Enhancing Monocular Depth Estimation with Multi-Source Auxiliary Tasks

Alessio Quercia [1,4] Erenus Yildiz [1] Zhuo Cao [1] Kai Krajsek [3]
Abigail Morrison [2,4] Ira Assent [1,5] Hanno Scharr [1]
[1] IAS-8 [2] IAS-6 [3] JSC, Forschungszentrum Juelich, Juelich, Germany
[4] Dept. of Computer Science, RWTH Aachen University, Aachen, Germany
[5] Dept. of Computer Science, Aarhus University, Aarhus, Denmark
{a.quercia,e.yildiz,z.cao,k.krajsek,a.morrison,i.assent,h.scharr}@fz-juelich.de

Abstract

Monocular depth estimation (MDE) is a challenging task in computer vision, often hindered by the cost and scarcity of high-quality labeled datasets. We tackle this challenge using auxiliary datasets from related vision tasks for an alternating training scheme with a shared decoder built on top of a pre-trained vision foundation model, while giving a higher weight to MDE. Through extensive experiments we demonstrate the benefits of incorporating various in-domain auxiliary datasets and tasks to improve MDE quality on average by $\sim 11\%$*. Our experimental analysis shows that auxiliary tasks have different impacts, confirming the importance of task selection, highlighting that quality gains are not achieved by merely adding data. Remarkably, our study reveals that using semantic segmentation datasets as Multi-Label Dense Classification (MLDC) often results in additional quality gains. Lastly, our method significantly improves the data efficiency for the considered MDE datasets, enhancing their quality while reducing their size by at least 80%. This paves the way for using auxiliary data from related tasks to improve MDE quality despite limited availability of high-quality labeled data. Code is available at* https://jugit.fz-juelich.de/ias-8/mdeaux.

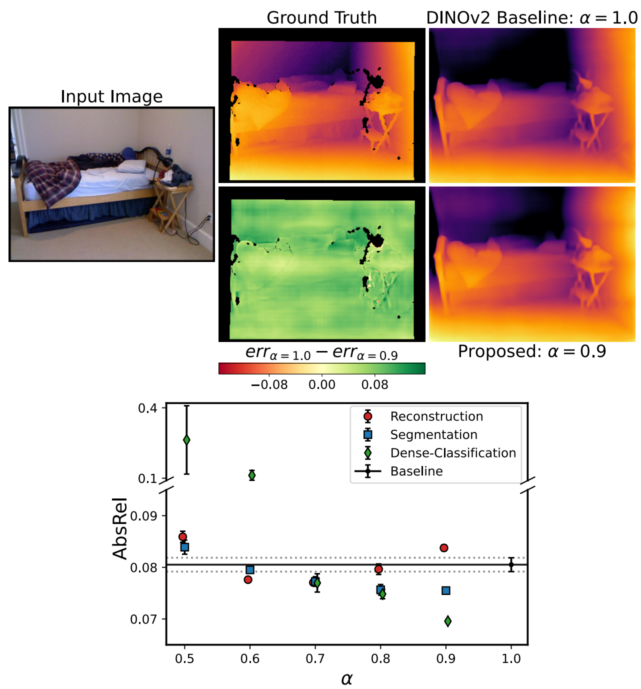

Figure 1. (Top) NYUv2 results with MIX6 auxiliary MLDC. From left to right: input, ground truth, error difference (w.r.t. ground truth) between the DINOv2 baseline and ours. Green indicates ours is better, while red vice versa. (Bottom) AbsRel (↓) for varying values of the task focusing parameter α with multiple tasks (markers). The solid and dashed lines represent the mean and standard error of DINOv2, respectively.

1. Introduction

Monocular depth estimation (MDE) is a well-established task in computer vision, with possible applications ranging from autonomous vehicles to augmented reality. It is inherently data-hungry, necessitating extensive, high-quality labeled datasets for effective training. Yet, the procurement of such datasets poses a significant challenge, often being a costly and time-consuming endeavor.

Recent advances in monocular depth estimation have been driven by suitable adoption of Vision Transformers (ViT) [18, 63]. In particular DPT [51] is the first attempt to relative MDE using ViT. Recent architectures, like MiDaS 3.1 [6] and ZoeDepth [4] extend DPT to multiple ViT backbones and metric monocular depth prediction, respectively. Furthermore, DINOv2 [46] is a vision foundation model with excellent zero-shot MDE capabilities when combined with a DPT decoder, and Depth Anything [69] fine-tunes this into an MDE foundation model using a large composition of labeled MDE and unlabeled vision data.

Vision transformers have been primarily employed for individual tasks or in self-supervised learning paradigms. Recently, their application in Multi-Task Learning (MTL) is a growing area of research. In particular, there is an increasing interest in using a single model to predict multiple dense tasks, often employing MTL approaches that propose either a unified architecture [5, 24, 27, 45, 48, 70] or methods to balance the loss functions during the training [3, 12, 23, 25, 39, 56, 67]. While MTL has been applied to MDE in prior studies [15, 29, 32, 60], we propose Multi-Label Dense Classification (MLDC) as an auxiliary task to improve MDE.

Motivated by the limited availability of high-quality labeled MDE data, we propose a data-efficient and resource-considerate approach that leverages related vision datasets without requiring extensive fine-tuning of a pre-trained foundation model. This strategy improves MDE performance while reducing the reliance on MDE labels.

Our approach leverages a frozen DINOv2 ViT Giant model [46] as a feature extractor, bypassing the need for fine-tuning. We jointly train a shared DPT decoder [51] with auxiliary datasets from related tasks to improve MDE. We illustrate the qualitative and quantitative improvements of our method over the DINOv2 baseline in Figure 1.

Our key contributions are summarized as follows:

- We propose an alternating training scheme leveraging auxiliary non-MDE datasets from related vision tasks to boost the MDE downstream task. This improves the MDE performance by weighting MDE steps more than auxiliary ones through our *task-focusing* parameter α;
- To the best of our knowledge, we are the first to use Multi-Label Dense Classification (MLDC) as an auxiliary task for MDE.
- We find that MLDC frequently outperforms semantic segmentation as auxiliary task, suggesting that classification aspects may be more beneficial than spatial details, especially with low-quality segmentation labels.
- We thoroughly test our method across various MDE datasets, using multiple auxiliary datasets and tasks. Our results show an average quality gain of 11% compared to the DINOv2 baseline on in-domain datasets, confirming the robustness of our method;
- We show that our method enhances the data efficiency of DINOv2, allowing for a reduction in MDE training data of 80-99%, while still improving over DINOv2;

These contributions represent a novel advancement in the field, both algorithmically and scientifically.

Notably, compared to the recent Depth Anything [69], which reports no improvements when jointly training their model (using DINOv2 ViT-L encoder) together with semantic segmentation data using task-specific DPT decoders, our method successfully leverages auxiliary tasks to enhance MDE. In particular, our approach keeps the larger DINOv2 ViT-G model frozen and jointly trains a single shared DPT decoder with an auxiliary task to improve MDE. In addition, our method has a reduced training cost by design as it does not fine-tune the backbone, while being bounded by its quality. In this sense, we see Depth Anything and methods that fine-tune the backbone as orthogonal to our work.

2. Related Work

Monocular Depth Estimation. Recent works on Monocular Depth Estimation [4, 30, 31, 35–37, 43, 46, 50, 52, 57, 65, 68] primarily use Vision Transformers [18, 63]. In particular DPT [51], also known as MiDaS 3.0, is the first attempt to relative monocular depth estimation using Vision Transformers, adapting the original MiDaS CNN-architecture [52]. DPT has been used as a baseline for recent MDE SOTA architectures, for instance, MiDaS 3.1 [6] which shows performance of multiple ViT backbones in the MiDaS model, and ZoeDepth [4], which combines DPT for relative depth estimation with a new module for metric depth prediction. Furthermore, DINOv2 [46] proposes a general-purpose vision foundation model that can be used together with DPT decoder to build a powerful zero-shot MDE predictor. Lastly, the recent Depth Anything [69] builds on top of DINOv2 and trains using a large combination of labeled MDE and unlabeled vision datasets. These works are task-specific and therefore miss the potential synergies between different vision tasks. In this paper, we show that the representations extracted by DINOv2 can be leveraged as a strong foundation for multi-task dense prediction, without further fine-tuning DINOv2.

Multi-Task Dense Prediction. Multi-Task Learning (MTL) [8,54,61,72] leverages the idea that related tasks can provide complementary insights and improve the overall model representations during training, without the need for training multiple separate models. Multi-task dense prediction pertains to the domain of MTL concentrated on dense vision tasks. Through this paper, the abbreviation MTL is employed to denote methods related to Multi-Task Dense Prediction. Research in MTL can be broadly classified into two primary categories. The first [5, 27, 33, 45, 67, 70] focuses on refining network architectures to optimize information sharing across tasks and to enhance task-specific representations. For instance, PolyMax [70] handles diverse prediction tasks effectively, showcasing adaptability and spatial data handling. Painter [67] leverages in-context learning with an innovative image-centric method. The second category [3, 12, 23–25, 39, 56] involves task-balancing optimization, where tasks synergistically learn a single model for solving multiple tasks. Sener et al.'s framework [56] fine-tunes task equilibrium, promoting synchronized learning. Kendall et al. [25] address task uncertainty via dynamic weighting mechanism based on confidence levels. Chen et al. [12] explore gradient contributions for adaptive loss balancing. Liu et al. [39] advocate an end-to-end approach, emphasizing integrated learning workflows. Guo et al. [23] propose dynamic task prioritization.

The methods discussed so far aim to learn multiple tasks simultaneously, whereas we propose an auxiliary learning approach [11, 17, 38, 53] where auxiliary tasks are used only in the learning phase to improve MDE performance and are discarded afterwards. This approach allows us to show the positive impact of auxiliary tasks on improving MDE without direct reliance on MTL comparisons.

Cross-Task Relations. The investigation into cross-task relations [1, 9, 19, 21, 47, 55, 59, 64, 66, 71] goes beyond understanding individual tasks in isolation, aiming to uncover how tasks mutually inform, and enhance each others. Zamir et al.'s Taskonomy [71], provides a systematic framework that maps relationships between different vision tasks. Fifty et al. introduced Task Affinity Grouping [21], which clusters related tasks to maximize learning efficiency and performance, providing a pragmatic approach to task integration. Wang et al. [66] emphasize how cross-task relations can be domain-specific and used for improved model performance. Saha et al. [55] present methodologies for models to effectively learn from multiple task interactions.

In contrast to these works, our focus is not on discovering general cross-task relations, but to leverage data from specifically chosen vision tasks, known to be complementary to MDE, to improve the MDE quality.

3. Proposed Method

Monocular Depth Estimation (MDE) is frequently impeded by limited availability of high-quality labeled datasets. Images in publicly available datasets often contain many invalid regions (represented in black in images throughout the paper), where the correct label is undefined or out of range. Additionally, datasets may come with their own pre-processing procedure, where, depending on the context, the distance is limited to specific ranges. MDE methods address this issue by mixing MDE datasets and scaling their size, while MTL methods attempt to train a single model to predict multiple tasks, improving on some or all of them. These methods may suffer from the "task interference" problem, where gradients from different tasks conflict and hinder each other's progress [8].

Motivated by these limitations, we propose a method that bridges the gap between conventional MDE techniques and the prevalent methodologies within the MTL domain. Differently from these works, we use datasets from related vision tasks to boost the performance of a frozen pre-trained foundation model on the MDE downstream task, disregarding the auxiliary outcomes. We select DINOv2 ViT-Giant [46] as backbone due our requirement for a high-quality and robust feature extractor. We jointly train a shared DPT [51] decoder by alternating MDE steps and auxiliary task steps, while prioritizing MDE. Differently from Depth Anything [69], we share a single DPT decoder, and only separate the smaller task-specific heads. This allows us to have an additional shared model component and to freeze the larger pre-trained backbone, while still benefiting from joint training with auxiliary datasets and tasks. Our training method is summarized in Figure 2.

Formally, let $D = (x_i, y_i)_{i=1}^N$ be an MDE training dataset with N samples, $A = (x_k, y_k)_{k=1}^M$ be an auxiliary training dataset from a related task with M samples, $V = (x_l, y_l)_{l=1}^L$ an MDE test dataset with L samples, where $x_.$ are input images and $y_.$ the respective task ground truths. Further, let f be a frozen pre-trained foundation model, g_θ a decoder with shared parameters θ and h_ϕ and h_ψ respectively the depth and auxiliary task heads with parameters ϕ and ψ, our goal is to improve the decoder quality on the main downstream task (MDE) by jointly training it with external auxiliary datasets from a related vision task. More specifically, let $d_.(x) = h_\phi(g_\theta(f(x))$ be the depth predictor with parameter states θ and ϕ, let $\mathcal{M}(V, d_.)$ be a set of MDE validation metrics to minimize on all test samples $(x_l, y_l) \in V$ using the depth predictor $d_.$. Then, let

$$m_b = \min \mathcal{M}(V, d_b) \tag{1}$$

be the smallest test metrics achieved by the DINOv2 baseline d_b, obtained with the following minimization problem:

$$d_b = \underset{\theta,\phi}{\arg\min}\, \mathcal{L}_D(\hat{y}_i, y_i) \tag{2}$$

where $\hat{y}_i = d_.(x_i)$ is a prediction on a depth training sample $(x_i, y_i) \in D$ with parameters θ and ϕ updated during training, and $\mathcal{L}_D$ is an MDE training loss.

Let $a_j = h_\psi(g_\theta(f(x))$ be the predictor on an auxiliary related vision task with parameter states θ and ψ, we look for model d_j that best performs on the MDE test dataset V with a set of metrics $\mathcal{M}$, while jointly trained on an auxiliary training dataset A from a related task

$$d_j = \underset{\theta,\phi,\psi}{\arg\min}(\mathcal{L}_D(\hat{y}_i, y_i), \mathcal{L}_A(\hat{y}_k, y_k)) \tag{3}$$
$$\text{s.t.} \;\; \mathcal{M}(V, d_j) \leq m_b$$

where $\hat{y}_k = a_j(x_k)$ is a prediction on an auxiliary training sample $(x_k, y_k) \in A$ with parameters θ and ψ updated during training, and $\mathcal{L}_A$ is an auxiliary task training loss.

In order to train (3), we define a joint global gradient step as two consecutive task-specific gradient steps. In particular, we first apply a depth gradient step

$$\theta_{t_D} = \theta_{t-1} - \alpha\eta_{g_\theta}\frac{1}{B_D}\sum_{i=1}^{B_D}\nabla_{\theta_{t-1},\phi_{t-1}}\mathcal{L}_D(\hat{y}_i, y_i) \tag{4}$$

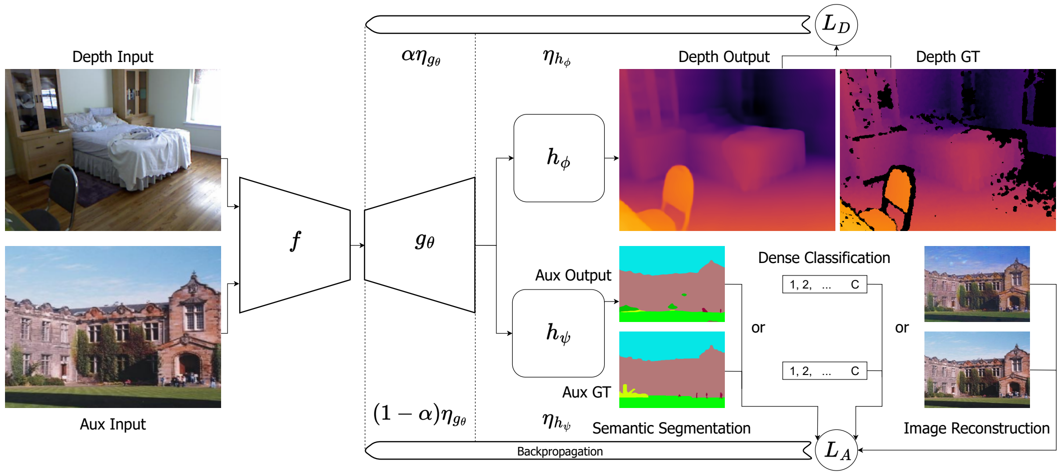

Figure 2. Overview of the proposed training pipeline. We use a frozen pre-trained DINOv2 ViT-G backbone (f) as a feature extractor and jointly train a DPT decoder (g_θ) with auxiliary datasets from related vision tasks (semantic segmentation, dense classification, or image reconstruction). We train 2 task-specific heads: MDE (h_ϕ) and auxiliary (h_ψ). We alternate MDE steps and auxiliary steps. During backpropagation, each head has its own learning rate (η_{h_ϕ} and η_{h_ψ}), while the decoder shares a common learning rate η_{g_θ}, scaled by α for MDE and $1-\alpha$ for the auxiliary task.

and then an auxiliary task gradient step

$$\theta_t = \theta_{t_D} - (1-\alpha)\eta_{g_\theta}\frac{1}{B_A}\sum_{k=1}^{B_A}\nabla_{\theta_{t_D},\psi_{t-1}}\mathcal{L}_A(\hat{y}_k, y_k) \quad (5)$$

where $t-1$, t_D and t respectively represent the latest global training step, the intermediate MDE step and the updated global step, B_D and B_A are respectively a batch of depth samples $(x_i, y_i) \in D$ and auxiliary task samples $(x_k, y_k) \in A$, η_{g_θ} is the overall decoder learning rate for a full gradient step, weighted by the *task-focusing* parameter α, deciding how much of the overall decoder gradient step focuses on the main and auxiliary tasks. This ensures that the total learning rate for the joint scheme is comparable to that of the baseline, which we tested to be optimal (see Section 5.1). When $\alpha = 1$, our training method is reduced to baseline MDE, and when $\alpha = 0$, it becomes solely an auxiliary task. Lastly, note that (4) and (5) omit the gradient step for the task-specific heads h_ϕ and h_ψ, which use their own learning rates, η_ϕ and η_ψ respectively (here $\eta_\phi = \eta_\psi$).

Auxiliary tasks. We consider three tasks in our evaluations. (1) Semantic segmentation, for which we use a CrossEntropy (CE) loss. (2) *Multi-Label Dense Classification* (*MLDC*), defined as the classification task that can be computed out of semantic segmentation outputs y_s and labels $\hat{y}_s$. We average along the spatial dimensions of y_s, obtaining a multi-label vector $y_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} y_{s_{i,j}}$, taking the unique classes $\hat{y}_c$ of $\hat{y}_s$ and computing BinaryCrossEntropy (BCE) loss between y_c and $\hat{y}_c$. We include this task to disentangle the contributions of the classification and positioning parts of the semantic segmentation task. (3) Lastly, we use image reconstruction by using 3 output channels (RGB) and computing the MeanSquaredError (MSE) loss w.r.t. the original input images. We include this task to understand whether task labels are needed at all.

4. Experimental Results

This section presents a comprehensive evaluation of our method. We begin by describing the training details, the studied auxiliary tasks, MDE and auxiliary datasets, and their pre-processing steps. We then evaluate the efficacy of our method, investigating three main areas: the impact of different auxiliary datasets in improving MDE quality; the effectiveness of jointly training across a variety of MDE datasets; and the task relevance, exploring the impact of the auxiliary task selection for the improvement of MDE.

Training details. We adopt the training procedure of DINOv2 [46], both for our main task (MDE) and for our auxiliary task. In practice, we duplicate the optimizer and learning rate scheduler, and then we scale the MDE DPT learning rate by α and the auxiliary task learning rate by $1-\alpha$, as depicted in Figure 2. In Figure 1 (right) we study the impact of the hyperparameter α on MIX6, a dataset we introduce below, see Table 1. Notably, the figure reveals that values greater than 0.6 are good candidates for achieving improvements on MDE. Throughout the experimental section, we adopt 0.9 as it showed the best performance. However, this may vary depending on the dataset, task and training procedure. The general optimal selection of α is complementary and not the main goal of this work. Lastly, note that for all experiments we report mean and standard error of 4

Table 1. Adopted auxiliary datasets. We name their union MIX6.

Name	Scene Type	Classes	Train Size
ADE20K [73]	Indoor, Outdoor	150	20K
SUN RGBD [58]	Indoor	37	5.2K
Cityscapes [14]	Driving	34	2.9K
COCO-Stuff [7]	Indoor, Outdoor	171	118K
Pascal VOC [20]	Indoor, Outdoor	21	8.5K
Pascal Context [41]	Indoor, Outdoor	59	5K

independent runs (with different random but fixed seeds). Further training details, and hardware details are provided in Appendices A and B, respectively.

Auxiliary Datasets. We consider the auxiliary datasets reported in Table 1. We use the pre-processing steps in the *mmsegmentation* framework [13]. Additionally, we compose a dataset out of these, which we name *MIX6*. This dataset is used in combination with dataset-specific prediction heads, to avoid the need to join the classes into a unique class space. Each composing dataset keeps its original pre-processing steps. We excluded Taskonomy from MIX6 as its relatively poor segmentation label quality might negatively impact the MDE performance [70].

MDE datasets. We use the MDE datasets reported in Table 2. In particular, for NYUv2 and SUN RGBD we use the pre-processing steps adopted by DINOv2 [46], while for the other indoor/outdoor datasets we use the NYUv2 pre-processing, without the NYUCrop. For DIODE Outdoor we limit the depth to 80 meters [69], while for Matterport3D we first resize the images to match the NYU images size (480x640). For Matterport3D and Taskonomy we report results on the validation set. To maintain a proportional scale between MDE and auxiliary datasets, we randomly select 10% of the images for Taskonomy.

For KITTI we use the configurations in the *Monocular Depth Estimation Toolbox* [34]. Note that KITTI evaluates a different case than the remaining datasets: images are taken from autonomous driving and street scenes and therefore are semantically and structurally different[1]. Nonetheless, we include it to observe the effects of using mainly out-of-domain auxiliary datasets on MDE performance.

4.1. Enhancement via Multi-Source Auxiliary Tasks

We study whether we can compensate a decrease in depth estimation information during training with an increase in information from an auxiliary vision task and an external dataset. In other words, if downscaled MDE steps can be compensated by complementary upscaled auxiliary

[1] wider images (352x704), large outdoor scenes, and poor label quality.

Table 2. MDE datasets used in this work, where SL=Structured Light; TOF=Time-of-Flight; SCS=Stereo Camera Sensing. For more information about the sensor names, we refer the reader to the work of Lopes et al. [40]. For MatterPort3D and Taskonomy we test on the validation split.

Name	Scene Type	Sensor Type	Train Size	Test Size
NYUv2 [42]	Indoor	SL	24.2K	654
SUN RGBD [58]	Indoor	SL, TOF	5.2K	5K
MatterPort3D [10]	Indoor	SCS, SL	144K	19.2K
Taskonomy [71]	Indoor	SL	3.2M	498K
DIODE [62]	Indoor	LiDAR	8.5K	325
	Outdoor		16.8K	446
KITTI [22]	Driving	SCS, LiDAR	23K	652

task steps from an external dataset. We conduct extensive experimental evaluations using the training method described in Section 3. Results show that not only can we compensate MDE information with an auxiliary task, but we also improve the overall MDE quality. This suggests that our method can be used as ”augmentation” for the main downstream task, while disregarding the auxiliary task outcomes. In the following paragraphs, we show that our method succeeds independently from the auxiliary dataset, consistently improving MDE performance when using in-domain auxiliary data. Lastly, we report observations on the auxiliary task relevance.

Auxiliary dataset choice. Figure 3 illustrates Absolute Relative Error on NYUv2, SUN RGBD and DIODE Outdoor of the baseline and our method with each auxiliary dataset listed in Table 1, including their mix, MIX6. The results confirm the enhancement of MDE quality on the three datasets, irrespective of the chosen auxiliary dataset. Notably, MIX6 consistently outperforms others, emerging as the most effective auxiliary dataset for both NYUv2 and SUN RGBD, leading us to use it subsequent experiments unless differently specified. The consistent outperformance of MIX6 suggests that the diverse combination of auxiliary datasets within MIX6 offers a set of cues to enhance MDE quality, with each dataset contributing unique information. We note that for DIODE Outdoor using Cityscapes in dense classification works even a bit better than MIX6.

Results across multiple MDE datasets. We extend our experiments to a variety of MDE datasets, covering indoor, outdoor and driving scenes, as reported in Table 2. We here use MIX6 as auxiliary dataset as it generally performed the best (Figure 3). From Table 3, we see that our method consistently improves MDE quality on indoor and outdoor datasets by an average of $\sim 11\%$, indicating that our method

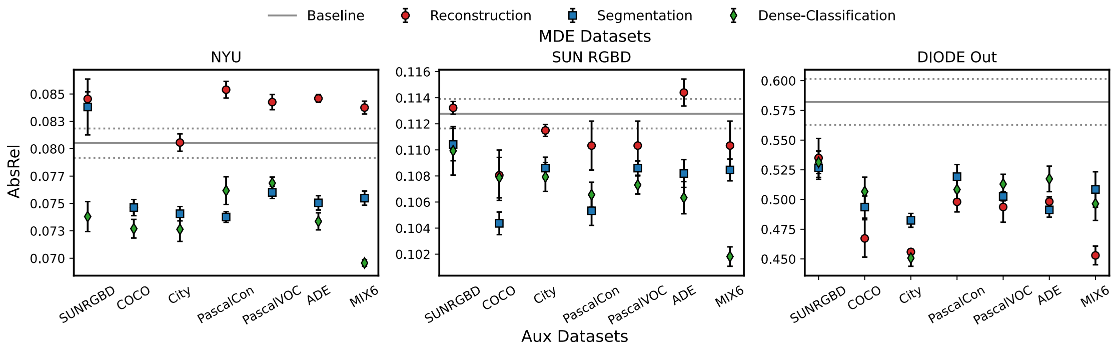

Figure 3. Absolute Relative Error (AbsRel) of MDE on NYUv2, SUN RGBD and DIODE Outdoor using the DINOv2 baseline and our method with multiple auxiliary datasets and tasks. Dots and bars depict the mean and standard error of AbsRel (↓), respectively.

is robust to the MDE dataset selection and that MIX6 provides useful information mainly but not exclusively when used as auxiliary training dataset for MLDC. Furthermore, Table 3 demonstrates that incorporating MIX6 as auxiliary dataset for KITTI does not yield beneficial results. We hypothesize that this outcome is due to two main factors: firstly, the inherent challenges presented by KITTI dataset (see Section 4), and secondly, the out-of-distribution characteristics of the MIX6 dataset, which is dominated by non-driving scenes. Consequently, we have chosen to exclude KITTI from further evaluations.

When also considering the out-of-distribution performance on KITTI, our method achieves an average $\sim 9\%$ AbsRel gain compared to the DINOv2 baseline.

Task relevance. We ask the question: "Is the contribution coming from the auxiliary task or simply from adding data?". As depicted in Figure 3 and detailed in Table 3, employing the same dataset with different tasks yields varying results, indicating that improvements in MDE quality do not merely stem from data addition. Remarkably, Table 3 shows the consistent superiority of MLDC as an auxiliary task, outperforming other tasks across datasets, with exceptions for DIODE Outdoor, where image reconstruction is more effective. Interestingly, MLDC often surpasses semantic segmentation, suggesting that identifying classes in an image, without the need for precise object positioning, is sufficient for auxiliary task effectiveness. We hypothesize that this holds true, especially when semantic segmentation labels lack positional precision. The broader implication is that semantic segmentation datasets can be repurposed as dense classification datasets to enhance various vision downstream tasks, showcasing the versatility and impact of repurposing datasets for auxiliary tasks. Given these findings, MLDC is selected as the auxiliary task for the subsequent experiments, unless explicitly specified otherwise. We additionally evaluate the method with single-label dense classification as an auxiliary task and show in Appendix C.1 that it also improves MDE quality compared to the baseline, while using multiple labels seems better.

4.2. Improved Data Efficiency Properties

Given challenges associated with collecting new high-quality labeled MDE datasets, we explore the data efficiency properties of our method by seeking an answer to the following question: "To what extent can we reduce the usage of MDE data while maintaining performance comparable to our full-data baseline?". Our investigations highlight the intrinsic data efficiency of the frozen DINOv2 ViT-G for MDE, up to a certain degree. As shown in Figure 4, this efficiency is amplified by our method on all tested MDE datasets, allowing from 80% to 99% less labeled MDE data usage without decreasing the performance. These promising results suggest that for future MDE dataset collection it may be possible to gather less data, while still being able to compensate for it using auxiliary datasets.

5. Ablation studies

We conduct ablation studies on the impact of learning rate scaling, and robustness to backbone configurations.

5.1. Tuning the Learning Rate with and without Auxiliary Task

In order to verify that the contribution comes from the auxiliary task, and not merely from adjusting the learning rate, we repeat the baseline experiments with learning rate scaled by a factor γ, as reported in Figure 5 (Baseline, γ). The figure shows that the adopted learning rate (scaled by $\gamma = 1$) is optimal for MDE on NYUv2, and that differently scaling it produces on par or worse performance.

Furthermore, we motivate our method design by reporting results for an alternative method, where only auxiliary steps are scaled. In other words we re-design our method such that the learning rates for the shared DPT decoder are η_{g_θ} and $\beta\eta_{g_\theta}$, respectively for MDE and the auxiliary task.

Table 3. AbsRel $\times 10^4$ ($\downarrow$) scores and percentage gain of the best task w.r.t. the DINOv2 baseline on various indoor and outdoor (top 6) and driving scenes (bottom 2) datasets, using different auxiliary tasks. The indoor-outdoor dominated MIX6 dataset with $\alpha = 0.9$ is used for all the experiments. The best and second-best results are highlighted in **bold** and *italic*, respectively.

MDE Datasets	In MIX6	DINOv2	Aux Tasks			Gain %
			Classification	Segmentation	Reconstruction	
NYUv2	✗	809 ± 10	**696 ± 3**	*755 ± 6*	838 ± 7	13.9
SUN RGBD	✓	1128 ± 11	**1024 ± 10**	*1069 ± 9*	1111 ± 5	9.2
Matterport3D	✗	1874 ± 19	**1728 ± 9**	*1793 ± 13*	1805 ± 6	7.8
Taskonomy	✗	*1506 ± 13*	**1481 ± 4**	1567 ± 12	1543 ± 6	1.7
DIODE In	✗	3588 ± 19	**3239 ± 26**	3451 ± 47	*3368 ± 31*	9.7
DIODE Out	✗	5820 ± 223	*4965 ± 162*	5085 ± 172	**4530 ± 91**	22.2
KITTI	✗	**605 ± 5**	*609 ± 7*	646 ± 3	642 ± 6	-0.6

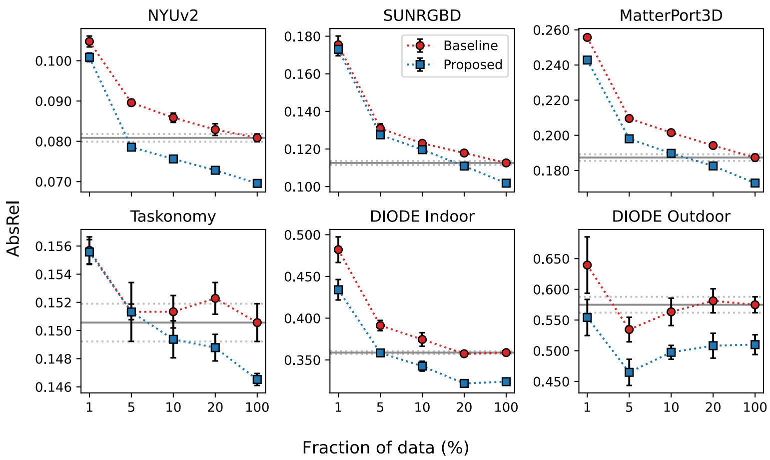

Figure 4. AbsRel of models trained with various fractions of the dataset. The dataset sizes are reported in Table 2. DINOv2 baseline (red circles) represents the model trained without auxiliary tasks, whereas Proposed (blue squares) depicts our method jointly trained with MIX6 MLDC auxiliary task with $\alpha = 0.9$.

Figure 5 (Proposed, β) shows that this method can still improve the MDE performance, while being slightly worse when compared to our proposed method, suggesting that using some auxiliary information from related datasets and tasks is beneficial for MDE training, while highlighting the importance of using a small weight for the auxiliary task, to prevent it from overtaking the MDE task.

5.2. Validation with Depth Anything Backbone

We validate our method using the recent Depth Anything model [69] as backbone, which is pre-trained on over 60 million samples, and therefore is a stronger MDE backbone than DINOv2. We follow its original training details, extending the training duration (same as DINOv2 [46]) to compensate for the frozen backbone and randomly initialized DPT decoder. Table 4 shows that our method improves when using this backbone, achieving an average 2.5% quality gain. This suggests that our training scheme can enhance MDE quality across various pre-trained backbones.

Table 4. AbsRel $\times 10^4$ ($\downarrow$) scores on various depth datasets using Depth Anything as baseline and our method with MLDC and $\alpha = 0.9$, and percentage gain of our method w.r.t. Depth Anything. The best results are highlighted in **bold**.

MDE Datasets	Depth Anything	Ours	Gain %
NYUv2	736 ± 15	**721 ± 10**	2.1
SUN RGBD	1028 ± 7	**1018 ± 15**	1
Matterport3D	1885 ± 9	**1843 ± 2**	2.3
Taskonomy	1741 ± 15	**1687 ± 13**	3.1
DIODE In	**1993 ± 14**	2073 ± 22	-0.4
DIODE Out	3835 ± 164	**3574 ± 66**	6.8

6. Qualitative Results

Figure 1 (top) reports a prediction on a NYUv2 validation sample using the baseline and our method using MIX6 MLDC auxiliary task with $\alpha = 0.9$. Note that black regions

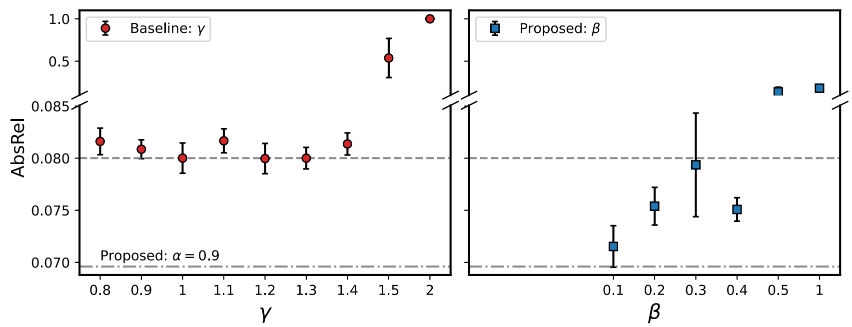

Figure 5. AbsRel (↓) when varying the learning rate. (Left) Learning rate tuning for the DINOv2 baseline by a factor γ. (Right) Our method re-designed to do unscaled depth steps and auxiliary steps scaled by a factor β, using MIX6 MLDC task. For both plots, dashed and dotted-dashed lines represent the performance of the baseline ($\gamma = 1$) and of our method with $\alpha = 0.9$, respectively.

represent invalid regions, i.e. pixels for which the label is unknown or out of range. We define $err_m = |\hat{y} - y|/\hat{y}$ as the (absolute relative) error map of the prediction y of a method m w.r.t. the ground truth $\hat{y}$. Note that rows in these figures should be considered independently. The figure shows $err_{\alpha=1.0} - err_{\alpha=0.9}$, i.e. the difference between the errors of the baseline and our method, respectively, where green depicts areas where our method is closer to the ground truth than the baseline, and red vice versa. In this figure, it is clear that the most distant area (right corner) is lighter and therefore closer to the ground truth. More qualitative results are provided in Appendix D. In general, our method's predictions are closer to the ground truth, even though not always these improvements are well visible by human eyesight.

7. Discussion and Limitations

Our method presents promising advancements, however it is essential to acknowledge its limitations.

The selection of auxiliary datasets and tasks plays a pivotal role in improving MDE. We show that MLDC is effective with various datasets. However, the observed task dependency highlights the significance of carefully choosing auxiliary tasks and datasets for optimal outcomes.

As mentioned in the results section, we report no improvement on KITTI. This can be attributed to several factors. First, as discussed in Section 4, driving images are pre-processed differently than other datasets and have poor and sparse labels. Second, the domain distribution of the KITTI dataset is specific to driving scenes, which is not adequately represented in our auxiliary datasets. Third, the quantity and variety of auxiliary datasets are crucial for enhancing the model performance. This is evident when comparing our results in Table 3 and the corresponding ablation in Table 6 in Appendix C.2, which shows a decrease in performance across all datasets when using the same dataset exclusively as auxiliary. These observations suggest that including of a broader and more diverse range of driving data could enhance the baseline performance for KITTI.

8. Conclusion

Monocular Depth Estimation poses challenges stemming from scarcity of high-quality labeled datasets [69]. We propose joint training of a shared decoder on top of a frozen pre-trained model using auxiliary vision datasets and tasks.

Through our experimental analysis, we show that our method is robust to the auxiliary dataset choice and that we can improve MDE performance on MDE datasets with in-domain auxiliary datasets by an average of 11% compared to the DINOv2 baseline.Our method demonstrates improved data efficiency for MDE, allowing for at least 80% MDE data reduction on the tested datasets. This suggests that it may enable quality gains even in scenarios where access to high-quality labeled data is limited. By leveraging auxiliary datasets and a frozen foundation model, our approach improves the quality without the need for extensive re-training. Moreover, our results highlight the influence of semantic segmentation datasets employed as MLDC task and that improvements are due to the selected auxilliary task, not only due to adding more data.

In conclusion, we introduce a compelling strategy for improving MDE by addressing its challenges and demonstrating the potential of using auxiliary datasets. We aim to encourage the exploration of auxiliary task balancing strategies and further studies on auxiliary data and task selection.

Acknowledgements

Alessio Quercia was funded by the Helmholtz School for Data Science in Life, Earth, and Energy (HDS-LEE). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS [26] at Jülich Supercomputing Centre (JSC).

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019. 3

[2] United States Environmental Protection Agency. Greenhouse gas equivalencies calculator, Mar 2024. 12

[3] Nitin Bansal, Pan Ji, Junsong Yuan, and Yi Xu. Semantics-depth-symbiosis: Deeply coupled semi-supervised learning of semantics and depth. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5828–5839, 2023. 2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 2

[5] Deblina Bhattacharjee, Sabine Süsstrunk, and Mathieu Salzmann. Vision transformer adapters for generalizable multitask learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19015–19026, 2023. 2

[6] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 1, 2

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 5

[8] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997. 2, 3

[9] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8001–8008, 2019. 3

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 5

[11] Hong Chen, Xin Wang, Chaoyu Guan, Yue Liu, and Wenwu Zhu. Auxiliary learning with joint task and data scheduling. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3634–3647. PMLR, 17–23 Jul 2022. 3

[12] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 2, 3

[13] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 5

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[15] Hanz Cuevas-Velasquez, Alejandro Galán-Cuenca, Robert B Fisher, and Antonio Javier Gallego. Efficient multi-task progressive learning for semantic segmentation and disparity estimation. *Pattern Recognition*, 154:110601, 2024. 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. Ieee, 2009. 12

[17] Lucio M. Dery, Paul Michel, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Aang: Automating auxiliary learning. *ArXiv*, abs/2205.14082, 2022. 3

[18] 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, 2

[19] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019. 3

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 5

[21] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021. 3

[22] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5

[23] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018. 2, 3

[24] Ankit Jha, Biplab Banerjee, and Subhasis Chaudhuri. S 3 dmt-net: improving soft sharing based multi-task cnn using task-specific distillation and cross-task interactions. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. 2

[25] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 2, 3

[26] Stefan Kesselheim, Andreas Herten, Kai Krajsek, Jan Ebert, Jenia Jitsev, Mehdi Cherti, Michael Langguth, Bing Gong,

Scarlet Stadtler, Amirpasha Mozaffari, et al. Juwels booster–a supercomputer for large-scale ai research. In *International Conference on High Performance Computing*, pages 453–468. Springer, 2021. 8

[27] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. *Advances in Neural Information Processing Systems*, 35:26295–26308, 2022. 2

[28] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 12

[29] Steven Landgraf, Markus Hillemann, Theodor Kapler, and Markus Ulrich. Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation. *arXiv preprint arXiv:2402.10580*, 2024. 2

[30] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[31] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9167–9176, 2019. 2

[32] Kunqian Li, Xiya Wang, Wenjie Liu, Qi Qi, Guojia Hou, Zhiguo Zhang, and Kun Sun. Learning scribbles for dense depth: Weakly-supervised single underwater image depth estimation boosted by multi-task learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[33] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18879–18889, 2022. 2

[34] Zhenyu Li. Monocular depth estimation toolbox. `https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox`, 2022. 5

[35] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2

[36] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 2

[37] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5162–5170, 2015. 2

[38] Huan Liu, Zhixiang Chi, Yuanhao Yu, Yang Wang, Jun Chen, and Jin Tang. Meta-auxiliary learning for future depth prediction in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5756–5765, January 2023. 3

[39] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 2, 3

[40] Alexandre Lopes, Roberto Souza, and Helio Pedrini. A Survey on RGB-D Datasets. *arXiv e-prints*, page arXiv:2201.05761, Jan. 2022. 5

[41] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

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

[43] Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shen, and Ian Reid. Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7101–7107. IEEE, 2019. 2

[44] Matthias Niessner. Matterport3d eula for academic use, Mar 2024. 12

[45] Jia Ning, Chen Li, Zheng Zhang, Chunyu Wang, Zigang Geng, Qi Dai, Kun He, and Han Hu. All in tokens: Unifying output space of visual tasks via soft token. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19900–19910, 2023. 2

[46] 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. 1, 2, 3, 4, 5, 7, 12

[47] Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2198, 2019. 3

[48] Sangjoon Park and Jong Chul Ye. Multi-task distributed learning using vision transformer with random patch permutation. *IEEE Transactions on Medical Imaging*, 2022. 2

[49] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. 12

[50] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 2

[51] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1, 2, 3

[52] 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

[53] Peter Rottmann, Thorbjörn Posewsky, Andres Milioto, Cyrill Stachniss, and Jens Behley. Improving monocular depth estimation by semantic pre-training. In *2021 IEEE/RSJ Interna-*

tional Conference on Intelligent Robots and Systems (IROS), pages 5916–5923, 2021. 3

[54] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2

[55] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8197–8207, 2021. 3

[56] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 2

[57] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2

[58] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5

[59] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020. 3

[60] Pardis Taghavi, Reza Langari, and Gaurav Pandey. Swin-mtl: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images. *arXiv preprint arXiv:2403.10662*, 2024. 2

[61] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3614–3633, 2021. 2

[62] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arXiv e-prints*, page arXiv:1908.00463, Aug. 2019. 5

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[64] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[65] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 541–550, 2020. 2

[66] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 3

[67] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2

[68] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3917–3925, 2018. 2

[69] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv:2401.10891*, 2024. 1, 2, 3, 5, 7, 8

[70] Xuan Yang, Liangzhe Yuan, Kimberly Wilber, Astuti Sharma, Xiuye Gu, Siyuan Qiao, Stephanie Debats, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, et al. Polymax: General dense prediction with mask transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1050–1061, 2024. 2, 5, 12

[71] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 3, 5

[72] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021. 2

[73] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 5