation, and pose estimation— often rely on pretrained representations [2, 15]. Naturally, self-supervised learning lends itself as a potential solution. Despite the impressive performance on object recognition and other high-level tasks, self-supervised representations for dense prediction tasks have not yet fully delivered. The representations trained on single-view images lack the correspondences required for 3D reasoning of our visual world [35]. Moreover, the joint-embedding-based objectives (SimCLR [8], MoCo [14], DINO [6]) that are often used on object-centric datasets utilize augmentations that do not preserve geometric pixel-wise information. In response, the general purpose representation learning method—masked image modeling and specifically masked autoencoders (MAE)—has become a popular default mechanism for such tasks [2, 15, 35]. Unfortunately, recent findings suggest that the representations learned by MAE are devoid of sufficient local information for tasks like depth estimation [35].

Based on these observations, we ask the following question: *What data do we need to learn useful representations for dense vision tasks*? We find a potential answer in cognitive science: 3D understanding of the physical world is one of the first visual skills emergent in infants; it plays a critical role in the development of other skills, like depth estimation, understanding surfaces, occlusions, etc [16]. Scientists hypothesize that 3D understanding emerges from infants learning the relationship between changes in visual stimuli in response to their self-motion [18], i.e. 3D awareness emerges by learning correspondences between appearances as the infant's vantage point changes [26].

Very recently, a machine learning paper proposed a variant of masked image modeling, named **cro**ss-view **co**mpletion (CroCo), which uses an objective that operationalizes learning representations in response to changes in self-motion [35]. Given a pair of multi-view images, CroCo reconstructs a masked view using the second view as support. Unfortunately, CroCo is a data-hungry objec-

[*] The authors contribute equally to this work.

tive. Its synthetic MULTIVIEW-HABITAT dataset of 1.8M multi-view images was curated using a method that requires ground truth 3D meshes to be annotated. Although CroCo shows promise, the lack of datasets with 3D annotations is a severe limitation, preventing its objective from scaling. If one could mine large-scale multi-view datasets, perhaps dense vision tasks could enjoy the success that the field of natural language processing has welcomed due to the availability of large-scale pretraining text [5].

In this work, we contribute MIMIC: a data-curation method for developing multi-view datasets that scale. Our method does not require any 3D meshes and can generate multi-view datasets from unannotated videos and 3D simulated environments. We leverage classical computer vision techniques, such as SIFT (Scale Invariant Feature Transform) keypoint detection [23], RANSAC [12], homography estimation [13], etc. to extract correspondences between frames in open-sourced unannotated videos (see Fig. 1). In other words, MIMIC produces a pretraining dataset for masked image modeling using image correspondences. Our work enables data curation from both real and synthetic sources and we hope it will help advance further research in large-scale dense representation learning.

We experiment with two scales: MIMIC-1M and MIMIC-3M, and show that they effectively train useful self-supervised (MAE and CroCo) representations. See **??** for example image pairs obtained from the MIMIC-3M dataset. Our experiments show the following: Most importantly, representations learned from MIMIC-3M, our automatically generated dataset, outperform those trained using ImageNet-1K [10], an expensive human-labeled dataset on dense geometric tasks: depth estimation (NYUv2 [25]) and surface normals (Taskonomy [40]); Second, MIMIC also trains better representations than MULTIVIEW-HABITAT [35], a baseline that uses 3D annotations to automatically generated dataset, on both dense geometric tasks, such as depth estimation (NYUv2) and surface normal prediction (Taskonomy), as well as on dense object-related tasks, such as semantic segmentation (ADE20K [41]) and pose estimation (MSCOCO [20]). Third, larger pretraining dataset (MIMIC-3M $>$ MIMIC-1M) improves performance, which is promising since our curation method can arbitrarily scale to produce even larger datasets. Finally, our representations demonstrate better few-shot performance on depth estimation (NYUv2) and semantic segmentation (ADE20K) compared to MULTIVIEW-HABITAT.

2. Related work

In this section, we discuss masked image modeling - a promising paradigm for self-supervised dense representation learning at scale and data curation methods for large-scale visual learning.

Masked Image Modeling. Amongst masked image modeling, BEiT [3] proposes the pre-training task of recovering the visual tokens from a corrupted image, MAE [15] learns by masking patches of an image and inpainting the masked patches; MultiMAE extends MAE to a multi-task formulation [2]. Their approach uses pseudo-labels– hence, MultiMAE is not fully self-supervised. CroCo [35] uses cross-view completion and ingests multi-view images. Their data curation method, though, uses 3D metadata and meshes of synthetic 3D environments; their dataset is also not publicly available. By contrast, MIMIC neither needs any pseudo labels extracted using supervised methods nor it needs any 3D meshes, point clouds, or camera parameters for dataset curation.

Data curation for large scale visual learning. Large-scale image datasets have incredibly accelerated progress in visual learning. ImageNet-1K, with 1.2M images annotated by crowdsourcing led to several breakthroughs and is still a standard dataset used for pretraining vision models. It was manually designed to cover a diverse taxonomy of object categories with sufficient representation of instances per category. Unfortunately, this approach is extremely costly, not scalable, and serves as an upper bound for what is possible with manual curation instead of our automatic curation. Moreover, the efforts so far have been focused on high-level semantic tasks like classification, and large-scale pretraining datasets for dense prediction tasks such as MULTIVIEW-HABITAT with synthetic image pairs mined using Habitat simulator [31] are not available publicly. MULTIVIEW-HABITAT uses annotations such as camera parameters and meshes to sample image pairs with a co-visibility threshold of 0.5. The use of such metadata for mining image pairs is a limiting factor as (1) it requires expensive sensors to obtain these annotations on real-world datasets (2) it cannot be scaled up to mine web-scale data sources where this information is not readily available. MVImgNet [38], a recent effort for building large-scale multiview dataset uses crowdsourcing to collect object-centric data. Infinigen [27] on the other hand, introduces a procedural generator of synthetic 3D scenes to create a training dataset. While these directions show promise, the need for manual intervention is a major limitation. Moreover, the use of synthetic pretraining data for real-world applications is still an open question. To address these challenges we propose a methodology for curating multi-view datasets using videos and synthetic 3D environments. We demonstrate that it is possible to use our data collection strategy and outperform on different dense vision tasks without making use of any explicit annotations.

3. MIMIC: Curating multi-view image dataset for dense vision tasks

While CroCo recently utilized MULTIVIEW-HABITAT, a multi-view dataset, their dataset creation process requires

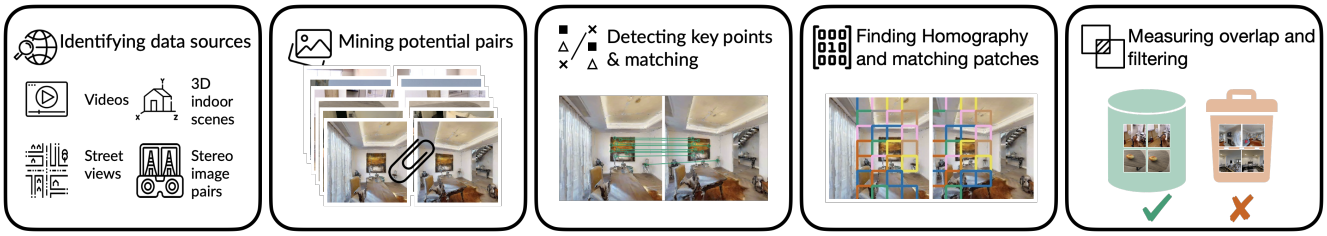

Figure 1. We introduce a data-curation method that generates multi-view image datasets for self-supervised learning. Our method identifies potential data sources, including videos of indoor scenes, people, and objects, 3D indoor environments, outdoor street views, and stereo pairs to mine potential multiview images. Next, we use classical computer vision methods such as SIFT keypoint detection and homography transformation to locate corresponding patches. Finally, we filter pairs based on a threshold for significant overlap, ensuring a substantial percentage of pixels match between a pair.

the availability of 3D mesh, point cloud, or camera pose information for each scene. This dependency imposes limitations on the range of data sources that can be used for crafting a multi-view dataset. Unfortunately, there is currently no large-scale publicly available dataset to address this void. To bridge this gap, we introduce MIMIC.

MIMIC can generate multi-view datasets from unannotated videos and 3D simulated environments. Any data source that contains multi-view information with static objects or at least with minimal object movement is a suitable data source. MIMIC works by cleverly combining traditional computer vision methods (Fig. 1). The only mechanism our curation process requires is a sampling mechanism $(I_1, I_2) \sim g(S)$, where S is some data source from which $g(\cdot)$ samples two images I_1 and I_2. For example, S can be a video from which $g(\cdot)$ samples two image frames. Or S can be a synthetic 3D environment from which $g(\cdot)$ navigates to random spatial locations and samples two random image renderings of the same scene.

Identifying data sources. We generate our MIMIC dataset from both real as well as synthetic data sources. We use DeMoN [34], ScanNet [9], ArkitScenes [4], Objectron [1], CO3D [30], Mannequin [19], and 3DStreetView [39] as real data sources. DeMoN is a dataset containing stereo image pairs. ScanNet and ArkitScenes contain videos from indoor environments. Objectron and CO3D are collections of videos containing objects. Mannequin provides a video dataset featuring individuals engaged in the mannequin challenge. 3DStreetView offers a collection of street images from multiple urban areas.

We also source data from 3D indoor scenes in HM3D [28], Gibson [36], and Matterport [7] datasets, using the Habitat simulator [31]. We initialize an agent randomly in the 3D environment and design $g(\cdot)$ to move the agent in random steps and directions. For each scene, the agent moves to numerous locations and captures various views. All our data sources with their distributions are visualized in Fig. 2.

Mining potential pairs. The primary characteristic of the image pairs in our dataset resides in their ability to capture the same scene or object from varying viewpoints while exhibiting a substantial degree of overlap. The dataset is designed to strike a balance: the overlap is not excessively large to the point of containing identical images, rendering the pre-training task trivial; nor is it excessively small, resulting in disjoint image pairs that offer limited utility, making the task only self-completion. Particularly, we discard the image pairs with a visual overlap of less than 50% and more than 70%. We base this design decision on empirical ablations performed in CroCo. Their experiments suggest that cross-view completion offers no advantage if the visual overlap is outside of this range.

In each video or scene, many image pairs can be generated. However, we focus on selecting a limited number of pairs that are more likely to meet our desired condition of having sufficient overlap. Nonetheless, not all of these candidate pairs may ultimately be chosen. For instance, when dealing with video data, a practical strategy involves creating a list of frames at regular time intervals, which depends on the video's speed. By selecting consecutive frames from this list, potential pairs are generated. Conversely, collecting potential pairs in 3D scenes such as HM3D [28] or Gibson [36] presents greater challenges. Therefore, inspired by CroCo, we employ the habitat simulator [31] to capture comprehensive environment views. The agent undergoes random rotations and movements, exploring the scene from various perspectives. By capturing images during these random walks, we generate potential pairs for further analysis. The selection process involves filtering based on a specified overlap range (50% to 70%) and ensuring the inclusion of pairs with diverse viewpoints. However, our approach does not rely on additional annotations and solely utilizes the available images.

Matching and measuring overlap. Given a potential image pair capturing a scene, we employ the widely used SIFT features to localize the key points in both images. Note that these features are used only to quantify the visual overlap and the actual learning happens during the pretraining

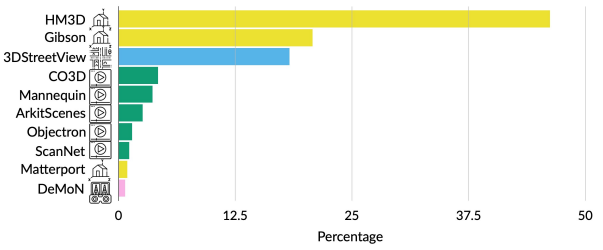

Figure 2. Distribution of Data Sources (%). Real data sources, including DeMoN, ScanNet, ArkitScenes, Objectron, CO3D, Mannequin, and 3DStreetView, contribute to 32% of MIMIC. The remaining portion consists of synthetic sources, namely HM3D, Gibson, and Matterport.

phase.

After obtaining the key points and descriptors, we apply a brute-force matching technique to establish correspondences between the key points in the first image and those in the second image. More efficient methods, such as FLANN matcher [24], may offer ($\approx 1.24\times$) speedups. However, our initial exploration shows that brute-force matching yields better matches; also, extracting pairs is a one-time process. We further utilize these matches to estimate the homography matrix, using the RANSAC (Random Sample Consensus) algorithm to eliminate outliers. Note that the homography transformation holds in three scenarios–(1) when capturing planar surfaces, (2) when capturing a distant scene, and (3) when a camera undergoes a pure rotation. In real-world videos, these assumptions may not always hold true. Regardless, homography serves as an approximation to the transformation. We further use this approximated matrix to filter out unwanted image pairs with no visual overlap.

We then partition each image into non-overlapping patches. For each patch in the first image, we search for the corresponding patch in the second image with the highest overlap. We randomly sample points within the first image and match them with their correspondences in the second image. Next, we map each patch in the first image to the patch with the highest number of matched correspondences in the second. Lastly, we measure visual overlap by calculating the total number of matched patches divided by all patches. Refer to the Appendix for more details.

Filtering out degenerate matches. In our approach, the selection of image pairs is guided by the objective of capturing shared 3D information while mitigating redundancy. Hence, the desired pairs consist of images that depict the same objects or scenes from different perspectives. This characteristic enables the learning model to acquire valuable insights about the underlying 3D structure. However, it is crucial to avoid including pairs where one image is a zoomed-in version of the other, as such pairs provide limited additional information.

To address this concern, we modify the overlap metric used in the pair selection process. Specifically, we incorporate a criterion that prevents the inclusion of patches from the first image that have exact correspondences in the second image. Therefore, in the counting, we consider all patches that have the same corresponding patch in the second image as a single entity.

Overall statistics. To understand the effect of data size we experiment with two scales. MIMIC-1M, comprises a total of $1,316,199$ image pairs, each capturing different scenes or objects from varying viewpoints. Among these pairs, $761,751$ are sourced from HM3D, $305,197$ from Gibson, $29,658$ from Matterport, $114,729$ from Mannequin, $22,184$ from DeMoN, $36,433$ from ScanNet, and $46,250$ from Objectron. We further expand the dataset to create a larger version, MIMIC-3M, to contain a total of $3,163,333$ image pairs. This expansion involves augmenting the HM3D dataset with an additional $699,322$ pairs, the Gibson dataset with $351,828$ pairs, and the inclusion of new datasets such as ArkitScenes with $81,189$ pairs, CO3D with $133,482$ pairs, and 3DStreetViews with $579,310$ pairs. By incorporating these new datasets, we further enrich the diversity and quantity of image pairs available in our dataset.

4. Training with MIMIC

We analyze the effectiveness of MIMIC, by training two models with masked image modeling objectives and evaluate the learned representations on downstream dense prediction tasks. We compare against existing pretraining dataset alternatives.

4.1. Pretraining

We use MAE [15] and CroCo [35] for pretraining. We follow the protocol from CroCo and use a ViT-B/16[11] as a backbone for all our experiments with input images sizes of 224×224. We train our models on 8 RTX A6000 GPUs for 200 epochs with a warmup of 20 epochs with a base learning rate of 1.5×10^{-4}, an AdamW [22] optimizer with a cosine learning rate schedule, a weight decay of 0.05, and an effective batch size of 4096. Lastly, we evaluate these pretrained representations on a series of downstream dense prediction tasks.

MAE pretraining. To understand the importance of including correspondences in the pretraining objective, we train MAE, which does not encode multi-view correspondences and treats each image in our image pairs independently. MAE masks out a large portion (75%) of the input patches of an image and uses an asymmetric encoder-decoder architecture to reconstruct the masked-out pixels. Specifically, it uses a ViT-based encoder to extract the latent representations of the masked view. Then it pads the output with the masked tokens and feeds it to a lightweight

decoder. The decoder's output reconstruction is optimized with an L2 loss. The reconstruction pixel targets are normalized by computing the mean and standard deviation of the image patches.

CroCo pretraining. Unlike MAE, CroCo aims to encode relationships between the two views of the same scene from different viewpoints and learns to reason about the illumination and viewpoint changes. CroCo reconstructs a masked image input similar to MAE but supports the reconstruction process through an unmasked second reference view. CroCo masks 90% of the first image. CroCo uses the same ViT encoder as MAE, with shared weights to encode both views. The decoding cross-attends over the second view while reconstructing the first masked view.

4.2. Baseline Datasets.

We compare MIMIC with: ImageNet-1K [10] and MULTIVIEW-HABITAT [35].

ImageNet-1K is a widely used large-scale dataset with 1.2M training images. It was manually designed to cover a taxonomy of a thousand object categories. The images were chosen to have sufficient instances per category. Therefore, ImageNet-1K serves as an example for what is possible with immense human data-curation effort.

MULTIVIEW-HABITAT comprises of synthetic renderings of indoor scenes collected using the 3D meshes available in the Habitat simulator [31]. It is derived from HM3D [28], ScanNet [9], Replica [32] and ReplicaCAD [33]. This dataset is not available publicly. So, we compare it against the released models trained on it. MULTIVIEW-HABITAT serves as our main baseline dataset since it is the only large-scale multi-view dataset that has been used for training use representations for dense vision tasks.

4.3. Downstream tasks, datasets, evaluation protocols.

We evaluate our models on two dense geometric tasks: depth estimation and surface normal estimation. We also evaluate on two dense object-related tasks: semantic segmentation, and pose estimation. Finally, we report object classification numbers for completion. We provide below the details of the datasets, metrics, and protocols used for fine-tuning and evaluations.

Depth Estimation involves estimating the depth of each pixel of an input image from the camera. For evaluation, we use the NYUv2 [25], a dataset of RGB images and their corresponding ground truth depth maps. It consists of 795 training and 654 test images of indoor scenes. We report the $\delta 1$ metric on the test images - which computes the percent of the pixels with error $max(\frac{y_{p_i}}{y_{g_i}}, \frac{y_{g_i}}{y_{p_i}})$ less than 1.25, where y_{p_i} is the depth prediction and y_{g_i} is the ground truth of the ith pixel of an image. We use DPT [29] head as in MultiMAE for finetuning.

Surface Normals, Edges, Depth, and Curvature Estimation are regression tasks that aim to estimate the orientation, edges, depth, and bend of a 3D surface respectively. We use a subset of Taskonomy [40] with 800 training images, 200 validation images, and $54,514$ test images. We use the L1 loss value on the test set as evaluation metric.

Semantic Segmentation entails assigning a class to each pixel of an image based on its semantic category. We use ADE20K [41], with $20,210$ training images and 150 semantic categories, Cityscapes with 19 classes, 2975 training images, 500 validation images, and 1525 test images, and NYUv2 dataset with 795 training and 654 test images.

We report the mIOU which quantifies the percentage overlap between the predictions and the ground truth annotations. We use a segmentation head based on ConvNext [21] adapter for finetuning.

Classification is a high-level semantic task that involves assigning a category to an image based on its content. We use ImageNet-1K[10] which contains 1.28M training images and 50K validation images. This task allows us to measure how large the gap is when models are pretrained for dense tasks in mind. We follow the linear probing protocol from MAE and report accuracy.

Pose Estimation involves detecting keypoints and their connections in an image. We use the MSCOCO [20] dataset for finetuning and report Average Precision and Average Recall on the validation set. Specifically, we adopt ViTPose-B [37] for finetuning.

5. Experiments with MIMIC-3M

We evaluate our pre-trained models on two dense geometric vision tasks – depth estimation and surface normal prediction. MIMIC-3M's dense representations outperform both tasks (Tab. 1)Next, we finetune our encoders for pixel-level tasks that also require object understanding – semantic segmentation, and pose estimation, and high-level semantic tasks – image classification. For these three tasks, our experiments demonstrate that models trained using our automatically generated data close the gap with models trained on ImageNet-1K (Tab. 2). We further experiment with the data size used for pretraining and showcase that more data leads to improvements on depth estimation and semantic segmentation tasks (Tab. 3). Unlike CroCo trained on MULTIVIEW-HABITAT, our pre-trained models do not saturate or degrade over time on depth estimation and semantic segmentation (Fig. 3 (a)). Our performance benefits also hold as we vary the number of fine-tuning data points available for both depth estimation and semantic segmentation (Fig. 3 (b)) Finally, we find that our models produce higher-quality reconstructions using the pretraining decoder (Tab. 4).

Table 1. CroCo pretrained with MIMIC-3M outperforms MAE, MultiMAE and DINO pretrained on ImageNet-1K obtained with expensive human annotations and ensured to have diverse categories as well as MULTIVIEW-HABITAT collected using 3D annotations. We report the results from the CroCo paper (marked with *) as well as those with our task-specific fine-tuning setup adopted from MultiMAE.

Model	Frozen	Dataset	NYUv2 (↑)	Taskonomy(↓)			
			depth est. $\delta 1$	surface norm. L1	Curv. L1	Edges L1	Depth L1
DINO	×	ImageNet-1K	81.45	65.64	48.25	43.41	38.60
MAE	×	ImageNet-1K	85.1	59.20	41.61	34.97	34.81
MultiMAE	×	ImageNet-1K	86.4	60.86	42.36	52.90	33.40
MAE	✓	MV-HABITAT	-	-	-	-	-
MAE	✓	MIMIC-3M	80.65	68.97	44.01	33.63	38.46
MAE	×	MV-HABITAT	79.00	59.76	-	-	-
MAE	×	MIMIC-3M	85.32	58.72	42.71	25.00	30.45
DINO	✓	MIMIC-3M	77.98	74.59	43.77	48.25	36.44
CroCo	✓	MV-HABITAT	85.20* (84.66)	64.58	42.66	28.44	34.86
CroCo	✓	MIMIC-3M	85.81	61.70	42.99	25.95	35.85
DINO	×	MIMIC-3M	78.67	65.43	43.37	37.79	36.44
CroCo	×	MV-HABITAT	85.60* (90.19)	54.13	**41.24**	**22.90**	32.82
CroCo	×	MIMIC-3M	**91.79**	**53.02**	41.35	23.96	**30.33**
			+1.60	-1.11	+0.11	+1.06	-2.49

Does MIMIC-3M improve dense geometric tasks? We finetune our trained models on two dense geometric tasks: NYUv2 depth estimation and Taskonomy surface normal prediction. We also finetune the CroCo models trained on MULTIVIEW-HABITAT using task-specific decoders adopted from MultiMAE and report their improved results.

Even though MIMIC-3M was generated automatically, without manual intervention, and uses no 3D annotations, representations pretrained on MIMIC-3M perform better on both dense geometric tasks (Tab. 1). These gains can be attributed to the inclusion of real sources–thanks to the flexibility of our method which allows us to use real-world videos of complex scenes as a data source.

We also validate the utility of multi-view correspondences by comparing MAE with CroCo models. CroCo offers significant gains over MAE on MIMIC-3M demonstrating the benefits of using correspondences during pretraining (Tab. 1). We observe that CroCo when trained on MIMIC-3M leads to the state-of-the-art $\delta 1$ of 91.79 NYUv2 depth and L1 of 53.02 on surface normals and performs equally well with model trained on MULTIVIEW-HABITAT using MIM pretraining.

Does MIMIC-3M pretraining improve dense semantic tasks? To understand the potential of MIMIC for dense tasks which also require object-level understanding, we evaluate MAE and CroCo pretrained with MIMIC-3M on ADE20K, NYUv2, Cityscapes semantic segmentation and MSCOCO pose estimation (Tab. 2). We observe gains in comparison to the MULTIVIEW-HABITAT on ADE20K and MSCOCO. We hypothesize that these improvements come from the real-world object-centric data from Objectron and Co3D. We also evaluate classification accuracy on ImageNet-1K. When compared to MULTIVIEW-HABITAT, MIMIC-3M reduces the linear probing performance gap by 7.36% with MAE and 2.64% with CroCo on manually curated, object-centric, and human-annotated ImageNet-1K.

Does scaling up MIMIC improve finetuning performance? We study the scaling effect of MIMIC by varying the data size. We experiment with two scales: the first MIMIC-1M with 1.3M image pairs and the second MIMIC-3M with 3.1M image pairs. We train CroCo with these two training sets and evaluate the performance on depth estimation (NYUv2), semantic segmentation (ADE20K), and surface normals (Taskonomy) (Tab. 3). We observe consistent improvements: $\delta 1$ by 2.33, mIOU on ADE20K by 3.73, and L1 loss by 4.10.

Do MIMIC-3M representations quality saturate with training iterations? In contrast to models trained on MULTIVIEW-HABITAT, we do not observe performance saturation or degradation with pretraining iterations (see Figure 6 in their paper [35]). Instead, the performance of both MIMIC-1M and MIMIC-3M improves on depth estimation and semantic segmentation (Fig. 3(a)) for an iterations-matched training run. This trend holds regardless of whether the representations are fine-tuned or kept frozen.

This suggests that more training with the cross-view completion objective helps build better representations for dense downstream tasks.

Does MIMIC-3M pretraining improve few-shot finetuning? We measure the label efficiency of the learned representations trained on MIMIC-3M by evaluating its few-shot performance on NYUv2 depth estimation and ADE20K semantic segmentation. We freeze the image en-

Table 2. MIMIC-3M, our automatically generated dataset shows improvements over MULTIVIEW-HABITAT on dense object-related tasks such as ADE20K semantic segmentation and MSCOCO pose estimation. It even improves on ImageNet-1K classification and further closes the gap with models pre-trained on ImageNet-1K, curated with expensive crowdsourcing.

Model	Pretraining dataset	ADE-20K(↑)	Cityscapes (↑)	NYU(↑)	MSCOCO pos. est.(↑)		ImageNet-1K cls.(↑)
		mIOU	mIOU	mIOU	AP	AR	% accuracy
DINO	MIMIC-3M	37.82	65.57	40.75	70.23	76.40	52.01
MAE	MV-HABITAT	40.30	-	-	-	-	32.50
MAE	MIMIC-3M	40.54	70.47	43.76	69.13	75.22	39.86
CroCo	MV-HABITAT	40.60* (41.33)	71.84	47.10	66.50	73.20	37.00
CroCo	MIMIC-3M	42.18	71.24	46.61	72.80	78.40	39.64
		+0.85	-0.60	-0.49	+6.30	+5.20	+2.64
MAE	ImageNet-1K	**46.10**	**73.98**	**49.12**	**74.90**	**80.40**	**67.45**

Table 3. MIMIC-3M shows improvements over MIMIC-1M on depth estimation (NYUV2), Semantic Segmentation (ADE20K), Surface Normals Estimation (L1)

Dataset	Frozen	NYUv2(↑)	ADE20K(↑)	Taskonomy(↓)
		$\delta 1$	mIOU	L1
MIMIC-1M	✓	82.67	27.47	67.23
MIMIC-3M	✓	**85.81**	**30.25**	**61.70**
		+3.14	+2.78	-5.53
MIMIC-1M	×	89.46	38.45	57.12
MIMIC-3M	×	**91.79**	**42.18**	**53.02**
		+2.33	+3.73	-4.10

coder and fine-tune the task-specific decoders by varying the number of training images. We run each k-shot finetuning at least 5 times and report the mean and the standard deviation of the runs. For depth estimation, we also experimented with k-shot regimes where k is less than 10. Overall the representations trained on our MIMIC-3M show better labeling efficiency than those trained using MULTIVIEW-HABITAT (Fig. 3(b)). These gains can be attributed to the diverse, and real-world training data during pretraining.

Does MIMIC-3M pretraining improve reconstruction quality? We analyze the quality of the reconstructions trained on MIMIC-3M versus MULTIVIEW-HABITAT. We use FID scores [17], which indicate how realistic the reconstructions are and the reconstruction error (L2 loss) in the original masked image modeling objective. We sample a test set of 500 images from the Gibson dataset. We ensure that these images are sampled from the scenes that are exclusive of MULTIVIEW-HABITAT and MIMIC-3M pretraining datasets. We mask 90% of each test image and then compare the quality of the reconstructions (Tab. 4). Our analysis shows that CroCo trained on MIMIC-3M improves the FID by 12.65 points and reduces the reconstruction loss on the test set (see Appendix for visualizations).

What are the effects of data sources on MIMIC-3M representations? To understand the effect of data sources we experiment with two subsets of MIMIC-3M.

Table 4. MIMIC-3M achieves better FID score and reduces the reconstruction loss on 500 test images from the Gibson dataset compared to MULTIVIEW-HABITAT

Model	Dataset	Reconst. loss (↓)	FID score (↓)
CroCo	MV-HABITAT	0.357	85.77
CroCo	MIMIC-3M	**0.292**	**73.12**

First, we train a CroCo model on a 1.8M subset of MIMIC-3M obtained from three synthetic sources such as HM3D, Gibson, and Matterport. Second, we train and evaluate on 1.3M subset of MIMIC-3M constituting real sources such as Scannet, Mannequin, Objectron, DeMon, Co3D, and ArKitScenes. Tab. 5 shows the results of these experiments. Interestingly NYUv2 with a model pretrained on 1.3M images achieves better δ_1 compared to the one pretrained on a larger 1.8M synthetic subset.

6. Discussion

We present MIMIC, an approach to curate large-scale pretraining datasets from real-world videos and synthetic environments, geared towards dense vision tasks. Our work aims to provide a holistic solution that requires no manual intervention and domain knowledge about the data sources. We discuss below the limitations and safety considerations regarding our dataset and lay out opportunities for future work.

Limitations. There are several limitations of our work. First, we pretrain CroCo on MIMIC-3M using a fixed-sized architecture ViT-B/16; model scaling experiments are outside the scope of this work. Second, our curated dataset primarily consists of static objects and does not involve dynamic scenes. Lastly, MIMIC-3M has a small amount of object-centric data, and its suitability for object-related tasks is limited. Including more object-centric sources may help bridge this gap.

Safety and ethical considerations. While MIMIC uses publicly available datasets for data curation, we acknowledge that the algorithm can be scaled up to scrape videos in the wild. We are aware of the privacy, and ethical is-

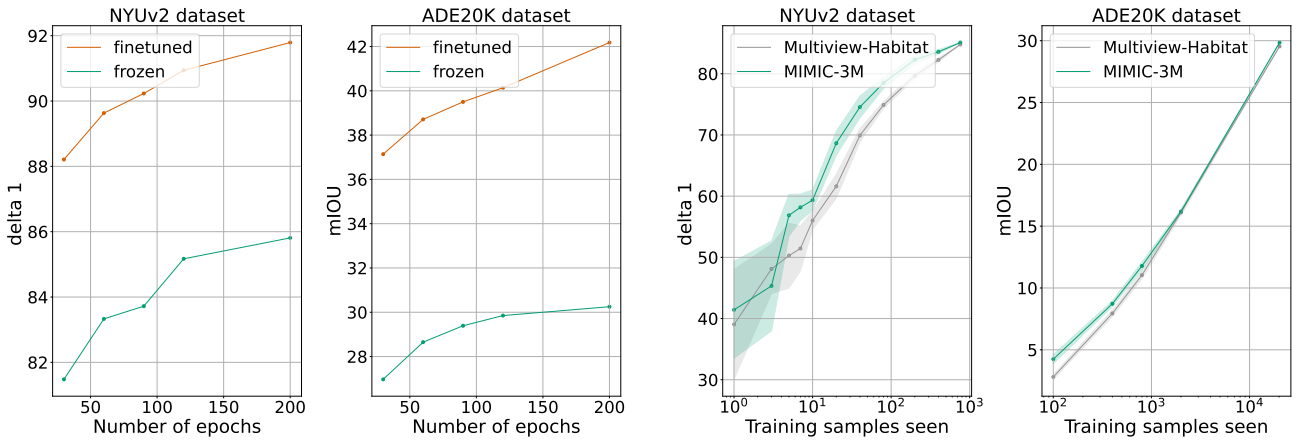

Figure 3. **(a)** CroCo pretrained on MIMIC shows an increasing trend with the number of training epochs. The figure on the left shows the trends for the fine-tuned and frozen versions of the encoder on NYUv2 depth estimation. The figure on the right shows the trend on the ADE20K dataset. **(b)** CroCo pretrained on MIMIC-3M achieves better few shot performance on CroCo pretrained on MULTIVIEW-HABITAT. The figure on the left shows the few shot performance on the NYUv2 dataset and the figure on the right shows the few shot performance on ADE20K.

Table 5. CroCo pretrained on smaller MIMIC-REAL achieved higher δ_1 on NYUv2 depth estimation compared to larger MIMIC-SYNTHETIC

Model	Dataset	Size	NYUv2 depth.est.(↑)	Taskonomy surf.norm.(↓)	ADE20K sem.seg.(↑)
CroCo	MULTIVIEW-HABITAT	1.8M	85.60* (90.19)	54.13	40.60* (41.33)
CroCo	MIMIC-3M	3.1M	91.79	53.02	42.18
CroCo	MIMIC-SYNTHETIC	1.8M	81.03	**60.05**	**37.52**
CroCo	MIMIC-REAL	1.3M	**84.80**	65.35	36.82

sues caused by models trained on large-scale datasets and the amplification of the biases these models may result in. As such, we ensure to limit our data sources to only open-sourced video datasets. Lastly, we recommend the use of face blurring and NSFW filtering before scraping internet videos.

Future work. We would like to design methodologies to mine dynamic videos where epipolar geometric constraints do not apply, design new objectives for pretraining on image pairs curated using MIMIC, and evaluate representations on more diverse tasks. The flexibility of MIMIC makes it suitable for further scaling it up to even larger pretraining datasets.

7. Acknowledgements.

This research is sponsored by grant from Amazon Technologies, Inc. as part of the Amazon-UW Science HUB. We thank Michael Wolf and Ariel Gordon for helpful discussions, Saygin Seyfioglu for helpful feedback, Mitchell Wortsman for help with pretraining, Wei-Chiu Ma and Zixian Ma for providing comments on earlier drafts of this paper. We also thank the UW-IT team: Stephen Spencer, Nam Pho, and Matt Jay.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7822–7831, 2021. 3

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer, 2022. 1, 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 2

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

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

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3, 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2, 4

[16] Richard Held and Alan Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 56(5):872, 1963. 1

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[18] Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[19] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4521–4530, 2019. 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 4

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[24] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications*, 2009. 4

[25] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5

[26] Nancy Rader, Mary Bausano, and John E Richards. On the nature of the visual-cliff-avoidance response in human infants. *Child development*, pages 61–68, 1980. 1

[27] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023. 2

[28] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 3, 5

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 5

[30] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 3

[31] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 5

[32] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[33] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 5

[34] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 3

[35] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 1, 2, 4, 5, 6

[36] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018. 3

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 5

[38] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9150–9161, 2023. 2

[39] Amir R Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3d representation via pose estimation and matching. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 535–553. Springer, 2016. 3

[40] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2, 5

[41] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 2, 5