Towards Depth-Guided Self-Supervised World Models

Juan Tarazona Rodríguez Ahmed Elgaradiny Hanqiu Li Cai Sergi Sánchez Orvay
ETH Zürich
{jtarazona, aelgaradiny, hlicai, sanchezs}@student.ethz.ch

1. Introduction

In recent years there have been significant advancements in the field of autonomous driving, led by improvements in perception, planning and control systems. Modern autonomous vehicles rely on complex sensor suites including cameras and LiDAR scanners to interpret their surroundings and make real-time decisions. A current major challenge lies in predicting future scenes, which is crucial to enable safe and reliable navigation.

Recent research trends in autonomous driving explore the usage of world models to perform the prediction of such future scenes [1]. These models take in sensor inputs from a number of frames in the past and infers the scene and trajectory of the ego vehicle for the next frames. Current state-of-the-art (SOTA) world models leverage different representations of the scene to perform this task, leading to the following pipeline categories:

Image-based These models leverage RGB data from cameras. Their main limitation lies in the inability to fully capture the details of the 3D environment.

BEV-based Bird's-Eye view (BEV) converts multi-modal sensor data into a top-down view in 2D. These models suffer from difficulty capturing fine-grained 3D geometries in certain scenarios.

OG-based Occupancy grids (OG) discretize the environment into a voxel grid, encoding high-fidelity 3D spatial information. The main limitation lies in the large memory and computational requirements.

PC-based PC (Point Cloud) based models use 3D scanning data from LiDAR sensors. Despite their precision, these models are computationally intensive due to data sparsity and high input dimensionality.

However, all these approaches treat geometry and appearance separately, potentially limiting their ability to capture the full complexity of driving environments.

2. Problem Statement

The original project proposal aimed to develop a unified multimodal world model modelled by a spatio-temporal transformer or a diffusion model that jointly predicts both point clouds and video frames. However, after the initial literature review phase on the topic, it was concluded that such a model would be infeasible to train given the limited compute available to us and short timeframe of the course. After discussing with the supervisor, it was decided that we would shift our focus towards improving existing self-supervised representation learning methods from RGB images by enhancing them with depth guidance, thus enabling them to encode implicit 3D information in their features.

As a baseline, we use I-JEPA [2], a self-supervised learning method that predicts masked regions in latent space, enabling the model to learn high-level semantics rather than pixel-level details. This has shown significant improvements in downstream tasks that require structural understanding of the scene, such as in autonomous driving. We explore two methods of adding depth guidance to I-JEPA's training: pixel-level guidance with ground truth metric labels, and feature-level guidance via latent representations from a SOTA monocular depth estimation model [3].

3. Related Work

We expand upon our original literature review adapting to the newly formulated task:

Self-supervised learning Self-supervised learning aims to capture meaningful and transferable representations from unlabeled visual data by designing pretext tasks that capture semantic or structural information. Recent advances in this area have leveraged ViTs [4] to achieve strong performance on downstream tasks.

SimCLR [5] introduced contrastive learning with strong data augmentations, training models to bring different views of the same image closer in representation space. DINO [6] proposed a self-distillation approach where a student network learns from a momentum-updated teacher, producing semantically aligned features without supervision. MAE [7] demonstrated the effectiveness of masked image modeling,

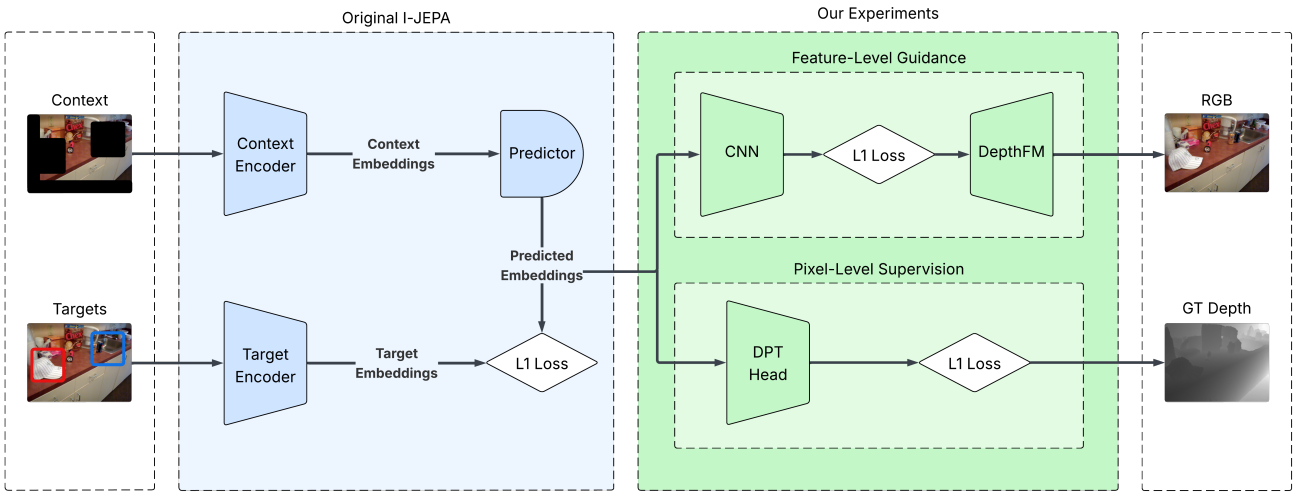

Figure 1. Overview of our architecture. The I-JEPA [2] training pipeline (blue) learns to predict target patch representations from partial context. We introduce two auxiliary depth supervision paths (green): pixel-level guidance from ground-truth depth and feature-level alignment using a pretrained DepthFM [3] model.

using a lightweight encoder to process visible patches and a decoder to reconstruct the masked ones. DINOv2 [8] scaled the DINO framework with improved architectures and training recipes, resulting in more robust and transferable visual features.

While these methods focus on visual similarity and low-level reconstruction, they often lack high-level structural understanding. I-JEPA [2], which forms the basis of our work, addresses this by predicting abstract representations of masked regions using context tokens—focusing on semantics over pixels and enabling richer spatial reasoning.

Monocular depth prediction Monocular depth estimation involves predicting the 3D structure of a scene from a single RGB image. This task is particularly relevant to us as it enables self-supervised learning from image-only data without relying on ground-truth depth labels. Modern methods can be largely classified between discriminative methods and generative methods. Discriminative methods directly predict depth from images using neural networks, often trained with supervised or self-supervised loss functions. Notable works include MiDaS [9], Depth Anything [10, 11] and Metric3D [12]. Generative methods instead leverage the rich knowledge embedded in large-scale vision foundation models to produce high-quality depth maps directly. Recent works such as Marigold [13] and GeoWizard [14] use Latent Diffusion Models (LDMs) to create high-fidelity, dense metric predictions. However, such methods suffer from slow sampling times even when using ODEs to approximate the underlying SDEs. Furthermore, the diffusion distribution relies on a Gaussian source distribution, which might not capture the natural relationships between images and their depth maps. Flow matching-based models [15] demonstrated faster sampling time and showed promise across multiple tasks. DepthFM [3] is the first work that leverages flow matching for depth prediction and achieves SOTA results in the task.

4. Methodology

The aim of the project is to inject geometric information into a self-supervised visual representation learning pipeline by integrating depth signals into the I-JEPA framework. Our goal is to guide the model towards learning features that not only capture high-level semantic content but also encode implicit 3D structure, which is essential for downstream tasks in autonomous driving and scene understanding.

We achieve this by extending the original I-JEPA model with depth supervision during training, leveraging both ground-truth depth maps and learned depth representations from a pretrained model. The core idea is to preserve I-JEPA's semantics-driven learning objective while introducing auxiliary depth-based losses to promote geometric awareness.

4.1. Architecture

The overall architecture of our approach is illustrated in Figure 1. It integrates the original I-JEPA pipeline [2] (highlighted in blue) with our proposed depth supervision module (shown in green). While I-JEPA focuses on learning semantic representations from visual context, our additions encourage the model to capture 3D geometric structure by

introducing depth guidance during training.

The original I-JEPA framework consists of three Vision Transformers (ViTs) [2, 4]: a context encoder, a target encoder and a predictor. The training objective is to predict latent representations of target image regions from partial context, entirely in feature space rather than at the pixel level. To enrich these representations with geometric awareness, we introduce a depth supervision module (green in Figure 1). We explore two complementary strategies:

- **Feature-level guidance:** A pretrained DepthFM [3] model provides depth-aware latent features from the input image, which are aligned with I-JEPA's predicted target embeddings to inject geometric structure into the learned features.

- **Pixel-level supervision:** A depth regression head is added to the target encoder, predicting metric depth maps that are compared to the ground truth depth map during training, guiding the model to internalize 3D spatial cues.

These depth-guided signals are used only during training. At inference time, only the context encoder is used and operates over the entire image rather than just a single block.

4.2. Model Details and Training Procedure

In this section, the details of the models we use, as well as the different training procedures of our depth-guided I-JEPA model are described. We first present the original I-JEPA model details, hyperparametrs and masking strategy [2], which serves as the foundation for both of our proposed extensions: feature-level guidance and pixel-level supervision. The datasets and final setup are further described in section 5.

4.2.1 I-JEPA Setup

Our I-JEPA implementation strictly follows the configuration and hyperparameters provided in the official public codebase I-JEPA [2], except for a change in masking strategy as one of the experiments during ground truth depth supervision. We use a ViT-H/14 [4] architecture for the context encoder, target encoder, and predictor, where each 14×14 image patch is embedded into a corresponding latent feature vector. At each training iteration, a single context block is sampled from the input image. This block has a unit aspect ratio and a scale uniformly drawn from the range (0.85, 1.0) of the full image size. Separately, four possibly overlapping target blocks are randomly selected, each with an aspect ratio in the range (0.75, 1.5) and scale in the range (0.15, 0.2) of the full image size. To ensure a non-trivial prediction task, any overlapping regions between the context and target blocks are removed from the context input before encoding.

The masked context block is passed through the context encoder to obtain a latent representation. For each of the four target blocks, the predictor takes this context representation along with a set of learned mask tokens (one per target patch) and predicts the corresponding latent features. Meanwhile, the target encoder, identical in architecture to the context encoder but updated via an exponential moving average (EMA) of its weights, processes the full unmasked image to compute the ground-truth latent features for the target blocks. The training objective is to minimize a smooth L1 loss between the predicted and ground-truth latent features for each patch in the target blocks, averaged over all targets. This encourages to learn abstract and spatially coherent representations by predicting the content of unseen regions from partial context.

This exact I-JEPA training configuration was used as the backbone for both of our depth supervision strategies. In the following subsections, we detail how we incorporate feature-level and pixel-level depth signals into this setup to enrich the learned representations with 3D structure.

4.2.2 Pixel-Level Guidance

One of the most straightforward approaches to incorporate depth information into the training process is through supervision using ground truth depth maps. In this framework, a dedicated depth estimation head is applied to the predicted target embeddings produced by the original I-JEPA loop, yielding a pixel-wise depth prediction. The predicted depth map is then compared against the ground truth via an L1 loss function. This is illustrated in the bottom depth supervision path in Figure 1. We investigate four different architectural designs for the depth decoding head under this supervision scheme:

- **MLP Head**: A simple multilayer perceptron consisting of two linear layers, projecting the embedding dimension to a depth prediction for each pixel within a patch.
- **CNN Head**: A convolutional decoder that first aggregates information along the embedding dimension, projects the aggregated features spatially, and subsequently applies two convolutional layers to produce the final depth estimate.
- **Deep CNN**: A deeper convolutional decoder comprising five convolutional layers. This model transforms the input embeddings into a single-channel depth map while progressively upsampling the spatial dimensions to match the desired output resolution.
- **DPT Head**: Based on the monocular depth estimation module proposed in the DINOv2 framework [8], this approach extracts intermediate features at multiple levels of the vision transformer (ViT) encoder. Notably, DINOv2 leverages a hierarchical ViT architecture [8], which pro-

duces features at varying resolutions across layers. In contrast, I-JEPA employs a flat ViT encoder [2], where features at different layers maintain the same spatial resolution but may encode varying levels of semantic information.

Depending on the masking strategy, we consider two variants for the supervision signal: predicting depth only over the target-patch regions, or generating a complete depth map over the entire image. The former approach requires no architectural modifications to I-JEPA, as predictions are made directly from the available target embeddings. However, producing a full-image depth map necessitates access to complete image context, rendering the task ill-posed when relying solely on target patches.

To address this, we modify the I-JEPA masking strategy to resemble that of masked autoencoders [7], where a significantly larger portion of the image is masked. In our implementation, we mask most of the image (80%, like MAE[7]), while selecting a single target as opposed to 4 like original I-JEPA; all remaining patches are treated as context. The loss is computed in the same manner as in the original I-JEPA framework, but applied solely to the single target. We then reconstruct a full image embedding via position-aware concatenation: the context embeddings are inserted at their original patch indices alongside the target embedding. This complete embedding is then fed into the depth prediction head.

For both supervision schemes (target-only or full-image prediction) and all depth heads, we conduct experiments under two training paradigms: (1) fine-tuning a pre-trained I-JEPA model and (2) training both the base model and the depth head from scratch. This dual approach allows us to evaluate whether depth supervision can meaningfully influence already strong pretrained representations or whether learning depth guidance must be incorporated from the outset. The detailed training procedures for both cases are described in Section 4.2.3 on DepthFM guidance.

4.2.3 Feature-Level Guidance

To inject 3D awareness into the representations learned by I-JEPA, a feature-level guidance strategy is introduced using DepthFM [3], a state-of-the-art monocular depth estimation model. DepthFM operates in latent space using a flow-matching framework [3], predicting depth not as explicit metric maps but as dense feature embeddings. These embeddings capture geometric structure implicitly and are therefore well suited for use as supervisory signals in representation learning. In this setup, the latent features predicted by DepthFM serve as reference targets for a convolutional projection head built on top of the I-JEPA architecture. The goal is to align the internal representations of I-JEPA with the geometry-aware features produced by DepthFM.

During I-JEPA training, representations are only computed for a limited subset of image patches—those corresponding to the context block and the predicted target blocks. To construct a dense feature map compatible with DepthFM supervision, a feature reconstruction step is performed. The embeddings from the context block and all target blocks are first concatenated. Since target blocks can overlap, embeddings at overlapping spatial positions are averaged. For all patches not present in the union of context and target blocks, zero-padding is applied. This results in a full spatial grid of patch embeddings with fixed dimensions. The reconstructed representation is subsequently passed through a small convolutional neural network (CNN) that projects I-JEPA features to the dimensionality of DepthFM features. A smooth L1 loss is then computed between the CNN output and the DepthFM latent features. This loss is propagated through both the CNN and the I-JEPA components, including the context encoder and the predictor. The loss is added to the original I-JEPA prediction loss [2], enabling the model to jointly optimize for semantic and geometric understanding.

Two training strategies are explored for this guidance setup:

Pre-trained checkpoint The I-JEPA context encoder, target encoder, and predictor are initialized from a publicly released pre-trained model [2] trained for 300 epochs on ImageNet-1k [16] (1000 classes, approximately 1300 images per class), using 16 NVIDIA A100 GPUs. During the initial phase of training, the I-JEPA modules are kept frozen, and only the CNN projection head is optimized. This enables the CNN to adapt to the pre-learned representations without disrupting the feature space learned during large-scale pre-training. After a warm-up period, all components are jointly fine-tuned. This strategy benefits from stronger initial representations, which can accelerate convergence and improve stability during early training. However, it may constrain the model's flexibility to adapt to the specific depth supervision signal due to the inherited inductive biases from pretraining on diverging objectives.

Training from scratch In this variant, the I-JEPA architecture and CNN projection head are trained jointly from random initialization. This approach allows the model to learn representations fully aligned with the depth-based supervision from the beginning, potentially leading to better integration of geometric cues. However, training from scratch may result in slower convergence and increased instability, especially when operating on smaller datasets with more limited semantic diversity.

In both settings, DepthFM is kept frozen and only used at inference time to provide target feature maps. This feature-

level guidance framework facilitates the incorporation of 3D structural cues into the self-supervised learning process of I-JEPA, enriching the learned representations without requiring access to explicit depth labels or metric reconstructions.

5. Results and Discussion

This section examines pixel-level guidance and feature-level depth supervision, evaluates their feasibility and robustness, and then presents the results of the selected method.

5.1. Pixel-Level Guidance

To evaluate the effectiveness of ground truth depth map supervision, and to compare training from scratch versus fine-tuning a pre-trained checkpoint, we conduct experiments on the densely labeled NYUv2 dataset [17], which contains 1449 annotated RGB-depth image pairs. All models described in Section 4.2.2 are trained for 100 epochs under identical hyperparameter settings.

In order to supervise I-JEPA with ground truth depth maps, it is necessary to decode a dense pixel-level depth prediction from the target patch embeddings. This prediction is then compared to the ground truth using an L1 loss to provide a training signal. However, across all experimental configurations—whether using a pre-trained I-JEPA checkpoint or training the entire model from scratch—we consistently found that it was not possible to predict coherent dense depth maps from the patch-level embeddings. Pre-training the decoder on the frozen I-JEPA checkpoint did not affect this outcome. This failure was observed across all decoder head architectures, resulting in random maps or structurally inconsistent predictions that failed to provide a meaningful supervisory signal. Consequently, the image features cannot incorporate depth information during training.

We hypothesize that these challenges arise due to the nature of the I-JEPA target embeddings, which are optimized to encode high-level semantic content rather than geometric structure. While I-JEPA representations have demonstrated strong performance on downstream tasks such as classification [2], they appear to be ill-suited for tasks that require fine-grained spatial or geometric reasoning, such as pixel-wise depth estimation.

5.2. Feature-Level Guidance

To evaluate the relative benefits of initializing from a pre-trained I-JEPA checkpoint versus training from scratch, an experiment on a small subset of ImageNet-100 (100 classes, 150 images per class) was conducted. Both variants were trained for 10 epochs with identical hyperparameters. For the pre-trained model, a 1000-step warm-up was applied during which only the CNN projection head was unfrozen; thereafter, the projection head and all ViT components (context encoder, target encoder, and predictor) were fine-tuned.

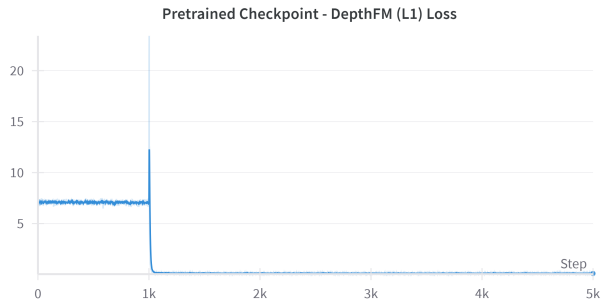

Figure 2. Depth projection head loss over the first 5000 steps of training from the Pre-trained Checkpoint.

Figure 2 illustrates the instant drop in the pre-trained model's depth-feature loss immediately after the 1000-step warm-up, settling around 0.1 within just a few steps in the first epoch. In contrast, the model trained from scratch reaches a similar loss level towards the end of training. At first glance, this rapid convergence suggests that the pre-trained model, benefiting from its prior semantic training, already encodes representations that closely align with depth features and can quickly "learn" depth.

Figure 3 presents the qualitative inference assessment of the predicted depth-feature maps from both variants against the DepthFM ground truth. Neither model perfectly reconstructs the true depth features, which is unsurprising given the limited number of training epochs. However, the checkpoint-initialized model produces a blurred, structureless map that bears little resemblance to the ground truth. On the other hand, the model trained from scratch, while not exact, retains an outline of the bird at the center of the image, demonstrating a clearer grasp of the underlying 3D shape. These results indicate that, checkpoint fine-tuning demonstrated shortcut learning, where the checkpoint model relies on semantic priors to lower the depth feature loss without learning actual 3D structure, while training from scratch compels the network to integrate semantic and geometric cues into its representations.

5.3. Final Training Results

Consequently, the final model selected was the feature-level guided I-JEPA trained from scratch. Due to limited computational resources, it was not feasible to train on the full dataset or match the training duration of the original I-JEPA [2] checkpoint to use it as a baseline. Therefore, both guided and unguided I-JEPA models were trained from scratch on 4 NVIDIA A100 GPUs for 120 epochs on a reduced version of the ImageNet-100 dataset, consisting of 50 classes with approximately 1300 images per class.

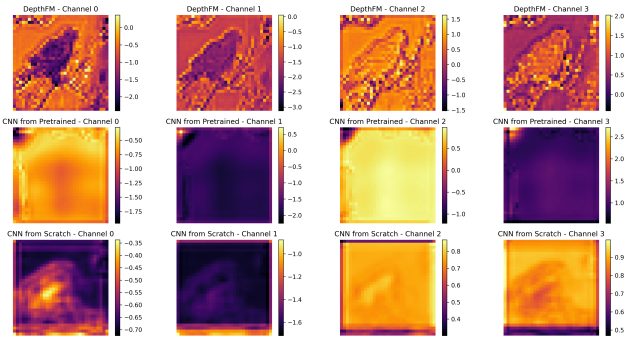

Figure 3. Depth latent feature maps: (top) original DepthFM [3], (middle) our model fine-tuned from a pre-trained checkpoint, and (bottom) our model trained from scratch.

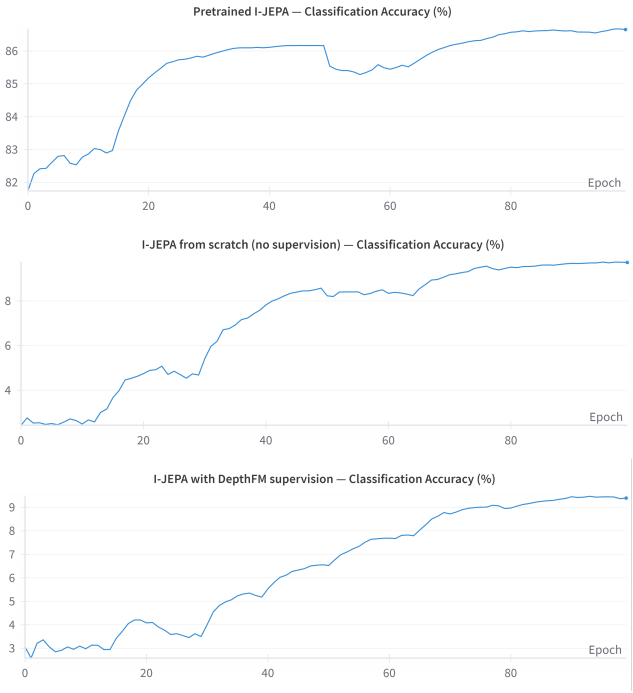

Figure 4. Linear classification accuracy over 100 epochs for three models: (top) the original I-JEPA checkpoint pretrained on ImageNet-1K, (middle) I-JEPA trained from scratch without supervision, and (bottom) I-JEPA trained from scratch with feature-level guidance using DepthFM.

To assess the impact of geometric distillation on representation quality, each of the two encoders was frozen following the training phase, and a lightweight MLP classification head was trained on top, following the linear evaluation protocol from the original I-JEPA [2] work. This evaluation used only the training dataset to ensure that improvements due to enhanced representational fidelity are not hindered by generalization, given the training dataset's limited scale.

Figure 4 presents the linear classification results for three models: the original I-JEPA [2] checkpoint, I-JEPA trained from scratch without depth supervision, and I-JEPA trained from scratch with feature-level depth guidance. A substantial drop in classification accuracy, from approximately 86% to below 10%, is observed when comparing the pretrained checkpoint to both models trained from scratch. This pronounced difference makes it difficult to draw reliable conclusions regarding the effect of depth supervision, as both scratch-trained models perform similarly poorly. The results highlight a key insight: the strong representational capacity of the original I-JEPA encoder stems primarily from training on a large and diverse dataset. Without such training, even models employing the same architecture and training procedure fail to develop meaningful representations.

6. Conclusion

This work investigated the integration of depth-based supervision into the I-JEPA [2] encoder using both pixel-level and feature-level guidance to enhance representational understanding.

Pixel-level depth supervision was infeasible, as the patch-wise semantic representations produced by I-JEPA lacked the spatial resolution and geometric information necessary for predicting dense, pixel-level depth maps. On the other hand, while feature-level guidance from a pre-trained checkpoint led to shortcut learning where the model minimized loss without acquiring true geometric understanding, training from scratch with depth guidance demonstrated potential for learning meaningful 3D structure.

Linear probing results further revealed that I-JEPA-like architectures rely heavily on large-scale and diverse datasets to develop strong representations. Models trained from scratch on the reduced ImageNet-100 subset, even without depth supervision (replicating I-JEPA), showed significantly lower performance compared to the original pretrained checkpoint. This constraint, driven by limited computational resources, hindered a conclusive evaluation of the effectiveness of depth-guided training.

To fully assess the value of geometric supervision in representation learning, future work must scale training to large datasets like the full ImageNet-1K dataset [16] and mimicking the original I-JEPA [2] training schedule. Only under such conditions can the potential of multimodal, geometry-aware pretraining be accurately measured.

References

[1] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving, 2025. 1

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15619–15629, 2023. 1, 2, 3, 4, 5, 6

[3] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 1, 2, 3, 4, 6

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

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

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

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 4

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2

[12] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 2

[13] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2

[14] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 2

[15] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4, 6

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