








# On the Viability of Monocular Depth Pre-training for Semantic Segmentation

Dong Lao<sup>1</sup>, Fengyu Yang<sup>2</sup>, Daniel Wang<sup>2</sup>, Hyungseob Park<sup>2</sup>, Samuel  
Lu<sup>1</sup>, Alex Wong<sup>2</sup>, and Stefano Soatto<sup>1</sup>

<sup>1</sup> UCLA Vision Lab, Los Angeles, CA 90024, USA  
{lao, soatto}@cs.ucla.edu, samuellu@ucla.edu

<sup>2</sup> Yale Vision Lab, New Haven, CT 06511, USA  
{fengyu.yang, daniel.wang.dhw33, hyungseob.park, alex.wong}@yale.edu

**Abstract.** The question of whether pre-training on geometric tasks is viable for downstream transfer to semantic tasks is important for two reasons, one practical and the other scientific. If the answer is positive, we may be able to reduce pre-training costs and bias from human annotators significantly. If the answer is negative, it may shed light on the role of embodiment in the emergence of language and other cognitive functions in evolutionary history. To frame the question in a way that is testable with current means, we pre-train a model on a geometric task, and test whether that can be used to prime a notion of “object” that enables inference of semantics as soon as symbols (labels) are assigned. We choose monocular depth prediction as the geometric task, and semantic segmentation as the downstream semantic task, and design a collection of empirical tests by exploring different forms of supervision, training pipelines, and data sources for both depth pre-training and semantic fine-tuning. We find that monocular depth *is* a viable form of pre-training for semantic segmentation, validated by improvements over common baselines. Based on the findings, we propose several possible mechanisms behind the improvements, including their relation to dataset size, resolution, architecture, in/out-of-domain source data, and validate them through a wide range of ablation studies. We also find that optical flow, which at first glance may seem as good as depth prediction since it optimizes the same photometric reprojection error, is considerably less effective, as it does not explicitly aim to infer the latent structure of the scene, but rather the raw phenomenology of temporally adjacent images. Code: <https://github.com/donglao/DepthToSemantic>.

**Keywords:** Depth estimation · semantic segmentation · pre-training

## 1 Introduction

We probe the following seemingly counter-intuitive hypothesis:

*Can pre-training on a geometric task benefit a downstream semantic task?*

Geometric inference is often viewed as a low-level vision task requiring little abstraction that is needed for semantics [26]. For example, depth can be acquired through minimizing reprojection error, *i.e.* from multi-view or videos, or

directly from range sensors. Both can be performed procedurally, without inductive learning, rendering depth a meaningless task for pre-training. However, induction is needed to infer one 3D scene among infinitely many compatible with the same 2D image. Therefore, if a model could solve this ill-posed problem, it would provide evidence of the viability of pre-training with little to no human intervention, which is important in specialized data domains for which little annotated data is publicly available.

In this paper, we focus on testing monocular depth as the pre-training geometric task, and semantic segmentation as the downstream semantic task. They are purposefully chosen: Training deep neural networks for semantic segmentation requires labor-intensive pixel-level annotation, so the choice of pre-training is essential to its performance. Existing studies have shown mixed results about the relationship between the two tasks. Taskonomy [57], a framework for measuring relationships between visual tasks, suggests that depth estimation is “far” from semantic segmentation, while recent work [17] shows that depth pre-training can beat “closer” tasks like image classification. Prior work [19, 23, 40, 42] has also shown improvement in semantic segmentation when incorporating depth. This mixed evidence motivates us to take a closer look at the underlying mechanism of how monocular depth may benefit semantic segmentation.

Another more subtle reason for exploring this hypothesis is that pre-training is often performed on heavily human-biased datasets [11, 38], where the photographer who framed the picture meant to convey a particular concept (say, a cup), and therefore took care to make sure that the manifestation of the concept (the image) prominently features the object by choice of vantage point, illumination, and (lack of) occlusion. This bias is mitigated if data is not purposefully organized into “shots.” Unfortunately, existing datasets are mostly composed of purposefully framed shots which could obfuscate the analysis. We note that depth can be inferred without any semantic interpretation [26] regardless of whether the data is captured purposefully or randomly. With monocular depth as the pre-training task, there are two ways of reducing the aforementioned human selective bias: The first involves directly pre-training within the specific domain of interest, leveraging the simplicity of data gathering; The second way is scaling up pre-training by incorporating diverse sources of data, which is made possible by recent developments [33, 41, 54] on relative depth estimation.

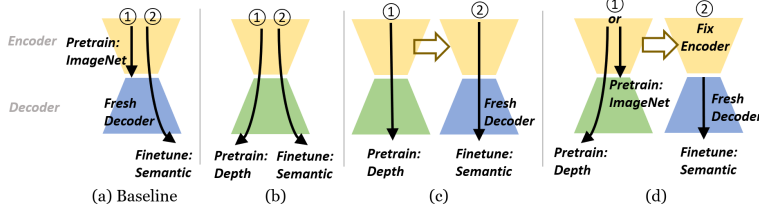
**Methods.** We formalize the main hypothesis in Sect. 3. Since it cannot be tested analytically without knowledge of the joint distribution of test images and labels, we propose an empirical testing protocol. We test on monocular depth models trained under multiple forms of supervision, including structure-from-motion, binocular stereo, and depth sensors. We then change the prediction head of the resulting network, either the final layer or the whole decoder, and fine-tune it for semantic segmentation (Fig. 1). We consider depth estimation as a *viable* pre-training option if it yields comparable improvements to downstream semantic segmentation tasks as other common pre-training practices, *e.g.* ImageNet classification. To this end, we design a series of controlled experiments to test the effect of choice of initialization (Tab. 1, Fig. 2), training with various datasets

sizes (Fig. 3), choice of network component to be frozen and fine-tuned (Fig. 4), effect of resolution of training images (Fig. 5). Conclusions are drawn from both quantitative and qualitative (Fig. 6) results.

**Findings.** Pre-training for depth estimation improves the performance of downstream semantic segmentation across different experimental settings. Particularly, we show that depth estimation is indeed a viable pre-training option as compared to existing methods (Tab. 4). For example, compared to classification, using depth on average improves by 5.8% mIoU and 5.2% pixel accuracy on KITTI. As a sanity check, we test both a depth network pre-trained from scratch and one trained after ImageNet initialization, and both outperform classification-based pre-training in downstream semantic segmentation. To control the effect of our choice of architecture, we used our pre-trained encoder to initialize a standard semantic segmentation network [5]. We observed similar findings on Cityscapes and NYU-V2 regardless of how depth training is supervised. Inferring depth without explicit supervision typically involves minimizing the prediction error, just like optical flow. Somewhat surprisingly, not only does pre-training for depth outperform optical flow, but the latter is often worse than random initialization (Fig. 7). One may also argue that observed improvements mainly come from the availability of in-domain pre-training data for depth. To test this conjecture, we fine-tune a depth model [55] trained on large-scale out-of-domain data. Improvements in semantic segmentation reveal that when trained at scale, depth models show strong transferability to unseen downstream data domains.

## 2 Related Work

Pre-training aims to learn a representation (function) of the *test* data that is maximally informative (sufficient), while providing some kind of complexity advantage. In our case, we measure complexity by the validation error after fine-tuning on limited amount of labeled data, which measures the inductive value of pre-training. The recent literature comprises a large variety of “self-supervised” methods that are purportedly task-agnostic. In reality, the task is specified indirectly by the choice of hand-designed nuisance transformations that leave the outcome of inference unchanged. Such transformations are sampled through data augmentation while the image identity holds constant (Contrastive Learning) [4,8,9,38], or reconstruction [3]. Group transformations organize the dataset into orbits, which contrastive learning tries to collapse onto its quotient, which is a maximal invariant. Such a maximal invariant is transferable to all and only tasks *for which the chosen transformation is uninformative*. For group transformation, the maximal invariant can, in theory, be computed in closed form [44]. In practice, contrastive learning are extended to non-group transformations, *e.g.* occlusions, as seen in language [2] and images [7]. All self-supervised methods boil down to hand-designed and quantized subsets of planar domain diffeomorphisms (discrete rotations, translations, scaling, reflections, *etc.*), range homeomorphisms (contrast, colormap transformations) and occlusion masks.



**Fig. 1: Diagram for different pre-training and fine-tuning setups.** (a) Common practice: pre-train the encoder, *e.g.* on ImageNet, attach a decoder, and fine-tune the network. (b) Our best practice: pre-train the network by monocular depth, and fine-tune for semantic segmentation. (c) Cross architecture: for fair comparisons with common practice, we pre-train by depth, replace the decoder, and fine-tune. (d) To test the quality of pre-trained encoders, we fix the encoders and fine-tune decoders only.

In our case, rather than hand-designing the nuisance transformations assumed to be shared among pre-training and fine-tuning tasks, we let the scene itself provide the needed supervision: images portend the same scene, either from the same timestamp (stereo) or adjacent in the temporal domain (video frames), so their variability defines the union of nuisance factors. These include domain deformations due to ego- and scene motion, range transformations due to changes in illumination, and occlusions. In addition to sharing nuisance variability, pre-training and fine-tuning tasks should ideally also share the hypothesis space. It may seem odd to choose a geometric task, where the hypothesis space is depth, to pre-train for a semantic task, where the hypothesis space is a discrete set of labels. However, due to the statistics of range images [21] and their similarity to the statistics of natural images [22], this is actually quite natural: A range map is a piecewise smooth function defined on the image domain, whereas a segmentation map is a piecewise constant function where the levels are mapped to arbitrary labels. As a result, the decoder for depth estimation can be easily modified for semantic segmentation. A discussion of this choice, specifically on the representational power of deterministic predictors, in Sect. 5. [19, 23] also utilize depth for semantic segmentation. [23] proposes pre-training on relative depth prediction, and [19, 20] utilizes self-supervised depth estimation on video sequences. Our experiments validate their findings. We further investigate whether features obtained purely from monocular depth improve semantic segmentation.

Monocular depth [16, 50] may use different supervision, either through additional sensors [14, 51, 52], or synthetic data [36, 47, 56], but none require human annotation. Some use regularizers with sparse seeds [35, 48, 49], or adopt pre-trainings [41]. We design experiments agnostic to how depth models are trained, but also make comparisons across different forms of depth supervision (Tab. 2).

### 3 Formalization

Let  $x : D \subset \mathbb{R}^2 \rightarrow \{0, \dots, 255\}^3$  be an image, where the domain  $D$  is quantized into a lattice,  $z : D \rightarrow \{1, \dots, Z\}$  a depth map with  $Z$  depth or disparity levels,

and  $y : D \rightarrow \{1, \dots, K\}$  a semantic segmentation map. In coordinates, each pixel in the lattice,  $(i, j) \in \{1, \dots, N\} \times \{1, \dots, M\}$  is mapped to RGB intensities by  $x(i, j)$ , a depth by  $z(i, j)$ , and a label by  $y(i, j)$ . Despite the discrete nature of the data and the hypothesis space, we relax them to the continuum by considering the vectors  $\mathbf{x} \in \mathbb{R}^{NM^3}$ ,  $\mathbf{y} \in \mathbb{R}^{NMK}$  and  $\mathbf{z} \in \mathbb{R}^{NM}$ . With a slight abuse of notation,  $y \in \{1, \dots, K\}$  denotes a single label and  $\bar{y} \in \mathbb{R}^K$  its embedding, often restricted to a one-hot encoding.

Now, consider a dataset  $\mathcal{D}_z = \{\mathbf{x}_t^i, \mathbf{z}_t^i\}_{i,t=1}^{V,T_i}$  comprised of  $V$  image sequences each of length  $T_i$ . For the simplicity of the notations, in the case of multi-view stereo, we also consider multiple 2D image inputs as a “sequence” without loss of generality. In the case of supervised depth estimation, synchronized depth maps  $z_t^i$ ’s are measured by a range sensor. Typical datasets supporting depth estimation may include just image sequences or both modalities.

Training for monocular depth estimation yields a mapping  $\phi_w : \mathbf{x} \mapsto \mathbf{z}$ , parametrized by weights  $w$  in a neural network, via

$$w = \arg \min_{w, g_t} \sum_{i,j,n,t} \ell(x_{t+1}^n(i, j), \hat{x}_t^n(i, j)) + \lambda \sum_{i,j,n,t} \ell(z_t^n(i, j), \hat{z}_t^n(i, j)) \quad (1)$$

where  $\hat{x}_t$  is the warping of an image  $x$  from  $t$  to  $t+1$  based on camera pose  $g_t$ . Here we consider a generic formulation for different modalities of depth estimation. The first term in Eq. (1) measures reprojection error across frames, and the second term measures the distance between estimated depth values and the ground-truth from the range sensor. In the case of unsupervised depth estimation from videos, *e.g.* [16], only the reprojection loss is considered; while in the case of supervised depth estimation from single images, *e.g.* [37],  $T_i = 1$  for all  $i$ ’s, so only the second term is minimized.

The goal is to use these representations as encodings of the data to then learn a semantic segmentation map. In practice, the representations above are implemented by deep neural networks, that can be truncated at intermediate layers thus providing embedding spaces larger than the respective hypothesis spaces. We refer to the parts before and after this intermediate layer *encoder* and *decoder*, respectively. We overload the notation and refer to the encoding as  $\mathbf{h} = \phi_w(\mathbf{x})$  for both depth estimation and other pre-training methods, presumably with weights  $w'$ , assuming they have the same encoder architecture. The goal of semantic segmentation is then to learn a parametrized map  $\psi_{w''} : \mathbf{h} \mapsto \mathbf{y}$  using a small but fully supervised dataset  $\mathcal{D}_s = \{\mathbf{x}^n, \mathbf{y}^n\}_{n=1}^N$ , by minimizing some loss function or (pseudo-)distance in the hypothesis space  $d(\mathbf{y}, \hat{\mathbf{y}})$ , where

$$w'' = \arg \min_w \sum_{n=1}^N d(\mathbf{y}^n, \psi_w(\mathbf{h}^n)) \quad (2)$$

plus customary regularizers. In the aggregate, we have a Markov chain:  $\mathbf{x} \longrightarrow \mathbf{h} = \phi_w(\mathbf{x}) \longrightarrow \mathbf{y} = \psi_{w''}(\mathbf{h}) = \psi_{w''} \circ \phi_w(\mathbf{x})$  for depth estimation, and  $\mathbf{x} \longrightarrow \hat{\mathbf{h}} = \phi_{w'}(\mathbf{x}) \longrightarrow \mathbf{y} = \psi_{w''} \circ \phi_{w'}(\mathbf{x})$  for other pre-training methods. A representation obtained through a Markov chain is optimal (minimal sufficient) only if the

**Table 1: Semantic segmentation accuracy on KITTI.** Unsupervised depth as pre-training improves semantic segmentation accuracy under all settings. Our best practice (in **blue**) improves common practice (in **purple**) by 7.53% mIoU and 4.68% pixel accuracy. Freezing the encoder with ImageNet pre-training (in **red**) is worse than no pre-training (random initialization). DeepLabV3<sup>†</sup>: with ResNet50 encoder.

	Fine-tune All						Freeze Encoder			
	ResNet18		ResNet50		DeepLabV3 <sup>†</sup>		ResNet18		ResNet50	
Pre-training	mIoU	P.Acc	mIoU	P.Acc	mIoU	P.Acc	mIoU	P.Acc	mIoU	P.Acc
None	41.35	70.75	44.66	73.37	21.93	52.32	41.24	70.52	37.72	67.38
ImageNet	45.15	72.39	44.65	73.06	<b>43.39</b>	<b>72.66</b>	<b>33.33</b>	<b>65.34</b>	<b>32.03</b>	<b>62.53</b>
Depth-Rand	46.00	72.43	49.90	76.28	43.43	71.34	43.02	72.38	45.79	<b>74.71</b>
Depth	<b>50.20</b>	<b>76.39</b>	<b>50.92</b>	<b>77.34</b>	<b>43.77</b>	<b>72.68</b>	<b>46.53</b>	<b>74.42</b>	<b>46.55</b>	74.48

intermediate variable  $\mathbf{h}$  or  $\hat{\mathbf{h}}$  reduces the Information Bottleneck [45] to zero. In general, there is information loss, so we **formalize the key question** as whether the two Information Bottleneck Lagrangians satisfy the following:

$$H(\mathbf{y}|\mathbf{h}) + \beta I(\mathbf{h}; \mathbf{x}) \stackrel{?}{\leq} H(\mathbf{y}|\hat{\mathbf{h}}) + \beta' I(\hat{\mathbf{h}}; \mathbf{x}) \quad (3)$$

where  $\beta$  and  $\beta'$  are hyperparameters that can be optimized as part of the training process, and  $I, H$  denotes the (Shannon) Mutual Information and cross-entropy respectively. If the above is satisfied, then pre-training for depth estimation is a viable option, or even better than pre-training with another method. It would be ideal if this question could be settled analytically. Unfortunately, this is not possible, but the formalization above suggests a protocol to settle it empirically.

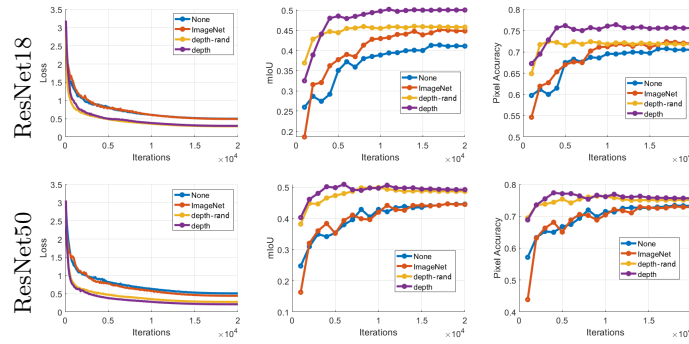
To test this empirically, we use the validation error on a supervised dataset  $\mathcal{D}_s$  as a proxy for residual information. We conduct fine-tuning under several configurations (Fig. 1): with respect to  $w''$  using  $\mathcal{D}_s$ , *i.e.* yielding a comparison of the raw pre-trained back-bones (encoders,  $w, w'$ ), or with respect to *both*  $w''$  and  $w$  (for depth estimation) or  $w'$  (for other pre-training methods). Finally, all four resulting models can be compared with one obtained by training from scratch by optimizing a generic architecture with respect to  $w''$  alone.

## 4 Experiments

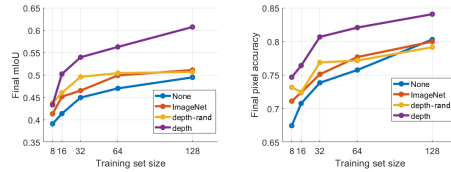
### 4.1 Controlled Experiments with Few-shot Fine-tuning

We first cover an extensive collection of controlled experiments and ablations to gain insights into the main hypothesis. We specifically conduct experiments under the *few-shot* setting, where only a small amount of labels are available for fine-tuning, to highlight the role of pre-training.

**KITTI [15]** contains 93000 video frames for depth training with 200 densely annotated images for semantic segmentation. Segmentation results are evaluated by mean IoU (mIoU) and pixel-level accuracy (P.Acc). We randomly choose a small training set of 16 images and limit data augmentation to horizontal



**Fig. 2: Comparison between different network initializations.** Models initialized by depth pre-training (unsupervised) train faster and achieve higher final accuracy.



**Fig. 3: Final accuracy vs different training set size.** Under all training set sizes, our best practice constantly outperforms ImageNet pre-trained. Encoder: ResNet 18.

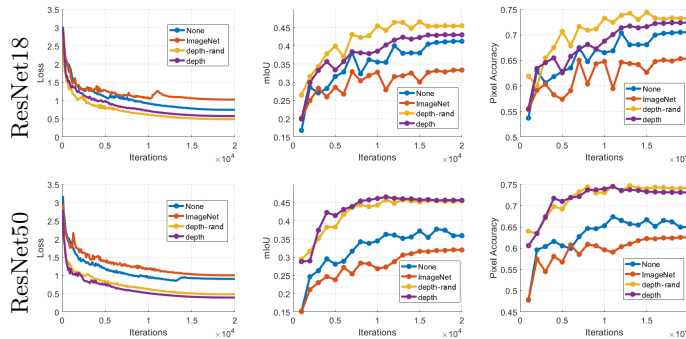
Depth Sup.	mIoU	P.Acc
Video	46.00	72.43
Stereo	49.11	74.58
Lidar	<b>52.78</b>	<b>77.17</b>

**Table 2: Forms of depth supervision matter.** Direct supervision with Lidar works the best, followed by stereo (with known camera pose), and monocular video (camera pose unknown).

flips to highlight the impact of pre-training, except for Fig. 3 where we test on different training set partitions. We use Monodepth2 [16] for depth pre-training. For semantic segmentation, we replace the last layer of the decoder with a fully connected layer, using the finest scale of the multi-scale output. We test on ResNet18 and ResNet50 encoders due to their compatibility with various network architectures and widely public-available pre-trained models. Fig. 1 summarizes our experimental setup and Tab. 1 summarizes the outcomes. In all cases, depth pre-training improves segmentation accuracy.

**Full model.** Fig. 2 shows the evolution of training loss and model accuracy. Depth pre-training outperforms ImageNet and random initialization. ImageNet pre-training slightly improves over random initialization on ResNet18, but shows almost identical performance to random initialization on ResNet50. Depth also speeds up training, taking  $\sim 5000$  iterations to converge, while ImageNet takes 15000 to 20000. Similar results on full-resolution are deferred to the Supp. Mat.

**Different training set size.** Fig. 3 shows that depth pre-training improves final segmentation scores over all dataset partitions (with ResNet 18). When training samples increase (*e.g.* 128), ImageNet and depth pre-training (from random initialization) are comparable to random initialization, but depth pre-training initialized with ImageNet yields the best results.



**Fig. 4: Frozen encoder results.** Using an encoder pre-trained by depth significantly outperforms one with random weights and one for ImageNet classification. Note that in this experiment, ImageNet pre-training performs worse than random initialization.

**Different forms of depth supervision.** Pre-training quality depends on the source of supervision. Training on monocular videos involves minimizing reprojection error, which requires joint estimation of depth and pose. Since pose estimation relies on sufficiently distinctive textures (large eigenvalues of the structure tensor of image gradients), the supervision signal is sparse. Conversely, with stereo images, one may omit the pose network when training. With depth sensors, training losses minimize error w.r.t. dense or semi-dense measured depth, offering stronger supervision. Tab. 2 shows that supervising with Lidar is the best, followed by stereo, and monocular video – all improving over ImageNet.

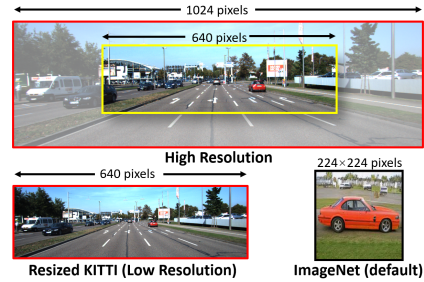
**Frozen encoder.** We freeze pre-trained encoders, and fine-tune the decoder only, testing the ability of features from pre-trained encoders to capture semantics. With both ResNet18 and ResNet50, encoder pre-trained for depth significantly outperforms random initialization and ImageNet pre-trained (Fig. 4). It is surprising that ImageNet pre-training is detrimental in this case (after a grid search over learning rates): worse than fixed random weights. This suggests that while classification is a semantic task, it removes semantic information about the *scene* due to the object-centric bias in datasets. ImageNet pre-training tends to favor image-level features, that may not capture object shape, making fine-tuning the decoder difficult for segmentation. We conjecture that these uncontrolled biases in ImageNet pre-training cause difficulties in directly predicting segmentation without fine-tuning the encoder.

**Initializing with a pre-trained encoder only.** To eliminate the effect of the depth-initialized decoder and only test the encoder, we replace the decoder by ‘fresh’ randomly initialized weights and fine-tune the whole network. Tab. 3 shows that depth pre-training outperforms ImageNet when the effect of pre-training is isolate to the encoder. This is also supported by neural activations in Fig. 6 where the regions activated after depth pre-training align well with semantic boundaries. Nonetheless, the decoder does play a role in segmentation



	Pre-training	mIoU	P.Acc
Res. 18	None	41.35	70.75
	ImageNet	45.15	72.39
	Depth (encoder only)	<b>46.69</b>	<b>75.04</b>
Res. 50	None	44.66	73.37
	ImageNet	44.65	73.06
	Depth (encoder only)	<b>46.99</b>	<b>73.57</b>
ViT-L	ImageNet	57.53	81.48
	Depth (encoder only)	<b>58.12</b>	<b>81.94</b>

**Table 3: Initializing with depth encoder and random decoder.** Initializing with the depth encoder and a random decoder outperforms ImageNet initialization, but is worse than initializing with both encoder and decoder from the depth network (see Tab. 1).



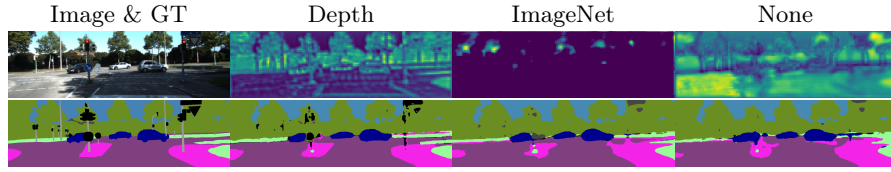
**Fig. 5: Mismatch between object scales.** ImageNet models are trained with a fixed input size and objects of interest tend to have a similar scale. However, objects in semantic segmentation dataset vary drastically in scale. Pre-training for depth provides robustness to scale change.

accuracy: Initializing the whole network with depth pre-training still performs the best, an advantage that is not afforded by a classification head.

**Results on Vision Transformers.** Given that ViTs [12] necessitate extensive pre-training, conducting comprehensive ablation studies on ViTs, as feasible with ResNet, becomes impractical due to data constraints. Thus, we conduct experiments using the DPT [41] depth model trained on a collection of datasets [33] (not including KITTI), and report the best results in Tab. 3 after an exhaustive search for the optimal learning rate. Notably, we identified the optimal learning rate for this experiment to be  $5e-8$ , with larger learning rates yielding suboptimal results. This suggests that DPT inherently provides robust representations sufficient for segmentation, requiring minimal fine-tuning in comparison to CNNs.

**Cross architecture.** To test whether depth pre-training favors our particular choice of architecture, we use the same encoders to initialize DeepLab V3 and follow the same fine-tuning procedure as common practice, *i.e.*, ImageNet initialization. Results are presented in Tab. 1. All pre-trainings significantly improve accuracy compared with random initialization. The depth model trained from scratch provides the same level of performance as ImageNet, while subsequent depth training after ImageNet pre-training leads to further improvements.

**Robustness to object scales.** Given a fixed resolution, “primary” objects in object-centric data tend to have a similar scale. For example, ImageNet models are trained on  $224 \times 224$ , thus cars typically have a size of 100 to 200 pixels. In KITTI, however, cars appear at different scales, varying from a few to a few hundred pixels. Fig. 5 illustrates the scale mismatch between datasets. On the other hand, depth pre-training can be done in the same domain as segmentation, hence having robustness to object scales. We examine such robustness: Pre-training on one resolution, fine-tuning on another. Since higher resolution images contain smaller scales (yellow box in Fig. 5), pre-training on them should still work on smaller images. Pre-trained on smaller resolutions, however, should work



**Fig. 6: Neural activation and semantic segmentation result.** We visualize the neural activation map for a shallow layer of ResNet 18 encoder trained from different initializations, and their corresponding segmentation results. Boundaries are better aligned to semantic boundaries in our model.

poorly on larger images. Our results validate this conjecture: The former achieves a final mIoU of 48.54 (ImageNet: 45.15) while the latter diverges during training.

**Neural Activation.** We visualize the activations of the ResNet18 encoder by Grad-CAM [43], which was originally designed for classification. We modify it for segmentation by inspecting the gradient of neural response to the summation of predicted labels for pixels instead of one single class label. We visualize shallow layers (before pooling) for high spatial resolution. Fig. 6 shows neural activation maps and segmentation outputs from depth pre-training align with semantic boundaries. This confirms not just the similarities between natural and range image statistics discussed in [22], but also the bias introduced by classification, as activations of ImageNet pre-trained encoder do not resemble object boundaries.

**Comparison with optical flow.** One hypothesis for the effectiveness of depth pre-training is that the process leverages the statistics of natural scenes where simply-connected components of the range map often correspond to semantically consistent regions. Thus, fine-tuning simply aligns the range of two piece-wise smooth functions. We challenge this hypothesis by trying optical flow, which also exhibits a piecewise-smooth range and is obtained by minimizing the same photometric error. We train optical flow on a siamese network with two shared-weight encoders. While both optical flow and depth capture multiply-connected object boundaries (Fig. 7), using encoders pre-trained for optical flow is detrimental. We conjecture that optical flow does not capture the stable inductive bias afforded by the static component of the underlying scene. Specifically, optical flow is compatible with an underlying 3D geometry only when the scene is rigid, but rigidity is not enforced when inferring optical flow. In contrast, depth forces recognition of rigidity and discards moving objects [29–31] as outliers, which then enables isolating them, also beneficial to semantic segmentation.

**Comparison with other pre-training methods.** In Tab. 4, for completeness, we report results on supervised pre-training by semantic segmentation on MS-COCO [34]. Unsurprisingly, pre-training on the same task with additional annotated data yields good performance. Depth estimation, despite not needing additional annotation, yields on-par performance. We also report results on masked autoencoding. We remove random rectangular regions from images, and the network aims to reconstruct the original image. Also known as ‘inpainting’ [1], it is considered an effective method for feature learning [39]. It yields

	All		Freeze	
	mIoU	↑	mIoU	↑
None	41.35	-	41.24	-
Flow	38.47	<b>-2.88</b>	32.19	<b>-9.05</b>
Depth-Rand	46.00	4.65	43.02	1.78
Depth	50.20	8.85	46.53	5.29

**Fig. 7: Depth helps, flow hurts.** Although both are pre-trained by minimizing a photometric reconstruction error, monocular depth outperforms optical flow. This stems from the fact that inferring depth from a single image is ill-posed so the network learns inductive priors that are rich in semantics over the structures within a scene. In contrast, any discriminative features will support the correspondence search, so the flow network is not constrained to learning semantics, yielding poor fine-tuning accuracy.

**Table 4: Comparison with different pre-trainings.** Encoder: ResNet50; Reconstruction: by inpainting randomly corrupted regions (masked autoencoding); Supervised Segmentation: trained on MS-COCO.

Pre-training	mIoU	P.Acc	Pre-training	mIoU	P.Acc
Supervised Segmentation	<b>51.28</b>	74.88	Contrastive (DINO)	44.19	71.36
Depth	50.92	<b>77.34</b>	Optical Flow	42.72	71.80
Reconstruction (MAE)	47.18	74.16	Contrastive (MOCO V2)	37.04	65.91

inferior performance compared with depth. We conjecture that this is due to artificial rectangular masking which does not respect the natural image statistics, while monocular depth estimation yields occluded regions that border on objects’ silhouettes. We also provide results initializing the encoder with contrastive learning (MOCO V2 [9] and DINO [4]). While they also remove bias in human annotation for classification, similar to supervised classification, they are still prone to the inherent inductive bias in pre-training data.

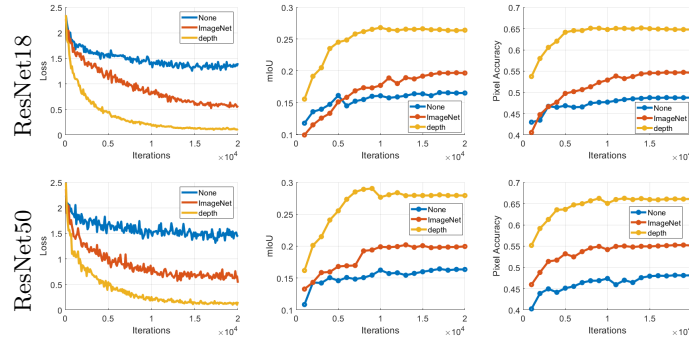
## 4.2 Full-scale Fine-tuning on the Whole Dataset

We now extend experiments to fine-tuning on full-scale datasets, aiming to validate the findings drawn from the controlled experiments conducted on KITTI.

**Cityscapes** [10] contains 2975 training and 500 validation images. Each image has a resolution of  $2048 \times 1024$  densely labeled into 19 semantic classes. The dataset also has 20000 unlabeled stereo pairs with a disparity map, converted to depth via focal length and camera baseline. Like KITTI, Cityscapes is also an outdoor driving dataset. Here we minimize an L1 loss between depth estimates and depth computed from stereo. We modify the prediction head of DeepLabV3 to train for depth, then re-initialize the last layer of the decoder for semantic segmentation. We are unable to reproduce the original numbers. For a fair comparison, we retrained [6] using the discussed pre-training methods and finetuned under the same training protocol *i.e.* batch size, augmentations, schedule, *etc.* Tab. 5 summarizes the outcomes. We present not only the most favorable results from an extensive training process but also results from a controlled approach

**Table 5: Segmentation accuracy on Cityscapes.** Similar to KITTI, pre-training for depth improves segmentation accuracy. Interestingly, under the controlled settings with limited data augmentation and fewer fine-tuning epochs, the model achieves higher performance when pre-training on cropped  $256 \times 256$  patches.

	Full				Controlled			
	Training		Validation		Training		Validation	
	mIoU	P.Acc.	mIoU	P.Acc.	mIoU	P.Acc.	mIoU	P.Acc.
None	76.10	95.41	63.97	93.76	73.82	95.26	60.43	93.07
ImageNet	81.84	96.46	70.41	95.75	76.70	95.71	61.80	93.40
Depth	83.46	96.80	<b>73.17 95.24</b>		77.42	95.82	62.57	93.46
Depth-cropped	<b>86.80 97.41</b>		72.22	95.01	<b>79.90 96.24</b>		<b>65.09 94.00</b>	



**Fig. 8: Results on NYU-V2.** Similar to KITTI, initializing with depth pre-trained weights trains faster and significantly improves semantic segmentation accuracy.

with restricted data augmentations and fewer training iterations. Remarkably, improvements remain consistent across both scenarios.

One may argue that since depth and semantic segmentation maps are both piece-wise smooth, adapting from depth is naturally easy if the model is aware of each pixel’s relative position in the image. In order to test this statement, instead of pre-training for depth on the full image, we train depth on randomly cropped  $256 \times 256$  patches, and the model has no spatial awareness of the position of a patch in the image, so depth is purely estimated by local information. This practice (Depth-cropped) surprisingly improves semantic segmentation results under controlled settings, showing that using depth as pre-training goes beyond a simple mapping from one smooth function to another. Interestingly, this approach significantly improves training accuracy in the full setting but leads to a slight reduction in validation accuracy, suggesting a potential issue of over-fitting. Future research is necessary to delve into this intriguing phenomenon. Another noteworthy observation is that when pre-trained using depth data, the model exhibits superior performance with a higher initial learning rate of 0.1, as opposed to 0.01 used for ImageNet initialization. Conversely, employing an initial learning rate of 0.1 with ImageNet weights can be detrimental and may

result in divergence. These findings suggest that pre-training the network with depth estimation may lead to a smoother local loss landscape.

**NYU-V2 [37]** is an indoor dataset that contains 795 densely annotated images for training and 654 for testing. There are also 407024 unannotated frames with synchronized depth images captured by a Microsoft Kinect. Since the main hypothesis is agnostic to how depth is learned, we pre-trained for depth using ground-truth as supervision. Unlike outdoor driving, which commonly features sky on top, and road and vehicles in the middle of the image with largely planar camera motion, indoor scenes are characterized by more complex layouts with 6 DoF camera motion, yielding images that are even less likely to resemble the object-centric ones commonly observed in classification datasets. This may be why initializing the model with depth pre-trained weights significantly improves semantic segmentation accuracy with both ResNet18 and ResNet50 (see Fig. 8). Note that pre-training by depth yields faster convergence, similar to KITTI.

### 4.3 Out-of-domain Transfer from Large-scale Pre-training

One may conjecture that the advantage of pre-training with monocular depth stems from training within the same domain as downstream semantic segmentation. In practice, in-domain data collection for depth pre-training is indeed ideal since it does not require human labeling. In a scientific context, we are interested in testing this conjecture by investigating the transferability of depth models to tasks outside their original domain. However, common depth datasets are considerably small in scale, making fair comparisons with popular pre-trained models, *e.g.* MAE [18] and DINO v2 [38] that are trained on millions or even billions of images, infeasible. Fortunately, recent methods [33, 41] suggest an alternative approach for learning monocular depth by training for relative depth instead of absolute depth, allowing for the integration of multiple data sources during training. Leveraging Depth Anything [55], a depth model trained on such scaled-up mixed datasets, we fine-tune for segmentation on three out-of-domain downstream datasets: ADE20k [59], PascalVOC [13], and CityScapes [10].

The PascalVOC dataset comprises 10,582 fully annotated images for training purposes and an additional 1,449 for testing, covering a variety of 20 foreground object classes. On the other hand, ADE20k, a sizable dataset, includes 20,210 training images and 2,000 testing images across 150 different classes. The resolution of PascalVOC and ADE20k is  $512 \times 512$  and  $896 \times 896$  respectively. On both datasets, we try both fine-tuning the whole network and linear probing with a frozen encoder. Reported in Tab. 6, the results align with the findings on ViTs on the KITTI dataset (in Tab. 3), in which case the pre-training data also originates from out-of-domain sources. Notably, consistent improvements in semantic segmentation performance are observed across all datasets compared to the baseline initialization DINO v2 [38]. It is anticipated that the improvement achieved with a frozen encoder is more substantial than when fine-tuning the entire network, which is consistent with our previous results. This trend underscores that, when trained at a scale, depth models exhibit robust transferability to novel downstream data domains, just as other pre-training methods.

**Table 6: Out-of-domain transfer with large-scale pre-trainings.** We compare Depth Anything with MAE and DINO v2 for semantic segmentation, reporting results (in mIoU) on both fine-tuning (ft) and linear probing (lin.). Depth Anything improves semantic segmentation across all datasets and settings. \*: with DINO v2 initialization.

Pre-training	# of pre-train data	ADE20k		PascalVOC		CityScapes	
		ft	lin.	ft	lin.	ft	lin.
MAE [18]	1.28 million	53.6	49.0	-	67.6	-	58.4
DINO v2 [38]	142 million	58.1	47.7	86.5	86.3	82.7	71.3
Depth Anything [55]	63.5 million*	<b>59.7</b>	<b>52.3</b>	<b>87.7</b>	<b>87.3</b>	<b>84.8</b>	<b>74.8</b>

## 5 Discussion

Inferring depth only requires multiple images (*e.g.* videos or multiple viewpoints) *of the same scene* [25] or range sensing, both do not require human-induced priors and bias, unlike semantic tasks that rely entirely on induction: We can associate a label to an image because that image has *something* in common with *some other* image, portraying a different scene, that some annotator attached a particular label to. That inductive chain has to go through the head of human annotators, who are biased in ways that cannot be easily quantified and controlled. Depth from binocular or motion imagery does not require induction and can be performed *ab-ovo*. Learning a monocular model from such supervision eliminates the implicit selective bias from human annotators, yet our findings validate the main hypothesis that the inductive bias learned from such a “human-free” process transfers well to the downstream semantic segmentation task. Of course, if different supervisions are available, either for semantic [11], segmentation [27], or both [28], we want to incorporate that information.

Our hypothesis and findings are agnostic to how depth is attributed to a single image: One can perform pre-training using monocular videos, stereo, structure light, LIDAR, or even human guidance [58]. One of our pre-training minimizes the photometric reprojection error, used by many predictive and generative approaches. However, an unstructured displacement field is in general not compatible with a rigid motion. Only if this displacement field has the structure of an epipolar transformation [44] is the prediction task forced to encode the 3D structure of the scene. This may explain why video prediction is not as effective for pre-training despite many attempts [24, 32, 46, 53].

One limitation of monocular depth estimation is that it may require a calibrated camera, so one cannot use generic videos harvested from the web. There is nothing in principle preventing us from using uncalibrated cameras, simply by adding the calibration matrix  $K$  to the nuisance variables. While the necessary conditions for full Euclidean reconstruction are rarely satisfied in consumer videos (for instance, they require cyclo rotation around the optical axis, not just panning and tilting), the degrees of freedom that cannot be reconstructed are moot as they do not affect the reprojection. Moreover, recent progress in training for relative depth from mixed data sources [33, 41, 55] shows the potential to unlock a virtually unlimited volume of training data, warranting future research.

**Acknowledgements.** This work was supported by ONR N00014-22-1-2252 and ARO W911NF-17-1-0304.

## References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 417–424 (2000)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
3. Caron, M., Bojanowski, P., Mairal, J., Joulin, A.: Unsupervised pre-training of image features on non-curved data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **40**(4), 834–848 (2017)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017)
7. Chen, M., Artières, T., Denoyer, L.: Unsupervised object segmentation by redrawing. *Advances in neural information processing systems* **32** (2019)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
9. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
13. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* **88**(2), 303–338 (Jun 2010)
14. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters* **4**(2), 1661–1668 (2019)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

16. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction (October 2019)
17. Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., et al.: Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *Advances in Neural Information Processing Systems* **36** (2024)
18. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 15979–15988 (2021)
19. Hoyer, L., Dai, D., Chen, Y., Köring, A., Saha, S., Van Gool, L.: Three ways to improve semantic segmentation with self-supervised depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11130–11140 (2021)
20. Hoyer, L., Dai, D., Wang, Q., Chen, Y., Van Gool, L.: Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *arXiv preprint arXiv:2108.12545 [cs]* (2021)
21. Huang, J., Lee, A.B., Mumford, D.: Statistics of range images. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. vol. 1, pp. 324–331. IEEE (2000)
22. Huang, J., Mumford, D.: Statistics of natural images and models. In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. vol. 1, pp. 541–547. IEEE (1999)
23. Jiang, H., Larsson, G., Shakhnarovich, M.M.G., Learned-Miller, E.: Self-supervised relative depth learning for urban scene understanding. In: *Proceedings of the european conference on computer vision (eccv)*. pp. 19–35 (2018)
24. Jin, B., Hu, Y., Tang, Q., Niu, J., Shi, Z., Han, Y., Li, X.: Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4554–4563 (2020)
25. Julesz, B.: Binocular depth perception without familiarity cues: Random-dot stereo images with controlled spatial and temporal properties clarify problems in stereopsis. *Science* **145**(3630), 356–362 (1964)
26. Julesz, B.: *Foundations of cyclopean perception*. (1971)
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4015–4026 (2023)
28. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* **128**(7), 1956–1981 (2020)
29. Lao, D., Hu, Z., Locatello, F., Yang, Y., Soatto, S.: Divided attention: Unsupervised multi-object discovery with contextually separated slots. *arXiv preprint arXiv:2304.01430* (2023)
30. Lao, D., Sundaramoorthi, G.: Minimum delay moving object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4250–4259 (2017)
31. Lao, D., Sundaramoorthi, G.: Extending layered models to 3d motion. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 435–451 (2018)



32. Lao, D., Zhu, P., Wonka, P., Sundaramoorthi, G.: Flow-guided video inpainting with scene templates. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14599–14608 (2021)
33. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341 (2019)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
35. Liu, T.Y., Agrawal, P., Chen, A., Hong, B.W., Wong, A.: Monitored distillation for positive congruent depth completion. In: European Conference on Computer Vision. pp. 35–53. Springer (2022)
36. Lopez-Rodriguez, A., Busam, B., Mikolajczyk, K.: Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. In: Proceedings of the Asian Conference on Computer Vision (2020)
37. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: ECCV (2012)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), <https://openreview.net/forum?id=a68SUt6zFt>
39. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
40. Ramirez, P.Z., Tonioni, A., Salti, S., Stefano, L.D.: Learning across tasks and domains. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8110–8119 (2019)
41. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
42. Saha, S., Obukhov, A., Paudel, D.P., Kanakis, M., Chen, Y., Georgoulis, S., Van Gool, L.: Learning to relate depth and semantics for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8197–8207 (2021)
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
44. Sundaramoorthi, G., Petersen, P., Varadarajan, V., Soatto, S.: On the set of images modulo viewpoint and contrast changes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 832–839. IEEE (2009)
45. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
46. Wang, Y., Wu, J., Long, M., Tenenbaum, J.B.: Probabilistic video prediction from noisy data with a posterior confidence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10830–10839 (2020)

47. Wong, A., Cicek, S., Soatto, S.: Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters* **6**(2), 1495–1502 (2021)
48. Wong, A., Fei, X., Hong, B.W., Soatto, S.: An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters* **6**(2), 3120–3127 (2021)
49. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters* **5**(2), 1899–1906 (2020)
50. Wong, A., Soatto, S.: Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5644–5653 (2019)
51. Wong, A., Soatto, S.: Unsupervised depth completion with calibrated backprojection layers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12747–12756 (2021)
52. Wu, Y., Liu, T.Y., Park, H., Soatto, S., Lao, D., Wong, A.: Augundo: Scaling up augmentations for monocular depth completion and estimation. In: *European Conference on Computer Vision*. Springer (2024)
53. Wu, Y., Gao, R., Park, J., Chen, Q.: Future video synthesis with object motion prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5539–5548 (2020)
54. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891* (2024)
55. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: *CVPR* (2024)
56. Yang, Y., Wong, A., Soatto, S.: Dense depth posterior (ddp) from single image and sparse range. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3353–3362 (2019)
57. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3712–3722 (2018)
58. Zeng, Z., Wang, D., Yang, F., Park, H., Soatto, S., Lao, D., Wong, A.: Wordepth: Variational language prior for monocular depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9708–9719 (2024)
59. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 633–641 (2017)