

Boosting Generalizability towards Zero-Shot Cross-Dataset Single-Image Indoor Depth by Meta-Initialization

Cho-Ying Wu, Yiqi Zhong, Junying Wang, and Ulrich Neumann

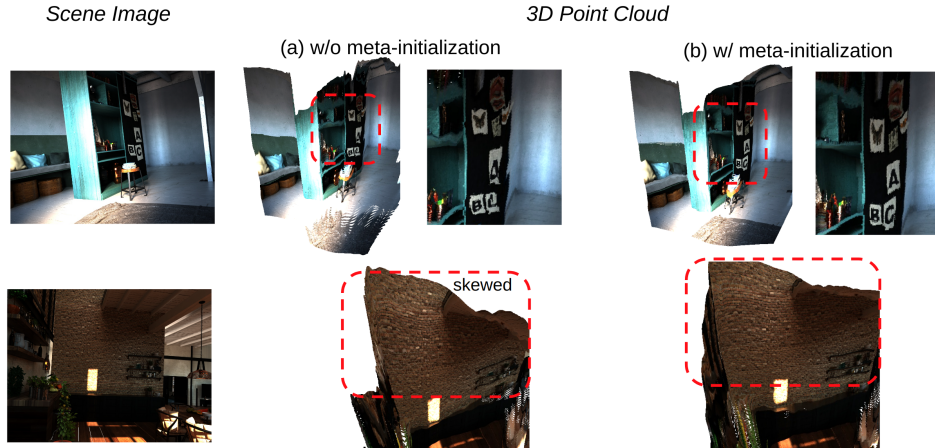


Fig. 1: Geometry structure comparison in 3D point cloud view. We back-project the predicted depth maps from images into textured 3D point cloud to show the geometry. The proposed meta-initialization has better domain generalizability that leads to more accurate depth prediction hence better 3D structures. (zoom in for the best view).

Abstract—Indoor robots rely on depth to perform tasks like navigation or obstacle detection, and single-image depth estimation is widely used to assist perception. Most indoor single-image depth prediction focuses less on model generalizability to unseen datasets, concerned with in-the-wild robustness for system deployment. This work leverages gradient-based meta-learning to gain higher generalizability on zero-shot cross-dataset inference. Unlike the most-studied meta-learning of image classification associated with explicit class labels, no explicit task boundaries exist for continuous depth values tied to highly varying indoor environments regarding object arrangement and scene composition. We propose fine-grained task that treats each RGB-D mini-batch as a task in our meta-learning formulation. We first show that our method on limited data induces a much better prior (max 27.8% in RMSE). Then, finetuning on meta-learned initialization consistently outperforms baselines without the meta approach. Aiming at generalization, we propose zero-shot cross-dataset protocols and validate higher generalizability induced by our meta-initialization, as a simple and useful plugin to many existing depth estimation methods. The work at the intersection of depth and meta-learning potentially drives both research to step closer to practical robotic and machine perception usage.

I. INTRODUCTION

Much research learns depth from single images to fulfill indoor robotic tasks like collision detection [9, 23], navigation [36, 37], grasping [31, 39], or it benefits 3D sensing or learning good 3D representations for AR/VR and view synthesis [7, 30]. However, generalization is still a major issue in robustly estimating depth on unseen scenes or datasets, especially indoor scenes, since their composition

varies widely, and objects are usually cluttered in the near field without an order. An intuitive solution is to learn from large-scale mixed datasets [27, 28, 43] or adopt pretrained auxiliary models as guidance [40], but they require extra information or exogenous models. Without those resources when training on data of limited *appearance and depth variation* (scene **variety** in this work), with an extreme case that only sparse and non-overlapping views are available, networks barely learn valid depth.

Inspired by meta-learning’s advantages on domain generalization and few-shot learning [8, 24], we dig into how meta-learning can be applied to pure¹ single-image depth prediction. Conventional meta-learning focuses on image classification and follows few-shot multitask learning, where a task represents a distribution to sample data from [17]. Instead, we study a more complex problem of scene depth estimation: the difficulties first come from estimating per-pixel and continuous range values, in contrast to global and discrete values for image classification. Next, indoor RGB-D captures vary greatly. Even within a sequence, an adjacent frame to a close view of cluttered objects can be large spaces without objects.

This observation indicates that pure single-image depth estimation lacks clear task boundaries as conventional meta-learning [15]. The problem also differs from online depth adaptation, which treats each video sequence as a task [47, 48]. To address the specific challenges, we propose to treat

¹This differs from either stereo or video online depth adaptation [47, 48], which assume other sequential frames exist and learn from affinity between nearby frames for online adaptation. Instead, we work on *pure* single-image problems without utilizing other frames.

*Authors are with the Department of Computer Science, University of Southern California, Los Angeles, CA, USA

each mini-batch as a **fine-grained task** (Sec. III-A).

This work follows gradient-based meta-learning, which adopts a meta-optimizer and a base-optimizer [8, 24]. The base-optimizer explores multiple inner steps to find weight-update directions. Online and task augmentations are adopted during the base-optimization. The former augments a fine-grained task sample using color jittering and left-right flip. The latter is after the inner steps, we further sample another fine-grained task and perform mix-up and channel shuffle. Then, the meta-optimizer updates the meta-parameters following the inner-step explored directions. With several epochs of dual-loops, a function θ^{prior} that maps images to depth is learned. θ^{prior} works as a better initialization for the subsequent supervised learning (Sec. III-B). Note that both meta-learning and supervised learning operate on the same training set without needing extra data. The improvements are explained via *progressive learning* in Sec. III-C.

A previous study [40] points out that indoor depth is especially hard to robustly resolve from images due to complex object arrangements in the near field with 6-DoF camera poses that can render scenes from nearly arbitrary viewing directions. Furthermore, surface textures may cause confusion. A depth estimator needs to separate **depth-relevant-irrelevant cues** from all extracted high-frequency features. The former indicates that depth values change when color and appearance change, such as boundaries between objects and background; the latter refers to surface textures, where depth is invariant to colors, such as paintings or material patterns on walls [4, 40].

We show that meta-learning induces a better prior with *higher generalization* to unseen scenes and can identify depth-relevant-irrelevant cues robustly. To validate generalization, we propose evaluation protocols using multiple popular indoor datasets [26, 29, 34, 40] for **zero-shot cross-dataset evaluation**. In contrast, most prior indoor works only train and test on a single dataset [1, 21, 44].

In experiments we show that either meta-learning on general-purpose architecture or meta-learning as plugins to several dedicated depth frameworks shows consistently superior performance. *We highlight the finding that by only modifying training schema but keeping the same data, loss function, and architecture, performances can still be improved with more robust and generalizable depth estimation for real-world usage such as indoor robots or AR/VR surrounding perception.* In addition to contributions in depth estimation, from meta-learning’s view, we introduce fine-grained task on a continuous, per-pixel, and real-valued regression problem to advance meta-learning’s study on practical problems. Fig. 1 shows sample results.

Contributions:

- The first method to apply meta-learning on pure single image depth prediction, which helps achieve higher generalization in depth estimation without multiple training datasets, side information, or other pretrained networks.
- A novel fine-grained task concept to overcome a pure single-image setting where no explicit task boundaries exist. The work further serves as an empirical study on a practical

and challenging problem in meta-learning research.

- A proposed protocol for zero-shot cross-dataset evaluation on indoor scenes. The protocol faithfully evaluates a model’s robustness and generalizability. Extensive experiments are shown to validate our strategy.

II. RELATED WORK

Depth from Single Indoor Images. The task has gained higher popularity [20, 21] while more high-quality datasets become available, such as Hypersim [29], Replica [34], HM3D [26], and VA [40]. Some methods adopt surface normal [42], plane constraints [19, 20], advanced loss functions [1], auxiliary depth completion [14], mixed-dataset training [2, 27, 28, 43], or modules customized for depth estimation [18, 21, 44]. In contrast, our work focuses on *designing a better learning scheme* without extra data or exogenous models: we adopt the fundamental regression loss and off-the-shelf networks without loss of generalization in meta-learning methodology. Most prior works train/test on the same dataset without validating cross-dataset performance. We instead devise zero-shot cross-dataset evaluation protocols using recent high-quality synthetic and real captures to validate generalization from meta-initialization.

Gradient-based Meta-Learning. Meta-Learning principles [16] illustrate an oracle about learning how to learn, especially useful on domain adaption, generalization, and few-shot learning. Popular gradient-based algorithms such as MAML [8] and Reptile [24] are formulated as bilevel optimization problems using a base- and meta-optimizer. MAML uses gradients computed on the query set to update the meta-parameters. Reptile does not distinguish support and query sets and simply samples data from a task distribution for inner-loop exploration. We refer readers to [17] for a survey on algorithms.

The majority of meta-learning studies in vision community focuses on image [5, 25, 50] or pixel-level classification [3, 13, 22]. One pioneer [10] investigates single-image pose regression that naively regresses rotation angles for one synthetic object in each image. The study may be far from real use.

Few works use meta-learning for depth but only on driving scenes with a much different problem setup [35, 38, 47, 48]. Many perform online learning and adaptation using stereo [47] or monocular videos [38, 48] by enforcing temporal consistency. They require affinity in nearby frames and meta-optimize within a single sequence. Our problem is arguably harder due to the pure single-image setting in highly diverse indoor structures. [35] works on single images but requires multiple driving datasets to build tasks and train on. We do not require multiple training sets, and we find their method is limited for indoor scenes by experiments.

III. METHODS

A. Single Mini-Batch as a Fine-Grained Task

Definition. Single-image depth prediction learns a function $f_\theta : \mathcal{I} \rightarrow \mathcal{D}$, parameterized by θ , to map from imagery to depth. A training set $(\mathbb{I}_{train}, \mathbb{D}_{train})$, containing images

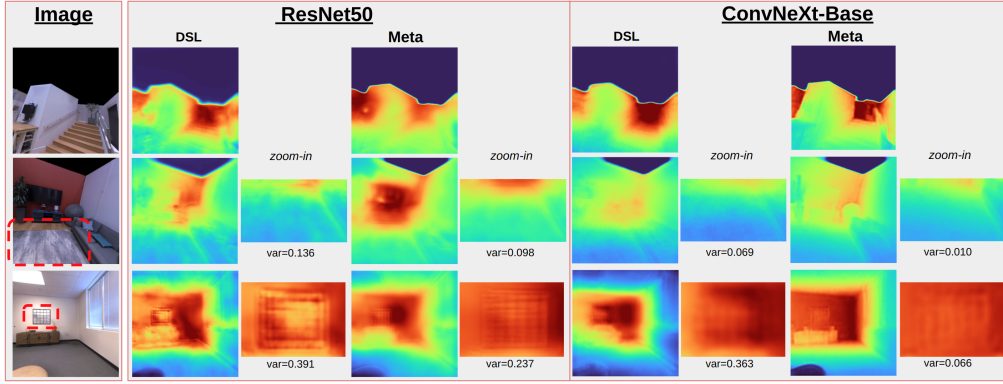


Fig. 2: Fitting to training environments. *var* shows variance for depth values in the highlighted regions. We show comparisons of fitting to training data between first-stage meta-learning (Meta) and direct supervised learning (DSL) using Replica Dataset that contains limited scene appearance and depth variation (scene variety). Meta produces smooth and more precise depth. Depth-irrelevant textures on planar regions can be resolved more correctly. In contrast, DSL produces irregularities affected by local high-frequency details, especially ResNet50. See Sec. IV-A for details and III-C for the explanation.

$I \in \mathbb{I}_{train}$ and associated depth maps $D \in \mathbb{D}_{train}$, is used to train a model. Each mini-batch with a size K , $(I_k, D_k) \forall k \in [1, K]$, is treated as a **fine-grained task**.

Difference with a task in meta-learning context. A fine-grained task is different from a task in the most-used meta-learning or few-shot learning context [8], where a task contains a data distribution, and batches are sampled from it. A fine-grained task does not contain a data distribution but is sampled from a meta-distribution, the whole RGB-D dataset. For example, to train a model, a navigating agent captures an RGBD dataset, which is the meta-distribution to sample each mini-batch as a fine-grained task.

Design. The design is motivated by the fact that appearance and depth variation can be arbitrarily high. A view looking at small desk objects and a view of large room spaces are highly dissimilar in contents and ranges, and thus mappings from their scene appearance to depth values are different. Still, they can be captured in the same environment or even in neighboring frames. This contrasts with image classification, where class samples share a common label. The observation explains why we treat each mini-batch as a fine-grained task instead of each environment.

B. Meta-Initialization on Depth from Single Image

Meta-Learning stage. In the first meta-learning stage, we adopt a meta-optimizer and a base-optimizer. In each meta-iteration, a fine-grained task B contains K samples that are sampled from the whole training set: $(I_k, D_k) \sim (\mathbb{I}_{train}, \mathbb{D}_{train}), \forall k \in [1, K]$. Then we take L steps to explore gradient directions that minimize the regression loss, \mathcal{L}_{reg} . We perform *online augmentation*, **Aug**, at each exploration step, including color jittering and left-right flip to craft multiple samples for a task. We get $(\theta_{expl}^1, \theta_{expl}^2, \dots, \theta_{expl}^L)$ from

$$\theta_{expl}^i \leftarrow \theta_{expl}^{i-1} - \alpha \frac{1}{K} \nabla_{\theta} \sum_{k \in [1, K]} \mathcal{L}_{reg}(\mathbf{Aug}(I_k), D_k; \theta_{expl}^{i-1}). \quad (1)$$

To avoid over-fitting specific to gradient-based meta-learning [41] and improve generalization, after the inner steps, we do *task augmentation*, including mix-up and channel shuffle. We sample another fine-grained task B' and linearly blend B and B' at the bottleneck $\phi = f_E(I)$ after encoder

f_E and interpolate the depth groundtruth.

$$\phi_k^m = \lambda_k f_E(I_k) + (1 - \lambda_k) f_E(I'_k), \quad D_k^m = \lambda_k D_k + (1 - \lambda_k) D'_k, \quad (2)$$

where $\forall k \in [1, K]$ and $\lambda_k \sim \mathbf{Beta}(0.5, 0.5)$ following the common mixup [46]. ϕ^m is passed into the decoder head supervised by D^m with loss \mathcal{L}_{reg} . We also perform channel shuffle at B 's bottleneck ϕ^B with a channel size of p by randomly choosing a subset of channels in ϕ^B to be replaced by the same channels in $\phi^{B'}$. After shuffles, we get ϕ^{cs} , which is passed into the decoder head and supervised by B 's depth groundtruth using the loss \mathcal{L}_{reg} :

$$\phi_k^{cs} = r_k \phi^B + (1 - r_k) \phi^{B'}, \quad r_k \sim \mathbf{Ber}(0.95), \quad (3)$$

which means on average $p/20$ channels are shuffled with B' by Bernoulli random variables, and we empirically find more shuffles can break down the representations of B . After inner loops and getting θ_{expl}^L , either channel shuffle or mix-up is chosen with equal chances to use and optimize to get θ_{aug}^L .

Subsequent to the inner steps and task augmentation, we update the meta-parameters in Reptile style [24], i.e., following the explored weight updates in the inner steps

$$\theta_{meta}^j \leftarrow \theta_{meta}^{j-1} - \beta (\theta_{meta}^{j-1} - \theta_{aug}^L), \quad (4)$$

where α and β are respective learning rates (*lr*), and i and j denote inner and meta-iterations. After the last meta-iteration, we obtain meta-learned weights as the depth prior θ^{prior} .

Compared with MAML [8], we find Reptile more suitable for training fine-grained tasks. In Reptile's paper, it is designed without support and query split, and thus it inherently does not require multiple samples in a task, which matches our fine-grained task definition. Next, first-order MAML computes gradients on the query set at the last inner step to update meta-parameters. Yet, without simple augmentation of color jittering and flipping, only one sample exists in each fine-grained task, and each fine-grained task differs greatly by random sampling from the dataset. Thus, if taking exploration on a support split and computing gradients on the query split, but the support and query samples are sampled from different scenes without any similarity, the gradients are nearly random and prevented from converging. This contrasts with video sequence as tasks, where they use MAML by exploiting affinity between frames [38, 48]. Reptile does not require

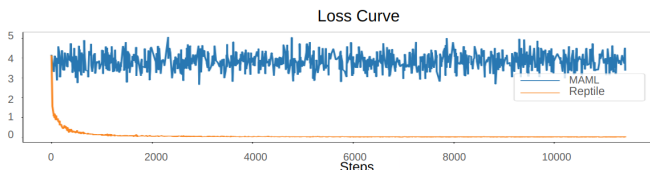


Fig. 3: Loss curve for MAML v.s. Reptile on ResNet50. We grid search MAML’s α and β in $[10^0, 10^{-4}]$ and cannot find parameters for convergence (a non-converging curve with $\alpha=0.001$ and $\beta=0.5$ is shown for clarity), and we plot Reptile with $\alpha=0.001$ and $\beta=0.5$ also by grid search.

affinity between samples and naturally stabilizes training fine-grained tasks. We show the loss curves in Fig. 3.

Supervised learning stage. We use θ^{prior} to initialize the subsequent supervised learning with stochastic gradient descent to minimize the regression loss, \mathcal{L}_{reg} , computed on the same training data ($\mathbb{I}_{train}, \mathbb{D}_{train}$).

$$\theta^* \leftarrow \min_{\theta} \mathcal{L}_{reg}(\mathbb{I}_{train}, \mathbb{D}_{train} | \theta^{prior}). \quad (5)$$

Last, a test set ($\mathbb{I}_{test}, \mathbb{D}_{test}$) is used to evaluate the depth estimation performance of θ^* . The full procedure is shown in Algorithm 1. The implementation only needs a few lines of codes as a plugin to existing depth estimation frameworks.

Algorithm 1 Meta-Initialization with fine-grained task

```

1: for epoch = 1 : N do
2:   for j = 1 : T (iterations) do
3:      $\theta_{expl}^j \leftarrow \theta_{meta}^j; (I_k, D_k) \sim (\mathbb{I}_{train}, \mathbb{D}_{train}), \forall k \in [1, K]$ .
4:     for i = 1 : L (steps) do
5:        $\theta_{expl}^i \leftarrow \theta_{expl}^{i-1} - \alpha \frac{1}{K} \nabla_{\theta} \sum_k \mathcal{L}_{reg}(Aug(I_k), D_k; \theta_{expl}^{i-1})$ .
6:     end for
7:     Mix-up or channel shuffle to get  $\theta_{aug}^L$  minimized by  $\mathcal{L}_{reg}$  with  $lr = \alpha$ .
8:      $\theta_{meta}^j \leftarrow \theta_{meta}^{j-1} - \beta(\theta_{meta}^{j-1} - \theta_{aug}^L)$ .
9:   end for
10: end for
11: Prior  $\theta^{prior} \leftarrow \theta_{meta}^T$  at epoch N.
12: Supervised learning by  $\theta^* \leftarrow \min_{\theta} \mathcal{L}_{reg}(\mathbb{I}_{train}, \mathbb{D}_{train} | \theta^{prior})$ .

```

C. Strategy Explanation

Meta-Initialization. For each meta-iteration, the base-optimizer explores its neighborhood with L steps. Compared to the usual single-step update, the meta-update first takes L -step amortized gradient descent with a lower learning rates to delicately explore local loss manifolds. Then it updates meta-parameters by direction from the inner steps but with a step size β towards θ^L . The meta-learned θ^{prior} may underfit a training set since it does not wholly follow optimal gradients but with a β for control. However, it also forces the inner exploration to reach a better understanding and avoid anchoring on seen RGB-D local cues.

Progressive learning perspective. The above strategy can be seen as progressive learning. The first-stage meta-learning avoids anchoring on seen local cues and gets coarse but smooth depth. Fig. 2 shows applying the first-stage meta-learning and direct supervised learning on data with limited scene variety. The meta-learning estimates smooth depth, but the direct supervised learning suffers from irregularity. The irregularity indicates the data cannot sufficiently demonstrate how images map to the depth domain due to the limited scene variety to learn smooth depth from global context, and thus only local high-frequency cues show up. See further analysis in Fig. 4. The irregularity occurs at surface textures

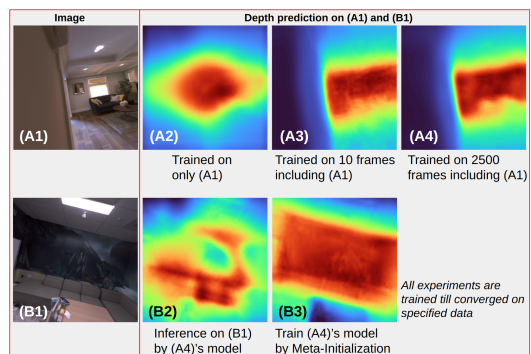


Fig. 4: Analysis on scene variety and model generalizability. (A) shows that limited scene variety hinders learning an image-to-depth mapping, with an extreme case (A2) for only one training image. (B) shows that though a model (A4) fits well on the training scenes, it still cannot generalize to an unseen scene, especially the wall painting contains many depth-irrelevant cues. Meta-initialization attains better model generalizability (Sec. III-C).

since those depth-irrelevant local cues are triggered but barely suppressed. In the second stage, a network learns finer depth based on the smoothness prior.

IV. EXPERIMENTS AND DISCUSSION

Aims. We validate our meta-initialization with four questions.

- Q1** Can meta-learning improve performances on limited scene-variety datasets (Sec. IV-A)?
- Q2** What improvements can meta-initialization bring compared with the most popular ImageNet-initialization (Sec. IV-B)?
- Q3** How does meta-initialization help zero-shot cross-dataset generalization (Sec. IV-C)?
- Q4** How does better depth helps learning 3D representations (Sec. IV-D)?

Datasets: We introduce the adopted datasets as follows.

- Hypersim [29] has high scene variety with 470 synthetic indoor environments, from small rooms to large open spaces, with about 67K training and 7.7K testing images.
- HM3D [26] and Replica [34] have 200K and 40K images, where we use data rendered in SimSIN [40]. HM3D has 800 scenes with high scene variety, and Replica has low scene variety with 18 overlapping scenes.
- NYUv2 [33] contains 464 real indoor scenes with 654 testing images but with limited camera viewing direction.
- VA [40] as a test set has 3.5K photorealistic renderings with arbitrary camera viewing directions.

Training Settings. ResNet and ConvNeXt are backbones to extract bottleneck features. We build a depth regression head following [12] that contains 5 convolution blocks with skip connection. Each convolution block contains a 3×3 convolution, an ELU activation, and a bilinear $2 \times$ upsampling layer. Channel-size of each convolution block is (256, 128, 64, 32, 16). Last, a 3×3 convolution with a sigmoid activation is used to get 1-channel output depth maps. We set $N = 5, L = 4, K = 32, (\alpha, \beta) = (10^{-3}, 0.5)$ for ResNet, and $(\alpha, \beta) = (5 \times 10^{-4}, 0.5)$ for ConvNeXt. At the supervised learning stage, we train models with a learning rate (lr) of 3×10^{-4} till convergence. Input size to the network is 256×256 . L_2 loss is used as \mathcal{L}_{reg} .

Metrics. Error metrics (in meters if having physical units): Mean Absolute Error (MAE), Absolute Relative Error (AbsRel), Root Mean Square Error (RMSE), Scale Invariant

TABLE I: Cross-Dataset generalization trained by different scene varieties. $a \rightarrow b$: training on a - and testing on b -dataset. Replica: lower scene variety; HM3D: higher scene variety. Meta-Learning’s gains are especially prominent in Replica \rightarrow VA.

Method	Replica \rightarrow VA				HM3D \rightarrow VA			
	MAE \downarrow	AbsRel \downarrow	RMSE \downarrow	SILog \downarrow	MAE \downarrow	AbsRel \downarrow	RMSE \downarrow	SILog \downarrow
Direct supervised learning	0.718	0.538	1.078	0.372	0.544	0.456	0.715	0.320
First-Stage meta-learning (K=32)	0.569	0.441	0.778	0.295	0.421	0.379	0.599	0.267
Improvement	-20.8%	-18.0%	-27.8%	-20.7%	-22.6%	-16.9%	-16.2%	-16.6%
First-Stage meta-learning (K=8)	0.587	0.455	0.819	0.311	0.429	0.386	0.610	0.274
First-Stage meta-learning (K=1)	0.617	0.475	0.847	0.325	0.445	0.398	0.635	0.280

Log Error (SILog). Accuracy: δ_C (percentage of correctness). Higher δ_C implies more structured and accurate depth. Correctness: the ratio of prediction and groundtruth is within 1.25^C , $C = [1, 2, 3]$.

A. Meta-Learning on Data with Limited Scene Variety

We first examine the first-stage meta-learning on data with limited scene variety. We train $N = 15$ epochs of the first-stage meta-learning with $(\alpha, \beta) = (10^{-4}, 0.5)$ and compare with the direct supervised learning by a learning rate $lr = 10^{-4}$ with an equivalent update epochs of $NL\beta = 30$, where we also use the early stopping to prevent explicit overfitting. We do not use the online and task augmentation in the meta-learning. ResNet50 is used. The other hyperparameters are the same as given in Training Settings. Fig. 2 shows fitting to training data. Replica with limited scene variety is used to verify gains in low-resource situations. From the figure, the meta-learning is capable of identifying near/far fields without irregularity, where the direct supervised learning struggles. Under this low-resource case, the meta-learning still induces a better image-to-depth mapping that delineates object shapes, separates depth-relevant/-irrelevant cues, and shows flat planes where rich depth-irrelevant textures exist. The observation follows the explanations in Sec. III-C.

We next numerically examine generalization to unseen scenes when training on data of different-level scene variety. HM3D (high-variety) and Replica (low-variety) are used as training sets, and VA is used for testing. Table I shows that models trained by the first-stage meta-learning substantially outperform the direct supervised learning with 16.2%-27.8% improvements. The advantage is more evident when trained on data with low scene variety. Table I further investigates the batch size K , and we find that smaller batch sizes lead to less generalization and higher variance in parameter update. Following Table I without augmentations, we study how the augmentations improve generalization in Table II. Applying the three techniques leads to the best results that further boost the performances of $K = 32$ meta-learning in Table I.

Comparison to other learning strategies to learn a prior.

We first compare meta-initialization with *simple pretraining with a strong weight decay* (WD-pre), which is conventionally used to smooth weight updates, and gradient accumulation (GA) with L steps, which also learns weight updates from multi-step forwards. WD-pre baselines: We replace the first-stage meta-learning with supervised learning using a stronger weight decay (wd) and the same learning rate as the second stage. The regular second-stage supervised learning follows. GA baseline: We replace the meta-learning dual loop with supervised learning using gradient accumulation, which accumulates gradient for $L = 4$ steps and then updates the

TABLE II: Augmentation studies including Online Augmentation (OnAug), mix-up (MX), and channel shuffle (CS). Discarding them all means not using augmentations in the first stage. Setting: Replica \rightarrow VA following Table I.

OnAug	MX	CS	MAE \downarrow	AbsRel \downarrow	RMSE \downarrow	SILog \downarrow
✓	✓	✓	0.569	0.441	0.778	0.295
✓	✓	✗	0.548	0.430	0.761	0.288
✓	✓	✗	0.528	0.425	0.740	0.283
✓	✓	✓	0.531	0.427	0.748	0.285
✗	✓	✓	0.524	0.429	0.744	0.284
✓	✓	✓	0.515	0.418	0.729	0.282

TABLE III: Meta-Initialization v.s. simple learning strategies. Stronger weight decay pretraining (WD-pre) and gradient accumulation (GA) are compared against. We use ResNet50 and train/ test on NYUv2.

NYUv2	AbsRel \downarrow	RMSE \downarrow	SILog \downarrow	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
w/o prior learning	0.131	0.480	0.175	83.6	96.4	99.0
w/ WD-pre ($wd=0.1$)	0.133	0.484	0.176	83.7	96.6	98.9
w/ WD-pre ($wd=0.05$)	0.130	0.479	0.174	83.6	96.3	98.8
w/ WD-pre ($wd=0.01$)	0.130	0.478	0.173	83.6	96.4	99.0
w/ GA	0.133	0.484	0.175	83.8	96.4	98.9
w/ Meta-Initialization	0.122	0.454	0.167	85.4	96.8	99.3

TABLE IV: Intra-Dataset evaluation. We evaluate full Algorithm 1 (meta-initialization, "+Meta") on NYUv2 and Hypersim and train/test on the same dataset. +Meta consistently outperforms ImageNet-initialization (no marks).

Hypersim	AbsRel \downarrow	RMSE \downarrow	SILog \downarrow	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
ResNet50	0.248	1.775	0.261	64.8	87.1	94.7
ResNet50+Meta	0.235	1.659	0.246	66.9	88.0	95.1
ResNet101	0.234	1.671	0.243	67.4	88.5	95.3
ResNet101+Meta	0.215	1.563	0.236	68.2	89.1	95.4
ConvNeXt-base	0.201	1.534	0.221	73.6	91.1	96.3
ConvNeXt-base+Meta	0.184	1.411	0.210	75.0	91.8	96.5

NYUv2	AbsRel \downarrow	RMSE \downarrow	SILog \downarrow	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
ResNet50	0.131	0.480	0.175	83.6	96.4	99.0
ResNet50+Meta	0.120	0.448	0.167	85.5	96.8	99.0
ResNet101	0.120	0.448	0.167	85.6	97.1	99.3
ResNet101+Meta	0.109	0.416	0.158	86.8	97.2	99.3
ConvNeXt-base	0.101	0.394	0.138	89.4	97.9	99.5
ConvNeXt-base+Meta	0.097	0.385	0.133	89.8	98.2	99.5

weights once, equivalent to larger-batch supervised learning to learn a prior. GA is trained with a learning rate of 0.0012, $L \times$ by the second-stage base learning rate. Then it is followed by the regular second stage. Table III shows the results. *w/o prior learning* only uses the second-stage supervised learning. *w/ WD-pre* ($lr=3 \times 10^{-4}$, $wd=10^{-2}$) is equivalent with longer second-stage training. The results reveal that without the interplay of dual loop optimization, simple tricks such as smoothing by weight decays or gradient accumulation do not learn better priors and show explicit gains.

B. Meta-Initialization v.s. ImageNet-Initialization

We next examine the full Algorithm 1, obtain higher-quality depth from the subsequent supervised stage, and **go beyond limited resources** to train on higher scene-variety datasets. Intuitively, higher scene variety may diminish meta-learning’s advantages in few-shot/low-resource learning [6]. However, such studies are necessary to validate meta-learning for a realistic purpose since depth estimators are practically trained on diverse environments. Comparison is drawn with baselines of the direct supervised learning without meta-initialization and begins from the standard ImageNet-initialization.

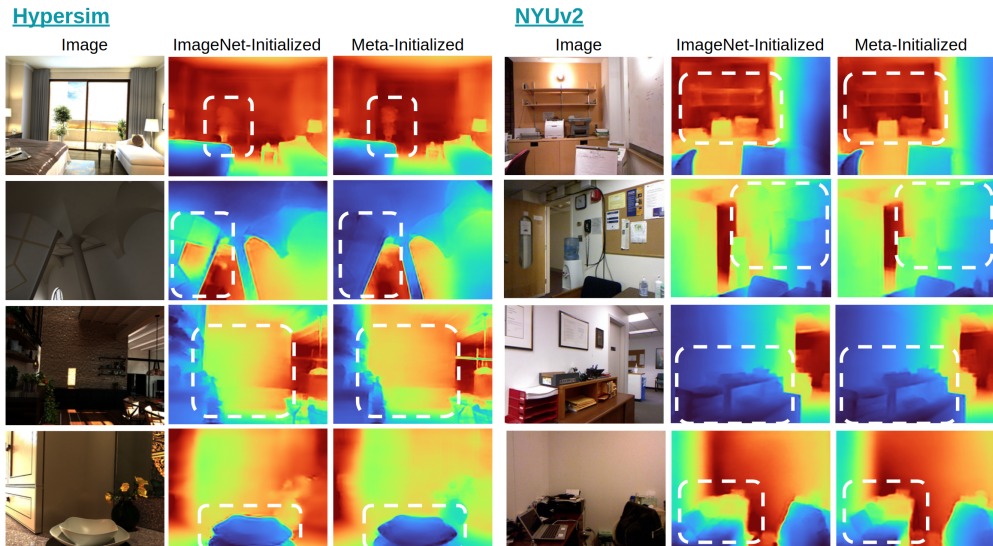


Fig. 5: Depth map qualitative comparison. Results of our meta-initialization have better object shapes with clearer boundaries. Depth-irrelevant textures are suppressed, and flat planes are predicted, as shown in Hypersim- Row 2 ceiling and 3 textured wall examples. Zoom in for the best view.

Table IV shows intra-dataset evaluation that trains/tests on each official data split. Hypersim/NYUv2 evaluation is capped for depth at 20m/10m, respectively. Meta-Initialization attains **consistently** lower errors and higher accuracy than the baselines, especially AbsRel (-7.2% on average). We display depth and 3D point cloud comparison in Fig. 5 and 1. The gain simply comes from better training schema without additional data, constraints, advanced loss, or model design.

C. Zero-Shot Cross-Dataset Evaluation

Protocol and evaluation. To faithfully validate a trained model in the wild, we design protocols for zero-shot cross-dataset inference that also aims for sim-to-real purposes. High scene-variety and large-size synthetic datasets, Hypersim and HM3D, are used as the training sets. Real or more photorealistic VA, Replica, and NYUv2 serve as the test set, and their evaluations are capped at 10m. D_{gt} and D_{pred} are groundtruth and predicted depth. We use median-scaling in the protocol to compensate for different camera intrinsics, following common unsupervised or zero-shot depth evaluation protocols [11, 12, 43, 45], which computes a ratio = $\text{median}(D_{gt}) / \text{median}(D_{pred})$ and multiply it with the prediction to align with groundtruth scale in the evaluation. In Table V, compared with the ImageNet-initialization, the meta-initialization **consistently** improves in nearly all the metrics, especially δ_1 (on average +1.74 points).

Meta-Initialization as plugins. To show wide applicability, we plug our meta-initialization into many recent high-performing architecture specialized for depth estimation, including BTS [19], DPT (hybrid and large size) [27], DepthFormer [21], AdaBins [1], GLPDepth [18], and NDDepth [32]. Comparison in Table VI shows using our meta-initialization consistently improves their original performances, giving them higher generalizability on zero-shot cross-dataset inference. Our meta-initialization is useful, especially since it only needs a few lines of code changes to modify learning schedules as a simple but effective piece.

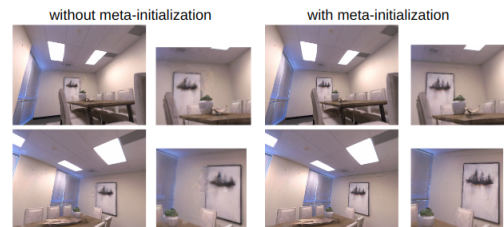


Fig. 6: Novel view comparison. Zoom in for the best view.

Comparison: single-image meta-learning depth. Next, we compare with [35], which also works on pure single images but requires multiple training datasets. They define tasks as different training set combinations and apply a consistency loss on the same scene but belonging to different tasks. We apply their strategy, train on NYUv2, HM3D, and Hypersim, define tasks as single or joint of two or three sets, compare with our meta-initialization trained on the three datasets, and test on VA. ConvNeXt-large is used. Results are shown in Table VII. We find the improvement by [35] is limited compared with HM3D→VA in Table V. We think it is because scene structures in HM3D, NYUv2, and Hypersim differ greatly. Even for a dataset, different sequences contain varying spaces. This contrasts with their original domain on driving scenes, where most scenes are structurally similar, such as upper sky, lower road planes, and objects of cars or trees. Thus, their tasks by grouping training sets do not benefit much due to the lack of similarity in a task.

Comparison: domain-adaptation depth. We further compare to an unsupervised domain adaptation method T²Net [51] and semi-supervised domain adaptation methods ARC [49] and S2R-DepthNet [4]. These works adopt the setting SUNCG→NYUv2. We follow the training scripts by S2R-DepthNet and show the results in Table VIII.

D. Better Depth Supervision in NeRF

We show that more accurate depth from meta-initialization can better supervise the distance d a ray travels in NeRF. d is determined by the volumetric rendering rule [7]. In addition to the vanilla pixel color loss, we use distance maps

TABLE V: Cross-Dataset evaluation on general architecture. Comparison: ImageNet-initialization (no marks) and meta-initialization (+Meta).

HM3D → VA	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ConvNeXt-small	0.180	0.389	0.166	74.6	91.0	96.1
ConvNeXt-small+Meta	0.159	0.340	0.145	78.0	93.2	97.3
ConvNeXt-base	0.176	0.385	0.155	76.1	91.1	95.4
ConvNeXt-base+Meta	0.161	0.352	0.144	78.1	92.6	96.7
ConvNeXt-large	0.170	0.357	0.147	78.1	91.5	95.7
ConvNeXt-large+Meta	0.156	0.322	0.142	79.0	92.3	96.4
HM3D → NYUv2	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ConvNeXt-small	0.213	0.728	0.255	69.2	88.7	95.8
ConvNeXt-small+Meta	0.202	0.706	0.242	70.8	89.1	96.0
ConvNeXt-base	0.208	0.717	0.252	70.1	89.4	96.0
ConvNeXt-base+Meta	0.197	0.686	0.246	71.5	89.8	96.2
ConvNeXt-large	0.192	0.690	0.244	72.0	90.4	96.4
ConvNeXt-large+Meta	0.186	0.654	0.239	73.3	90.7	96.6
HM3D → Replica	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ConvNeXt-small	0.138	0.321	0.095	84.5	93.9	96.6
ConvNeXt-small+Meta	0.124	0.282	0.089	85.7	95.7	98.0
ConvNeXt-base	0.134	0.316	0.097	84.6	94.2	96.6
ConvNeXt-base+Meta	0.116	0.275	0.086	87.2	96.5	98.5
ConvNeXt-large	0.137	0.307	0.099	84.3	94.0	96.6
ConvNeXt-large+Meta	0.115	0.270	0.084	87.2	96.5	98.7
Hypersim → VA	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ConvNeXt-small	0.215	0.404	0.205	68.5	90.8	96.7
ConvNeXt-small+Meta	0.207	0.398	0.196	70.4	91.3	97.0
ConvNeXt-base	0.201	0.393	0.188	71.3	91.8	97.3
ConvNeXt-base+Meta	0.194	0.365	0.173	72.8	92.8	97.8
ConvNeXt-large	0.198	0.369	0.175	73.0	92.0	97.1
ConvNeXt-large+Meta	0.183	0.355	0.164	74.6	93.5	97.8
Hypersim → NYUv2	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ConvNeXt-small	0.165	0.598	0.225	75.7	94.3	98.5
ConvNeXt-small+Meta	0.155	0.575	0.208	77.8	95.1	98.8
ConvNeXt-base	0.150	0.549	0.200	79.6	95.6	98.9
ConvNeXt-base+Meta	0.141	0.524	0.192	80.3	96.0	99.0
ConvNeXt-large	0.149	0.542	0.199	79.8	95.6	98.8
ConvNeXt-large+Meta	0.140	0.517	0.187	81.2	96.2	99.1
Hypersim → Replica	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ConvNeXt-small	0.189	0.417	0.169	72.4	92.1	97.5
ConvNeXt-small+Meta	0.178	0.404	0.150	74.5	92.7	97.5
ConvNeXt-base	0.185	0.409	0.154	74.1	92.6	97.4
ConvNeXt-base+Meta	0.173	0.399	0.142	75.6	93.3	97.9
ConvNeXt-large	0.172	0.394	0.145	75.8	93.2	97.7
ConvNeXt-large+Meta	0.165	0.380	0.132	77.0	94.0	98.1

TABLE VI: Cross-Dataset evaluation using dedicated depth estimation networks. Our meta-initialization (+Meta) can be plugged into several methods to stably improve them.

Hypersim → Replica	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
BTS-ResNet101 [19]	0.214	0.488	0.211	69.9	89.8	96.3
BTS-ResNet101+Meta	0.194	0.463	0.177	71.2	90.5	96.8
DepthFormer [21]	0.185	0.415	0.166	72.9	92.3	97.4
DepthFormer+Meta	0.171	0.394	0.144	74.5	92.8	97.6
DPT-hybrid [27]	0.197	0.455	0.175	71.2	90.7	96.7
DPT-hybrid+Meta	0.165	0.391	0.154	75.7	93.5	97.8
DPT-large [27]	0.172	0.401	0.158	75.4	93.3	97.6
DPT-large+Meta	0.161	0.374	0.134	77.4	94.4	98.2
AdaBins [1]	0.210	0.445	0.195	70.2	90.1	96.6
AdaBins+Meta	0.193	0.427	0.177	72.0	92.1	97.4
GLPDepth [18]	0.188	0.418	0.165	72.9	92.4	97.6
GLPDepth+Meta	0.170	0.396	0.149	74.8	93.1	97.7
NDDepth [32]	0.165	0.380	0.144	76.1	93.5	98.0
NDDepth+Meta	0.159	0.371	0.136	76.7	93.8	97.9
Hypersim → NYUv2	AbsRel↓	RMSE↓	SILog↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
BTS-ResNet101 [19]	0.187	0.641	0.243	72.3	90.8	95.8
BTS-ResNet101+Meta	0.170	0.618	0.226	74.5	92.8	97.5
DepthFormer [21]	0.169	0.608	0.225	75.1	93.9	98.2
DepthFormer+Meta	0.151	0.572	0.204	78.1	94.3	98.3
DPT-hybrid [27]	0.149	0.580	0.204	78.9	94.8	98.3
DPT-hybrid+Meta	0.135	0.550	0.182	81.5	96.6	99.1
DPT-large [27]	0.136	0.530	0.180	82.3	96.2	98.8
DPT-large+Meta	0.130	0.507	0.176	83.1	96.6	99.0
AdaBins [1]	0.188	0.642	0.241	72.6	91.2	96.6
AdaBins+Meta	0.174	0.622	0.231	74.1	92.7	97.5
GLPDepth [18]	0.169	0.604	0.223	75.3	93.9	98.2
GLPDepth+Meta	0.155	0.577	0.208	77.8	94.2	98.4
NDDepth [32]	0.141	0.557	0.195	80.4	96.0	98.8
NDDepth+Meta	0.134	0.545	0.184	81.3	96.3	98.8

d^* , converted from monocular predicted depth by camera intrinsics, to supervise the training by $\mathcal{L}_D = |d^* - d|$. The

TABLE VII: Comparison to meta-learning on single-image depth. We aggregate Hypersim, HM3D, and NYUv2 as the training sets and test on VA.

Multiple → VA	AbsRel↓	RMSE↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Learn to adapt [35]	0.318	0.133	79.5	92.4	96.4
Our Meta-Initialization	0.280	0.109	82.2	93.4	97.4

TABLE VIII: Cross-Dataset comparison to domain adaptation. We follow the settings from S2R-DepthNet [4] and compare with their reported benchmark.

SUNCG → NYUv2	AbsRel↓	RMSE↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
T ² Net [51]	0.203	0.738	67.0	89.1	96.6
ARC [49]	0.186	0.710	71.2	91.7	97.7
S2R-DepthNet [4]	0.196	0.662	69.5	91.0	97.2
Our Meta-Initialization	0.177	0.635	72.8	92.8	97.8

TABLE IX: Results on depth-supervised NeRF. We test on Replica 'room-0', 'room-1', 'room-2', 'office-0', 'office-1', and 'office-2' environments.

Environment	w/o meta-initialization		w/ meta-initialization	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Room-0	29.99	0.818	30.92	0.837
Room-1	34.55	0.928	34.87	0.931
Room-2	36.68	0.956	37.46	0.961
Office-0	38.67	0.963	39.29	0.968
Office-1	36.20	0.943	36.87	0.946
Office-2	42.65	0.964	42.67	0.965

experiment is conducted on Replica's office-0 environment with 180 training views. After 30K training steps, we obtain NeRF-rendered views and calculate the commonly-used image quality metrics (PSNR and SSIM, the higher the better). We use ConvNeXt-base to predict d^* . Table IX shows a comparison made between with and without meta-initialization, and Fig. 6 shows a visual comparison.

V. CONCLUSION AND DISCUSSION

This work investigates how meta-learning helps in pure single-image depth estimation and closely analyzes improvements. Meta-learning can learn smooth depth from global context (IV-A). It is a better initialization to obtain higher model generalizability verified on intra-dataset, zero-shot cross-dataset, and 3D representation evaluations (IV-B, IV-C, IV-D). From depth's perspective, this work studies an effective learning scheme to gain generalizability and validates with the proposed cross-dataset protocols. From meta-learning's perspective, this work proposes fine-grained task for a challenging pure single-image setting and studies a complex and practical goal of pixel-level real-valued regression.

Discussion: large foundation models. Foundation models generally require a large corpus of pretrained data and still need a finetune set to adapt to the downstream task. In this case, another RGB-D dataset is required for adaptation. In contrast, we only require an RGB-D set. If the size of a finetune set is not large enough, a large model may easily overfit by overparameterization and show worse generalization. We experiment finetuning foundation models using ConvNeXt-XXLarge and ViT-L/14 and compare with our meta-initialization. They are CLIP weights as initial parameters and further tuned on ImageNet22K. Finetuning those models does not win over our +Meta on Replica → VA and only show marginal gain on HM3D → VA.

	ConvNeXt-XXLarge+finetuning	ViT-L/14+finetuning	ConvNeXt-Base+Meta
AbsRel↓			
Replica → VA	0.437	0.434	0.430
HM3D → VA	0.160	0.162	0.163

REFERENCES

- [1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2, 6, 7

- [2] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. [2](#)
- [3] Zhiying Cao, Tengfei Zhang, Wenhui Diao, Yue Zhang, Xiaode Lyu, Kun Fu, and Xian Sun. Meta-seg: A generalized meta-learning framework for multi-class few-shot semantic segmentation. *IEEE Access*, 2019. [2](#)
- [4] Xiaotian Chen, Yuwang Wang, Xuejin Chen, and Wenjun Zeng. S2r-depthnet: Learning a generalizable depth-specific structural representation. In *CVPR*, 2021. [2](#), [6](#), [7](#)
- [5] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaocong Wang. Meta-baseline: exploring simple meta-learning for few-shot learning. In *ICCV*, 2021. [2](#)
- [6] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaocong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *ICCV*, 2021. [5](#)
- [7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022. [1](#), [6](#)
- [8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. [1](#), [2](#), [3](#)
- [9] Fabrizio Flacco, Torsten Kröger, Alessandro De Luca, and Oussama Khatib. A depth space approach to human-robot collision avoidance. In *ICRA*, 2012. [1](#)
- [10] Ning Gao, Hanna Ziesche, Ngo Anh Vien, Michael Volpp, and Gerhard Neumann. What matters for meta-learning vision regression tasks? In *CVPR*, 2022. [2](#)
- [11] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. [6](#)
- [12] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. [4](#), [6](#)
- [13] Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In *CVPR*, 2021. [2](#)
- [14] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *CVPR*, 2021. [2](#)
- [15] Xu He, Jakub Sygnowski, Alexandre Galashov, Andrei A Rusu, Yee Whye Teh, and Razvan Pascanu. Task agnostic continual learning via meta learning. *arXiv preprint arXiv:1906.05201*, 2019. [1](#)
- [16] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *ICANN*, 2001. [2](#)
- [17] Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. [1](#), [2](#)
- [18] Doyeon Kim, Woonghyun Ga, Pyunghwan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. [2](#), [6](#), [7](#)
- [19] 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](#), [6](#), [7](#)
- [20] Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, and Wenxian Yu. Structdepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *ICCV*, 2021. [2](#)
- [21] 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](#), [6](#), [7](#)
- [22] Xinyu Luo, Jiaming Zhang, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelhagen. Towards robust semantic segmentation of accident scenes via multi-source mixed sampling and meta-learning. In *CVPRW*, 2022. [2](#)
- [23] Hugo Nascimento, Martin Mujica, and Mourad Benoussaad. Collision avoidance interaction between human and a hidden robot based on kinect and robot data fusion. *IEEE Robotics and Automation Letters*, 6(1):88–94, 2020. [1](#)
- [24] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. [1](#), [2](#), [3](#)
- [25] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, 2020. [2](#)
- [26] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. *NeurIPS Datasets and Benchmarks Track*, 2021. [2](#), [4](#)
- [27] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. [1](#), [2](#), [6](#), [7](#)
- [28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. [1](#), [2](#)
- [29] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *ICCV*, 2021. [2](#), [4](#)
- [30] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, 2022. [1](#)
- [31] Philipp Schmidt, Nikolaus Vahrenkamp, Mirko Wächter, and Tamim Asfour. Grasping of unknown objects using deep convolutional neural networks based on depth images. In *ICRA*, 2018. [1](#)
- [32] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Ndddepth: Normal-distance assisted monocular depth estimation. In *ICCV*, 2023. [6](#), [7](#)
- [33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. [4](#)
- [34] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. [2](#), [4](#)
- [35] Qiyu Sun, Gary G Yen, Yang Tang, and Chaoqiang Zhao. Learn to adapt for monocular depth estimation. *arXiv preprint arXiv:2203.14005*, 2022. [2](#), [6](#), [7](#)
- [36] Lei Tai, Jingwei Zhang, Ming Liu, and Wolfram Burgard. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *ICRA*, 2018. [1](#)
- [37] Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. Depth-aware vision-and-language navigation using scene query attention network. In *ICRA*, 2022. [1](#)
- [38] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *CVPR*, 2019. [2](#), [3](#)
- [39] Ulrich Viereck, Andreas Pas, Kate Saenko, and Robert Platt. Learning a visuomotor controller for real world robotic grasping using simulated depth images. In *CoRL*, 2017. [1](#)
- [40] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuo Chen Su. Toward practical monocular indoor depth estimation. In *CVPR*, 2022. [1](#), [2](#), [4](#)
- [41] Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, et al. Improving generalization in meta-learning via task augmentation. In *ICML*, 2021. [3](#)
- [42] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. [2](#)
- [43] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. [1](#), [2](#), [6](#)
- [44] Weihao Yuan, Xiaodong Gu, Zuo Zhou Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *CVPR*, 2022. [2](#)
- [45] Chi Zhang, Wei Yin, Gang Yu, Zhibin Wang, Tao Chen, Bin Fu, Joey Tianyi Zhou, and Chunhua Shen. Robust geometry-preserving depth estimation using differentiable rendering. In *ICCV*, 2023. [6](#)
- [46] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2018. [3](#)
- [47] Zhenyu Zhang, Stéphane Lathuilière, Andrea Pilzer, Nicu Sebe, Elisa Ricci, and Jian Yang. Online adaptation through meta-learning for stereo depth estimation. *arXiv preprint arXiv:1904.08462*, 2019. [1](#), [2](#)
- [48] Zhenyu Zhang, Stéphane Lathuilière, Elisa Ricci, Nicu Sebe, Yan Yan, and Jian Yang. Online depth learning against forgetting in monocular videos. In *CVPR*, 2020. [1](#), [2](#), [3](#)
- [49] Yunhan Zhao, Shu Kong, Daeyun Shin, and Charless Fowlkes. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In *CVPR*, 2020. [6](#), [7](#)
- [50] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021. [2](#)
- [51] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *ECCV*, 2018. [6](#), [7](#)