Enhancing 2D Representation Learning with a 3D Prior

Mehmet Aygün[1][*]   Prithviraj Dhar[2]   Zhicheng Yan [2]   Oisin Mac Aodha[1]   Rakesh Ranjan[2]

[1] University of Edinburgh   [2] Meta Inc.

Abstract

Learning robust and effective representations of visual data is a fundamental task in computer vision. Traditionally, this is achieved by training models with labeled data which can be expensive to obtain. Self-supervised learning attempts to circumvent the requirement for labeled data by learning representations from raw unlabeled visual data alone. However, unlike humans who obtain rich 3D information from their binocular vision and through motion, the majority of current self-supervised methods are tasked with learning from monocular 2D image collections. This is noteworthy as it has been demonstrated that shape-centric visual processing is more robust compared to texture-biased automated methods. Inspired by this, we propose a new approach for strengthening existing self-supervised methods by explicitly enforcing a strong 3D structural prior directly into the model during training. Through experiments, across a range of datasets, we demonstrate that our 3D aware representations are more robust compared to conventional self-supervised baselines.

1. Introduction

The visual stimuli processed by a binocular, actively moving, human observer provides direct information about the 3D world around them [21]. As a result, humans have a remarkable ability to perceive useful 3D shape cues, enabling them to interact and navigate adeptly in complex environments. Most impressively, the power of the human visual system is not understood to be a resulting property of supervised learning, i.e., it has developed thanks largely to 'self-supervision' [62].

While great advances have been made in the past decade in developing computer vision systems, their success can be mostly attributed to large-scale *supervised* representation learning. Moreover, current artificial vision systems are not yet nearly as robust as the human equivalent [19]. For example, existing commonly used architectures are known to heavily rely on texture cues, which results in sub-optimal generalization performance [19, 46]. Encouragingly, neural
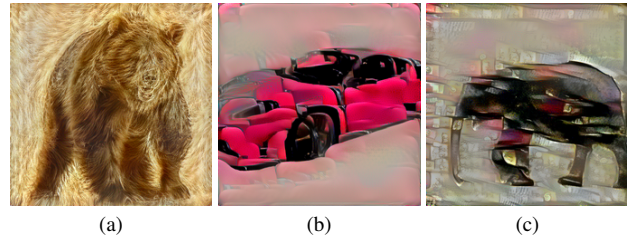
(a)   (b)   (c)

Figure 1. Humans have no difficulty in recognizing the categories depicted in the above images, even though the texture of the objects has been perturbed. This is thought to be in large part due to our reliance on shape, as opposed to texture, cues [19, 42, 64]. However, an automated recognition system built on top of a state-of-the-art self-supervised representation learning approach (i.e., DINOv2 [50]) classifies these examples as dog, chair, and knife respectively, as the texture of the images resembles those object classes. We introduce a new approach to improve the robustness of self-supervised methods using a proxy 3D reconstruction task which encourages representations that emphasize shape cues more. As a result, our approach correctly predicted these examples as bear, car, and elephant.

networks that also make use of more shape cues have also been observed to be more robust to different types of image distortions [19].

These observations point to two important questions that are potentially hindering our artificial vision systems: (i) how do we reduce the over-reliance on supervised labeled data and (ii) how do we encourage models to make greater use of shape information to improve their robustness? Thankfully, great progress has been made on the first question as we now have methods for obtaining effective visual representations through self-supervision alone, e.g., [2, 8, 31, 50, 70]. While methods exist for extracting shape-adjacent information in the form of depth using self-supervision from collections of image pairs [23] or video sequences [77], these approaches tend to require strong assumptions about the scenes they are trained on (e.g., smooth camera motion, static scenes, limited visual diversity, *etc*.). As a result, the current most effective approaches for predicting depth require explicit depth supervision during training [56]. Moreover, even when depth supervision is available, it is not trivial to use it to improve the performance on other tasks [65, 73].

In this work, we attempt to address these combined challenges by proposing a new method to improve existing self-supervised representation learning approaches by enforcing these models to reason about object/scene shape during training. We build on recent advances in 3D generative modeling [6, 61] to develop a self-supervised reconstruction method that generates a 3D representation of the input image. Our model is trained with a self-supervised reconstruction objective, starting from an already trained self-supervised network (e.g., [50]). Given an input image, we first extract a global feature representation using a pre-trained backbone network and then predict a 3D representation of the scene depicted in the image. Then we reconstruct appearance and depth maps using volume rendering from the predicted 3D representation. We use the difference between the reconstructed image and the original input image, and the difference between the predicted depth map and its pseudo ground truth as our training objectives. We do not utilize any manual labels during training as we only require an unordered (i.e., not from videos or stereo pairs) collection of monocular images and their corresponding estimated depth from a previously trained depth prediction model [4] as input. To minimize the training loss, the learned image representation needs to capture details about the shapes of the objects depicted in the input scenes.

While conceptually simple, the advantage of our approach is that it works with monocular image collections and does not make strong assumptions about the types of images it is trained on. As a result, we can train it using standard representation learning datasets such as ImageNet [57]. Quantitative and qualitative results illustrate that our shape-aware representations are more robust compared to variants that are not shape aware on a variety of downstream tasks. See Figure 1 for a qualitative example.

In summary, we make the following contributions: (i) We explore the role of 3D information when performing self-supervised learning on unordered monocular image collections. (ii) We propose a new method that enhances self-supervised learned representations via a proxy task that explicitly encodes 3D knowledge during training. (iii) When applied to a range of robustness tasks, our approach obtains superior performance compared to baselines that do not make use of 3D information at training time.

2. Related work

In this section, we discuss related work in self-supervised learning, monocular shape understanding, and the role of shape in visual recognition.

Self-supervised learning. Recent approaches for deep learning-based self-supervised learning (SSL) in computer vision can be categorized into two groups: (i) predictive methods, where the learning objective depends solely on the input image, and (ii) discriminative approaches, which use additional images as inputs.

Predictive tasks include context prediction (e.g., patch or pixel prediction from a masked input image) [2, 7, 13, 31, 76], colorization of grayscale input images [75], inpainting of randomly selected areas [52], predicting image rotation [22], or object counting in the input image [48]. In contrast, discriminative approaches aim to learn representations that make the input image, and an augmented version of it, more similar to each other compared to other randomly selected images [8, 14, 26, 29, 49, 50, 70, 74]. Regularization to prevent trivial solutions [26, 50, 74] and selecting challenging negative examples [9, 30] are important considerations for these methods. It is worth noting that some of the above methods make use of both types of losses. For a more comprehensive overview of SSL approaches, we direct the reader to [27, 37].

One limitation of the above approaches is that their focus is on 2D representation learning. In this work, we aim to enhance the robustness of self-supervised networks by utilizing a 3D proxy task during training. Recently, [72] introduced a new dataset consisting of common everyday objects containing multiple images, from different camera viewpoints, for each object instance. Their dataset is significantly larger than existing comparable multi-view datasets (e.g., [34]). They use this data to perform view-consistent self-supervised fine-tuning and show that this pseudo-3D supervision results in better downstream image classification performance on their dataset. However, multi-view data of this form is still very cumbersome and time-consuming to collect and thus current datasets are still limited in their scope.

Recently, a new synthetic dataset named Photorealistic Unreal Graphics (PUG) [5] was introduced. It could be used as a source of multi-view data as the images are rendered using 3D assets. However, the images lack realism and the diversity of objects is still not on par with large scale 2D image collections. In this work, we show that it is possible to inject 3D information into a self-supervised model by training on single-view (i.e., not multi-view) image collections alone.

Single-view 3D understanding. Our approach uses a proxy monocular 3D reconstruction task during training to enhance SSL performance. There is also a body of work that aims to estimate 3D shape from monocular images where their focus is on generation and not representation learning.

Example existing works estimate partial 3D shape in the form of depth maps, i.e., per-pixel continuous depth predictions. These methods either use pseudo ground truth depth supervision during training [4, 56] or are trained without depth supervision via image reconstruction losses [18, 23, 24, 77]. Another line of work attempts to estimate the full 3D geometry of objects using 3D category priors using ex-

plicit representations like meshes [38], implicit representations like surface maps [28], or with skeletons [69]. The disadvantage of these methods is that they require strong category shape priors (e.g., a 3D deformable model of a human). More recently, there have been some category-centric works that attempt to relax the need for strong shape priors [1, 44]. However, these are still category focused and are thus limited to specific classes of objects that have well-defined shapes (e.g., animals or humans).

The specific choice of 3D representation (e.g., volume, mesh, or points) used by these methods can have a big impact on the quality of the 3D generated outputs and the computation required to train the model. In the last few years, implicit 3D representations parameterized via neural networks have become widely adopted for a range of 3D tasks [43, 47, 71]. However, conventional implicit networks can be very slow to train, which hinders their applicability to large-scale SSL. To address this, in this work, we make use of efficient implicit representations popularized by methods that perform 3D generative modeling from single image collections [6, 61].

Shape and semantics. Finally, we review work that utilizes shape information for visual recognition. It is well established, especially in the early years of cognitive development, that infants more heavily rely on shape cues compared to other cues such as texture during early category learning [42, 63, 64]. However, computational methods like CNNs [19] and Vision Transformers [20, 46] do the opposite. With more data, and bigger models, there is some evidence to suggest that this over reliance on texture may decrease [12], but it still does not fully disappear.

Prior to the wide adoption of deep-learning methods in computer vision, there were a large number of works that utilized (2D) shape information for recognition tasks. Examples include seminal works such as pictorial structures [17] and deformable templates [11, 16, 36, 53]. Subsequently, end-to-end trained approaches that did not use any structure or shape overtook these methods. However, recently a new set of methods have been developed that illustrate the benefit of using explicit shape information when combined with end-to-end learning methods for tasks like tracking [54] and action recognition [55].

Furthermore, recent studies have employed alternative forms of training data, such as styled images [19] and edge maps [45], to enhance shape awareness, albeit in a supervised context. In this work, we take inspiration from human cognition to add more shape information into our models by developing a proxy 3D reconstruction task to enhance SSL. To solve the resulting 3D reconstruction task, our model needs to learn more about the shape, and not just the texture, of objects during training.

3. Method

The aim of visual representation learning is to learn a function that can map an input image into a representation. This is achieved by optimizing an objective function on a set of training data. For self-supervised learning (SSL), the objective function is optimized without using human-provided supervision [2, 8, 31, 50, 70]. However, given the lack of large-scale and semantically diverse datasets containing 3D information, current self-supervised methods are typically limited to using 2D unordered image collection during training. As a result, the learned representations that emerge from models trained on 2D images are not necessarily fully capable of capturing all the properties of the 3D world [20]. In this work, we aim to improve these learned representations by using an additional proxy 3D task during training. Our aim is not to learn a new function from scratch, but to instead improve an existing pre-trained one.

3.1. Background

Our goal is to learn a representation function $f(.)$, represented as a neural network, that can map an input image I into a representation $h = f(I)$, where $h \in \mathcal{R}^d$. This will be achieved by optimizing an objective function $\mathcal{L}$ on a set of training data, without using any manually labeled data. We want to improve networks that are trained without supervision, since it has been shown in [25] that SSL-based backbones like DINOv2 [50] outperform supervised counterparts as a global image representation.

There are a large number of publicly available self-supervised models that can extract useful representations from 2D images. Therefore, instead of learning a new representation function from scratch, we utilize a backbone that is already pre-trained with 2D self-supervised methods such as DINOv2 [50], and improve it by training it on a new proxy 3D task.

3.2. 3D aware robust representation learning

We use 3D reconstruction as our proxy 3D task, i.e., given an input monocular image, at training time, the network is tasked with reconstructing the 3D scene/objects depicted in the image. The intuition behind this is that for the network to successfully perform reconstruction, it must also learn 3D aware features from the input images. As we want the network to learn image representations that transfer well to a large variety of scenes, the input images should be visually diverse and the 3D representation should be able to model complex scenes with multiple objects and diverse backgrounds. Moreover, the reconstruction task should be relatively computationally efficient to enable large-scale training. While there are alternative approaches for generating 3D predictions from 2D images [33, 51], motivated by the need for efficiency we opted to use triplanes [6] as our 3D
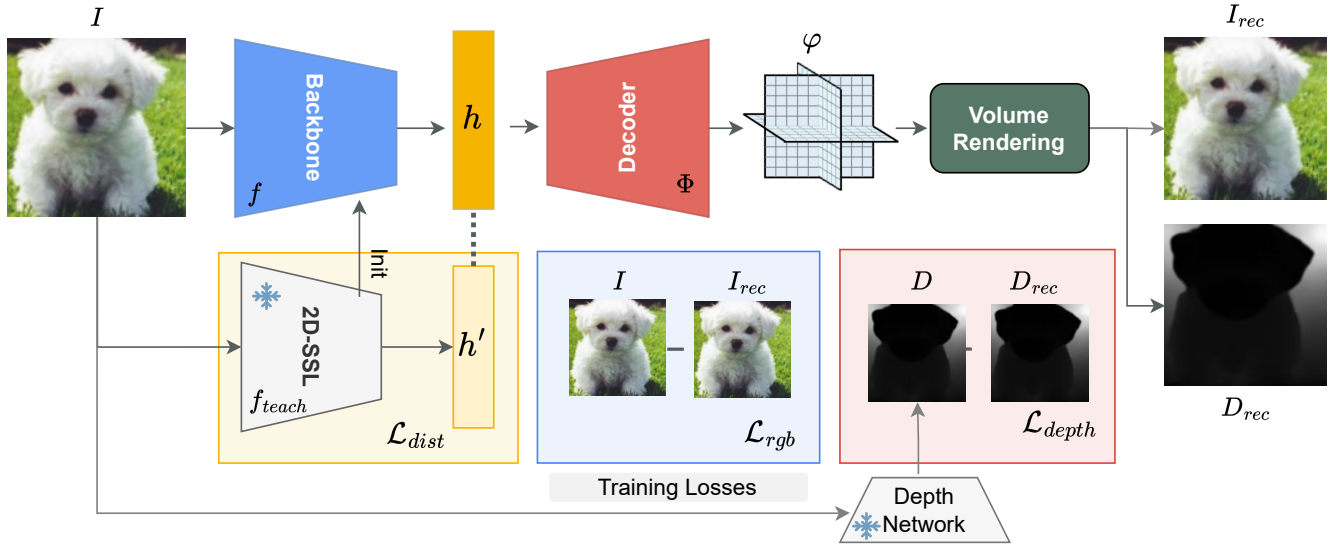
Figure 2. Overview of our self-supervised single-view 3D reconstruction approach. Given an input image, I, we first extract a representation of the image using an encoder network, $h = f(I)$. Then using a decoder network, Φ, we generate triplane features [6, 61]. Using volume rendering [43], conditioned on a fixed camera location, we reconstruct the input image, I_{rec}, and its depth D_{rec}. We optimize all networks using a combination of reconstruction losses on the input image, $\mathcal{L}_{rgb}$, and estimated depth, $\mathcal{L}_{depth}$, along with a distillation loss, $\mathcal{L}_{dist}$, from a frozen 2D self-supervised learning model to prevent the forgetting of already learned informative representations.

representation. Triplanes explicitly encode latent network features on axis-aligned planes. These features can then be aggregated via lightweight implicit feature decoders to perform efficient volume rendering for 3D reconstruction. Recently, [61] showed that triplanes can be used to generate 3D depictions of images for various types of scenes from visually diverse datasets such as ImageNet [57].

Formally, given an input image I, we first extract a global image representation $h = f(I)$ using a backbone feature extractor network $f(.)$. This backbone can be pre-trained using a 2D self-supervised method. Then we use a decoder $\Phi(.)$ to generate triplane features from the representations of the input image. Note that we only require the decoder and triplane at training time and they can be discarded at inference as we only need to retain the backbone.

This decoder takes the backbone features as input and produces triplane features $\varphi = \Phi(h), \varphi \in \mathcal{R}^{H \times W \times C \times 3}$. The decoder consists of two components, a set of learnable triplane embeddings $\xi \in \mathcal{R}^{(h \cdot w \cdot 3) \times D}$ and upsampling blocks. As done in [58], we first apply cross-attention between the triplane embeddings and the image representation to obtain low resolution triplane features, $\varphi' = \text{cross}(\xi, h), \varphi' \in \mathcal{R}^{h \times w \times C \times 3}$. Then we apply upsampling layers, which consist of bilinear upsampling and convolution operations, to obtain full resolution triplane features $\varphi = \text{upsample}(\varphi')$. Different than [58], we do not employ any quantization or style mapping [40] as our goal is not to learn an unconditional generator, but to estimate 3D representation from the input image.

To perform volume rendering, we compute the radiance field using a simple two-layer MLP similar to [6, 61] using features from the triplane at specified 3D points. Note that as [6] and [61] are generative methods, and not representation learning approaches, they generate triplanes from random codes. In contrast, our approach generates triplanes conditioned on the input image's representation h which is obtained from the backbone network. Using volume rendering from the triplane features, we produce the reconstructed image, I_{rec}, and its corresponding depth map, D_{rec},

$$I_{rec}, D_{rec} = \Pi(\Gamma(\varphi, \pi)), \tag{1}$$

where Γ is the function that queries the radiance fields from triplanes conditioned on camera pose π which contains extrinsic and intrinsic parameters, and Π is volume a rendering operation [43]. We use a fixed camera pose in our experiments since we want to learn a viewer-centered 3D representation, which is shown to be more generalizable compared to object-centric representations [59, 66]. Moreover, as we reconstruct the whole scene with potentially multiple objects, the canonical pose is ambiguous.

Given that 3D reconstruction from a 2D image is an ill-posed problem, like 3DGP [61], we make use of 2D depth information to produce plausible 3D predictions. As ground truth depth maps are not available for large-scale, in-the-wild, datasets like ImageNet [57], we use pseudo ground truth depth maps obtained from off-the-shelf monocular depth methods such as ZoeDepth [4]. Different from 3DGP [61], we do not modify the depth reconstruction with

an adapter, but use it as it is for computing a depth reconstruction loss.

Given the input image I^i and its pseudo depth D^i, we simply train the decoder such that it generates a plausible 3D prediction that is capable of reconstructing them both using the following losses

$$\mathcal{L}_{rgb} = \frac{1}{B} \sum_{i=1}^{B} ||I^i - I^i_{rec}||_2, \qquad (2)$$

$$\mathcal{L}_{depth} = \frac{1}{B} \sum_{i=1}^{B} ||D^i - D^i_{rec}||_2, \qquad (3)$$

where B is the batch size. Here, $\mathcal{L}_{rgb}$ and $\mathcal{L}_{depth}$ represent mean squared error losses, that depend on the input image and its depth. An illustration of our overall pipeline is depicted in Figure 2. In addition to the reconstruction losses, we apply a L1 normalization loss for the density values the radiance fields, $\mathcal{L}_{norm} = \sum_{i=1}^{B} ||\Delta^i||_1$, where Δ^i is the set of density values that calculated for all the queried 3D points for the image I^i.

3.3. Preventing forgetting

The benefit of our approach is that we can apply it to any self-supervised network that has already been pre-trained using 2D objectives. However, our 3D reconstruction objective might inadvertently bias the model towards the 3D task and force it to 'forget' the useful representation that it has already encoded. To prevent this, we add a knowledge distillation loss [35],

$$\mathcal{L}_{dist} = \frac{1}{B} \sum_{i=1}^{B} ||f(I^i) - f_{teach}(I^i)||_2. \qquad (4)$$

Here, $f(.)$ is the representation function that we are optimizing and f_{teach} is a frozen backbone that is already trained using a 2D self-supervised objective.

Our final overall training objective consists of a combination of four losses

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{dist}\mathcal{L}_{dist} + \lambda_{norm}\mathcal{L}_{norm}, \qquad (5)$$

where the λ values are weights for each of the respective loss terms.

3.4. Implementation details

Backbone: We implement our approach using different variants of standard Vision Transformers (ViTs) [15]. Unless otherwise stated, for each experiment we start with a backbone network that has been pre-trained with the state-of-the-art SSL DINOv2 [50] method. We first extract class and patch tokens from the last four layers of the backbone networks, and concatenate these features to obtain a global

image representation which is of size $16 \times 16 \times D$, where D depends on the backbone architecture.

Decoder: To generate the triplane features, we combine self-attention with learnable embeddings and 2D upsampling convolution layers similar to [58]. We learn $16 \times 16 \times 3 \times D$ triplane embeddings, and by applying cross attention and upsampling blocks we obtain our final triplane features that are of size $128 \times 128 \times 3 \times D$. Finally, we reconstruct the image and depth map using volume rendering [3], which results in an output resolution of $256 \times 256 \times 3$ and 256×256, for the reconstructed image and depth respectively. For volume rendering, for each pixel we first sample a ray that passes through pixel location and sample 16 points along the ray to obtain 3D points that we want to estimate the radiance field. Similar to previous works [3, 61], we utilize importance sampling. Then, we bilinearly sample the feature of each 3D points from the triplane, and calculate the radiance field (occupancy and rgb color) using a two layer MLP and pass these values to the final rendering operation to obtain the final reconstruction.

Training: For our training dataset, we use ImageNet-1k [57] which contains approximately 1.2 million images depicting 1,000 object classes. We extract pseudo depth maps for all training images using ZoeDepth [4] with the DPT backbone. We optimize all of the components of the reconstruction network (e.g., backbone, decoder, and triplane embeddings) in an end-to-end manner using the Adam [41] optimizer for 10 epochs with a fixed learning rate of $1e^{-4}$. We set $\lambda_{rgb} = 0.1, \lambda_{depth} = 1, \lambda_{dist} = 1$ and $\lambda_{norm} = 1e^{-3}$ in all of our experiments. During training, we only use basic random crops and horizontal flip augmentations. We provide more details about our model and training in the supplementary document.

4. Experiments

The main goal of our proposed method is to enhance the robustness of existing representation learning methods. We first show how our method results in improved performance on several robustness benchmarks such as ImageNet-Rendition (Im-R) [32], ImageNet-Sketch (Im-Sketch) [68], and Photorealistic Unreal Graphics (PUG) [5]. We also perform experiments on conventional tasks like image recognition [57], fine-grained image classification [67], and depth estimation [60]. This is to illustrate that our approach does not decrease performance for other tasks at the expense of improving robustness.

After training the network with the proxy 3D task, we discard the 3D estimation components of the network and use the backbone representation function to extract global image representation from images $h = f(I)$. For evaluation, we train per-task decoder networks using a fixed representation function $y = \psi(h)$, where the form of y and $\psi(h)$ depends on the specific downstream task. In each ex-

perimental section, we provide details about the decoding function and training details. In all of our experiments, the representation functions are frozen, unless stated otherwise.

4.1. Robustness

Datasets. We present experimental results on benchmarks that are designed to test the robustness of methods in the face of various appearance related shifts. ImageNet-Rendition (Im-R) [32] contains 30,000 images of art, cartoon, graffiti, *etc*. from 200 ImageNet classes. ImageNet-Sketch (Im-Sketch) [68] contains 50 sketch images for each of the 1000 original ImageNet classes. These two datasets contain examples where the texture of the objects is significantly different compared to real in-the-wild photographs, which leads to a significant drop in performance for previous SSL methods [50].

Photorealistic Unreal Graphics (PUG) [5] is a dataset that is designed to evaluate the robustness of visual recognition models. It contains synthetically generated examples from 3D assets by controlling for factors like object texture, background, lighting *etc*. It has been demonstrated that state-of-the-art visual recognition models obtain inadequate performance on this dataset due to changes in appearance factors like object texture and size [5]. We report performance on these two main factors in our experiments.

The 3D Common Corruptions (ImageNet-3DCC) [39] dataset was created with synthetic corruptions with varying levels of difficulty using ImageNet validation images. Compared to previous datasets, it was constructed with synthetic augmentations, but it contains real images and realistic corruptions such as low lighting, flash, and motion blur which reflect real-world challenges for visual recognition models.

Protocol. All of the experiments that we present here are designed to measure the robustness of classifiers that are trained on ImageNet [57] classification data. Given this, we first train a linear classifier on top of various frozen backbones from DINOv2 [50] that are either enhanced via our method (denoted as '+ 3D-Prior') or not, using 1k ImageNet classes from the original training set. We then test the respective linear classifiers on various robustness datasets.

Results. In Table 1, we observe that our proposed method ('3D-Prior') improves the robustness of SSL methods on all robustness benchmarks tested, irrespective of architecture type. For instance, we improve the performance for the different backbone architectures on both ImageNet-Rendition and ImageNet-Sketch datasets, which contain highly challenging out-of-distribution examples. In particular, the performance of DINOv2 [50] using the ViTB/14 architecture is improved by 2% on both benchmarks. Furthermore, for the PUG benchmark, our method improved the performance of the models for object size and texture variation in all cases.

We also present results on ImageNet-3DCC dataset for various synthetic corruption types in Table 2. For each

Method	Im-R	Im-Sketch	PUG-Texture	PUG-Size
ViT-S/14	53.7	41.2	20.7	26.8
ViT-S/14 + 3D-Prior	**54.6**	**41.8**	**21.2**	**26.9**
ViT-B/14	63.3	50.6	25.3	32.2
ViT-B/14 + 3D-Prior	**65.9**	**52.4**	**26.2**	**33.4**
ViT-L/14	74.4	59.3	34.5	42.7
ViT-L/14 + 3D-Prior	**75.9**	**59.5**	**36.4**	**43.2**

Table 1. Robustness evaluation using frozen backbone features from DINOv2 [50] and their enhanced versions from our method ('+ 3D-Prior'). Here we report top-1 accuracy for all benchmarks. Irrespective of backbone architecture type, our 3D-Prior method improves performance on across all datasets. For the PUG experiments we re-run the DINOv2 baselines with our evaluation setting.

level, there is 5 different corruption levels, for simplicity we report averaged the top-1 accuracy for each corruption type. We observe slight improvement for corruption types like far focus, xy motion blur and z motion blurs. However, for other factors like low light, iso noise, we achieve performance that is comparable to the baselines.

We also present qualitative results in Figure 3. We illustrate some top-5 predictions from linear classifiers that were trained on top of representations from DINOv2, either with or without our method. For instance, the top left example is misclassified as a 'starfish' by the DINOv2-based classifier due to the color of the input image while our *shape-aware* approach correctly identifies the images as containing a 'goldfish' due to improved shape-bias.

4.2. Downstream tasks

Tasks and datasets. We present results on additional downstream tasks to show that our method does not lead to worse performance for other tasks at the expense of improved robustness. We report results for visual recognition on ImageNet [57], fine-grained classification using the iNaturalist 2021 [67] and depth estimation on NYU-DepthV2 [60].

Protocol. We follow the same evaluation protocol as in DINOv2 [50]. For ImageNet and iNat21 experiments we froze the backbone, and trained a single-layer classifier using the respective training sets and reported the top-1 validation accuracy. For depth estimation on NYU-DepthV2, we trained two different decoders on top of frozen backbone features, a single linear layer and a more complex DPT [56] decoder and followed the same training recipe from DINOv2 [50]. We also compare to the non-3D baseline numbers from [50].

Results. Results are presented in Table 3. For the visual recognition task on ImageNet-1k, we observe that linear classification performance is slightly improved for all the backbone architectures evaluated and the performance of the models on fine-grained classification for iNat21 is maintained. Furthermore, the performance on depth estimation is improved compared to the baselines, especially when we use a high-resolution DPT decoder on top of our learned representation.

Method	color quant	far focus	flash	fog 3d	iso noise	low light	near focus	xy motion blur	z motion blur
ViT-B/14	72.5	71.6	**60.8**	62.5	**63.9**	72.3	75.5	58.5	58.3
ViT-B/14 + 3D-Prior	**72.6**	**72.0**	60.7	**62.7**	63.3	72.3	**75.8**	**59.0**	**58.7**

Table 2. Robustness evaluation using frozen backbone features from DINOv2 [50] and their enhanced versions from our method on the ImageNet-3DCC dataset [39] using a ViT-B/14-based architecture. While our method improves robustness for corruptions such as 'motion blur' and 'far focus', there are cases such as 'flash' where we are slightly worse.
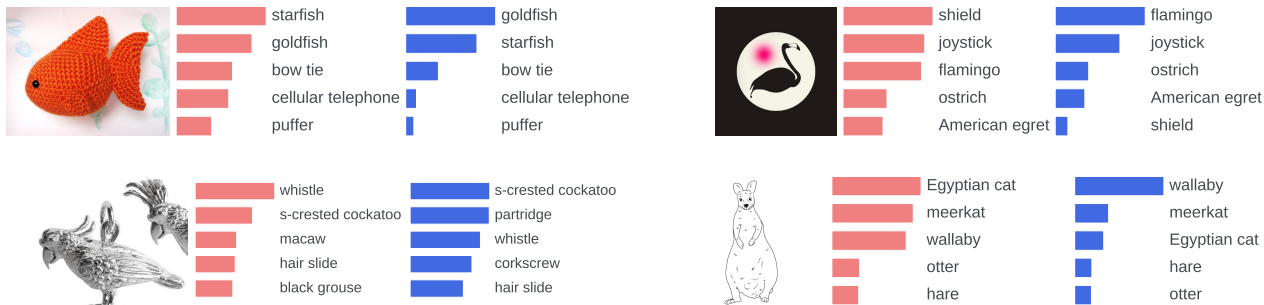

Figure 3. Here we compare top-5 predictions from linear classifiers that are trained on original DINOv2 [50] backbone features (shown in red) and our 3D enhanced approach (shown in blue) on various challenging examples from ImageNet-Rendition [32] and ImageNet-Sketch [68]. Our method results in more shape information being encoded in the representation and hence leads to classifiers that are more robust for these challenging out-of-distribution examples.

Method	ImageNet-1k	iNat21	NYU-DepthV2 ↓	
			linear	DPT
ViT-S/14	81.1	**74.2**	0.499	0.356
ViT-S/14 + 3D-Prior	**81.4**	73.6	**0.438**	**0.346**
ViT-B/14	84.5	81.1	0.399	0.317
ViT-B/14 + 3D-Prior	**85.1**	**82.0**	**0.398**	**0.300**
ViT-L/14	86.3	85.1	**0.384**	0.293
ViT-L/14 + 3D-Prior	**86.5**	**85.2**	0.389	**0.286**

Table 3. Downstream task evaluation using frozen backbone features on various tasks using DINOv2 [50] with and without our 3D-Prior method. We report top-1 accuracy for the ImageNet-1k [57] and iNat21 [67] datasets (higher is better), and RMSE for NYU-DepthV2 [60] dataset (lower is better). Our method leads to improvements in visual recognition performance on ImageNet and for depth estimation on NYU-DepthV2, and does not negatively impact performance on the fine-grained iNat21 dataset.

4.3. Shape bias

Similar to humans, we want our visual recognition models to pay more attention to shape cues compared to texture. As our proxy 3D task requires learning more shape-oriented representations, our hypothesis is that it should lead to representations that have more shape bias. We use the same experimental protocol and dataset from [19] to measure the shape bias of different models. The dataset contains various synthetically generated examples, where the shape of the object comes from one class and the texture of the object comes from another.

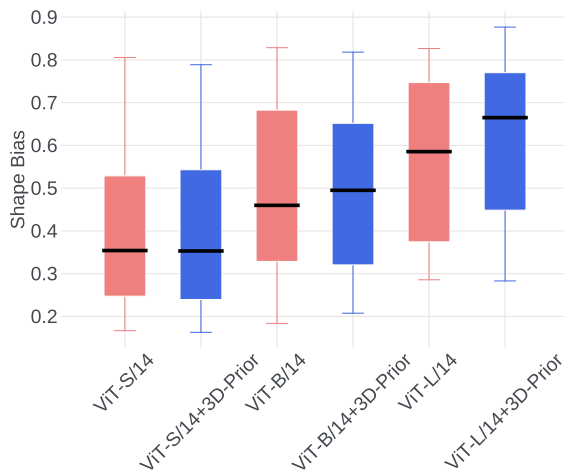We measure the shape bias of representations from DI-NOv2 [50] before and after it is trained with our proxy 3D

Figure 4. Quantification of the shape bias of different DI-NOv2 [50] representations with and without our 3D-Prior method. We calculate the shape bias using the data and protocol from [19]. Our approach increases the shape bias of visual recognition models and we observe that with larger backbones, the difference grows.

objective. The results are visualized in Figure 4. We observe that our method improves the shape bias of the original representations, which is the objective of our shape-centric 3D reconstruction task. Qualitative examples, where we compare predictions of models with and without our method, can be seen in Figure 1.

These results, combined with the robustness experiments, show that the hypothesis of improving shape bias to obtain more robust representations is valid. Furthermore,

Method	Im-1k	Im-R	Im-Sketch	PUG-Texture	PUG-Size
DinoV2	84.5	63.3	50.6	25.3	32.2
Ours	85.1	65.9	52.4	26.2	33.4
Ours w/o triplane	84.3	63.1	50.7	24.9	31.1
Ours w/o $\mathcal{L}_{dist}$	70.7	34.2	22.9	12.1	16.4
Ours from scratch	14.4	8.5	7.4	0.2	0.1

Table 4. Ablation experiments on various robustness benchmarks using a DinoV2 ViT-B/14 model. We investigate the importance of using an explicit 3D representation during training (i.e., w/o triplane), disabling our distillation loss (i.e., w/o $\mathcal{L}_{dist}$), and we evaluate if we can learn reasonable representations when training from scratch without using a pre-trained backbone or distillation loss (i.e., from scratch).

these results may further encourage future lines of work in SSL to develop methods that are designed to explicitly consider 3D representations during training.

4.4. Ablations

To quantify the importance of individual components of our model, we present ablation experiments on the robustness tasks in Table 4.

Removing the triplane. First we investigate if using an 3D representation in the form of a triplane with volume rendering is necessary or if training a basic depth and image decoder network on top of representations is sufficient. Here, we added a decoder which consists of multiple upsampling and convolution layers to predict depth and images. We observe a drop in performance on all benchmarks, but with a smaller drop on ImageNet. This experiment indicates that using an explicit 3D representation is crucial to improve the robustness of the learned representations.

Removing distillation. Next, we try to understand what happens if we do not employ a distillation loss. Without distillation, the model is free to forget useful representation that are already encoded in the 2D SSL backbone. To test this, we simply trained a separate model without the distillation loss. Removing distillation leads to a significant drop in performance across all benchmarks. This experiment shows that preventing the forgetting of already learned useful representations is essential.

Training from scratch. We also investigate if we can learn a global image representation using only the proxy 3D task. Here, we initialized the backbone network randomly (i.e., they are no longer pre-trained) and trained the network using only the image and depth reconstruction losses. The experimental results show that the learned representation is not meaningful and, by itself, the 3D reconstruction task is not a sufficient way to learn a global image representation.

Amount of data. Finally, we explore the impact of varying the size of the training dataset that is used for the 3D proxy task. [10] showed that 2D-based SSL methods benefit from being trained on larger unlabeled datasets, but that there are diminishing returns after a certain amount for the methods
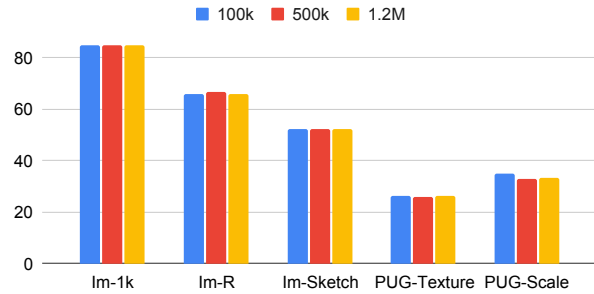

Figure 5. We compare the performance of our approach using different amounts of training data from that same source for the 3D proxy task with DinoV2 ViT-B/14 backbones. Surprisingly, we observe that more data does not change the performance drastically, which shows that our method is data efficient.

they tested. Similarly, we quantify how efficient our method is in terms of the training data size. For this experiment, we randomly selected 100k and 500k images from the ImageNet training set, and trained different instances of our model on these subsets using the same number of iterations as the full model. We report the results in Figure 5. Interestingly, compared to 2D self-supervised methods [10], the performance of our 3D enhanced models are not significantly impacted by the reduction in training data.

4.5. Limitations

One of the restrictions of our approach is the requirement for pseudo depth maps for each input image during training. Existing pre-trained monocular depth estimation models are used, e.g., [4, 56], and these pre-trained depth models only provide depth supervision and do not provide any semantic signal. Hence, these depth maps are free and easy to obtain for large-scale 2D image collections in an automatic way. More importantly, these types of depth estimation models have been demonstrated to be robust and generalizable to various settings.

5. Conclusion

We presented a new approach to enhance the robustness of visual representations from 2D self-supervised networks. Our method utilizes a conceptually simple single-view 3D reconstruction task to encourage learning more shape-aware 3D centric representations. One of the distinct advantages of our approach is that it can be applied to unordered single image collections as it does not impose any stringent assumptions on the types of images it is trained on. We show that incorporating shape-aware knowledge into the representation learning process enhances robustness when compared to alternatives that are not shape aware across a range of visual understanding benchmarks. We hope that our results will encourage a new line of self-supervised works that are designed to consider 3D representations during training.

References

[1] Mehmet Aygün and Oisin Mac Aodha. SAOR: Single-View Articulated Object Reconstruction. In *CVPR*, 2024. 3

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 1, 2, 3

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 5

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, 2023. 2, 4, 5, 8

[5] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S Morcos. Pug: Photo-realistic and semantically controllable synthetic data for representation learning. *arXiv:2308.03977*, 2023. 2, 5, 6

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 2, 3, 4

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *ICML*, 2020. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 3

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020. 2

[10] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *CVPR*, 2022. 8

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3

[12] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 3

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2

[14] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *NeurIPS*, 2014. 2

[15] 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 *ICLR*, 2021. 5

[16] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 3

[17] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *Transactions on Computers*, 1973. 3

[18] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 2

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019. 1, 3, 7

[20] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *NeurIPS*, 2021. 3

[21] James J Gibson. The perception of the visual world. 1950. 1

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

[23] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1, 2

[24] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 2

[25] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, et al. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *NeurIPS*, 2023. 3

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020. 2

[27] Jie Gui, Tuo Chen, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey of self-supervised learning from multiple perspectives: Algorithms, theory, applications and future trends. *arXiv:2301.05712*, 2023. 2

[28] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 3

[29] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 3

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu,

Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 5, 6, 7

[33] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, 2019. 3

[34] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *CVPR*, 2021. 2

[35] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 5

[36] Anil K Jain and Stan Z Li. *Handbook of face recognition*. 2011. 3

[37] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *PAMI*, 2020. 2

[38] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[39] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *CVPR*, 2022. 6, 7

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 4

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[42] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 1988. 1, 3

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 3, 4

[44] Tom Monnier, Matthew Fisher, Alexei A Efros, and Mathieu Aubry. Share with thy neighbors: Single-view reconstruction by cross-instance consistency. In *ECCV*, 2022. 3

[45] Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hendrik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions? In *ICLR*, 2021. 3

[46] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *NeurIPS*, 2021. 1, 3

[47] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 3

[48] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *ICCV*, 2017. 2

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2

[50] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 2, 3, 5, 6, 7

[51] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. 2021. 3

[52] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[53] Bojan Pepik, Peter Gehler, Michael Stark, and Bernt Schiele. 3D2PM – 3D Deformable Part Models. In *ECCV*, 2012. 3

[54] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *CVPR*, 2022. 3

[55] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *CVPR*, 2023. 3

[56] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *PAMI*, 2022. 1, 2, 6, 8

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2, 4, 5, 6, 7

[58] Bokui Shen, Xinchen Yan, Charles R Qi, Mahyar Najibi, Boyang Deng, Leonidas Guibas, Yin Zhou, and Dragomir Anguelov. Gina-3d: Learning to generate implicit neural assets in the wild. In *CVPR*, 2023. 4, 5

[59] Daeyun Shin, Charless C Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In *CVPR*, 2018. 4

[60] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 6, 7

[61] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *ICLR*, 2023. 2, 3, 4, 5

[62] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005. 1

[63] Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 1990. 3

[64] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 2007. 1, 3

[65] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *ICML*, 2020. 1

[66] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 4

[67] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, 2021. 5, 6, 7

[68] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 5, 6, 7

[69] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *CVPR*, 2023. 3

[70] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1, 2, 3

[71] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 3

[72] 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 *CVPR*, 2023. 2

[73] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1

[74] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021. 2

[75] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

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

[77] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1, 2