Improving Equivariance in State-of-the-Art Supervised Depth and Normal Predictors

Yuanyi Zhong, Anand Bhattad, Yu-Xiong Wang, David Forsyth
University of Illinois Urbana-Champaign
{yuanyiz2, bhattad2, yxw, daf}@illinois.edu

Abstract

Dense depth and surface normal predictors should possess the equivariant property to cropping-and-resizing – cropping the input image should result in cropping the same output image. However, we find that state-of-the-art depth and normal predictors, despite having strong performances, surprisingly do not respect equivariance. The problem exists even when crop-and-resize data augmentation is employed during training. To remedy this, we propose an equivariant regularization technique, consisting of an averaging procedure and a self-consistency loss, to explicitly promote cropping-and-resizing equivariance in depth and normal networks. Our approach can be applied to both CNN and Transformer architectures, does not incur extra cost during testing, and notably improves the supervised and semi-supervised learning performance of dense predictors on Taskonomy tasks. Finally, finetuning with our loss on unlabeled images improves not only equivariance but also accuracy of state-of-the-art depth and normal predictors when evaluated on NYU-v2.

1. Introduction

Depth regression [2, 14, 24, 29, 38, 40, 42, 43, 69, 71, 72] and surface normal regression [1, 12, 22, 57] are image-to-image dense prediction tasks that involve predicting an output image of the same size as the input image. This contrasts with image classification, where only one or a few category labels are predicted per image. A shared feature among depth and normal prediction tasks is that they naturally require equivariance, such that a geometric transform (e.g., random cropping) applied to the input image results in the same transform to the output image [8, 15, 28, 30], when the effect (scale of the depth prediction) of camera intrinsic change due to cropping is accounted for. This is because the relative depths and normals are derived from the underlying geometrical and physical properties of the scene that are not affected by viewport changes. Consequently, a good depth or normal predictor must have the equivariant property.

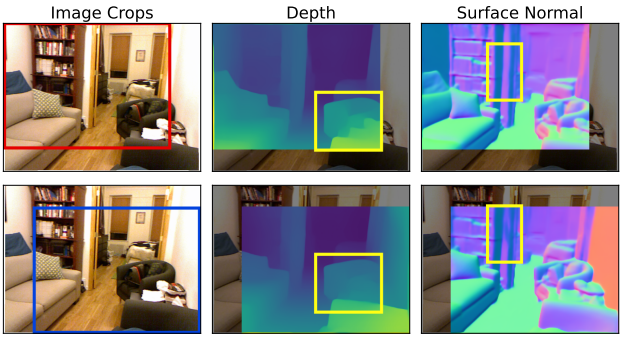

Figure 1. State-of-the-art depth and surface normal predictors fail to capture equivariance, while we know equivariance needs to hold for an ideal depth/normal predictor (when adjusted for prediction scale and offset). We crop and resize two patches (red & blue) from the same scene, then extract depths/normals with pre-trained models from [42] (MiDaS-v3) and [1]. We notice clear discrepancies between the predictions of the two crops, as highlighted by the yellow boxes. The same issue exists in other depth predictors (Figure 2) and dense prediction tasks as well (supplementary).

To our surprise, we find state-of-the-art well-engineered depth and normal predictors often fail at equivariance. We investigate two recent models: the MiDaS CNN-based (v2.1) and Transformer-based (v3.0) depth predictors from [42,43] and the uncertainty-guided CNN-based surface normal predictor from [1]. We generate a pair of resized crops of the same test image from NYU-v2 [49], extract predictions with the networks, and measure equivariance by comparing and computing the mean errors between the predictions of the two crops. A more equivariant network would produce smaller errors from this procedure. We discover that the examined depth and surface normal predictors do not handle equivariance to cropping very well, as shown in Figure 1. There are prominent, sometimes structural, inconsistencies in the predictions of the two crops. For this particular scene, the mean error induced by cropping is significant – as large as 12.6% absolute relative error (AbsRel) between crops for depth prediction, making it compara-

ble to the overall AbsRel error to ground truths (13.7%). Given the widespread use of such dense predictors, for example, MiDaS-v3 for the depth-guided inference in Stable Diffusion-v2 [45], it is imperative to solve such an issue.

Data augmentation is a widely-used strategy to promote the equivariance of models during training. In each mini-batch, instead of seeing the original images, the network sees random resized crops of them. The network is implicitly trained to cope with the variations caused by random crops in a straightforward data-driven manner. However, the problem persists even when randomly resized cropping augmentation is used during training. In fact, the state-of-the-art models we tested, for example, the MiDaS depth networks [42, 43], are already trained on random crops. This suggests that augmentation alone is not a sufficient solution to the equivariance issue. Other methods to enforce equivariance include invariant inputs and equivariant architectures, but they involve a nontrivial additional effort to construct and do not apply to the cropping transform we are concerned about. Therefore, we compare our approach to the data augmentation strategy as our primary baseline.

In this paper, we propose an equivariant regularization approach built on top of data augmentation to improve equivariance in dense depth and normal prediction networks. Our approach consists of two parts: an equivariant averaging step of the outputs of random crops, and an equivariant loss between the crop outputs and the average output. The averaging step is based on the key observation that the full output average of all possible transforms of a transformation group guarantees equivariance to that group. Our sampling version is effectively an unbiased estimate of the full average. The equivariant loss enforces self-consistency and promotes equivariance explicitly rather than implicitly as in data augmentation. Thanks to the flexible formulation, our approach can be applied to any layer of popular network architectures (e.g., CNN or Vision Transformer [13]), and with unlabeled images – both are beyond what data augmentation can do. Meanwhile, our approach retains the benefit of data augmentation, as it imposes no extra cost during testing because the network architecture and the inference procedure are not changed in any way.

Empirically, we demonstrate the effectiveness of our equivariant regularization approach in supervised, semi-supervised and unsupervised learning settings. In the supervised setting, we benchmark our approach against the no-augmentation and augmentation baselines on edge detection, depth prediction, and surface normal prediction tasks of the Taskonomy dataset [75]. We find that our approach overcomes the ineffectiveness of using data augmentation alone. In the semi-supervised setting, we show our approach benefits from unlabeled data, improving the sample efficiency further. Finally, in the unsupervised setting, we demonstrate the capability to adapt the state-of-the-art depth and surface normal models to the NYU-v2 dataset [49] (which these models are not trained on), and improve their accuracy and equivariance, without using any ground truth labels.

To summarize, our contributions are the following:

- We point out an obvious but overlooked issue: The state-of-the-art depth and normal prediction networks fail at equivariance to cropping.
- We propose an equivariant regularization approach to learn more equivariant networks effectively.
- We show empirical successes of our approach in a range of settings, and improve the equivariance and accuracy of the state-of-the-art depth and normal models.

2. Related Work

Equivariance in ML. Equivariance is tied closely to geometry and symmetry. The entire subject of physics is founded on concepts surrounding symmetry. A wide range of natural phenomena admits equivariance inherently since the underlying mechanism is oftentimes geometrical. As a consequence, a lot of data that machine learning deals with has the equivariance property. For example, camera photography follows simple 3D geometry rules, thus a shift in camera position leads to a shift in the photograph; the molecules and point clouds have translation and rotation symmetry in 3D, thus an SE(3) transform should not change any property. Therefore, it is natural to consider equivariance in developing machine learning models.

Equivariance in 2D computer vision. Convolutional neural networks (CNN) for 2D images are shown to have the approximate translation equivariance property due to the nature of convolution [30]. There is a line of work developing rotation equivariant 2D CNNs [8, 36, 59, 60, 62]. The transformation group for 2D rotation is the Special Orthogonal group SO(2), and the Special Euclidean group SE(2) if the translation is allowed. The derivation of group equivariance constraint typically results in steerable filters constructed from 2D harmonic bases. The convolution filter weights are parameterized as a linear combination of the harmonic bases.

Equivariance can also be achieved by parameter sharing of the neural net weights [44]. However, this approach is only possible for limited kinds of groups, such as 90-degree rotations. 2D scale equivariant CNN has been studied [35, 50, 61]. This is typically done by applying the same convolution kernel on several scales or constructing steerable filters from the bases. Equivariant network design method can be generalized to other groups [15,28,39,46,70] and has rich theory in math and physics [7, 9, 20]. Equivariance can also be achieved by transforming the data to canonical coordinate systems [17,41,52]. In particular, [41]

transforms the data to key canonical frames of the group and averages over those frames, while we average over a random sample of the cropping transform. Transformers are the current state-of-the-art neural net architecture [13]. People have sought to combine Transformer and equivariance, resulting in Lie-Transformer [23].

In terms of applications, there is good evidence that equivariance benefits image semantic segmentation [6, 37, 51], object detection to shifting [34] and rotation of images [18]. Equivariance is also useful for generative modeling, for example, for normalizing flow-based generative models [27,48], and variational autoencoders [26]. Equivariance to rotation is beneficial in digital pathology [56]. Extension to time-equivariance for video is also possible [25].

Equivariance in self-supervised learning. Equivariance and invariance are useful in self-supervised learning. The popular contrastive learning algorithm relies on the invariance of representations between augmented views of the same image [3, 5, 19]. More recently, people are exploring ways to use equivariance in contrastive learning [63]. Leveraging equivariance to cropping transform results in dense contrastive learning at pixel-level: PixelPro [64], DenseCL [58], and at region-level: RegionCL [67], DetCon [21]. Equivariance to 4-way rotation can be jointly used with the image-level contrastive objective to improve performance [11]. Self-supervised learning from equivariance between flow transformations of the input image is also effective [66] and between matching points for landmark representation learning [54].

These works are especially successful for downstream segmentation and detection tasks. However, the advancements in these work have yet to be thoroughly explored in the state-of-the-art depth or normal predictors to the best of our knowledge [1, 42, 43], where the dominant paradigm is still supervised training. Inspired by prior work in SSL and segmentation, our work brings in the powerful idea of equivariance to improve state-of-the-art supervised depth and normal predictors.

3. Background

We give some background on the issue of equivariance and how people typically approach equivariance in the literature.

Definition 1 (Equivariance)**.** Formally, a function $f : X \to Y$ is equivariant under the action of a group G on X and a group of G' on Y if for any $t \in G$ there exists $t' \in G'$ such that $f \circ t(x) = t' \circ f(x)$. More commonly, it is true that $G = G'$, i.e., the transformation on both X and Y domains is the same, and the condition becomes

$$f \circ t(x) = t \circ f(x). \tag{1}$$

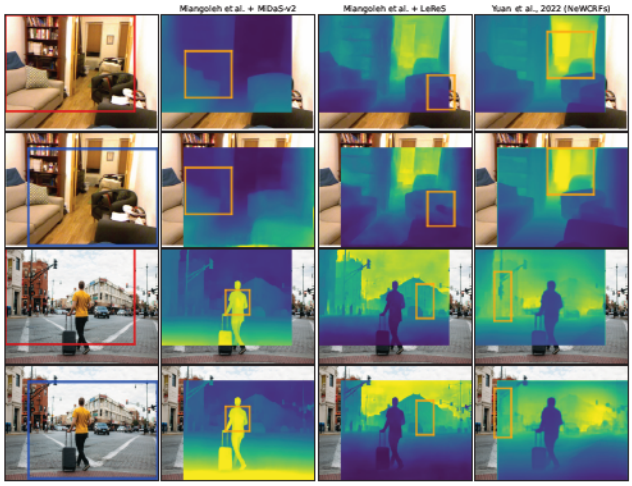

Figure 2. Equivariance failures in more depth predictors [38,73] than Figure 1, suggesting the issue is *prevalent*. The prediction values of the two crops are aligned with the least square. The 2nd column is disparity, others are depth. Top-down, left-right: Notice the blurry/sharp edge, missing object on the stand, wall, pattern on the person's back, vertical line on the building, and inconsistent traffic light.

It essentially states that transform t commutes with f and changes the input and output in the same way.

Invariance can be regarded as a special case of equivariance where g' is always the identity operation. In other words, invariance means $f \circ t(x) = f(x)$ for any action $t \in G$. For example, equivariance is useful for modeling transform-aware phenomena, while invariance is useful for modeling classification tasks.

Non-equivariance issue in depth and normal predictors. Convolutional neural networks possess a certain degree of translation equivariance, but for a broader class of transformations, such as resized cropping, rotation, and scaling, they are not designed to capture equivariance. More recent networks such as Transformers [13] have little inductive biases built-in, they likely do not possess much equivariance on their own as well, and need to see a large number of training examples to learn equivariance in a purely data-driven manner.

Figure 1 shows the failure of equivariance of depth [42] and surface normal predictors [1]. The issue is not unique to these two methods. We examined two more recent approaches [38,73] in Figure 2. [38] is especially interesting, because it similarly merges small crops to reduce the error of pre-trained depth predictors at inference time.

While [38] and we both use the average idea, we conduct averaging at training time instead of inference time. The fact there are still structural changes with cropping suggests that inference-time averaging does not completely solve the issue. On the other hand, a network trained with our approach improves equivariance without extra inference costs.

Additionally, we focus on reducing inconsistent predictions between (often large) crops, while [38] focuses on improving depth details with lots of *small* crops, not necessarily improving equivariance.

Another point to note is that camera intrinsic change (of center and scale) caused by random cropping cannot explain the discrepancies in Figures 1, 2. Camera intrinsic change may lead to overall shifting or rescaling of predicted depths, as noticed and fixed by [14]. However, the failures we observe are structural and related to the content, like missing or creating non-existent objects, even with large crops that only mildly affect intrinsics. What we observe is a separate non-equivariance issue that needs to be solved.

3.1. Existing Approaches

We recognize three types of methods to introduce equivariance into machine learning models. They have different advantages, disadvantages, and suitable application domains. Unfortunately, data augmentation has been the only approach that works for the random resized cropping transform in the dense prediction tasks we study here, which is still inadequate.

Data Augmentation. Data augmentation is the simplest way to encourage the equivariance of ML models [5,19,68]. As long as the transformation function is available, we can artificially create more training examples by transforming the original data randomly. In the case of invariance, we only augment the input data, e.g., the input images for image classification, where the output of the machine learning model is trained to be invariant to the transformation. In the case of equivariance, we can augment the input and the output simultaneously, e.g., the RGB images and depth maps. Commonly used data augmentations include random color jittering, random resizing, and random cropping. The benefit of this approach is simplicity. One can keep the training pipeline and the modeling part the same. However, the downside is potential inefficiency, as we also see with state-of-the-art depth and normal networks in Figure 1. The model may need to see a very large quantity of augmented data examples to learn the equivariance property in a data-driven manner.

Invariant Inputs. The second type of approach converts the data into a format that is invariant or equivariant to the specific transformation. An example of this approach is the distance matrix when dealing with molecular data [16, 48]. People turn the Cartesian coordinates of points (atoms) into a relative distance matrix between pairs of points. It is easy to verify that the distance matrix is invariant to 3D translation and rotation. If the model only depends on the invariant inputs, it is guaranteed to be equivariant or invariant to any input transformation. Another example is the alignment procedure in 3D point cloud/data processing [4, 31, 32, 47, 55], where one can align the points according to their principle canonical axes either globally or locally. This approach works well when the invariant inputs exist, contain sufficient information for the task, and are easy to compute. However, the usage is limited when these requirements are not met. For example, it is not immediately clear how to come up with invariant inputs for standard image augmentations including the crop-and-resize in dense prediction tasks.

Equivariant Architecture. A rich line of research focuses on building equivariance property into the ML model in a "hard-wired" manner [7, 8, 9, 15, 20, 28, 36, 46, 59, 60, 62]. They typically start from a group theory and symmetry standpoint and derive functional forms that satisfy equivariance (relatively) precisely with math and physical science flavor. For example, 2D convolution can be derived for the planer translation group with a Fourier basis. Mirroring constraints on convolutional kernels can be derived for the left-right mirroring group. Convolutions with spherical harmonics can be derived for SO(3) groups. The advantage of this type of approach is that it is principled, exact, and sample-efficient. As rewriting the functional form with equivariance in mind restricts the size of the function class and introduces strict inductive biases, searching for the right hypothesis from data may become easier, and the learning may be accelerated. However, the disadvantages are that one has to modify the model architecture, and deriving the analytical solution for the equivariance basis might be complicated or even impossible, such as for the randomly resized transform in our dense prediction case.

4. Our Approach

Our approach is equivariant regularization. Equivariant property can be imposed by a regularization loss in a "soft" manner together with data augmentation.

We will first describe the mathematical intuition of our approach. We start with the definition of equivariance, then introduce the equivariant average operator as a core technique. The average operator has nice properties, such as being able to turn a non-equivariant function into an equivariant one. We leverage such properties to build our equivariant regularization technique. We introduce a differentiable equivariant loss between the average and individual predictions, which can be minimized to encourage equivariance.

Now we consider the following average operator.

Definition 2 (Equivariant average operator)**.** Let $P(t)$ be a uniform distribution over group elements $t \in \mathcal{T}$. We define the equivariant average of an arbitrary function f as

$$\bar{f}(x) = \mathbb{E}_{t\sim P(t)}\left[t^{-1} \circ f \circ t(x)\right]. \tag{2}$$

The intuition behind this definition is variance reduction. Each summand in the expectation is an estimator of the predicted quantity, with some variance. Taking crop transform as an example, each t takes a particular cropped view of the input image, f makes the predictions for this view, and t^{-1} transforms the predicted image back to the original coordinate frame. Now, each t might lead to a different type of error in the prediction, but averaging (or summing) over all of them will make the differences disappear. This intuition is formally described in the following properties.

Proposition 1. The averaged $\bar{f}$ is equivariant to $\mathcal{T}$.

Proof. For any $t \in \mathcal{T}$, it is straightforward to verify that

$$\begin{aligned}
\bar{f} \circ t(x) &= \mathbb{E}_{t_1 \sim P(t)} \left[t_1^{-1} \circ f \circ t_1 \circ t(x) \right] \\
&\qquad \text{(definition of } \bar{f}\text{)} \\
&= \mathbb{E}_{t_2 \sim P(t)} \left[(t_2 \circ t^{-1})^{-1} \circ f \circ t_2(x) \right] \\
&\qquad \text{(let } t_2 = t_1 \circ t\text{, associativity)} \\
&= \mathbb{E}_{t_2 \sim P(t)} \left[t \circ t_2^{-1} \circ f \circ t_2(x) \right] \qquad (3) \\
&= t \circ \mathbb{E}_{t_2 \sim P(t)} \left[t_2^{-1} \circ f \circ t_2(x) \right] \\
&\qquad \text{(linearity of expectation)} \\
&= t \circ \bar{f}(x) \qquad \text{(definition of } \bar{f}\text{)}
\end{aligned}$$

which is the definition of equivariance. □

Proposition 2. The equivariant average operator preserves the function f if f is already equivariant. As a corollary, the operator is idempotent, namely, $\bar{\bar{f}} = \bar{f}$.

Proof. Use the equivariance definition of f and the associativity of function composition,

$$\begin{aligned}
\bar{f}(x) &= \mathbb{E}_{t \sim P(t)} \left[t^{-1} \circ f \circ t(x) \right] = \mathbb{E}_{t \sim P(t)} \left[t^{-1} \circ t \circ f(x) \right] \\
&= \mathbb{E}_{t \sim P(t)} \left[f(x) \right] = f(x).
\end{aligned} \qquad (4)$$

From Proposition 1, we know that $\bar{f}$ is always an equivariant function, therefore $\bar{\bar{f}} = \bar{f}$. □

Proposition 1 and 2 are practically useful. They together justify treating the equivariant average as a normalization operation because (1) it can turn an arbitrary non-equivariant function into an equivariant one, (2) applying it twice has no further effect than applying it only once.

Once we have the equivariant average, we can use it as a training target to achieve higher equivariance. Specifically, we construct the following loss function based on the equivariant average operator to encourage equivariant property on a trainable function f. This f can be the output of a dense prediction network or any intermediate features.

Definition 3 (Equivariant loss)**.** We define the Equivariant loss as the mean L2 error between the individual prediction $f \circ t(x)$ and the averaged prediction $\bar{f}(x)$:

$$\xi(f) = \frac{1}{Z(\bar{f})} \mathbb{E}_{t \sim P(t)} \left[\| f \circ t(x) - t \circ \bar{f}(x) \|_2^2 \right] \qquad (5)$$

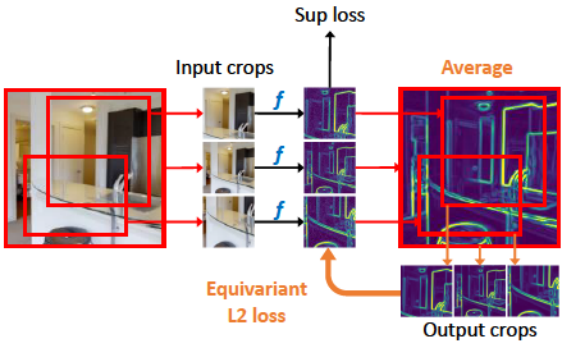

Figure 3. Illustration of our equivariant regularization approach with 3 crops for the 2D texture edge detection task. We generate 3 crops of the input image and pass them through the network f to get 3 outputs. We perform the equivariant average to register them together and get the averaged output. Next, we crop the averaged output to obtain 3 output crops. They correspond to the same image regions as the input crops. We use them as training targets (gradient-stopped) for the individual crop's outputs. A standard supervised loss would use the ground truth as a target, whereas our approach uses the averaged output as a target.

where Z is the normalizing constant: $Z(\bar{f}) = \|\bar{f}(x)\|_2^2$ assuming $\bar{f}$ is not everywhere 0.

The normalizing constant Z is a technical trick to normalize the scale of the equivariant loss. Without Z, simply multiplying f with a scalar will enlarge the equivariant loss, which is undesired. With $Z(\bar{f})$, since $Z(\alpha \bar{f}) = \alpha^2 Z(\bar{f})$, we can show that

$$\xi(\alpha f) = \frac{1}{\alpha^2 Z(\bar{f})} \mathbb{E}_{P(t)} \left[\alpha^2 \| f \circ t(x) - t \circ \bar{f}(x) \|_2^2 \right] = \xi(f). \qquad (6)$$

In practice, it is often computationally infeasible to enumerate and average over all possible transforms t's, as there are too many. This is the case for the commonly-used random resized cropping augmentation we care about. The random cropping induces a combination of *continuous* rigid transformation and scaling groups. We can circumvent this issue by Monte Carlo estimation, i.e., sample a couple of t's and compute the empirical average of $\bar{f}$ and loss. The sample size K is a hyper-parameter to be studied empirically that trades off accuracy and computation efficiency. The following equations state the sampling version:

$$\widehat{\bar{f}(x)} = \frac{1}{K} \sum_k \left[t_k^{-1} \circ f \circ t_k(x) \right], \qquad (7)$$

$$\widehat{\xi(f)} = \frac{1}{Z(\bar{f})} \frac{1}{K} \sum_k \left[\| f \circ t_k(x) - t_k \circ \widehat{\bar{f}(x)} \|_2^2 \right]. \qquad (8)$$

We can attach the equivariant loss onto any layer of a neural net, and train the network with the linear combination of the task loss and the equivariant loss. Formally, assume the task loss is ℓ and the equivariant loss is imposed

on the l-th layer f_l with loss coefficient λ_l, the total loss writes as

$$\ell_{\text{total}}(f) = \ell(f(x), y) + \lambda_l \widehat{\xi(f_l)}. \quad (9)$$

Figure 3 illustrates our equivariant regularization approach regarding the random resized crop transform.

4.1. Discussion

The difference between our approach and the equivariant architecture is that we do not emphasize exact equivariance in this case. Once trained, the model is allowed to have a certain degree of non-equivariance than a strict equivariant model but is expected to possess a higher degree of equivariance than a baseline model without any special equivariance treatment.

Our approach builds on top of data augmentation. It strikes a balance between the equivariant architecture and the data augmentation approaches. It improves upon pure data augmentation by introducing explicit learning signals for equivariance and does not require the complicated derivation or architecture modification of a strict equivariant model. In fact, there should be no additional overhead to a regular model at inference time. We also have the flexibility to adjust the regularization strength by tuning loss coefficients, when the task is not perfectly equivariant or to strive for better overall performance. Our approach can also extend naturally with unlabeled data since the procedure does not involve ground truth labels.

5. Experiment

5.1. Datasets, Models, and Tasks

We evaluate our approach on three data labeling settings with increasing difficulty: supervised, semi-supervised and unsupervised.

Supervised setting. For a supervised setting, we use the Taskonomy dataset [75] standard Tiny splits for experimentation. Taskonomy contains RGB-D scans of indoor building scenes. Several dense prediction tasks are derived from the scans. We focus on 2D texture detection, a low-level vision task; and depth z-buffer prediction and surface normal prediction, two related geometric vision tasks. The Tiny split has 24 training buildings (250K images; originally 25 buildings, 1 building is removed due to data corruption) and 5 validation buildings (52K images).

The model involved in the supervised setting is the standard U-Net from XTaskConsistency [74]. This U-Net has 6 downsampling, 6 upsampling blocks, and the skip connections between corresponding downsampling and upsampling stages. The supervised loss is the L1 loss between the outputs and ground truth targets. For depth, we use inverse depth (i.e., disparity) following [43]. We apply our equivariant regularization loss technique on the second to last convolutional layer of the network. The loss location will be ablated. The loss coefficient is set to 1e-4. For each image, we generate $K = 3$ random crops with scale variation uniformly sampled from 0.4-1.0, aspect ratio from 3/4-4/3, allowing at most 20% padding length, and common color jittering (brightness = contrast = saturation = 0.4, hue = 0.1). In practice, we also employ a weighting window with smooth edges when computing the equivariant average to suppress the boundary effects. We train all models with the AdamW optimizer [33], with learning rate cosine annealed from 1e-3 to 0, and weight decay 1e-4, for 78K gradient steps with batch size 32 distributed on 4 GPUs. Input resolution is 256x256. To maintain *fair comparison*, the supervised baseline is also trained with $K = 3$ crops per image, therefore the wall-clock time of all experiments is roughly the same.

The standard evaluation metrics are L1 error for edge; the percentage of pixels with a relative depth error larger than 1.25 (δ>1.25), mean absolute relative error (AbsRel) for depth; and mean angular error for surface normal [74, 75]. Since depth predictor is usually not aligned to metric depth, i.e., they output arbitrary scale, we align predicted depth to ground truth with least square regression following MiDaS [42, 43]. The detail is described in Supp. C.

The essential question we want to study is whether our approach performs better than the usual data augmentation approach in achieving equivariance and accuracy.

Semi-supervised setting. For this, we concentrate on the depth prediction task. We use 6 or 12 buildings out of 24 buildings in the training set as the labeled portion, and use the rest of the buildings as the additional unlabeled data. The model, hyper-parameters, and optimization schedule are the same as above. During training, we sample two equal-sized mini-batches (2×32 images $\times$ 3 crops) from the labeled and unlabeled data streams, respectively. We impose the supervised loss only on the labeled batch and our equivariant loss on both batches.

This setting is to test whether our approach provides additional benefits from unlabeled data, which is not possible with the simple data augmentation approach.

Unsupervised setting. We focus on unsupervised finetuning of pre-trained state-of-the-art models on the NYU-v2 dataset [49]. The NYU-v2 dataset contains RGB-D scans of 464 indoor scenes, of which 249 scenes (795 images) are used for training and 215 scenes (654 images) for testing. The resolution is 480x640.

We consider the MiDaS-v2.1 and v3.0 depth predictors from [42,43] with CNN and Vision Transformer backbones, respectively. According to their paper, these models are trained on random augmented crops of length 384; therefore, we set the input shape as 384x288 in our unsupervised

Table 1. Supervised setting: Taskonomy Edge2D, Surface Normal, and Depth-ZBuffer. Equivariant regularization on U-Net improves validation performance. Sup baseline refers to baseline without data augmentation, Aug refers to with augmentation, EqLoss refers to our equivariant loss approach. Ang error is the mean angular error in degrees, δ>1.25 is the percentage of pixels with a relative depth error larger than 1.25.

Task	**Edge2D**	**Normal**	**Depth-Z**
Metric	L1 error ($\times 10^{-3}$)↓	Ang error (°)↓	δ>1.25 (%)↓
Sup baseline	8.14	6.72	27.8
+ Aug	7.35 (-9.7%)	6.55 (-2.5%)	27.0 (-2.9%)
+ EqLoss (ours)	**6.35 (-22%)**	**6.47 (-3.7%)**	**25.0 (-10%)**

Table 2. Semi-supervised setting: Taskonomy Depth-ZBuffer. Equivariant regularization with additional unlabeled data improves more. We treat 1/4, 1/2 of the original data as labeled images; the rest as unlabeled images. Sup + EqLoss refers to using only the labeled part and our loss. Semi-sup + EqLoss refers to applying our equivariant loss on both the labeled and unlabeled images. δ>1.25 is the percentage of pixels with a relative depth error larger than 1.25, AbsRel is the mean absolute relative error.

Labeled portion	**1/4**	**1/2**	**All**
#Buildings	6	12	24
#Images	58,783	123,496	248,148
		δ>1.25 (%)↓	
Sup + Aug	43.4	30.4	27.0
Sup + EqLoss (ours)	42.0	29.8	**25.0**
Semi-sup + EqLoss (ours)	**41.0**	**29.3**	**25.0**
		AbsRel (%)↓	
Sup + Aug	25.2	20.3	18.9
Sup + EqLoss (ours)	24.9	20.2	**18.0**
Semi-sup + EqLoss (ours)	**24.8**	**19.6**	**18.0**

finetuning experiments. We also consider the pre-trained uncertainty-guided surface normal predictor from [1]. This network is based on the convolutional EfficientNet backbone [53]. We set the input shape as 640x480 for surface normal to match their training setting. We use AdamW optimizer for 800 steps, with a small learning rate of 1e-5 for depth and 1e-4 for surface normal, as we find them work the best. Two loss functions are involved in finetuning: the first is the supervised loss between the outputs and the pseudo labels generated from the pre-trained checkpoints, and the second is our equivariant loss on the output of the network. We set the equivariant loss coefficient to 1e-4 as well. We sample $K = 3$ random crops per image with scale variation 0.7-1.0, at most 10% padding and common color jittering.

Note that all the pre-trained checkpoints investigated here are *not* trained on NYU-v2. We want to see if our approach can boost the performance of state-of-the-art pre-trained models on this new dataset, by encouraging equivariance alone, *without* using any ground truth labels.

Table 3. Unsupervised setting: Adaptation of state-of-the-art pre-trained depth networks to NYU-v2. We finetune the network with images in NYU-v2, pseudo labels and our equivariant loss (EqLoss row), but without ground truth labels. EqLoss reduces validation errors, while also reducing the validation equivariant loss (EqLoss column), suggesting the network becomes more equivariant. The results compare favorably to other recent methods dedicated to NYU-v2.

Model	δ>1.25(%)↓	AbsRel(%)↓	EqLoss↓
Models trained only on NYU-v2			
Big-to-Small [29]	11.0	11.5	-
Yin et al. [71]	10.8	12.5	-
Huynh et al. [24]	10.8	11.8	-
TransDepth [69]	10.6	10.0	-
Models trained on mix datasets transfer to NYU-v2			
MiDaS-2.1 CNN [43]	8.71	9.68	7.10e-3
MiDaS-2.1 CNN + EqLoss	**7.82**	**8.92**	**3.77e-3**
MiDaS-3.0 DPT [42]	8.32	9.16	7.86e-3
MiDaS-3.0 DPT + EqLoss	**7.75**	**8.91**	**3.04e-3**

Table 4. Unsupervised setting: Adaptation of a state-of-the-art pre-trained surface normal network to NYU-v2. Our unsupervised equivariant finetuning strategy (EqLoss row) reduces the validation mean and median angular errors while reducing the validation equivariant loss (EqLoss column). 11.25° refers to the percentage of pixels with an error larger than 11.25°. All other models here are trained on ScanNet [10] and evaluated on NYU-v2 directly.

Model	Mean°↓	Median°↓	11.25°↑	EqLoss↓
FrameNet [22]	18.6	11.0	50.7	-
VPLNet [57]	18.0	9.8	54.3	-
TiltedSN [12]	16.1	**8.1**	**59.8**	-
Bae et al. [1]	16.03	8.47	58.8	1.26e-2
Bae et al. + EqLoss	**15.71**	**8.30**	**59.4**	**8.80e-3**

5.2. Results

Equivariant regularization improves edge, depth, and normal dense prediction tasks in the supervised setting. The results are organized in Table 1. Comparing the first row to the second, we confirm that data augmentation is better than no data augmentation, which is known widely. This suggests that the implicit encouragement of equivariance from augmentation is helpful. Comparing the second row to the third, we find that our approach brings noticeable gains on top of data augmentation. We achieve as large as 22%, 3.7%, and 10% error reduction for the edge, normal, and depth predictions relative to the supervised baseline without augmentation in their respective metrics. The results indicate that our approach is a more effective way to enforce equivariance during training than data augmentation and that by doing so, the accuracy is also improved.

Equivariant regularization enables unlabeled data in the semi-supervised setting. Our equivariant regularization approach naturally extends to the semi-supervised

learning setting, where the model can learn from additional unlabeled scene images. Apart from the usually labeled data stream, we train with an additional equivariant loss on the unlabeled data stream. In Table 2, we show that, although Sup + EqLoss already brings decent improvements, Semi-sup + EqLoss yields more improvements. These results demonstrate the capability of our approach to leverage unlabeled data to achieve higher label efficiency, which is impossible with the standard augmentation approach.

Unsupervised equivariant finetuning improves state-of-the-art depth and surface normal predictors. Another advantage (and important application) of our equivariant regularization approach over data augmentation is the ability to perform an unsupervised finetuning of state-of-the-art dense predictors to downstream datasets without any ground truth labels. Recall that the SoTA dense prediction networks do not preserve equivariance very well as in Figure 1. In this part, we demonstrate improvements in the equivariance and accuracy of the state-of-the-art MiDaS-v2.1 (CNN-based model), v3.0 (DPT-Large, Dense Prediction Transformer) depth predictors [42, 43] in Table 3, and the uncertainty-guided normal predictor [1] on the NYU-v2 dataset [49] in Table 4. Note that none of the finetuned models have seen any NYU-v2 ground truth. Our approach consistently boosts their performance of them.

Quantitatively, we observe that not only the accuracy metrics of the depth and normal predictors are increased, but also the EqLoss column in Table 3 and 4, which measures the equivariant loss on validation images, is reduced by our approach in both cases. Reducing the EqLoss means that the expected error magnitude of prediction inconsistency coming from different crops of the same image is reduced. This suggests improvements of the equivariance of these predictors. A more thorough and direct evaluation of the finetuned predictors is in the supplementary material.

Qualitatively, our equivariant finetuning approach significantly alleviates the non-equivariant issue of state-of-the-art models. Figure 4 visualizes the predictions before and after finetuning on NYU-v2. We can clearly see that the inconsistency between predictions of two crops is lessened after finetuning.

5.3. Ablation Study

We ablate hyper-parameters in the supervised Taskonomy Depth setting, and provide additional comparisons.

Number of crops (Figure 5). The optimal number of crops per image K (appears in Eq. 7) for our approach is around 3. We choose 3 in our experiments. Note that in the figure, a single stddev is estimated for all K as the error bars. We also control each run to take roughly the same wall-clock time, which means K=3 yields the best trade-off between extra computing and performance under the fixed computation time budget. The depth error of our approach is almost always below that of the data augmentation alone baseline, indicating the higher efficiency of our approach.

Table 5. Ablation study on the equivariant loss coefficient.

Coefficient	1e-5	3e-5	**1e-4**	3e-4
δ>1.25 (%)	25.5	25.2	**25.0**	25.8

Table 6. Ablation study on which layer to apply equivariant loss.

Layer	L	**L-1**	up0	up1	up2	up3
Dimension	1	16	16	32	64	128
δ>1.25 (%)	25.7	**25.0**	**25.0**	25.2	25.1	26.1

Table 7. Inference-time equivariant averaging yields only small depth error reduction on NYU-v2, compared to the reduction from our *training-time* equivariant finetuning.

δ>1.25, AbsRel(%)	No averaging	3 crops averaging
MiDaS-v2.1 pre-trained	8.71, 9.68	8.63 (-0.08), 9.61 (-0.07)
MiDaS-v2.1 + EqLoss	7.82, 8.92	7.78 (-0.04), 8.86 (-0.06)

Equivariant loss coefficient. In Table 5, 3e-5 or 1e-4 performs well; the latter is slightly better. Table 6 studies where to put the equivariant loss. L means applying the loss on the final output, L-1 means the penultimate Conv layer (which we use in experiments), up0-3 means the progressively earlier upsampling block of the U-Net. Applying the loss around L-1 seems to be working well, while going deeper into earlier layers yields worse results. This could be due to the lower resolution of early-stage feature maps.

Inference-time equivariant averaging (Table 7). We tested both the MiDaS pre-trained network and our finetuned network on NYU-v2. In both cases, inference-time averaging offers a small reduction of depth prediction error (smaller than our equivariant finetuning), which suggests the non-equivariant issue cannot be simply addressed by it. Note that inference-time averaging increases latency–small improvement at the cost of the multiplied running time. The benefit of our approach is that the workload of averaging is offloaded to training, so that the inference procedure is unchanged and efficient (1 forward pass).

Comparison to contrastive learning. In Table B.1 of the supplementary material, we initialize from DenseCL [58] or PixelPro [65] pre-trained ResNet, and Table B.2 where we use DenseCL loss as regularization during training. In summary, B.1 suggests that DenseCL indeed provides superior performance than random/supervised initialization, but it does not completely resolve the equivariance issue–and our method can further improve upon DenseCL initialization; B.2 suggests that our method (K=3) is stronger than pairwise DenseCL regularization during training both depth and normal tasks.

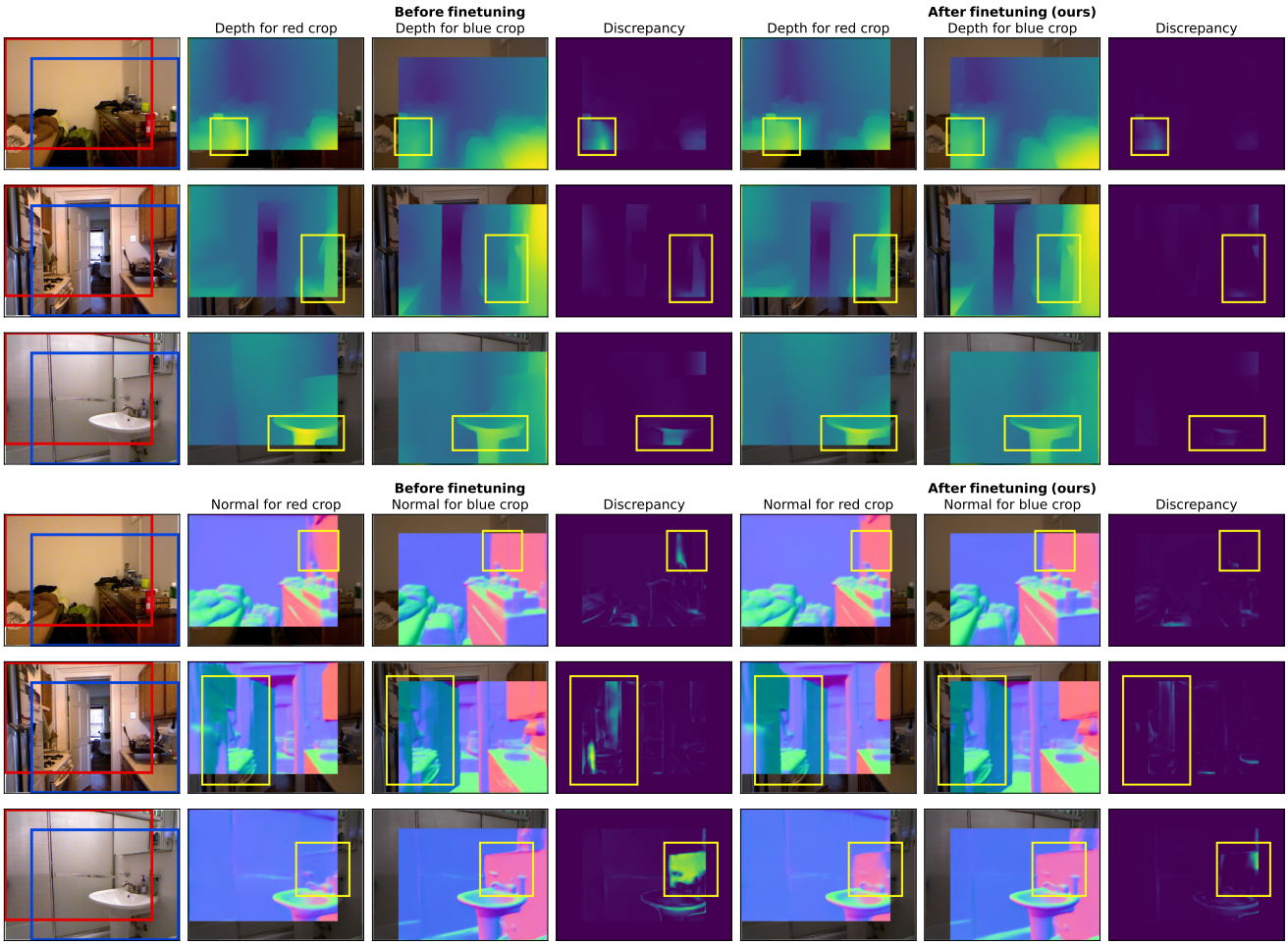

Figure 4. Visualization of predictions of a state-of-the-art depth network (MiDaS-v3.1 DPT Large [42]) and a surface normal network [1] on two crops (red and blue) of the same image, *before and after our unsupervised equivariant finetuning*. The yellow boxes highlight the regions where the pre-trained models struggle with. The discrepancies in those regions are suppressed considerably after finetuning.

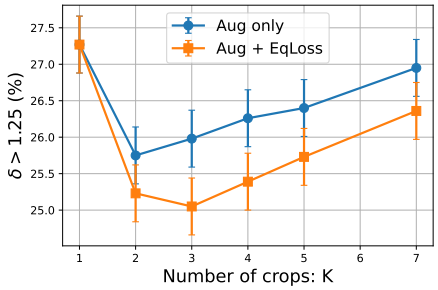

Figure 5. Study on the number of random crops per image.

6. Conclusion

This paper reveals a salient problem in state-of-the-art depth and normal predictors – that they are not equivariant to cropping, and proposes an equivariant regularization approach to address it. We demonstrate the usefulness of our approach in supervised, semi-supervised and unsupervised settings. We substantially improve equivariance and accuracy of state-of-the-art pre-trained models on NYU-v2 test set without using ground-truth labels. We hope future work can explore the powerful idea of equivariance in other dense prediction tasks and with transformations beyond cropping.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, the NCSA Fellows program, the Illinois-Insper Partnership, and the Amazon Research Award. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014 and CIS230012 from the ACCESS program.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 1, 3, 7, 8, 9

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *ECCV*, 2022. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerg-

ing properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[4] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *CVPR*, 2021. 4

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3, 4

[6] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, 2021. 3

[7] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *ICML*, 2019. 2, 4

[8] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016. 1, 2, 4

[9] Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In *NeurIPS*, 2019. 2, 4

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 7

[11] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. In *ICLR*, 2022. 3

[12] Tien Do, Khiem Vuong, Stergios I Roumeliotis, and Hyun Soo Park. Surface normal estimation of tilted images via spatial rectifier. In *ECCV*, 2020. 1, 7

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

[14] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: Camera-aware multi-scale convolutions for single-view depth. In *CVPR*, 2019. 1, 4

[15] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *ICML*, 2020. 1, 2, 4

[16] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. In *NeurIPS*, 2020. 4

[17] Kanchana Vaishnavi Gandikota, Jonas Geiping, et al. A simple strategy to provable invariance via orbit mapping. In *ACCV*, 2022. 2

[18] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *CVPR*, 2021. 3

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 4

[20] Lingshen He, Yiming Dong, Yisen Wang, Dacheng Tao, and Zhouchen Lin. Gauge equivariant transformer. In *NeurIPS*, 2021. 2, 4

[21] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, 2021. 3

[22] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas J Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In *ICCV*, 2019. 1, 7

[23] Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *ICML*, 2021. 3

[24] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *ECCV*, 2020. 1, 7

[25] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *ICCV*, 2021. 3

[26] T Anderson Keller and Max Welling. Topographic vaes learn equivariant capsules. In *NeurIPS*, 2021. 3

[27] Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *ICML*, 2020. 3

[28] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *ICML*, 2018. 1, 2, 4

[29] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1, 7

[30] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *CVPR*, 2015. 1, 2

[31] Feiran Li, Kent Fujiwara, Fumio Okura, and Yasuyuki Matsushita. A closer look at rotation-invariant deep point cloud analysis. In *ICCV*, 2021. 4

[32] Jiaxin Li, Yingcai Bi, and Gim Hee Lee. Discrete rotation equivariance for point cloud recognition. In *ICRA*, 2019. 4

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[34] Marco Manfredi and Yu Wang. Shift equivariance in object detection. In *ECCV*, 2020. 3

[35] Diego Marcos, Benjamin Kellenberger, Sylvain Lobry, and Devis Tuia. Scale equivariance in cnns with vector fields. *arXiv preprint arXiv:1807.11783*, 2018. 2

[36] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *ICCV*, 2017. 2, 4

[37] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *CVPR*, 2021. 3

[38] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *CVPR*, 2021. 1, 3, 4

[39] R Murphy, B Srinivasan, V Rao, and B Riberio. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *ICLR*, 2019. 2

[40] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *CVPR*, 2022. 1

[41] Omri Puny, Matan Atzmon, Edward J Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *ICLR*, 2021. 2

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 1, 2, 3, 6, 7, 8, 9

[43] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 1, 2, 3, 6, 7, 8

[44] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *ICML*, 2017. 2

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[46] David Romero, Erik Bekkers, Jakub Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks. In *ICML*, 2020. 2, 4

[47] Rahul Sajnani, Adrien Poulenard, Jivitesh Jain, Radhika Dua, Leonidas J Guibas, and Srinath Sridhar. Condor: Self-supervised canonicalization of 3d pose for partial shapes. In *CVPR*, 2022. 4

[48] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *ICML*, 2021. 3, 4

[49] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2, 6, 8

[50] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. In *ICLR*, 2019. 2

[51] M Naseer Subhani and Mohsen Ali. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In *ECCV*, 2020. 3

[52] Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Equivariant transformer networks. In *ICML*, 2019. 2

[53] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 7

[54] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *ICCV*, 2019. 3

[55] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 4

[56] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, 2018. 3

[57] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *CVPR*, 2020. 1, 7

[58] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021. 3, 8

[59] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. In *NeurIPS*, 2019. 2, 4

[60] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *CVPR*, 2018. 2, 4

[61] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *NeurIPS*, 2019. 2

[62] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *CVPR*, 2017. 2, 4

[63] Yuyang Xie, Jianhong Wen, Kin Wai Lau, Yasar Abbas Ur Rehman, and Jiajun Shen. What should be equivariant in self-supervised learning. In *CVPR*, 2022. 3

[64] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 3

[65] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 8

[66] Yuwen Xiong, Mengye Ren, Wenyuan Zeng, and Raquel Urtasun. Self-supervised representation learning from flow equivariance. In *ICCV*, 2021. 3

[67] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Regioncl: Exploring contrastive region pairs for self-supervised representation learning. In *ECCV*, 2022. 3

[68] Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. In *NeurIPS*, 2019. 4

[69] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 2021. 1, 7

[70] Dmitry Yarotsky. Universal approximations of invariant maps by neural networks. *Constr. Approx.*, 2022. 2

[71] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 1, 7

[72] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. 1

[73] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *CVPR*, 2022. 3

[74] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020. 6

[75] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 2, 6