MeSa: Masked, Geometric, and Supervised Pre-training for Monocular Depth Estimation

Muhammad Osama Khan
New York University
osama.khan@nyu.edu

Junbang Liang
Amazon
junbanl@amazon.com

Chun-Kai Wang
Amazon
ckwang@amazon.com

Shan Yang
Amazon
ssyang@amazon.com

Yu Lou
Amazon
ylou@amazon.com

Abstract

Pre-training has been an important ingredient in developing strong monocular depth estimation models in recent years. For instance, self-supervised learning (SSL) is particularly effective by alleviating the need for large datasets with dense ground-truth depth maps. However, despite these improvements, our study reveals that the later layers of the SOTA SSL method are actually suboptimal. By examining the layer-wise representations, we demonstrate significant changes in these later layers during fine-tuning, indicating the ineffectiveness of their pre-trained features for depth estimation. To address these limitations, we propose MeSa, a unified framework that leverages the complementary strengths of masked, geometric, and supervised pre-training. Hence, MeSa benefits from not only general-purpose representations learnt via masked pre-training but also specialized depth-specific features acquired via geometric and supervised pre-training. Our CKA layer-wise analysis confirms that our pre-training strategy indeed produces improved representations for the later layers, overcoming the drawbacks of the SOTA SSL method. Furthermore, via experiments on the NYUv2 and IBims-1 datasets, we demonstrate that these enhanced representations translate to performance improvements in both the in-distribution and out-of-distribution settings. We also investigate the influence of the pre-training dataset and demonstrate the efficacy of pre-training on LSUN, which yields significantly better pre-trained representations. Overall, our approach surpasses the masked pre-training SSL method by a substantial margin of 17.1% on the RMSE. Moreover, even without utilizing any recently proposed techniques, MeSa also outperforms the most recent methods and establishes a new state-of-the-art for monocular depth estimation on the challenging NYUv2 dataset.

1 Introduction

Monocular depth estimation is an important computer vision problem, with applications ranging from self-driving cars to augmented reality and robotics. Initially, supervised learning methods [10, 11, 26] were developed to tackle this problem, utilizing annotated depth data for training the models. However, collecting diverse real-world datasets with precise ground-truth depth is extremely challenging. Hence, self-supervised methods were developed that learn depth from stereo image pairs [12, 14] or

monocular videos [50] without relying on depth annotations. Particularly, self-supervised monocular depth estimation from videos is especially appealing for real-world applications, as it only requires a single camera for data collection. These self-supervised methods typically employ view synthesis as the main supervision signal and are trained via the photometric loss [50].

In recent years, pre-training has emerged as an important factor in developing strong depth estimation models. For instance, Ranftl et al. [35] noticed that randomly initialized models tend to be about 35% worse than the same models pre-trained on ImageNet. Building on this, Xie et al. [42] recently obtained SOTA results by directly fine-tuning a SimMIM [43] pre-trained network for depth estimation. SimMIM [43], a masked pre-training [21, 43] method, learns self-supervised representations via the following pretext task: given a partially masked input image, the network reconstructs the masked portions of the image. Interestingly, this relatively simple pre-training task obtains SOTA performance on depth estimation without utilizing any geometric priors.

Despite these pre-training algorithms yielding significant improvements, we demonstrate that the last layers of the SOTA SSL method [42] are actually suboptimal. To investigate this, we compare the similarity of the pre-trained representation and fine-tuned representation for each layer (Figure 2). Intuitively, a higher similarity means that the pre-trained representation is well-suited for the downstream task and hence requires minimal updates during the fine-tuning process and vice versa [30]. Our analysis reveals significant changes in these later layers during fine-tuning, indicating the ineffectiveness of the pre-trained features for depth estimation.

Based on the intuition that different pre-training strategies capture distinct types of features, we propose a unified pre-training framework, MeSa, that addresses the aforementioned limitations. MeSa leverages the complementary strengths of three types of learning strategies: masked, geometric, and supervised pre-training. Via this synergy, our framework benefits from not only general-purpose representations learnt via masked pre-training but also specialized depth-specific features acquired via geometric and supervised pre-training.

Our layer-wise analysis confirms that our pre-training strategy indeed produces improved representations for the later layers, thereby successfully overcoming the drawbacks of the SOTA SSL method. Furthermore, via experiments on the NYUv2 and IBims-1 datasets, we demonstrate that these enhanced representations translate to performance improvements in both the in-distribution and out-of-distribution settings. Additionally, we investigate the impact of the pre-training dataset and demonstrate the superiority of pre-training on the LSUN dataset, resulting in significantly improved pre-trained representations. Moreover, via benefiting from the 3D projective geometry during pre-training, our method helps eliminate the artifacts observed in the SOTA SSL model that lacks geometric priors.

To sum up, our main contributions include:

- A qualitative analysis of pre-training effectiveness based on layer-wise feature similarities that uncovers the insight that the later layers of the SOTA SSL method are not optimally pre-trained.

- A novel pre-training pipeline, MeSa, that effectively pre-trains the entire network by leveraging the complementary strengths of masked, geometric, and supervised pre-training, thereby benefiting from both general-purpose as well as specialized depth-specific features.

- Our pre-trained model surpasses the SOTA masked pre-training method by a substantial margin of 17.1% on the RMSE. Moreover, even without utilizing any recently proposed techniques, it also outperforms the most recent methods and establishes a new state-of-the-art for monocular depth estimation on the challenging NYUv2 dataset.

2 Related Work

2.1 Monocular Depth Estimation

Depth estimation was initially treated as a supervised learning problem [10, 11, 26], requiring ground-truth depth to train the networks. Since collecting ground-truth depth is a challenging problem in itself, some methods explored using synthetic data [29]. However, it is also not trivial to generate large varied synthetic datasets containing diverse scenes. A promising alternative is to train depth estimation networks using self-supervised learning, where view synthesis forms the main supervision
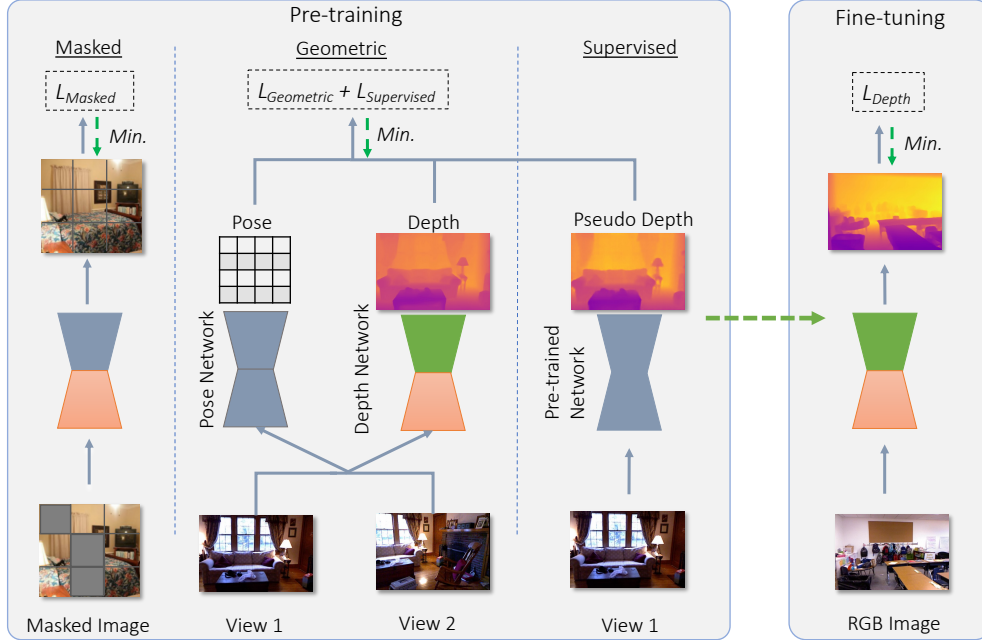
Figure 1: Our proposed framework, MeSa, effectively leverages both general-purpose as well as depth-specific features via utilizing the complementary strengths of three pre-training strategies – 1) Masked, 2) Geometric, and 3) Supervised pre-training. The pre-trained models are fine-tuned in a supervised manner on the downstream dataset.

signal and depth is synthesized as an intermediate step. Zhou et al. [50] developed one of the first monocular self-supervised algorithms, where they trained joint depth estimation and pose estimation networks using monocular videos. Since then, several methods [17, 45, 7, 14, 28] have been proposed to improve various components. For instance, Monodepth2 [15] proposed a simple model to elegantly handle occlusions as well as reduce visual artifacts.

SC-DepthV1 [4] proposed a geometry consistency loss in order to learn scale-consistent depth maps across the video. SC-DepthV2 [3] built on top of this and proposed an auto-rectify network to alleviate training instabilities due to rotation of handheld cameras, thereby enabling better depth estimation on indoor scenes. Following this, SC-DepthV3 [38] used an external pre-trained depth estimation network to generate pseudo-depth, which was used to improve the depth estimation performance in dynamic scenes. Remarkably, Xie et al. [42] recently achieved SOTA performance by directly fine-tuning a masked pre-trained SimMIM [43] model for depth estimation without utilizing any geometric constraints used in previous methods. In this paper, we analyze the layer-wise representations of this SOTA model and leverage the 3D projective geometry in order to overcome its drawbacks, thereby designing a strong self-supervised learning algorithm for monocular depth estimation.

2.2 Self-supervised Learning

Self-supervised learning (SSL) is a powerful approach for learning representations from unlabeled data without requiring human annotation. SSL learns by leveraging various pretext tasks such as relative patch prediction [9], rotation prediction [13], and colorization [48]. A popular class of SSL methods is contrastive learning [40, 20, 6, 5, 16, 41] which learns transformation invariant representations via maximizing the similarity of positive pairs and minimizing the similarity of negative pairs. Masked image modeling (MIM) has recently gained traction in vision [21, 43] following the success of masked pre-training in NLP [8, 27]. Although these MIM approaches have obtained SOTA results in several tasks, a comprehensive understanding of the pre-trained representations is still lacking, making the development of new SSL methods challenging. Recently, a few methods [42, 32, 33] have attempted to understand these pre-trained presentations. Our work

follows in this direction and uses the insights from our analysis to design a better self-supervised learning algorithm that effectively pre-trains the entire network for depth estimation.

3 Method

As illustrated in Figure 1, our framework consists of three pre-training strategies – 1) Masked, 2) Geometric, and 3) Supervised pre-training. By synergistically integrating these three strategies, MeSa facilitates effective representation learning that benefits from both general-purpose representations (via masked pre-training) as well as depth-specific features (via geometric and supervised pre-training). Hence, MeSa effectively pre-trains the entire network including the later layers, thereby addressing the limitations of the SOTA SSL method. In the following subsections, we briefly introduce masked, geometric, and supervised pre-training in Sections 3.1, 3.2, and 3.3 respectively. This is then followed by the fine-tuning and implementation details in Section 3.4. Lastly, Section 3.5 outlines the techniques used for the layer-wise analysis to qualitatively understand the representations learnt via the three pre-training strategies.

3.1 Masked Pre-training

Masked pre-training learns representations by masking a region of the input image and reconstructing the masked region based on the partially masked input. In this work, we utilize the SimMIM framework proposed by Xie et al. [43] for masked pre-training. However, other pre-training methods could be used as well since our framework is agnostic to the exact choice of pre-training method. We use a simple masking strategy and randomly mask out some patches of the input image. The decoder consists of a single linear layer to predict the pixels of the masked regions. Finally, the network is trained using a plain ℓ_1 loss applied to the masked regions:

$$L = \frac{1}{|\mathbf{x}_M|} \left\| \mathbf{x}_M - g(f(\hat{\mathbf{x}}))_M \right\|_1 \tag{1}$$

where f is the encoder, g is the decoder, $\mathbf{x}$ is the raw image, and $\hat{\mathbf{x}}$ is the partially masked image input to the encoder. The subscript M denotes the masked subset of pixels whereas $|\cdot|$ denotes the number of pixels. After training, the decoder g is usually discarded and the encoder f is used for downstream tasks.

3.2 Geometric Pre-training

Geometric pre-training learns representations via utilizing the 3D projective geometry. In this work, we use the SC-DepthV2 [3] framework for geometric pre-training, where depth and pose networks are trained jointly on monocular videos. Given consecutive image pairs I_a, I_b from a video, we first generate the predicted depths D_a, D_b of the two views and the relative camera pose P_{ab} between them:

$$D_a = g'(f(I_a)), \quad D_b = g'(f(I_b)), \quad P_{ab} = h(I_a, I_b) \tag{2}$$

where f and g' are the depth estimation encoder and decoder respectively whereas h is the pose estimation network.

View synthesis forms the main supervision signal (i.e., given one view, the task is to predict a view of the same scene from a different viewpoint). Given the predicted depth D_a, relative camera pose P_{ab}, and the source view I_b, differentiable bilinear interpolation [22] is used to generate a prediction of the reference view $I_{a'}$. Hence, depth is actually synthesized as an intermediate step and photometric loss between the generated $I_{a'}$ and actual I_a images is used to train the network:

$$L_P = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \left(\lambda \left\| I_a(p) - I'_a(p) \right\|_1 + (1 - \lambda) \frac{1 - \text{SSIM}_{aa'}(p)}{2} \right) \tag{3}$$

where $\mathcal{V}$ is the set of valid points projected from I_a to I_b and $\text{SSIM}_{aa'}$ is the similarity between I_a and $I_{a'}$ computed via the SSIM function [39].

Moreover, the geometry consistency loss is used to ensure that the two depth maps (D_a and D_b) are consistent in terms of 3D geometry:

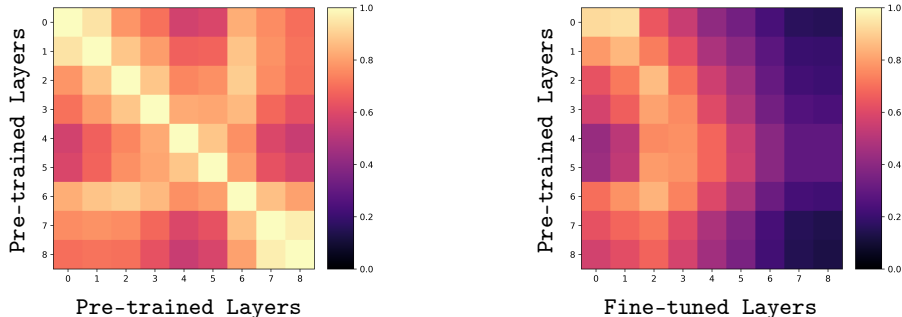$$L_G = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} D_{\text{diff}}(p) \tag{4}$$

4

Figure 2: Layer-wise analysis of the SOTA masked pre-trained model. **Left**: We discover a distinctive U-shaped pattern (top row) in representations of the pre-trained model as we delve deeper into the network. **Right**: We observe that the later layers (5-8) undergo significant changes during fine-tuning, as depicted by the lower similarities between pre-trained and fine-tuned features along the diagonal line. These results indicate that the pre-trained representations of the later layers are not effectively utilized for the downstream depth estimation task.

where D_{diff} is the depth inconsistency map between D_a and D_b. For further details, please refer to SC-Depth [4].

3.3 Supervised Pre-training

Supervised pre-training learns representations via using off-the-shelf supervised pre-trained networks for depth estimation. In this work, we use the SC-DepthV3 [38] framework for supervised pre-training, which uses the LeReS [44] pre-trained network to generate pseudo-depth. In contrast to standard supervised pre-training, the pre-trained network here is only used for pseudo-depth estimation [38], which allows to improve performance on dynamic objects via the confident depth ranking loss (L_{CDR}) and on object boundaries via the egde-aware relative normal loss (L_{ERN}). Hence, the overall loss for joint training via geometric and supervised pre-training is given by:

$$L = \alpha L_P^M + \beta L_G + \gamma L_N + \delta L_{\text{CDR}} + \epsilon L_{\text{ERN}} \tag{5}$$

where L_P^M is the weighted photometric loss, L_G is the geometry consistency loss, L_N is the normal matching loss, L_{CDR} is the confident depth ranking loss, and L_{ERN} is the edge-aware relative normal loss.

3.4 Implementation Details

We pre-train via the three learning strategies sequentially in order to avoid training instability associated with concurrent training [18]. Moreover, this also allows us to significantly reduce the pre-training time since we can leverage existing pre-trained models, thereby eliminating the need for training the entire pipeline from scratch whenever a new pre-training method is added.

In the masked pre-training phase, we employ a Swin-v2-L network as the encoder f and a single linear layer as the decoder g. For the subsequent geometric and supervised pre-training stages, we keep the pre-trained encoder f and replace the decoder with a DispNet [50] g'. Additionally, we utilize a ResNet18 [19] backbone in the pose network h to compute the relative camera pose between the concatenated input images.

We use the LSUN [46] dataset for masked pre-training whereas we utilize the training split of the NYUv2 [37] dataset for geometric pre-training. For supervised pre-training, we leverage the pre-trained network provided by LeReS [44] to generate the pseudo-depth. We follow the training details outlined in the official methods (SimMIM [43] and SC-Depth [38]), with the exception of using an image size of 480×480 instead of 256×320 for geometric and supervised pre-training on NYUv2.

To evaluate all the pre-trained methods, we follow the SOTA SSL method [42] and fine-tune them on the training split of the NYUv2 dataset [37]. We fine-tune for 75 epochs on 8 A100 GPUs, using a polynomial learning rate schedule with a 0.9 factor and a min lr of 10^{-5} and a max lr of 10^{-4}.

5

3.5 Layer-wise Analysis

Following previous work [42, 30], we use CKA (centered kernel alignment) [25] for our analysis, which is a metric that allows us to compare the representations of various layers within a network. CKA similarity values range from 0 to 1, with increasing values meaning that the two representations are more similar. We compare the layer-wise representations of nine layers within the Swin-v2-L architecture. For convenience, we refer to them as layers 0-8, with the exact layers detailed in the supplementary material.

We perform two types of analyses to evaluate the efficacy of the pre-trained layer-wise representations. Firstly, we compare the representations of layers 1-8 in the network to layer 0 (Figure 2 left) in order to study the changes in representations as we delve deeper into the network. Secondly, for each layer, we compare the similarity of its pre-trained representation to its fine-tuned representation (Figure 2 right) in order to understand how each layer evolves during the fine-tuning process. Intuitively, a higher similarity means that the pre-trained representation is well-suited for the downstream task and hence requires minimal updates during the fine-tuning process and vice versa [30].

4 Experiments

Firstly, we analyze the layer-wise representations of the SOTA SSL model in order to illustrate the suboptimal pre-trained representations of the later layers (Section 4.1). To address this shortcoming, we propose a novel pre-training pipeline, MeSa, that not only achieves improved quantitative performance in both in-distribution (Section 4.2) and out-of-distribution (Section 4.3) settings but also learns better layer-wise representations (Section 4.4). Lastly, we investigate the impact of different pre-training datasets on the layer-wise features (Section 4.5) and conclude with a comprehensive comparison against the SOTA depth estimation models (Section 4.6).

4.1 Layer-wise Analysis of the SOTA SSL Model

Although the SOTA SSL model yields excellent performance, it is unclear which parts of the pre-trained model contribute to this improvement. In order to understand this, we analyze the layer-wise representations of this masked pre-trained model using the techniques discussed in Section 3.5.

Based on the analysis in Figure 2 (left), we discover a U-shaped pattern in the representation space, where the representations initially start getting farther apart but then become more similar to the initial layers towards the end of the network. The top row of Figure 2 (left) compares the representation of layer 0 against the representations of layers 0-8. The similarity of layer 0's representation with itself is 1 as expected. Thereafter, for layers 1-4, the representations begin to diverge from layer 0's representation, as indicated by the decreasing similarities (darker colors). However, interestingly, past layer 4, the similarity increases (lighter colors) suggesting that the representations are becoming more similar to the layer 0 representation. This U-shaped pattern is likely a result of the pre-training objective, which is to reconstruct masked portions of the image. Hence, the later layers end up having similar representations to the first few layers. A similar pattern has also been observed in [32, 33] in the context of speech processing.

Whereas this U-shaped pattern is useful for solving the self-supervised pretext task, it is not obvious if it aids in learning effective representations for downstream tasks. To investigate this, we compare the similarity of the pre-trained and fine-tuned representations of each layer in Figure 2 (right). From

Table 1: Geometric and supervised pre-training have complementary benefits to masked pre-training and result in significant improvements. MP: masked pre-training, GP: geometric pre-training, SP: supervised pre-training.

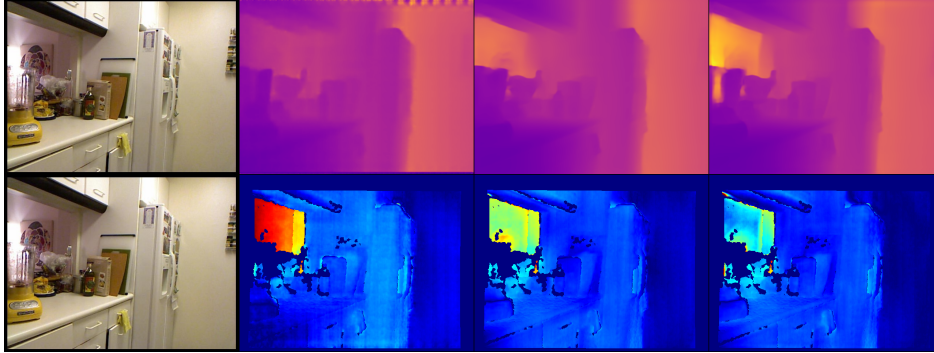| Pre-training Strategy	RMSE $\downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL $\downarrow$	log 10 $\downarrow$
MP [42]	0.287	0.949	0.994	0.999	0.083	0.035
MP + GP	0.269	0.951	0.993	0.998	0.076	0.033
MP + GP + SP (MeSa)	**0.265**	**0.954**	**0.995**	**0.999**	**0.074**	**0.032**
Relative Improvement (%)	7.64	0.54	0.06	0	10.2	8.05

Figure 3: Qualitative comparisons of the three pre-training strategies on NYUv2. Left→right: Image, MP, MP+GP, MP+GP+SP (MeSa). **Top**: predicted depth maps, **bottom**: error maps (blue→lower error; red→higher error).

the diagonal line (top-left to bottom-right), we observe that the similarities are quite small for the later layers (5-8). Lower similarities imply that these layers undergo significant changes during fine-tuning, indicating that they are not effective for the downstream depth estimation task [30].

4.2 MeSa

To address the aforementioned drawbacks, we propose a novel pre-training pipeline, MeSa, that utilizes the complementary benefits of masked, geometric, and supervised pre-training.

Table 1 shows the depth estimation results of the three pre-training strategies on the NYUv2 dataset. The results demonstrate significant performance improvements achieved by all three learning strategies, highlighting their complementary strengths. Overall, our approach surpasses the SOTA SSL model, which relies solely on masked pre-training, by 10.2% on absolute relative error (0.083→0.074) and 7.64% on RMSE (0.287→0.265). Figure 3 presents qualitative results of the three pre-training strategies on the NYUv2 dataset, demonstrating significant improvements through the integration of geometric and supervised pre-training. When relying solely on masked pre-training, as in the SOTA SSL method, artifacts appear at the top and bottom of the predicted depth maps due to the absence of ground-truth depth (during fine-tuning) in these regions. By utilizing geometric pre-training, our method effectively eliminates these artifacts since the photometric loss utilizes 3D projective geometry and learns accurate depth across the entire image without relying on ground-truth. Moreover, geometric and supervised pre-training also help enhance the sharpness of the predicted depth maps.

4.3 Out-of-distribution Performance

In this section, we evaluate the performance of the three pre-training strategies in the out-of-distribution (OOD) setting. To this end, we evaluate the models trained on the NYUv2 dataset directly on the IBims-1 dataset without any additional fine-tuning.

Table 2 presents the OOD results on the IBims-1 dataset. In addition to the overall depth estimation accuracy (measured via AbsRel), it is also important to improve the accuracy of depth boundaries

Table 2: In addition to improving the overall depth estimation accuracy, geometric and supervised pre-training also enhance the accuracy of depth boundaries and planar regions in the OOD setting. MP: masked pre-training, GP: geometric pre-training, SP: supervised pre-training.

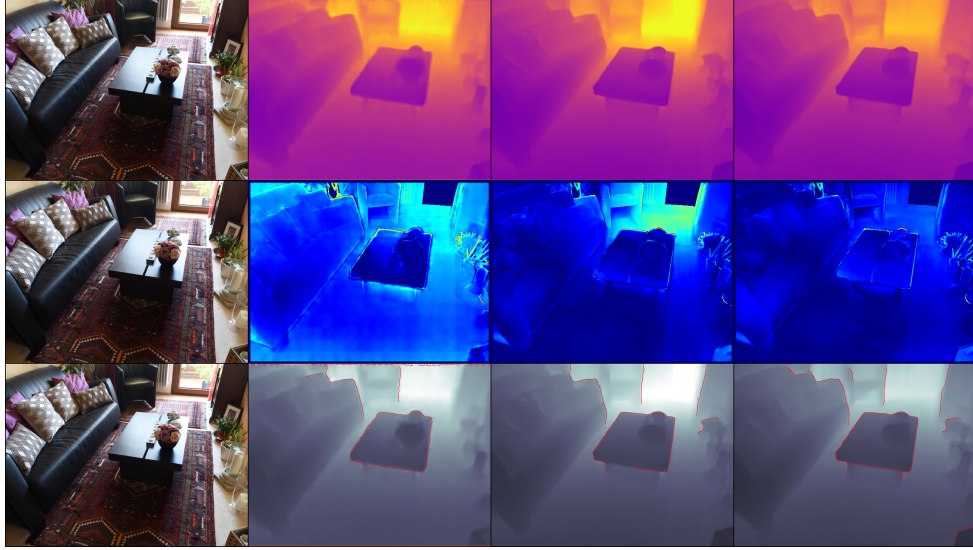| Pre-training Strategy	$\varepsilon_{\text{DBE}}^{\text{acc}} \downarrow$	$\varepsilon_{\text{DBE}}^{\text{comp}} \downarrow$	$\varepsilon_{\text{PE}}^{\text{plan}} \downarrow$	$\varepsilon_{\text{PE}}^{\text{orie}} \downarrow$	AbsRel $\downarrow$
MP [42]	2.40	30.05	3.26	8.16	0.089
MP + GP	2.25	24.69	2.46	**6.29**	**0.082**
MP + GP + SP (MeSa)	**2.16**	**19.31**	**2.17**	6.44	0.083
Relative Improvement (%)	10.3	35.8	33.4	22.9	8.4

Figure 4: Qualitative comparisons of the OOD performance of the three pre-training strategies on IBims-1. Left→right: Image, MP, MP+GP, MP+GP+SP (MeSa). **Top**: predicted depth maps, **middle**: error maps (blue→lower error; red→higher error), **bottom**: edge maps.
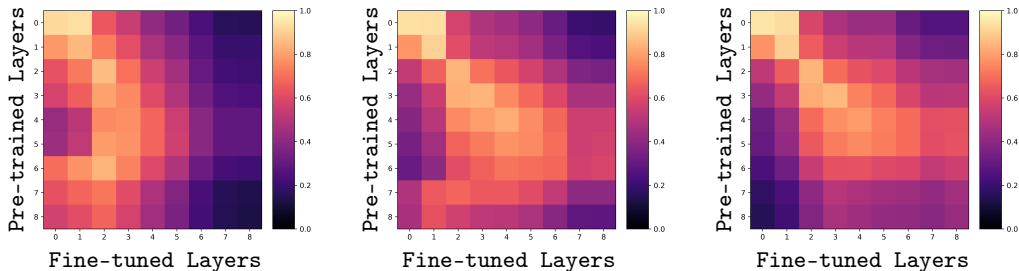

Figure 5: Layer-wise analysis of the three pre-training strategies, comparing the representations of the pre-trained model to the fine-tuned model. MeSa effectively pre-trains the entire network, including the last few layers. Left→right: MP, MP+GP, MP+GP+SP (MeSa).

(measured via $\varepsilon_{\mathrm{DBE}}^{\mathrm{acc}}$ and $\varepsilon_{\mathrm{DBE}}^{\mathrm{comp}}$) as well as planar regions (measured via $\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}$ and $\varepsilon_{\mathrm{PE}}^{\mathrm{orie}}$), both of which are critical for many real-world applications. Our results highlight that we surpass the SOTA masked pre-training method on all five metrics in the OOD setting. Notably, we achieve substantial improvements of 35.8% (30.05→19.31) and 33.4% (3.26→2.17) on depth boundary completeness ($\varepsilon_{\mathrm{DBE}}^{\mathrm{comp}}$) and planarity ($\varepsilon_{\mathrm{PE}}^{\mathrm{plan}}$), respectively. For a detailed description of all metrics, please refer to IBims-1 [24].

Figure 4 visualizes the OOD performance of the three pre-training strategies. From the results, we observe a significant improvement in the sharpness of depth maps when leveraging geometric and supervised pre-training. Moreover, the localization of depth boundaries is also greatly enhanced.

4.4 Layer-wise Analysis of MeSa

In this section, we analyze layer-wise representations of the MeSa pre-trained network to ascertain that our pre-training strategy indeed produces improved representations for the later layers, thereby successfully overcoming the drawbacks of the SOTA SSL method.

Similar to Section 4.1, for each layer, we compare the similarity of its pre-trained representation to its fine-tuned representation (Figure 5). As mentioned earlier, when using masked pre-training (MP) alone, the later layers (5-8) are not particularly beneficial for depth estimation since they undergo significant changes during fine-tuning, illustrated by the low similarity values. On the other hand, incorporating geometric pre-training (GP) alongside MP (i.e., MP+GP) allows for more effective

Table 3: Comparison with SOTA depth estimation methods on the NYUv2 dataset. MeSa outperforms previous methods across all metrics.

Method	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	REL ↓	log 10 ↓
BTS [23]	0.392	0.885	0.978	0.995	0.110	0.047
AdaBins [1]	0.364	0.903	0.984	0.997	0.103	0.044
DPT [36]	0.357	0.904	0.988	0.998	0.110	0.045
P3Depth [34]	0.356	0.898	0.981	0.996	0.104	0.043
NeWCRFs [47]	0.334	0.922	0.992	0.998	0.095	0.041
Xie et al. [42]	0.287	0.949	0.994	**0.999**	0.083	0.035
AiT [31]	0.275	0.954	0.994	**0.999**	0.076	0.033
ZoeDepth [2]	0.270	0.955	**0.995**	**0.999**	0.075	0.032
VPD [49]	0.254	**0.964**	**0.995**	**0.999**	0.069	0.030
MeSa	**0.238**	**0.964**	**0.995**	**0.999**	**0.066**	**0.029**

representations to be learned deeper into the network, as shown by the higher similarities (lighter colors) until layer 6. However, layers 7-8 are still not optimally utilized. By leveraging all three pre-training strategies (i.e., MP+GP+SP), even these last two layers have comparatively higher similarities. Hence, MeSa effectively pre-trains the entire network, including the last few layers, for downstream depth estimation. Appendix D shows more CKA [25] analyses comparing the layer-wise representations within the MeSa pre-trained network (i.e., Pre-trained Layers vs. Pre-trained Layers).

4.5 Impact of Different Pre-training Datasets on Layer-wise Features

In this section, we study the impact of different pre-training datasets on the layer-wise pre-trained features. To investigate this, we compare the representations of two networks pre-trained on LSUN and ImageNet respectively. LSUN is less diverse (featuring mostly indoor scenes) in comparison to ImageNet, allowing us to understand the impact of pre-training on these two datasets of varying diversities. To ensure a fair comparison, we use the same size of pre-training dataset (1200K) in both cases. For ImageNet, we use the pre-trained model provided by Xie et al. [43, 42] whereas for LSUN, we pre-train a Swin-v2-L network using SimMIM [43] on a subset of the LSUN dataset. The subset consists of 300K images each from the bedroom, dining room, living room, and kitchen categories.

Similar to the analysis in Section 4.1, Figure 9 (in the Appendix) compares the representations of various layers within the pre-trained networks, whereas Figure 10 compares the representations of the pre-trained networks and the (depth estimation) fine-tuned networks. The analysis in Figure 9 (top row) reveals that the intermediate layers of the LSUN pre-trained network exhibit lower similarities (darker colors) with layer 0 compared to the similarities of the corresponding layers of the ImageNet pre-trained network. Since earlier layers (e.g., layer 0) capture low-level features, this suggests that pre-training with LSUN learns more high-level representations in the middle layers that are beneficial for tasks requiring higher-level reasoning, as illustrated by the superior depth estimation results on NYUv2. This is further corroborated by Figure 10, which demonstrates that the LSUN pre-trained features are more effective for the downstream task than the ImageNet pre-trained features, as evidenced by the higher similarities (lighter colors) between the pre-trained and fine-tuned features along the diagonal line for LSUN compared to ImageNet. Since the images in LSUN have a smaller diversity compared to those in ImageNet, we hypothesize that pre-training on LSUN forces the network to reconstruct masked portions of similar images, thereby encouraging it to utilize more layers for higher level reasoning than for RGB reconstruction (Figure 9).

4.6 Comparison with SOTA Methods

Table 3 compares MeSa against state-of-the-art methods on the NYUv2 dataset. As in recent works, we primarily focus on RMSE since performance on the other metrics is highly saturated. Our method is most directly comparable to [42], which we surpass by a large margin of 17.1%. Moreover, even without utilizing any recently proposed techniques, such as soft token and mask augmentation [31], metric bins module [2] or pre-trained text-to-image diffusion model [49], our approach also outperforms the most recent methods. Overall, we surpass the current best approach [49] by 6.3%, establishing a new SOTA on the challenging NYUv2 dataset.

4.7 Future Work

In this work, we adopt a sequential pre-training regimen in order to mitigate issues related to training instability [18] and to capitalize on the benefits of leveraging existing pre-trained models. Nevertheless, it would be interesting to explore how concurrent pre-training impacts the learnt representations. Moreover, in addition to masked image modeling (MIM), it would be insightful to study the effect of alternative SSL approaches on shaping the learnt feature space. Lastly, it is intriguing to investigate the representations learnt by recent depth estimation methods such as soft token and mask augmentation [31], metric bins module [2] and pre-trained text-to-image diffusion models [49]. An exploration of how these methods might be effectively integrated with the MeSa framework to benefit from their complementary strengths is a promising avenue for future work.

5 Conclusion

We propose MeSa, a novel pre-training pipeline, to overcome the limitations of the SOTA SSL method, which fails to optimally pre-train the later layers. Via unifying the complementary strengths of masked, geometric, and supervised pre-training, MeSa learns effective layer-wise representations across the entire network. Moreover, MeSa leads to a substantial improvement of 17.1% on the RMSE compared to the SOTA SSL method and establishes a new state-of-the-art for monocular depth estimation on the challenging NYUv2 dataset.

References

[1] S. F. Bhat, I. Alhashim, and P. Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[2] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[3] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid. Auto-rectify network for unsupervised indoor depth estimation. *IEEE PAMI*, 2021.

[4] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid. Unsupervised scale-consistent depth learning from video. *IJCV*, 2021.

[5] Y. Cao, Z. Xie, B. Liu, Y. Lin, Z. Zhang, and H. Hu. Parametric instance classification for unsupervised visual feature learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020.

[7] Y. Chen, C. Schmid, and C. Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, pages 7063–7072, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[10] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

[11] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018.

[12] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*. Springer, 2016.

[13] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[14] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[15] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth prediction. In *ICCV*, 2019.

[16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[17] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020.

[18] Z. Guo, N. U. Islam, M. B. Gotway, and J. Liang. Discriminative, restorative, and adversarial learning: Stepwise incremental pretraining. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 66–76. Springer, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[22] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015.

[23] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019.

[24] T. Koch, L. Liebel, F. Fraundorfer, and M. Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[25] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[26] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[28] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.

[29] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126:942–960, 2018.

[30] B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[31] J. Ning, C. Li, Z. Zhang, Z. Geng, Q. Dai, K. He, and H. Hu. All in tokens: Unifying output space of visual tasks via soft token. *arXiv preprint arXiv:2301.02229*, 2023.

[32] A. Pasad, J.-C. Chou, and K. Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE, 2021.

[33] A. Pasad, B. Shi, and K. Livescu. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[34] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022.

[35] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE PAMI*, 2020.

[36] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[38] L. Sun, J.-W. Bian, H. Zhan, W. Yin, I. Reid, and C. Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *arXiv preprint arXiv:2211.03660*, 2022.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image Quality Assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4), 2004.

[40] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[41] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[42] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022.

[43] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[44] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, pages 204–213, 2021.

[45] Z. Yin and J. Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[46] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[47] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022.

[48] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

[49] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023.

[50] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

A NYUv2 Visualizations

Figure 6 shows more visualizations of the three pre-training strategies on the NYUv2 dataset.

Figure 6: Qualitative comparisons of the three pre-training strategies on NYUv2. Left→right: Image, MP, MP+GP, MP+GP+SP (MeSa). **Top**: predicted depth maps, **bottom**: error maps (blue→lower error; red→higher error).

B   IBims-1 Visualizations

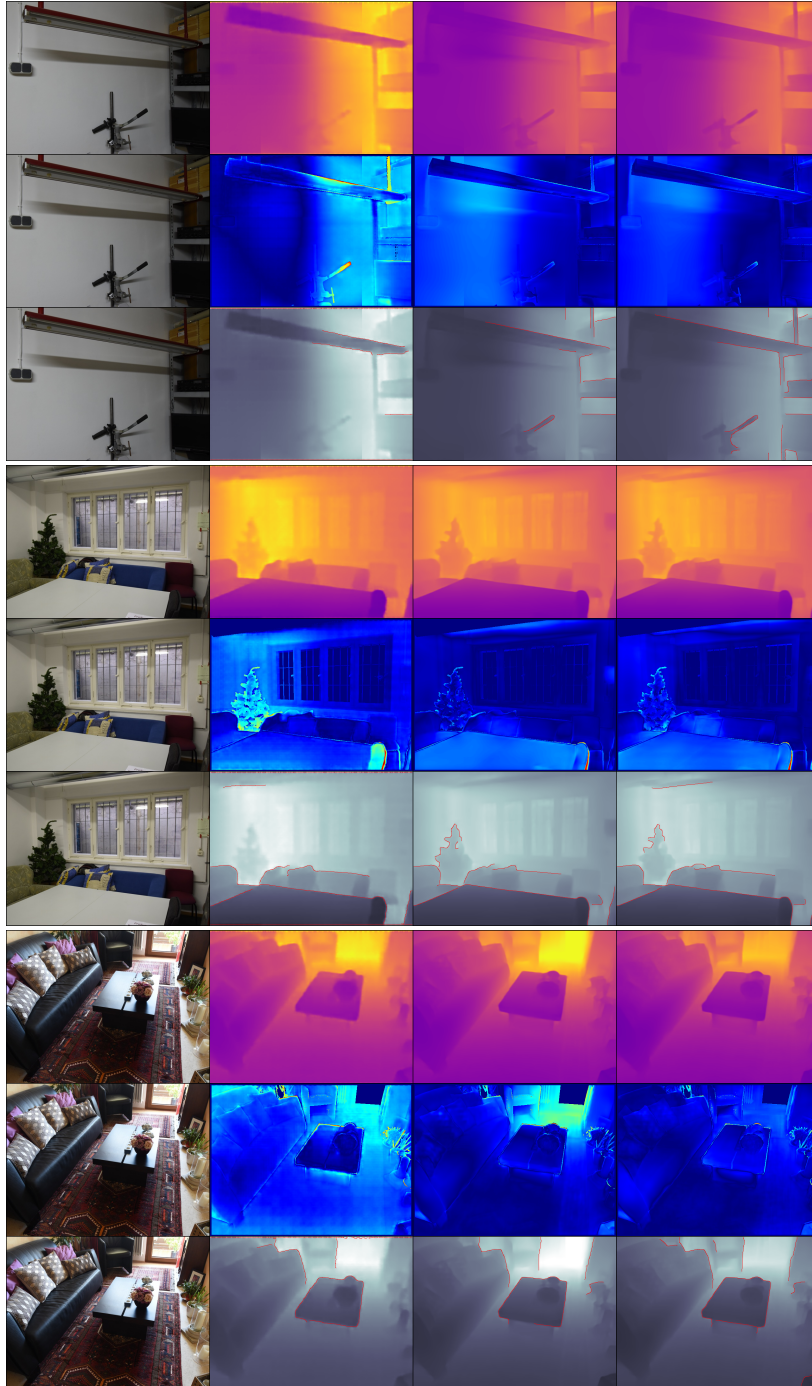Figure 7 shows more visualizations of the three pre-training strategies in the OOD setting on the IBims-1 dataset.

Figure 7: Qualitative comparisons of the OOD performance of the three pre-training strategies on IBims-1. Left→right: Image, MP, MP+GP, MP+GP+SP (MeSa). **Top**: predicted depth maps, **middle**: error maps (blue→lower error; red→higher error), **bottom**: edge maps.

C CKA Layers

Similar to previous studies [42, 30], we adopt the CKA (centered kernel alignment) [25] metric to compare the representations learnt via the various pre-training methods. The CKA similarity values range from 0 to 1, where higher values indicate greater similarity between the representations. In our investigation, we examine the layer-wise representations of nine layers within the Swin-v2-L architecture. For convenience, we refer to them as layers 0-8, with the exact layers detailed in Table 4.

Table 4: Nine Swin-v2-L layers used for CKA analysis.

Identifier	Layer in Swin-v2-L
Layer 0	layers.0.downsample.norm
Layer 1	layers.1.downsample.norm
Layer 2	layers.2.blocks.3.norm1
Layer 3	layers.2.blocks.6.norm1
Layer 4	layers.2.blocks.9.norm1
Layer 5	layers.2.blocks.12.norm1
Layer 6	layers.2.blocks.15.norm1
Layer 7	layers.2.downsample.norm
Layer 8	norm3

D MeSa Layer-wise Analysis

Here, we analyze the layer-wise representations of the MeSa pre-trained network to show that our pre-training strategy effectively pre-trains the entire network, including the later layers. To this end, we conduct a comparison between the layer 0 representation and the representations of layers 1-8, as illustrated in Figure 8 (top row). When utilizing masked pre-training (MP) alone, the later layers (5-8) do not provide significant benefits for downstream depth estimation since they become specialized for the reconstruction task, as evidenced by the increasing similarity values (lighter colors) past layer 4 in the top row. However, incorporating geometric pre-training in addition to masked pre-training (i.e., MP+GP) enables more effective representations to be learned deeper into the network, as indicated by the decreasing similarities (darker colors) up to layer 6. Nevertheless, layers 7-8 are still not fully utilized. By leveraging all three pre-training strategies (i.e., MP+GP+SP), we observe a monotonous decrease in similarities (darkening of colors) up to layer 8, indicating that the entire network learns distinct and high-level representations compared to layer 0. Hence, MeSa effectively pre-trains the entire network, including the later layers.
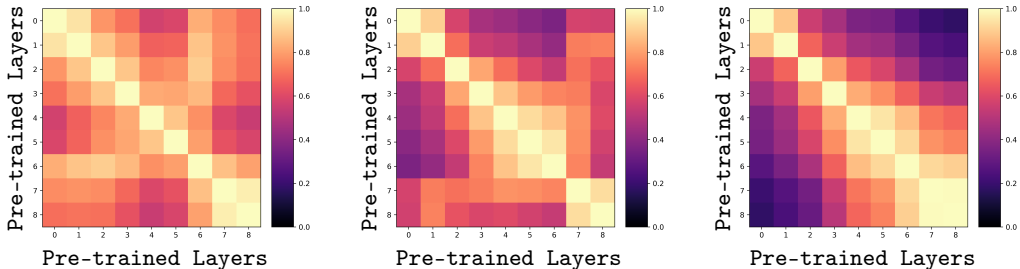

Figure 8: Layer-wise analysis of the three pre-training strategies, comparing the pre-trained representations as we delve deeper into the network (top row). MeSa effectively pre-trains the entire network, including the later layers. Left→right: MP, MP+GP, MP+GP+SP (MeSa).

E   Layer-wise Analysis of Different Pre-training Datasets

In this section, we study the impact of different pre-training datasets on the layer-wise pre-trained features. For detailed configurations of the ImageNet and LSUN pre-trained networks, please refer to Section 4.5 in the main paper.

Similar to the analysis in Section 4.1, Figure 9 compares the representations of various layers within the pre-trained networks, whereas Figure 10 compares the representations of the pre-trained networks and the (depth estimation) fine-tuned networks. The analysis in Figure 9 (top row) reveals that the intermediate layers of the LSUN pre-trained network exhibit lower similarities (darker colors) with layer 0 compared to the similarities of the corresponding layers of the ImageNet pre-trained network. Since earlier layers (e.g., layer 0) capture low-level features, this suggests that pre-training with LSUN learns more high-level representations in the middle layers that are beneficial for tasks requiring higher-level reasoning, as illustrated by the superior depth estimation results on NYUv2. This is further corroborated by Figure 10, which demonstrates that the LSUN pre-trained features are more effective for the downstream task than the ImageNet pre-trained features, as evidenced by the higher similarities (lighter colors) between the pre-trained and fine-tuned features along the diagonal line for LSUN compared to ImageNet. Since the images in LSUN have a smaller diversity compared to those in ImageNet, we hypothesize that pre-training on LSUN forces the network to reconstruct masked portions of similar images, thereby encouraging it to utilize more layers for higher level reasoning than for RGB reconstruction (Figure 9).
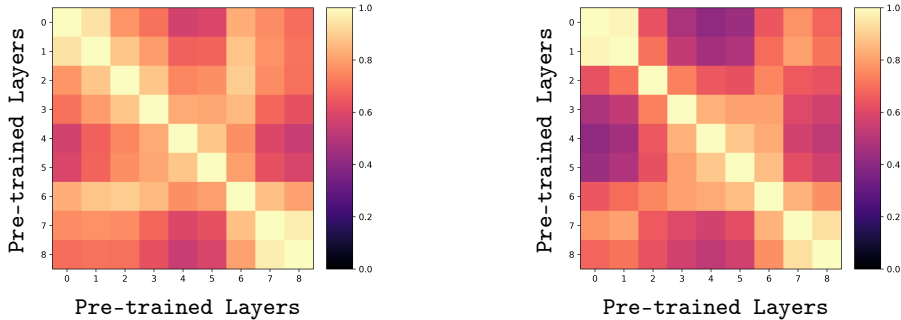

Figure 9: Layer-wise analysis comparing the representations of various layers within the pre-trained network. Left: ImageNet, right: LSUN. Pre-training with LSUN learns more effective higher-level features in the intermediate layers, as indicated by the lower similarities (darker colors) along the top row.
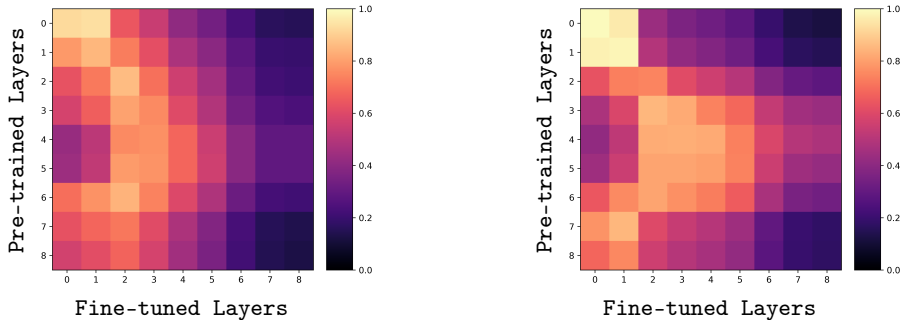

Figure 10: Layer-wise analysis comparing the pre-trained features to the fine-tuned features. Left: ImageNet, right: LSUN. LSUN pre-trained features are more effective for the downstream task, as indicated by the higher similarities (lighter colors) between the pre-trained and fine-tuned features along the diagonal line.