

# Kick Back & Relax++: Scaling Beyond Ground-Truth Depth with SlowTV & CribsTV

Jaime Spencer<sup>1\*</sup>, Chris Russell<sup>2</sup>, Simon Hadfield<sup>1</sup>,  
Richard Bowden<sup>1</sup>

<sup>1</sup>CVSSP, University of Surrey.

<sup>2</sup>Oxford Internet Institute.

\*Corresponding author(s). E-mail(s): [j.spencermartin@surrey.ac.uk](mailto:j.spencermartin@surrey.ac.uk);  
Contributing authors: [chris.russell@oii.ox.ac.uk](mailto:chris.russell@oii.ox.ac.uk); [s.hadfield@surrey.ac.uk](mailto:s.hadfield@surrey.ac.uk);  
[r.bowden@surrey.ac.uk](mailto:r.bowden@surrey.ac.uk);

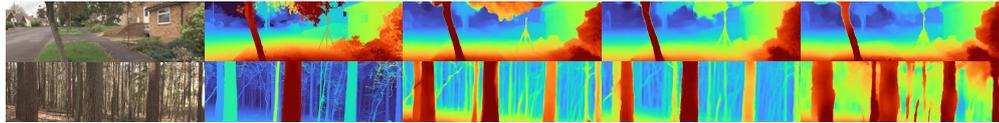
## Abstract

Self-supervised learning is the key to unlocking generic computer vision systems. By eliminating the reliance on ground-truth annotations, it allows scaling to much larger data quantities. Unfortunately, self-supervised monocular depth estimation (SS-MDE) has been limited by the absence of diverse training data. Existing datasets have focused exclusively on urban driving in densely populated cities, resulting in models that fail to generalize beyond this domain.

To address these limitations, this paper proposes two novel datasets: SlowTV and CribsTV. These are large-scale datasets curated from publicly available YouTube videos, containing a total of 2M training frames. They offer an incredibly diverse set of environments, ranging from snowy forests to coastal roads, luxury mansions and even underwater coral reefs. We leverage these datasets to tackle the challenging task of zero-shot generalization, outperforming every existing SS-MDE approach and even some state-of-the-art supervised methods.

The generalization capabilities of our models are further enhanced by a range of components and contributions: 1) learning the camera intrinsics, 2) a stronger augmentation regime targeting aspect ratio changes, 3) support frame randomization, 4) flexible motion estimation, 5) a modern transformer-based architecture. We demonstrate the effectiveness of each component in extensive ablation experiments. To facilitate the development of future research, we make the datasets, code and pretrained models available to the public at [https://github.com/jspenmar/slowtv\\_monodepth](https://github.com/jspenmar/slowtv_monodepth).

**Keywords:** Monocular Depth Estimation, Self-Supervised Learning, Zero-shot Generalization, Large-scale Dataset Curation



(a) Image (b) Ground-truth (c) KBR++(Ours) (d) KBR (Ours) (e) Baseline [1]

**Fig. 1: Zero-shot generalization.** Our SS-MDE models generalize to a wide range of complex environments. These models are trained on the novel large-scale SlowTV and CribsTV datasets and can match or even surpass supervised SotA.

## 1 Introduction

Reliably reconstructing the 3-D structure of the world is a crucial component in many real-world applications, such as autonomous driving, robotics, camera relocalization or augmented reality. While traditional depth estimation algorithms relied on correspondence estimation and triangulation, recent research has shown it is possible to train a neural network to reconstruct a scene from only a single image. Despite being an ill-posed task due to the scale ambiguity, monocular depth estimation (MDE) has become of great interest due to its flexibility and applicability to many fields.

Recent supervised MDE [2–5] approaches have achieved impressive results, but are limited both by the availability and quality of annotated datasets. LiDAR data is expensive to collect and frequently exhibits boundary artifacts due to motion correction. Meanwhile, Structure-from-Motion (SfM) is computationally expensive and can produce noisy, incomplete or incorrect reconstructions.

Self-supervised learning (SSL) should be able to scale to much larger data quantities, since only monocular or stereo video is required for training. These models instead leverage photometric constraints, using the predicted depth and motion to warp adjacent frames and synthesize the target image. However, in practice, existing SS-MDE models [6–9] rely exclusively on automotive datasets [10–12]. This lack of variety significantly impacts their generalization capabilities and results in failures when applied to natural or indoor scenes. Moreover, despite being fully convolutional, these models struggle to generalize to different image sizes.

We argue that the lack of diversity is due to the challenges of data collection, with new datasets aiming to provide high-quality ground-truth annotations that can be used for testing [13, 14]. Whilst this is important to accurately evaluate the performance of models, it also places strong limitations on the achievable scale for the training splits. In this paper, we instead focus on creating datasets that specifically target self-supervised learning, exploiting the fact that no ground-truth annotations are required. Combined with our additional contributions, we train self-supervised models capable of zero-shot generalization beyond the automotive domain. Our models significantly outperform all existing SS-MDE approaches and can even match or outperform State-of-the-Art (SotA) *supervised* techniques.

A preliminary version of this work [15] was published at the International Conference on Computer Vision. This paper introduced the **SlowTV** dataset, composed of 1.7M frames from 40 curated YouTube videos. These videos featured a wide diversity of

settings, including seasonal hiking, scenic driving and scuba diving. SlowTV provided our models with the general foundation for natural scenes, such as forests, mountainous terrains or deserts. Our dataset was combined with Mannequin Challenge [16] and Kitti [10], which targeted indoor scenes with humans and urban driving.

Our preliminary work introduced several contributions that further maximized zero-shot performance, whilst not increasing the complexity and computational requirements of the model. For instance, predicting the camera intrinsics prevented performance drops resulting from training with inaccurate intrinsics estimated via SfM. Meanwhile, the aspect ratio augmentation (AR-Aug) diversified the distribution of image shapes seen during training and facilitated transfer across datasets.

This paper presents several significant extensions to Kick Back & Relax (KBR) [15] and thus introduces KBR++. Despite performing on par with several supervised SotA models, there was still a gap when evaluating on indoor datasets due to the exclusive reliance on Mannequin Challenge as a source of indoor data. Following the design philosophy of our original work, we introduce the **CribsTV** dataset. This is an extension to SlowTV consisting of 330k images from curated YouTube real estate virtual tours. As such, this new dataset focuses on bedrooms, living rooms and kitchens and is complemented by gardens, swimming pools and aerial outdoor shots. This further reduces the gap between supervision and self-supervision in indoor settings.

We complement this novel dataset with additional augmentation strategy experiments. Since its inception [6, 7, 17], SS-MDE has restricted itself to simple augmentations, such as color jittering and horizontal flipping. However, recent contrastive SSL [18–20] research has shown the benefits of more aggressive augmentation schemes. We demonstrate this is also the case in SS-MDE and incorporate RandAugment [21] and CutOut [22] into the pipeline. Finally, we modernize the depth network architecture with the transformer-based backbones from DPT [3] and perform several new ablation experiments that give further insight into the performance of our models. Our updated models outperform all (self-)supervised approaches, except DPT-BEiT, despite not requiring any ground-truth annotations.

The contributions of our works can be summarized as:

1. We introduce a novel SS-MDE dataset of SlowTV YouTube videos and complement it with CribsTV, resulting in a total of 2M training images. This dataset features an incredibly diverse set of environments, including worldwide seasonal hiking, scenic driving, scuba diving and real estate tours.
2. We leverage SlowTV and CribsTV to train zero-shot models that generalize across multiple datasets. We additionally apply these models to the task of map-free relocalization, demonstrating their applicability to real-world settings.
3. We introduce a range of contributions and best-practices that further maximize generalization. This includes: camera intrinsics learning, an aspect ratio augmentation, stronger photometric augmentations, support frame randomization, flexible motion estimation and a modernized depth network architecture. We demonstrate the effectiveness of these contributions in detailed ablation experiments.
4. We close the performance gap between supervision and self-supervision, greatly furthering the SotA in SS-MDE. We share these developments with the community, making the datasets, pretrained models and code available to the public.

## 2 Related Work

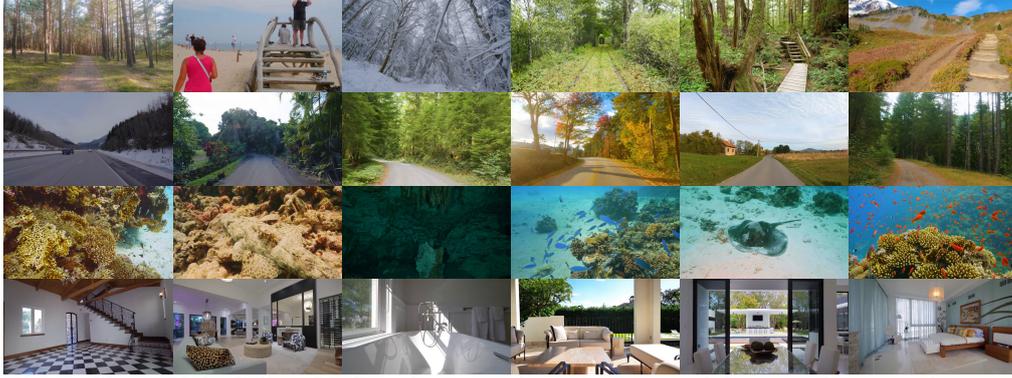
Instead of using ground-truth depth annotations from LiDAR or RGB-D sensors, self-supervised monocular depth estimation relies exclusively on photometric consistency constraints. The seminal approach by Garg *et al.* [6] combined the known baseline between stereo pairs with the predicted depth to obtain correspondences and perform view synthesis. Monodepth [17] complemented this with the virtual stereo consistency loss. Performance was further improved by introducing differentiable bilinear interpolation [23] and a reconstruction loss based on SSIM [24]. 3Net [25] extended the virtual stereo consistency into a trinocular setting.

To extend this formulation to the monocular domain, it is necessary to replace the fixed stereo baseline with a network to predict the relative pose between frames. This was first proposed by SfM-Learner [7] and extended by DDVO [26], which introduced a differentiable DSO module [27]. Purely monocular approaches are sensitive to dynamic objects, as their additional motion is not accounted-for by the relative pose estimation. This results in incorrect correspondences, which further lead to inaccurate depth predictions. Therefore, future research aimed to minimize this impact via predictive masking [7], uncertainty estimation [28–30], optical flow [31–33] or motion masks [34–36].

Several works have instead focused on improving the robustness of the photometric loss. One notable example is Monodepth2 [8], which introduced the minimum reconstruction loss and static-pixel automasking. FeatDepth [37] applied the same view synthesis to dense feature descriptors, which should be invariant to viewpoint and illumination conditions. DeFeat-Net [38] learned the feature descriptors simultaneously, while Shu *et al.* [39] used intermediate autoencoder representations. Others complemented the photometric loss with semantic segmentation [40–42] or geometric constraints [43–45]. Finally, it is also common to introduce proxy depth label regression obtained from SLAM [28, 46], synthetic data [47] or hand-crafted disparity [48, 49].

The encoder network architecture has been improved by introducing 3-D (un)packing blocks, positional encoding [50], transformers [51] or high-resolution networks [52]. Updated decoders have focused on sub-pixel convolutions [53], self-attention [52, 54, 55] and progressive skip connections [9]. Akin to supervised MDE developments [4, 56], Johnston *et al.* [54] and Bello *et al.* [50, 57] obtained improvements by representing depth as a discrete volume.

So far, these contributions have only been tested on automotive datasets, such as Kitti [10], CityScapes [58] or DDAD [12]. Recent benchmarks and challenges [1, 59, 60] have shown that these models fail to generalize beyond this restricted training domain. Meanwhile, recent supervised models [2, 3, 60] have leveraged collections of datasets to improve zero-shot generalization. In this paper, we aim to close the gap between supervised and self-supervised MDE in the challenging task of zero-shot generalization. This is achieved by greatly increasing the diversity and scale of the training data by leveraging unlabeled videos from YouTube, without requiring manual annotation or expensive pre-processing.



**Fig. 2: SlowTV & CribsTV.** Sample images from the proposed datasets, featuring diverse scenes for hiking, driving, scuba diving and real estate. The datasets consist of 45 videos curated from YouTube with a total of 2M training frames. Diversifying the training data allows our SS-MDE models to generalize to unseen datasets. We make the list of URLs and tools to process publicly available.



**Fig. 3: SlowTV Map.** We show the map location for each sequence in SlowTV. The distribution of locations ensures that the training data is highly diverse. **Green**=Natural, **Red**=Driving, **Blue**=Underwater.

### 3 Datasets

The proposed **SlowTV** and **CribsTV** datasets consist of 45 videos curated from YouTube, totaling more than 140 hours of content and 2 million training images. As shown in Table 1, this is an order of magnitude more data than any commonly used SS-MDE dataset. SlowTV contains three main outdoor categories (hiking, driving and scuba diving), while CribsTV focuses exclusively on real estate properties. When combined, these datasets provide an incredibly diverse set of training scenes for our models, allowing us to tackle the challenging task of zero-shot generalization.

**Table 1: Datasets Comparison.** The top half shows commonly used SS-MDE training datasets. SlowTV and CribsTV greatly diversify the training environments and scale to much larger data quantities. The bottom half summarizes the testing datasets used in our *zero-shot generalization* evaluation.

	Urban	Natural	Scuba	Indoor	Depth	Acc	Density	#Img
Kitti [10, 61] <sup>†</sup>	✓	✗	✗	✗	LiDAR	High	Low	71k
DDAD [12]	✓	✗	✗	✗	LiDAR	Mid	Low	76k
CityScapes [11]	✓	✗	✗	✗	Stereo	Low	Mid	88k
Mannequin [16] <sup>†</sup>	✓	✗	✗	✓	SfM	Mid	Mid	115k
<b>SlowTV (Ours)<sup>†</sup></b>	✓	✓	✓	✗	✗	✗	✗	<b>1.7M</b>
<b>CribsTV (Ours)<sup>†</sup></b>	✓	✗	✗	✓	✗	✗	✗	<b>330k</b>
Kitti [10, 61]	✓	✗	✗	✗	LiDAR	High	Low	652
DDAD [12]	✓	✗	✗	✗	LiDAR	Mid	Low	1k
Sintel [62]	✗	✓	✗	✗	Synth	High	High	1064
SYNS-Patches [1, 13]	✓	✓	✗	✓	LiDAR	High	High	775
DIODE [63]	✓	✗	✗	✓	LiDAR	High	High	771
Mannequin [16]	✓	✗	✗	✓	SfM	Mid	Mid	1k
NYUD-v2 [64]	✗	✗	✗	✓	Kinect	Mid	High	654
TUM-RGBD [65]	✗	✗	✗	✓	Kinect	Mid	High	2.5k

<sup>†</sup>Datasets used to train our networks.

Hiking videos target natural scenes, such as forest, mountains, deserts or fields, which are non-existent in current datasets. Our driving split seeks to complement existing automotive datasets, which tend to focus on urban driving in densely populated cities [10–12, 66–69]. The proposed split instead features videos from scenic routes, traversing forest, mountainous or coastal roads with sparse traffic. We also feature a variety of weather and seasonal conditions. Underwater scuba diving represents yet another previously unexplored domain, which further increases data diversity. Finally, real estate properties are a natural counterpart to the previous outdoor data. They also complement the Mannequin Challenge [16], which primarily focuses on human beings in indoor settings, rather than the indoor scenes themselves.

The proposed videos were collected from a diverse set of locations and conditions, as illustrated in Figure 3. This includes the USA, Canada, the Balkans, Eastern Europe, Indonesia and Hawaii, and conditions such as rain, snow, autumn and summer. Since CribsTV contains a large number of individual properties, it is challenging to obtain accurate information about each of their locations. However, they are predominantly located in the USA. Figure 2 shows sample frames from each of the available categories, illustrating the dataset’s incredible diversity.

Videos were downloaded at HD resolution ( $720 \times 1280$ ) and extracted at 10 FPS to reduce storage, while still providing smooth motion and large overlap between adjacent frames. In the case of SlowTV, only 100 consecutive frames out of every 250 were retrained. This reduces the self-similarity between training samples and keeps the dataset size tractable. The final SlowTV contains a total of 1.7M images, composed of 1.1M natural, 400k driving and 180k underwater.

Since we target SS-MDE, the only annotations required are the camera intrinsic parameters, which can be estimated using COLMAP [70]. However, as discussed in

Section 4.2, it is possible to let the network jointly optimize camera parameters alongside depth and motion. We find that this is more robust and improves performance compared to training with potentially inaccurate COLMAP intrinsics.

CribsTV instead consists of highly-produced cinematic house tours. In practice, this means they include cuts between shots of different viewpoints or each room. Some of these shots may also be unsuitable for SS-MDE, containing zooming/focusing/blurring effects or static shots. As such, it was first necessary to split each video into its individual scenes using an off-the-shelf scene detector [71]. We then performed a quick manual check to filter out potentially invalid scenes based on their first frame. The final dataset contains 330k training frames.

CribsTV also presents additional challenges when estimating the camera intrinsics. Due to the short shot duration (on average 5 seconds) and lack of overlap between scenes, COLMAP is unable to produce any reconstructions. This again motivates the need for a more flexible depth estimation framework, capable of estimating camera intrinsics. This reduces the complexity of dataset collection and allows us to train with much larger quantities of diverse data.

## 4 Methodology

Monocular depth estimation aims to reconstruct the 3-D structure of the scene using only a single 2-D image projection. However, additional support frames are required in order to synthesize the target view and compute the photometric reconstruction losses that drive optimization. In the case where only stereo pairs are used [6, 17], the predicted depth is combined with the known stereo baseline to perform the view synthesis. However, if only monocular video is available, such as YouTube videos from SlowTV or CribsTV, it becomes necessary to incorporate an additional pose network to estimate the relative motion between adjacent frames [7]. A key difference between these forms of supervision is that stereo approaches can estimate metric depth, while monocular approaches are only accurate up to unknown scale and shift factors.

**Depth.** The the depth estimation network  $\Phi_D$  can be formalized as  $\hat{\mathbf{D}}_t = \Phi_D(\mathbf{I}_t)$ , where  $\hat{\mathbf{D}}_t$  is the predicted sigmoid disparity and  $\mathbf{I}_t$  is the target image. Note that the disparity map must be inverted into a depth map and appropriately scaled in order to warp the support images.

**Pose.** Similarly, the pose estimation network is represented as  $\hat{\mathbf{P}}_{t+k} = \Phi_P(\mathbf{I}_t \oplus \mathbf{I}_{t+k})$ , where  $\oplus$  is channel-wise concatenation,  $\mathbf{I}_{t+k}$  is the support frame at time offset  $k \in \{-1, +1\}$  and  $\hat{\mathbf{P}}_{t+k}$  is the predicted motion as a translation and axis-angle rotation.

### 4.1 Losses

Supervised approaches such as MiDaS [2], DPT [3] or NewCRFs [5] require ground-truth depth annotations, in the form of LiDAR, depth cameras, SfM reconstructions or stereo disparity estimation. SS-MDE [6, 8, 9, 52] instead relies on the photometric consistency across the target and support frames.

Using the predicted depth  $\mathbf{D}_t$  and motion  $\hat{\mathbf{P}}_{t+k}$ , pixel-wise correspondences between these images can be obtained as

$$\mathbf{p}'_{t+k} = \mathbf{K}\hat{\mathbf{P}}_{t+k}\mathbf{D}_t(\mathbf{p}_t)\mathbf{K}^{-1}\mathbf{p}_t, \quad (1)$$

where  $\mathbf{K}$  represents a camera’s intrinsic parameters,  $\mathbf{p}_t$  are the 2-D pixel coordinates in the target image and  $\mathbf{p}'_{t+k}$  are its 2-D reprojected coordinates onto the corresponding support frame. The synthesized support frame is then obtained via  $\mathbf{I}'_{t+k} = \mathbf{I}_{t+k} \langle \mathbf{p}'_{t+k} \rangle$ , where  $\langle \cdot \rangle$  represents differentiable bilinear interpolation [23].

The reconstruction loss is then given by the weighed combination of SSIM+L<sub>1</sub> [17], defined as

$$\mathcal{L}_{ph}(\mathbf{I}, \mathbf{I}') = \lambda \frac{1 - \mathcal{L}_{ssim}(\mathbf{I}, \mathbf{I}')}{2} + (1 - \lambda) \mathcal{L}_l(\mathbf{I}, \mathbf{I}'). \quad (2)$$

As is common, the loss balancing weight is set to  $\lambda = 0.85$ .

It is well known that purely-monocular approaches [7] are sensitive to artifacts caused by dynamic objects. This is due to the additional motion of the object being unaccounted for in the correspondence estimation procedure from (1). Recent research [34–36] aimed to solve these challenges using motion masks or semantic segmentation maps. Whilst effective, the additional annotations and labeling required makes these contributions unsuitable for the proposed framework. Instead we opt for the simple, yet effective, contributions from Monodepth2 [8]. This includes the minimum reconstruction loss and static-pixel automasking.

The minimum reconstruction loss reduces the impact of occlusions by assuming only support frames contains a correct correspondence. This frame is obtained by finding the minimum pixel-wise error across all support frames, defined as

$$\mathcal{L}_{rec} = \sum_{\mathbf{p}} \min_k \mathcal{L}_{ph}(\mathbf{I}_t, \mathbf{I}'_{t+k}), \quad (3)$$

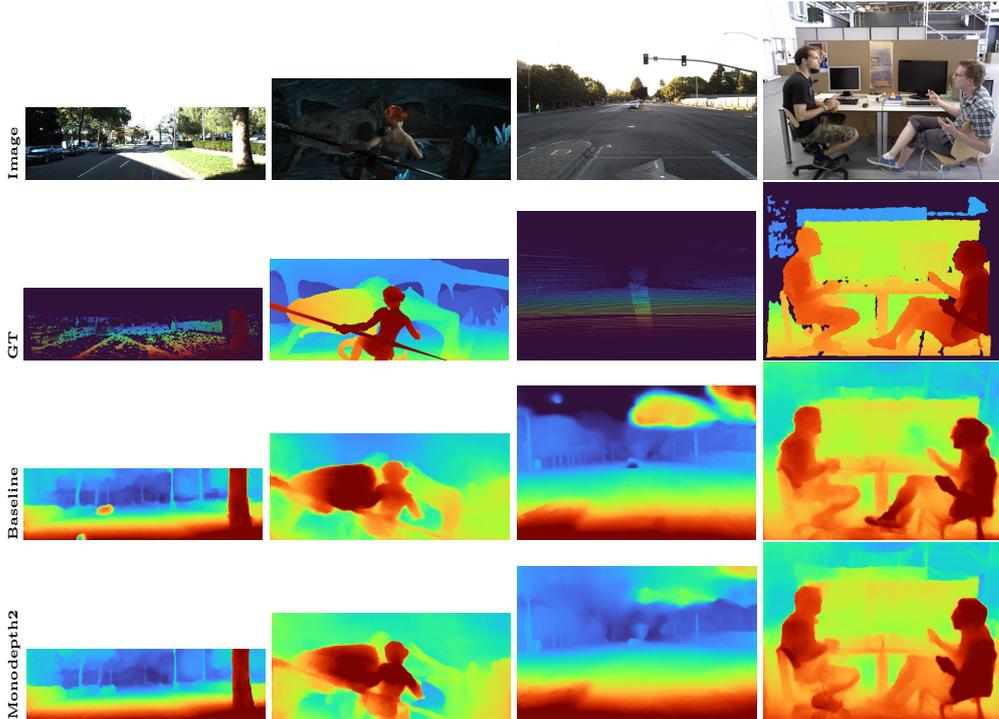
where  $\sum$  indicates averaging over a set.

Automasking instead reduces the effect of static frames and objects moving at the same speed as the camera. These objects remain static across frames, giving the impression of an infinite depth. Automasking simply removes pixels from the loss where the original *non-warped* support frame has a lower reconstruction error than the synthesized view. This is computed as

$$\mathbb{M} = \left[ \left[ \min_k \mathcal{L}_{ph}(\mathbf{I}_t, \mathbf{I}'_{t+k}) < \min_k \mathcal{L}_{ph}(\mathbf{I}_t, \mathbf{I}_{t+k}) \right] \right], \quad (4)$$

where  $\llbracket \cdot \rrbracket$  represents the Iverson brackets.

Whilst being simple to implement, Figure 4 demonstrates the effectiveness of incorporating these contributions. Finally, we complement the reconstruction loss with the common edge-aware smoothness regularization [17]. These networks and losses constitute the core baseline required to train the desired zero-shot depth estimation models. The following sections describe additional contributions that help to further maximize performance and generalization capabilities.



**Fig. 4: Dynamic Object Robustness.** Incorporating the contributions from Monodepth2 [8] mitigates artifacts from dynamic objects and improves the sharpness of depth discontinuities. This is achieved in a cost-effective manner, without requiring complex consistency losses or additional annotations.

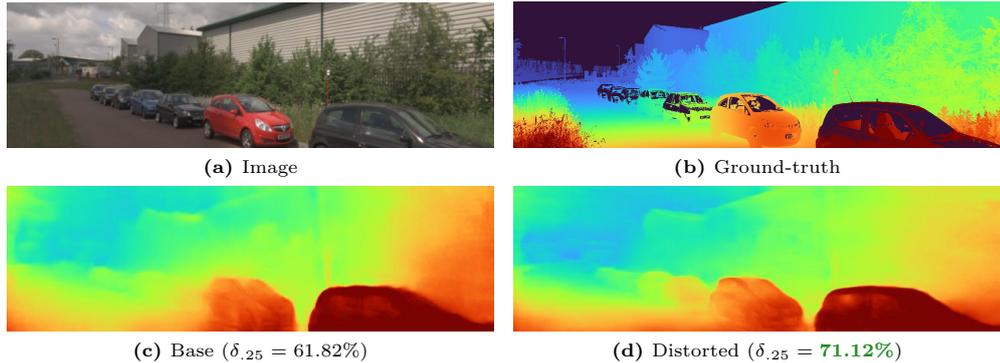
## 4.2 Learning Camera Intrinsic

Many datasets provide accurately calibrated camera intrinsic parameters. Unfortunately, in crowd-sourced [72], photo-tourism [73] or internet-curated datasets (such as SlowTV or CribsTV) these parameters are not freely available. Instead, it is common practice to rely on SfM reconstructions obtained from COLMAP [70]. Unfortunately, these reconstructions may be incorrect, incomplete or sometimes impossible to obtain. As such, especially in the case of internet-curated datasets that may continuously change or scale up in size, it would be extremely beneficial to omit these pre-processing requirements.

We take inspiration from [34, 74] learn depth, pose and camera intrinsics simultaneously. To achieve this, two additional branches are incorporated into the pose estimation network as

$$\hat{\mathbf{P}}_{t+k}, \mathbf{f}_{xy}, \mathbf{c}_{xy} = \Phi_P(\mathbf{I}_t \oplus \mathbf{I}_{t+k}), \quad (5)$$

where  $\mathbf{f}_{xy}$  and  $\mathbf{c}_{xy}$  represent the focal lengths and principal point, respectively. These parameters are predicted as normalized and scaled accordingly based on the input image size. The branch predicting the focal lengths uses a softplus activation to ensure



**Fig. 5: Image Shape Overfitting.** The same model can produce significantly different results with different resolution images. Distorting the image to the original training resolution can improve performance, despite introducing stretching/squashing artifacts. Note, for instance, the improved boundary sharpness in (d).

a positive output. The principal point instead uses sigmoid, under the assumption that it will lie within the image plane.

Incorporating these decoders results in a negligible 2MParam increase in the pose estimation network, with no additional computation required for the losses. Instead, we simply modify (1) to use the predicted intrinsics instead of the ground-truth ones. Despite this, our ablations in Section 5.4 show that this increases performance compared to training with intrinsics estimated by COLMAP.

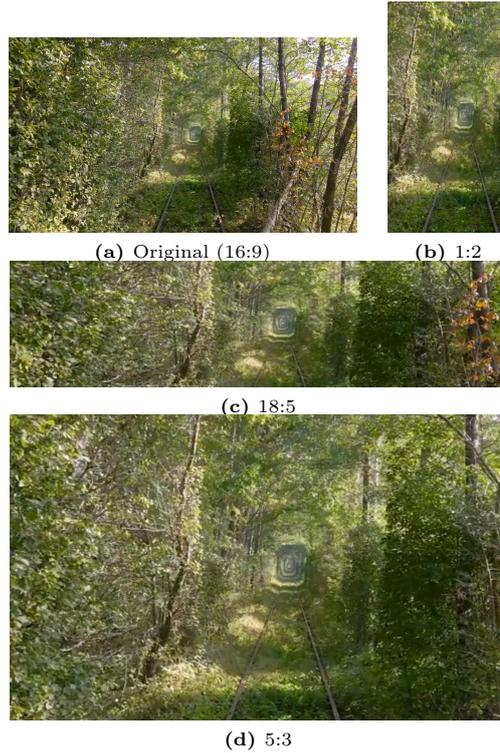
### 4.3 Augmentation Strategies

Existing research [21, 75, 76] has shown the importance of incorporating more sophisticated augmentation strategies. This is especially the case in SSL, which relies exclusively on the the diversity of the available data. However, existing SS-MDE (SS-MDE) approaches use only traditional augmentations such as color jittering and horizontal flipping. This section describes the additional augmentations incorporated into our training regime to further boost the generalization capabilities of our final models.

#### 4.3.1 Aspect Ratio

Dense predictions networks, such as the depth network used in this paper, can process images of arbitrary shape. However, when trained only on a single dataset with a fixed image size, it is common for them to overfit to this size, resulting in poor performance on out-of-dataset examples. An example of this effect can be seen in Figure 5, where first resizing the image to match the training aspect ratio results in better performance, despite introducing stretching or squashing artifacts.

We overcome this by introducing an augmentation that randomizes the image sizes and aspect ratios seen by the network during training. The proposed aspect ratio augmentation (AR-Aug) consists of two components: cropping



**Fig. 6: AR-Aug.** Sample training images generated using the proposed augmentation strategy. This augmentation prevents overfitting to image shapes and increases the diversity of images seen by the network.

and resizing. The first stage uniformly samples from a set of predefined common aspect ratios<sup>1</sup>. A random crop is generated using this aspect ratio, covering 50-100% of original image height or width. The resizing stage ensures that the final crop has roughly the same number of pixels as the original input image. Figure 6 shows training samples obtained using this procedure.

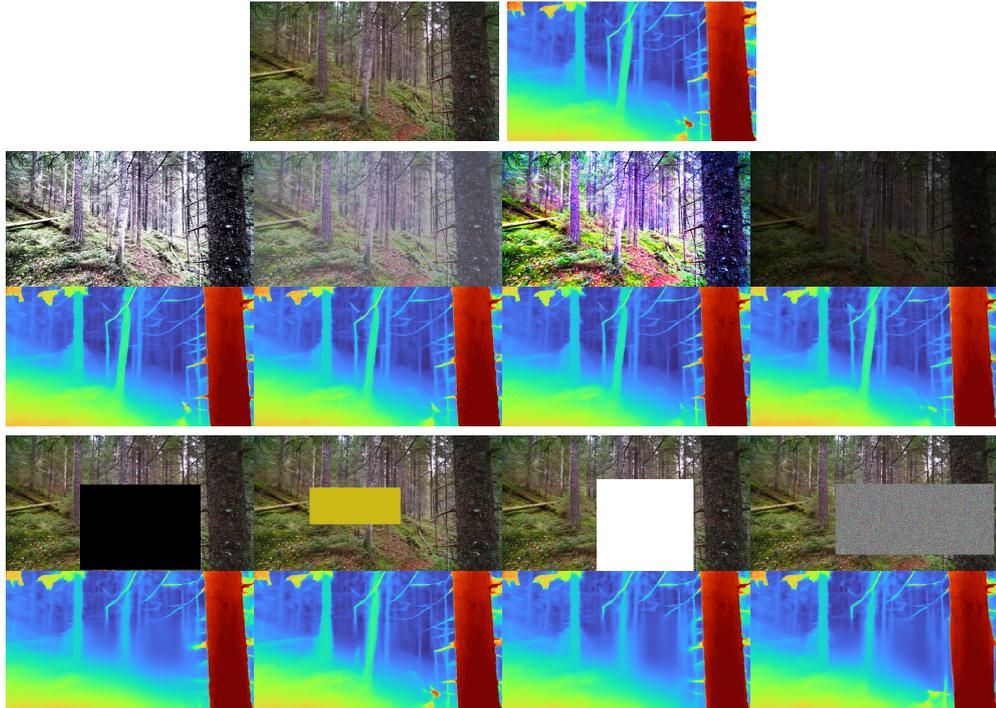
This augmentation is applied at the mini-batch level to ensure all images are the same shape. If using ground-truth intrinsics, these are rescaled accordingly. AR-Aug has the effect of drastically increasing the diversity of shapes and sizes seen by the network and prevents overfitting to a single shape.

### 4.3.2 RandAugment

We additionally proposed to complement color jittering with RandAugment [21]. This strategy sequentially applies a random combination of photometric and geometric augmentations. Since MDE requires accurate re-projections across a sequence of images we remove the geometric augmentations (*e.g.* translate, rotate and shear) and focus

---

<sup>1</sup>Portrait: 6:13, 9:16, 3:5, 2:3, 4:5, 1:1. Landscape: 5:4, 4:3, 3:2, 14:9, 5:3, 16:9, 2:1, 24:10, 33:10, 18:5.



**Fig. 7: Photometric Augmentations.** Sample augmentations produced by RandAugment [21] and CutOut [22]. The resulting model is highly robust to changing illumination conditions and is capable of filling in masked regions from the surrounding context. This can be seen in large regions of the ground-plane and connecting the tree trunk.

purely on photometric ones. The set of possible augmentations is thus reduced to: identity (*i.e.* no augmentation), auto-contrast, equalization, sharpness, brightness, color and contrast. At each training iteration, a random subset of three augmentations is chosen and applied to both the target and support frames. Sample augmented images using this strategy can be found in Figure 7.

### 4.3.3 CutOut

Inspired by the recent success of transformer token-masking augmentations [20, 77, 78], we additionally propose to re-introduce CutOut augmentations [22]. While CutOut was originally used to boost holistic tasks like classification, it can also be applied to dense prediction tasks. In this case, the objective is to teach the network to predict the depth for a missing region in the image, based only on the context surrounding it. As such, these models should learn to incorporate additional context cues and be more robust to test-time artifacts such as reflections or highlights.

To further increase the variability of the augmentations, we implement various fill modes for the masked-out regions: white, black, grayscale, RGB and random. A

**Table 2: Model Complexity.** KBR retains the architecture from Monodepth Benchmark [1], while KBR++ matches DPT [3]. Despite being of equivalent complexity, our models greatly outperforms the SSL baselines and can close the gap to supervised performance.

	Backbone	MParam↓	FPS↑
<b>KBR (Ours) [15]</b>	ConvNeXt-B [79]	<b>92.65</b>	<b>61.50</b>
<b>KBR++ (Ours)</b>	BEiT-L [20]	345.01	9.60
MiDaS [2]	ResNeXt-101 [80]	<u>105.36</u>	<u>51.38</u>
DPT [3]	ViT-L [81]	344.06	14.54
DPT [3]	BEiT-L [20]	345.01	9.60
NewCRFs [82]	Swin [83]	270.44	21.61

different fill mode is randomly selected at each training iteration. Figure 7 shows examples of applying this augmentation and the robustness of the model to it.

## 5 Results

We carry out extensive evaluations to demonstrate the effectiveness of the techniques and datasets proposed in this paper. This includes the zero-shot experiments for the final models, as well as detailed ablations on each proposed component. We additionally evaluated our models in the challenging task of map-free relocalization [72] and the MDEC-2 challenge [59, 60].

Since both KBR and KBR++ were trained exclusively on monocular data, it is necessary to first align the predictions to the ground-truth metric scale. This alignment is obtained using the least-squares procedure proposed by MiDaS [3] and is applied equally to every baseline.

### 5.1 Baselines

The SotA self-supervised models were obtained from the Monodepth Benchmark [1], which use a pretrained ConvNeXt-B backbone. These models were trained exclusively on the Kitti dataset [7, 84] and are therefore also zero-shot on all other datasets.

We also compare our frameworks to current SotA *supervised* models, which require accurate ground-truth annotations to train. MiDaS [2] and DPT [3] were trained on a collection of 10/12 supervised datasets that do not overlap with our testing set (unless otherwise specified). As such, these models are also evaluated in the challenging zero-shot setting. We use the pretrained models provided in the PyTorch Hub.

NewCRFs [5] instead provides separate outdoor/indoor models trained on Kitti and NYUD-v2 respectively. We evaluate the corresponding model in a zero-shot manner depending on the dataset category. Even though NewCRFs [5] should be capable of predicting metric depth, we apply the least-squares alignment procedure to ensure that all results are comparable.

### 5.2 Implementation Details

The proposed models were implemented in PyTorch [85] and based on the Monodepth Benchmark [1]. The original KBR [15] used a ConvNeXt-B backbone [79, 86] and a

DispNet decoder [17, 58]. The pose network instead used ConvNeXt-T for efficiency. As such, these models are comparable to the SSL baselines [1].

Table 2 shows a comparison between the computational complexity of the proposed models and the supervised SotA. As seen, these models use larger transformer-based backbones [20, 81, 83]. In order to make our results more comparable, KBR++ incorporates the same architecture used by DPT [3, 20]. Our ablation experiments in Section 5.4 show the impact of different backbone architectures.

In our experiments and ablations, each model is trained using three different random seeds and we report average performance. This improves the reliability of the results and reduces the impact of non-determinism. We make the datasets, pre-trained models and training code available at [https://github.com/jspenmar/slowtv\\_monodepth](https://github.com/jspenmar/slowtv_monodepth).

The final KBR++ models were trained on a combination of SlowTV (1.7M), CribstV (330k), Mannequin Challenge (115k) and Kitti Eigen-Benchmark (71k). To make the duration of each epoch tractable and balance the contribution of each dataset, we fix the number of images per epoch to 30k, 15k, 15k and 15k, respectively. The subset sampled from each dataset varies with each epoch to ensure a high data diversity.

The models were trained for 60 epochs using AdamW [87] with weight decay  $10^{-3}$  and a base learning rate of  $10^{-4}$ , decreased by a factor of 10 for the final 20 epochs. Empirically, we found that linearly warming up the learning rate for the first few epochs stabilized learning and prevented model collapse. When training with DPT backbones, finetuning the pretrained encoders at a lower learning rate was also found to be beneficial. We use a batch size of 4 and train the models on a single NVIDIA GeForce RTX 3090.

SlowTV, CribstV and Mannequin Challenge use a base image size of  $384 \times 640$ , while Kitti uses  $192 \times 640$ . We apply horizontal flipping and color jittering, along with the proposed RandAugment [21] and CutOut [22] augmentations, each with 30% probability. AR-Aug is applied with 70% probability, sampling from 16 predefined aspect ratios previously described.

Since existing models are trained exclusively on automotive data, most of the motion occurs in a straight-line and forward-facing direction. It is therefore common practice to force the network to always make a forward-motion prediction by reversing the target and support frame if required. Handheld videos, while still primarily featuring forward motion, also exhibit more complex motion patterns. As such, removing the forward motion constraint results in a more flexible model that improves performance.

Similarly, existing models are trained with a fixed set of support frames—usually previous and next. Since SlowTV and Mannequin Challenge are mostly composed of handheld videos, the change from frame-to-frame is greatly reduced. We make the model more robust to different motion scales and appearance changes by randomizing the separation between target and support frames. In general, we sample such that handheld videos use a wider time-gap between frames, while automotive has a small time-gap to ensure there is significant overlap between frames. As shown later, this leads to further improvements and greater flexibility.

**Table 3: Learning Camera Intrinsic.** Performance when training on a single dataset (Kitti or Mannequin Challenge) and learning camera intrinsics. If the cameras are not perfectly calibrated, learning the intrinsics can improve accuracy.

	Kitti Eigen-Zhou			Mannequin			
	Rel↓	F↑	$\delta_{.25}$ ↑	Rel↓	F↑	$\delta_{.25}$ ↑	
Baseline	<a href="#">5.69</a>	<a href="#">60.88</a>	<a href="#">95.89</a>	Baseline	<a href="#">16.66</a>	<a href="#">14.20</a>	<a href="#">77.18</a>
Learn <b>K</b>	<a href="#">5.68</a>	<a href="#">60.81</a>	<a href="#">95.90</a>	Learn <b>K</b>	<a href="#">16.12</a>	<a href="#">14.77</a>	<a href="#">78.40</a>

### 5.3 Evaluation Metrics

We follow the original evaluation procedure outlined in [15] and report the following metrics per dataset:

**Rel.** Absolute relative error (%) between target  $y$  and prediction  $\hat{y}$  as  $\text{Rel} = \sum |y - \hat{y}| / y$ .

**Delta.** Prediction threshold accuracy (%) as

$$\delta_{.25} = \sum (\max(\hat{y}/y, y/\hat{y}) < 1.25).$$

**F.** Pointcloud reconstruction F-Score [88] (%) as  $F = (2PR) / (P + R)$ , where  $P$  and  $R$  are the Precision and Accuracy of the 3-D reconstruction with a correctness threshold of 10cm.

We additionally compute multi-task metrics to summarize the performance across all datasets:

**Rank.** Average ordinal ranking order across all metrics as  $\text{Rank} = \sum_m r_m$ , where  $m$  represents each available metric and  $r$  is the ordinal rank.

**Improvement.** Average relative increase or decrease in performance (%) across all metrics as  $\Delta = \sum_m (-1)^{l_m} (M_m - M_m^0) / M_m^0$ , where  $l_m = 1$  if a lower value is better,  $M_m$  is the performance for a given metric and  $M_m^0$  is the baseline’s performance.

### 5.4 Ablation

To demonstrate the effectiveness of each proposed component, we carry out a series of ablation studies. These experiments generally use a more efficient architecture (ConvNeXt-Tiny) and a smaller training dataset.

**Learning K.** Table 3 shows the benefits of learning the camera intrinsics, as outlined in Section 4.2. We train models on either Kitti or Mannequin Challenge and test them on the same dataset. If the dataset provides accurately calibrated intrinsics (Kitti), this procedure provides comparable performance. However, in the case where these were estimated by COLMAP [70], learning **K** results in a slight performance boost. This highlights the flexibility of this contribution, which requires only a negligible increase of 2MParams in the pose network, yet allows us to train without ground-truth intrinsics. This further simplifies the process of data collection and results in a framework the requires only uncalibrated monocular video to train.

**Table 4: Ablations.** We carry out ablations for each component in our framework. From top to bottom: KBR [15] contributions, network architecture, augmentations and datasets. In each case, the proposed contributions improve the zero-shot generalization of our models.

	In-Distribution						Outdoor						Indoor							
	Multi-task		Kitti		Mannequin		DDAD		DIODE		Sintel		SYNS		DIODE		NYUD-v2		TUM	
	Rank↓	$\Delta\uparrow$	Rel↓	F↑	Rel↓	F↑	Rel↓	F↑	Rel↓	$\delta_{.25}\uparrow$	Rel↓	F↑	Rel↓	F↑	Rel↓	$\delta_{.25}\uparrow$	Rel↓	$\delta_{.25}\uparrow$	Rel↓	$\delta_{.25}\uparrow$
<b>KBR</b>	<b>2.37</b>	<b>0.00</b>	<b>6.84</b>	<b>56.17</b>	14.39	17.67	<b>12.63</b>	<b>20.21</b>	<b>33.49</b>	<b>57.08</b>	33.34	40.81	<b>22.40</b>	<b>18.50</b>	14.91	80.77	<b>11.59</b>	<b>87.23</b>	<b>15.02</b>	<b>80.86</b>
Fwd $\hat{P}$	<b>2.79</b>	<b>-0.64</b>	7.27	54.61	14.36	17.52	<b>12.43</b>	<b>19.76</b>	33.52	56.97	<b>32.25</b>	41.05	<b>22.32</b>	18.45	<b>14.89</b>	80.54	11.68	87.14	15.50	80.29
$k = \pm 1$	3.16	-1.18	<b>6.79</b>	<b>55.92</b>	<b>14.17</b>	<b>17.92</b>	13.66	19.18	33.75	56.56	32.30	<b>41.14</b>	23.05	17.90	15.00	<b>80.82</b>	11.73	86.94	15.66	79.91
Fixed <b>K</b>	4.00	-2.76	7.09	55.19	14.95	17.11	14.30	18.15	<b>33.49</b>	<b>57.10</b>	33.39	<b>41.43</b>	22.56	<b>18.67</b>	15.41	79.98	12.06	86.14	15.75	80.07
No AR-Aug	3.53	-3.86	8.32	50.15	<b>14.32</b>	<b>17.87</b>	14.75	16.90	34.49	55.82	33.25	40.51	23.38	17.33	<b>14.58</b>	<b>81.69</b>	<b>11.24</b>	<b>87.77</b>	<b>14.61</b>	<b>81.76</b>
None	5.16	-7.59	8.66	48.33	14.62	17.01	18.46	15.17	34.38	55.62	<b>31.88</b>	40.27	23.32	17.97	15.15	80.30	11.88	86.35	15.55	79.81
ConvNeXt-B	3.47	0.00	<b>6.84</b>	<b>56.17</b>	14.39	17.67	<b>12.63</b>	<b>20.21</b>	33.49	57.08	33.34	40.81	<b>22.40</b>	<b>18.50</b>	14.91	80.77	11.59	87.23	15.02	80.86
ViT-B-384	4.68	-3.17	7.41	54.31	14.80	17.05	14.47	16.31	33.37	56.99	33.00	39.15	23.08	17.55	14.19	83.09	11.58	86.83	15.41	81.65
ViT-L-384	<b>2.53</b>	<b>0.41</b>	7.29	54.70	<b>14.02</b>	<b>18.10</b>	14.38	16.88	33.02	58.18	<b>30.68</b>	<b>42.01</b>	22.51	17.83	<b>13.72</b>	<b>83.98</b>	<b>10.60</b>	<b>88.96</b>	<b>14.07</b>	<b>83.24</b>
BEiT-B-384	2.79	0.32	6.93	55.28	14.23	17.98	<b>13.50</b>	<b>17.84</b>	<b>32.52</b>	<b>58.33</b>	32.41	<b>41.34</b>	22.83	17.75	13.87	83.53	11.02	88.05	14.55	82.28
BEiT-L-384	<b>1.53</b>	<b>3.91</b>	<b>6.91</b>	<b>55.60</b>	<b>12.78</b>	<b>20.37</b>	14.39	17.75	<b>32.57</b>	<b>58.73</b>	<b>30.21</b>	41.22	<b>21.76</b>	<b>18.92</b>	<b>13.53</b>	<b>84.65</b>	<b>9.63</b>	<b>91.05</b>	<b>13.60</b>	<b>85.81</b>
No Aug	3.68	0.00	6.39	<b>56.55</b>	17.12	14.28	21.54	9.85	35.62	52.92	35.38	<b>39.54</b>	24.90	15.88	<b>16.69</b>	76.55	14.41	80.27	17.50	76.18
ColorJitter	3.00	1.30	<b>6.31</b>	<b>56.38</b>	<b>16.95</b>	14.36	19.72	10.71	35.65	53.08	<b>35.16</b>	38.73	<b>24.47</b>	<b>16.16</b>	<b>16.63</b>	<b>76.93</b>	14.25	80.59	17.58	75.62
CutOut	3.68	1.55	6.51	55.93	17.25	14.35	<b>18.98</b>	<b>11.86</b>	35.73	52.81	35.75	38.86	25.17	15.92	16.71	<b>76.61</b>	14.09	80.76	17.43	75.86
RandAugment	<b>2.37</b>	<b>2.09</b>	<b>6.36</b>	56.34	<b>16.89</b>	<b>14.61</b>	19.25	11.05	<b>35.36</b>	<b>53.30</b>	<b>34.11</b>	<b>39.40</b>	24.82	15.83	16.91	76.34	<b>13.98</b>	<b>81.16</b>	<b>16.98</b>	<b>76.65</b>
All	<b>2.26</b>	<b>2.97</b>	6.36	56.30	17.09	<b>14.36</b>	<b>17.17</b>	<b>11.66</b>	<b>35.52</b>	<b>53.44</b>	35.41	39.12	<b>24.58</b>	<b>16.31</b>	16.80	76.54	<b>13.61</b>	<b>81.96</b>	<b>17.08</b>	<b>76.48</b>
Base	<b>2.47</b>	<b>0.00</b>	<b>6.49</b>	<b>55.87</b>	16.62	14.95	<b>18.59</b>	<b>12.09</b>	<b>35.15</b>	<b>54.00</b>	36.01	<b>39.42</b>	<b>24.69</b>	<b>16.20</b>	16.44	77.08	13.52	82.54	16.87	<b>77.22</b>
No Kitti	3.05	-14.82	17.98	23.73	<b>16.02</b>	<b>15.57</b>	21.42	10.85	36.04	52.89	<b>34.93</b>	<b>39.39</b>	27.32	14.50	16.41	77.35	<b>13.19</b>	<b>83.43</b>	<b>16.36</b>	<b>78.12</b>
No Mannequin	3.89	-12.62	6.83	55.20	24.42	9.12	<b>17.07</b>	<b>11.60</b>	35.73	53.12	37.04	33.11	<b>24.67</b>	15.95	17.87	73.60	18.36	70.35	22.49	63.12
No SlowTV	3.79	-8.87	7.15	54.72	16.51	14.59	30.15	6.43	36.48	51.85	35.85	37.47	27.07	13.92	<b>16.30</b>	<b>77.67</b>	13.65	82.62	17.13	77.20
With CribsTV	<b>1.79</b>	<b>0.58</b>	<b>6.67</b>	<b>55.33</b>	<b>16.36</b>	<b>15.09</b>	19.41	<b>12.59</b>	<b>35.07</b>	<b>54.25</b>	<b>34.68</b>	38.90	24.84	<b>16.02</b>	<b>16.12</b>	<b>77.79</b>	<b>12.76</b>	<b>84.21</b>	<b>16.81</b>	77.14

Highlighted cells are NOT zero-shot results. S=Stereo, M=Monocular, D=Ground-truth Depth.

**KBR.** Table 4 (1<sup>st</sup> block) shows results when removing each component proposed by the original KBR [15]. *Fwd  $\hat{\mathbf{P}}$*  represents a network forced to always make a forward-motion prediction.  $k = \pm 1$  uses a fixed set of support frames, instead of the randomization proposed in Section 5.2. *Fixed  $\mathbf{K}$*  removes the learned camera intrinsics, while *No AR-Aug* removes the proposed aspect ratio augmentation. As expected, the model with the full set of contributions performs best, while the model without any contributions is worse by 7.6%. It is interesting to note that both learning the intrinsics and AR-Aug provide the biggest performance boost. Furthermore, it is worth remembering that none of the components (except learning  $\mathbf{K}$ ) results in a increase in model complexity. However, when combined together, they significantly improve the zero-shot generalization capabilities of our models.

**Network Architecture.** To make our models more comparable with the supervised SotA, we modify our architecture to match the one from DPT [3]. These results are shown in Table 4 (2<sup>nd</sup> block). All models start from an encoder pretrained on ImageNet. Interestingly, we find that most transformer-based architectures do not significantly improve upon the baseline (ConvNeXt-B [79]), which is much more efficient. However, the largest version of BEiT [20] provides the best performance.

**Augmentations.** Table 4 (3<sup>rd</sup> block) shows the results of incorporating the more advanced augmentation strategies from Section 4.3. All variants (except *No Aug*) additionally include horizontal flipping. As shown, the default color jittering augmentation used by most existing models is slightly better than using no augmentations. However, both CutOut and RandAugment provide larger improvements. Furthermore, the final row (All) demonstrates that these improvements are cumulative and that the model benefits from combining multiple augmentation strategies.

**Datasets.** The final ablation experiment explores the effect of removing or adding each training dataset, shown in Table 4 (4<sup>th</sup> block). As expected, removing each dataset results in a significant drop in performance. Whilst SlowTV seems to impact performance the least, we believe this is due to the lack of natural data within our evaluation set. This is supported by the fact that SlowTV has the most impact on SYNS-Patches, which is the only dataset with natural scenes. Finally, incorporating the CribsTV dataset proposed in this paper slightly increases the overall performance, especially in indoor scenes. It is worth remembering this variant also includes Mannequin Challenge, which results in a less drastic improvement. Furthermore, training on more varied data is likely to be more beneficial when combined with larger models with better generalization capacity.

## 5.5 In-distribution

We compare our final models to the (self-)supervised SotA on the two datasets from our training set with available ground-truth, namely Kitti and Mannequin Challenge. This represents the evaluation procedure commonly employed by most papers, where the test data is drawn from the same distribution as the training data.

These results can be found in Table 5 (*In-Distribution*). Both of our models outperform every (self-)supervised baseline on both datasets, excluding NewCRFs [5] on

Kitti. This shows that SlowTV provides complementary driving data that can generalize across datasets. Furthermore, the improved KBR++ increases performance on Mannequin Challenge due to the additional indoor data provided by CribsTV.

## 5.6 Zero-shot Generalization

The core of our evaluation takes place in a *zero-shot* setting, meaning that models are not fine-tuned on the target datasets. This tests the capability to adapt to out-of-distribution examples and previously unseen environments. Existing SS-MDE publications sometimes include zero-shot evaluations. However, this is frequently limited to CityScapes [11] and Make3D [89], which contain low-quality ground-truth and represent an urban automotive domain similar to the training Kitti. We instead opt for a much more challenging collection of datasets, constituting a mixture of urban, natural, synthetic and indoor scenes. Please refer to Table 1 for details regarding the evaluations datasets and their splits.

**Outdoor.** These results can be found in Table 5 (*Outdoor*). Both of our models outperform the SSL baselines by a large margin. This is even the case on DDAD [12], which is also an urban automotive dataset. Meanwhile, our model is capable of generalizing to urban [12, 63], synthetic [62] and natural [1, 13] datasets, performing on par with the SotA supervised baselines. It is interesting to note that NeWCRFs generalizes across automotive datasets and provides the best performance on DDAD. However, it fails when evaluated in alternative domains and provides only minimal improvements over the SSL baselines. Finally, even the more efficient KBR provides impressive performance that matches more complex transformer-based backbones.

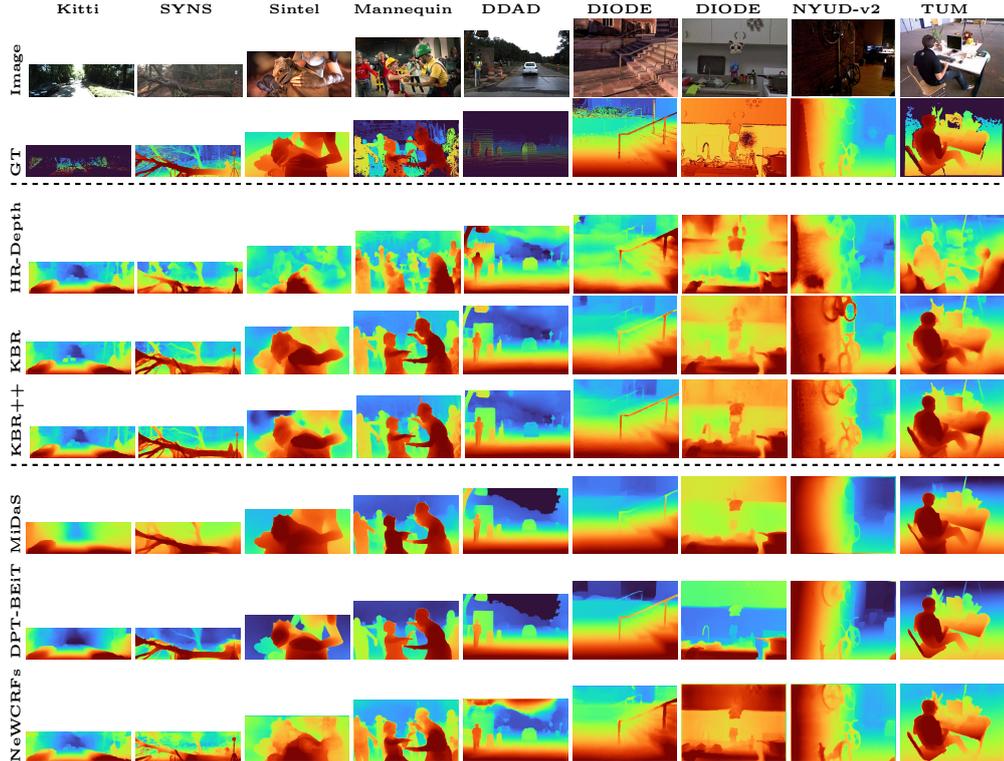
**Indoor.** Table 5 (*Indoor*) shows performance on all indoor datasets. Note that NeWCRFs was trained exclusively on NYUD-v2, while DPT uses it as part of its training collection. As such, this subset of results is *not* zero-shot. Our models outperform the SSL baselines by an even larger margin, due to the large shift in distribution when moving from outdoor to indoor scenes. In this case, KBR++ provides a noticeable improvement over KBR, thanks to the additional indoor training data provided by CribsTV. This helps to further close the gap w.r.t. supervised approaches. Once again, our model is now capable of performing on-par with all supervised models except DPT-BEiT, despite requiring no ground-truth annotations.

**Overall.** To summarize, Table 5 (*Multi-task*) reports the multi-task metrics across all datasets. Our models outperform the updated SS-MDE baselines from [1] by over 35%. Meanwhile, the contributions from this paper further improve our original model [15] by 4.5%. What’s more, KBR++ is the second-best model overall, outperforming supervised baselines such as NeWCRFs and DPT-ViT. It is worth emphasizing once again that our model is *entirely self-supervised*, relying exclusively on the photometric consistency across frames. Thus, we present the first approach demonstrating the true capabilities of SS-MDE, which can leverage much larger and more diverse collections of data freely available on YouTube.

**Table 5: Results.** *Outdoor* and *Indoor* represent zero-shot evaluations. We outperform all SS-MDE baselines [1] (top block). Our original model (KBR) performs on par with the supervised SotA, while the updated model from this paper (KBR++) outperforms every model except DPT [3]. Our models do not required ground-truth annotations for training and instead leverage large-scale YouTube data.

	Train	Multi-task		In-Distribution				Outdoor						Indoor							
		Rank↓	Δ↑	Kitti		Mannequin		DDAD		DIODE		Sintel		SYNS		DIODE		NYUD-v2		TUM	
		Rel↓	F↑	Rel↓	F↑	Rel↓	F↑	Rel↓	F↑	Rel↓	δ <sub>.25</sub> ↑	Rel↓	F↑	Rel↓	F↑	Rel↓	δ <sub>.25</sub> ↑	Rel↓	δ <sub>.25</sub> ↑	Rel↓	δ <sub>.25</sub> ↑
Garg [6]	S	7.58	-38.52	7.65	53.28	27.63	9.08	26.93	7.80	39.60	44.15	39.41	31.93	26.05	15.17	19.18	70.54	22.49	59.60	23.53	62.82
Monodepth2 [8]	MS	7.74	-38.34	7.90	50.50	27.44	7.97	24.31	8.25	39.53	44.71	40.09	29.49	25.31	14.83	19.40	70.42	22.41	60.09	23.50	62.36
DiffNet [52]	MS	7.05	-36.84	7.98	49.60	27.46	7.76	23.03	9.43	38.87	46.14	39.93	28.77	25.09	14.64	19.11	70.94	21.82	61.30	23.21	63.08
HR-Depth [9]	MS	5.95	-35.16	7.70	51.49	27.01	8.39	23.13	9.94	39.09	45.60	38.82	30.90	25.07	15.48	18.93	71.19	21.74	61.18	23.18	63.50
<b>KBR (Ours) [15]</b>	M	3.37	0.00	6.84	56.17	14.39	17.67	12.63	20.21	33.49	57.08	33.34	40.81	22.40	18.50	14.91	80.77	11.59	87.23	15.02	80.86
<b>KBR++ (Ours)</b>	M	2.95	4.58	6.77	56.57	12.95	20.08	13.10	17.87	32.81	57.77	30.46	41.89	21.92	18.79	14.26	82.94	9.24	91.85	12.71	85.86
MiDaS [2]	D	5.68	-11.84	13.71	33.44	16.96	12.62	16.00	15.41	32.72	59.04	30.95	39.55	26.94	14.69	10.71	88.42	10.48	89.59	14.43	82.35
DPT-ViT [3]	D	3.84	-1.74	10.98	40.56	15.52	14.46	15.49	18.25	32.59	59.82	25.53	43.57	23.24	17.44	9.60	91.38	10.10	90.10	12.68	86.25
DPT-BEiT [3]	D	2.11	11.12	9.45	44.22	13.55	16.58	10.70	22.63	31.08	61.51	21.38	46.46	21.47	17.73	7.89	93.34	5.40	96.54	10.45	89.68
NeWCRFs [5]	D	3.84	1.03	5.23	59.20	18.20	15.17	9.59	23.02	37.01	49.66	39.25	32.43	24.28	16.76	14.05	84.95	6.22	95.58	14.63	82.95

Highlighted cells are NOT zero-shot results. S=Stereo, M=Monocular, D=Ground-truth Depth.

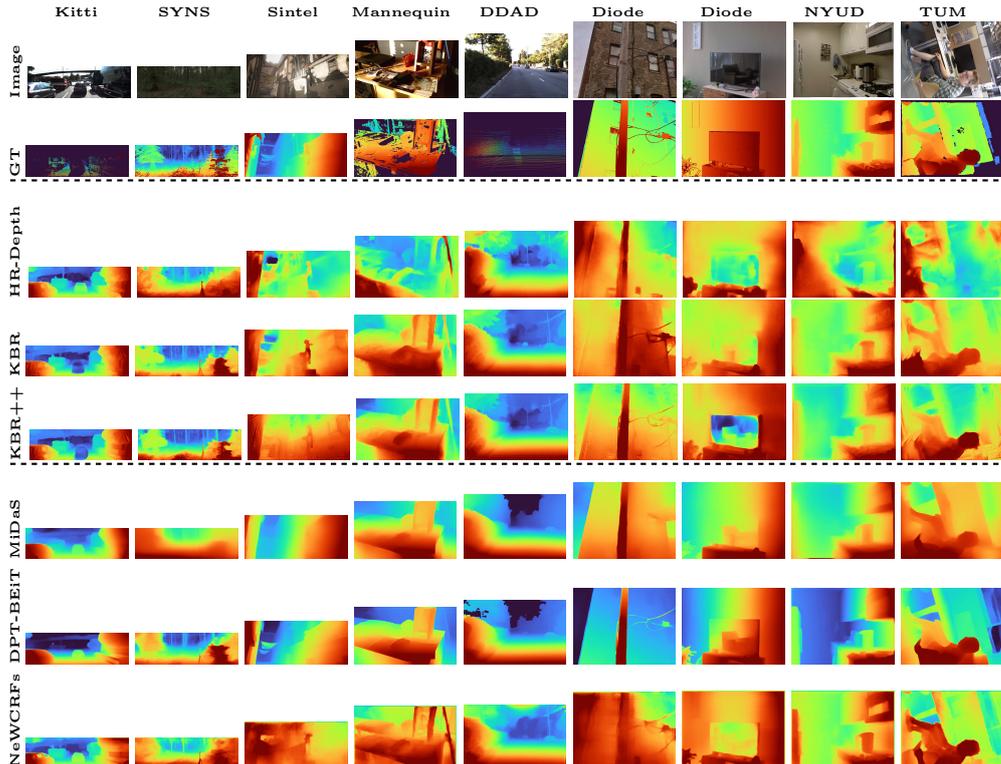


**Fig. 8: Zero-shot SS-MDE.** The proposed KBR models generalize to a wide range of scene types, greatly improving upon the SSL baselines. The contributions from this paper further improve robustness to indoor scenes and the accuracy at thin structures. *Middle=Self-Supervised* – *Bottom=Supervised*.

## 5.7 Qualitative Results

**Visualizations.** Sample predictions for each model and dataset can be found in Figure 8. Our models are significantly more robust than the best SSL baseline [9], which fails on all domains except the automotive. This is most noticeable in indoor settings, where it treats human faces as background. Meanwhile, our model generalizes across all datasets and environments, providing high-quality predictions. It can also be seen how KBR++ improves over the base KBR, especially in thin structures (DIODE Outdoors) and depth boundaries (TUM).

**Failure Cases.** Despite being a significant step forward for SS-MDE, our model still has a few limitations and failure cases. We show these examples in Figure 9. The main one is the lack of explicit modeling for dynamic objects. Whilst the minimum reconstruction loss and automasking [8] can reduce their impact, there are still cases where vehicles in front of the camera are predicted as holes of infinite depth. Another



**Fig. 9: Failure Cases.** The proposed model occasionally produces holes of infinite depth or texture-copy artifacts. However, complex regions such as foliage or object boundaries tend to be challenging for all approaches. Finally, the upright prior in training data makes the model sensitive to strong rotations. *Middle=Self-Supervised – Bottom=Supervised.*

common failure case is texture-copying artifacts, where textures from the original image are incorrectly predicted as changes in depth. This can happen on objects such as textured walls or pavements made with bricks or text on shirts and signs. Finally, another interesting failure case are reflective or transparent surfaces, as they do not violate the photometric constraints during training. However, these are also challenging for supervised methods, as the data cannot be correctly captured using LiDAR either.

## 5.8 Map-Free Relocalization

Following our preliminary publication [15], we report our results on the MapFreeReloc [72] benchmark. Map-free relocalization aims to regress the 6DoF pose of a target frame based only on the known pose of a single reference frame. This is contrary to traditional localization pipelines, which typically contain a map-building or network-training phase that requires large-scale captures for each specific scene.

**Table 6: Map-free Relocalization [72].** The feature-matching MapFreeReloc baselines [72] can be improved with our novel SS-MDE models. These results are zero-shot, without finetuning on the target dataset (which consists of portrait images). The improved variants from this paper (KBR++) perform on-par with DPT-BEiT and outperform every other (self-)supervised approach.

	Train	Multi-task		Pose				VCRE		
		Rank↓	Δ↑	Trans↓	Rot↓	P↑	AUC↑	Error↓	P↑	AUC↑
Garg [6]	S	10.50	-21.84	2.96	52.57	5.43	17.15	188.20	24.84	51.61
Monodepth2 [17]	MS	11.25	-22.32	2.95	52.92	5.50	17.22	189.67	24.38	50.63
DiffNet [52]	MS	10.88	-21.70	2.97	53.19	5.65	17.71	188.80	24.78	51.24
HR-Depth [9]	MS	9.38	-21.07	2.94	52.95	5.67	17.95	187.83	25.06	51.52
<b>KBR (Ours)</b>	M	4.50	0.00	<u>2.63</u>	<u>49.01</u>	<u>11.54</u>	<u>32.02</u>	<u>181.21</u>	<u>29.96</u>	<u>58.89</u>
<b>KBR++ (Ours)</b>	M	<u>2.00</u>	<u>4.86</u>	<u>2.58</u>	<u>46.22</u>	<u>12.50</u>	<u>33.76</u>	<u>178.26</u>	<u>31.93</u>	<u>61.48</u>
MiDaS [2]	D	4.00	0.35	2.60	46.92	<u>11.39</u>	30.44	<u>180.64</u>	30.45	59.72
DPT-ViT [3]	D	3.38	1.09	<u>2.56</u>	<u>45.62</u>	11.27	<u>30.92</u>	181.34	<u>30.60</u>	<u>60.03</u>
DPT-BEiT [3]	D	<b>1.75</b>	<b>5.32</b>	<b>2.49</b>	<b>44.99</b>	<b>12.56</b>	<b>32.48</b>	181.67	<b>32.46</b>	<b>62.03</b>
NeWCRFs [5]	D	8.00	-16.98	2.89	51.92	6.69	20.77	184.63	25.89	52.93
DPT-NYUD [72]	D+FT	6.50	-6.61	2.67	47.66	9.17	26.46	184.53	28.68	56.87
DPT-Kitti [72]	D+FT	5.88	-3.05	2.66	49.21	10.86	29.99	<b>178.49</b>	28.37	56.86

*Trans=meters, Rot=deg, VCRE=px, Precision=%, AUC=%.*

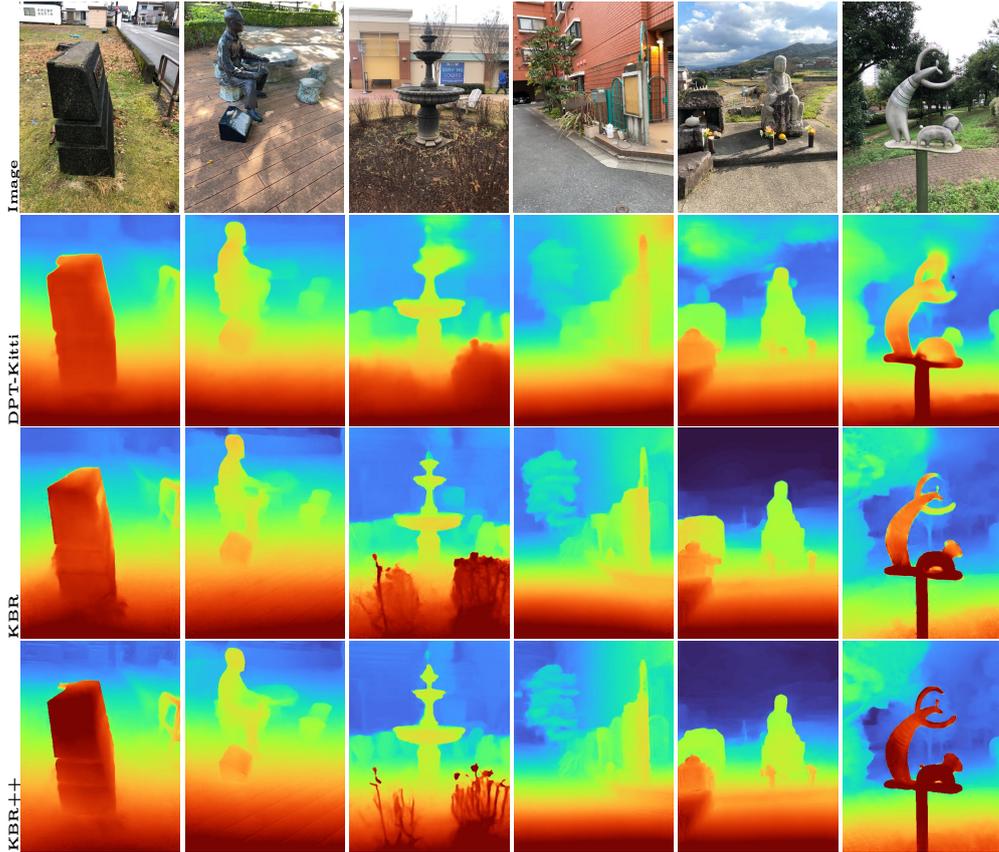
Recent research [72, 90] has shown that the scale ambiguity in map-free relocalization feature-matching pipelines can be resolved by incorporating SotA MDE predictions.

The models are evaluated on the validation split of the benchmark, which consists of 37k images from 65 small-scale landmarks. The data was crowd-sourced and collected using mobile phones, meaning that it features an uncommon portrait aspect ratio. This makes the task of zero-shot transfer even more challenging.

The feature-matching baseline [72] uses LoFTR [91] correspondences, a PnP solver and DPT [3] fine-tuned on either Kitti or NYUD-v2. Metric depth for all evaluated models is obtained by aligning the predictions to the baseline fine-tuned DPT predictions. We report the evaluation metrics provided by the benchmark authors. This includes translation (meters), rotation (deg) and reprojection (px) errors. Pose Precision/AUC were computed with an error threshold of 25 cm & 5°, while Reprojection uses a threshold of 90px.

The results can be found in Table 6, along with visualizations in Figure 10. The updated models from this paper (KBR++) outperform every (self-)supervised model, except DPT-BEiT. This demonstrates the effectiveness of training with our diverse datasets, as well as the improved robustness to image aspect ratios provided by AR-Aug. Furthermore, it showcases the ability to incorporate SS-MDE into real-world problem pipelines.

We find that the original DPT models perform better than their fine-tuned counterparts, despite using these as the metric scale reference. This suggests that the finetuning procedure of [72] may provide metric scale at the cost of generality. However, this highlights the need for models that predict accurate metric depth, rather than only relative depth.



**Fig. 10: MapFreeReloc [72] Predictions.** Our models are capable of adapting to the challenging portrait images of the MapFreeReloc benchmark, allowing us to outperform other (self-)supervised methods. Our predictions are sharper and more detailed than those provided by the baseline, which requires a large collection of ground-truth annotations during training.

## 6 Conclusion

This paper has introduced KBR and KBR++, the first SS-MDE models that match and even outperform SotA *supervised* algorithms. We demonstrated this in our challenging zero-shot experiments, which showcase the robustness and generalization capabilities of our models. This was made possible due to our approach to data collection, focusing on the scale of the training set and leveraging the lack of annotations needed for self-supervised learning. We curated two novel large-scale YouTube datasets, SlowTV and CribsTV, with a total of 2M training frames. These datasets contain an incredibly diverse set of environments, ranging from hikes in snowy forests, to luxurious houses and even underwater caves.

Performance and generalization were further maximized by introducing stronger augmentation regimes (AR-Aug, RandAugment and CutOut), simultaneously learning

the camera’s intrinsics, making the training more flexible and modernizing the network architecture. Our extensive ablations demonstrated the benefits of introducing each respective component.

The main limitation of the current models is their sensitivity to dynamic objects. Whilst the contributions from Monodepth2 [8] alleviate some of these artifacts, a more explicit motion model may be required to handle these scenarios. Introducing optical flow constraints may be the most feasible way to achieve this in a self-supervised manner. However, it is worth noting the increased computational requirements resulting from training a new network and computing the required consistency losses.

Finally, estimating metric depth (for generic scenes) without ground-truth annotations is an open research problem that could further increase the applicability of SS-MDE to real-world tasks. By making the datasets and code freely available to the public, we hope to further drive the SotA in SS-MDE and inspire future research that addresses these challenging problems.

## Acknowledgements

This work was partially funded by the EPSRC under grant agreements EP/S016317/1 & EP/S035761/1.

## References

- [1] Spencer, J., Russell, C., Hadfield, S., Bowden, R.: Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *Transactions on Machine Learning Research* (2022). Reproducibility Certification
- [2] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* (2020)
- [3] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188 (2021)
- [4] Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4009–4018 (2021)
- [5] Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3906–3915 (2022). <https://doi.org/10.1109/CVPR52688.2022.00389>
- [6] Garg, R., Kumar, V., Carneiro, G., Reid, I.: Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In: *European Conference on Computer Vision*, pp. 740–756 (2016)

- [7] Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised Learning of Depth and Ego-Motion from Video. Conference on Computer Vision and Pattern Recognition, 6612–6619 (2017) <https://doi.org/10.1109/CVPR.2017.700>
- [8] Godard, C., Aodha, O.M., Firman, M., Brostow, G.: Digging Into Self-Supervised Monocular Depth Estimation. International Conference on Computer Vision **2019-October**, 3827–3837 (2019) <https://doi.org/10.1109/ICCV.2019.00393>
- [9] Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation. AAAI Conference on Artificial Intelligence **35**(3), 2294–2301 (2021)
- [10] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal of Robotics Research **32**(11), 1231–1237 (2013) <https://doi.org/10.1177/0278364913491297>
- [11] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
- [12] Guizilini, V., Ambrus, A., Pillai, S., Raventos, A., Gaidon, A.: 3D packing for self-supervised monocular depth estimation. Conference on Computer Vision and Pattern Recognition, 2482–2491 (2020) <https://doi.org/10.1109/CVPR42600.2020.00256>
- [13] Adams, W.J., Elder, J.H., Graf, E.W., Leyland, J., Lugtigheid, A.J., Murry, A.: The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. Scientific Reports **6**(1), 35805 (2016) <https://doi.org/10.1038/srep35805>
- [14] Koch, T., Liebel, L., Fraundorfer, F., Körner, M.: Evaluation of CNN-Based Single-Image Depth Estimation Methods. In: European Conference on Computer Vision Workshops, pp. 331–348 (2018)
- [15] Spencer, J., Russell, C., Hadfield, S., Bowden, R.: Kick Back & Relax: Learning to Reconstruct the World by Watching SlowTV. In: International Conference on Computer Vision (2023)
- [16] Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Mannequinchallenge: Learning the depths of moving people by watching frozen people. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4229–4241 (2020)
- [17] Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised Monocular Depth Estimation with Left-Right Consistency. Conference on Computer Vision and Pattern Recognition, 6602–6611 (2017) <https://doi.org/10.1109/CVPR.2017.699>

- [18] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PMLR
- [19] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)
- [20] Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022). <https://openreview.net/forum?id=p-BhZSz59o4>
- [21] Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
- [22] DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
- [23] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial Transformer Networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
- [24] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004) <https://doi.org/10.1109/TIP.2003.819861>
- [25] Poggi, M., Tosi, F., Mattoccia, S.: Learning Monocular Depth Estimation with Unsupervised Trinocular Assumptions. In: International Conference on 3D Vision, pp. 324–333 (2018). <https://doi.org/10.1109/3DV.2018.00045>
- [26] Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning Depth from Monocular Videos Using Direct Methods. *Conference on Computer Vision and Pattern Recognition, 2022–2030* (2018) <https://doi.org/10.1109/CVPR.2018.00216>
- [27] Engel, J., Koltun, V., Cremers, D.: Direct Sparse Odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(3), 611–625 (2018) <https://doi.org/10.1109/TPAMI.2017.2658577>
- [28] Klodt, M., Vedaldi, A.: Supervising the New with the Old: Learning SFM from SFM. In: European Conference on Computer Vision, pp. 713–728 (2018)
- [29] Yang, N., Stumberg, L., Wang, R., Cremers, D.: D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. In: *Conference on Computer Vision and Pattern Recognition*, pp. 1278–1289 (2020). <https://doi.org/10.1109/CVPR42601.2020.00127>

- [30] Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the Uncertainty of Self-Supervised Monocular Depth Estimation. In: Conference on Computer Vision and Pattern Recognition, pp. 3224–3234 (2020). <https://doi.org/10.1109/CVPR42600.2020.00329>
- [31] Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1983–1992 (2018)
- [32] Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12240–12249 (2019)
- [33] Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts ++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(10), 2624–2641 (2020) <https://doi.org/10.1109/TPAMI.2019.2930258>
- [34] Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8977–8986 (2019)
- [35] Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8001–8008 (2019)
- [36] Dai, Q., Patil, V., Hecker, S., Dai, D., Van Gool, L., Schindler, K.: Self-supervised object motion and depth estimation from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
- [37] Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.M.: Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. Conference on Computer Vision and Pattern Recognition, 340–349 (2018) <https://doi.org/10.1109/CVPR.2018.00043>
- [38] Spencer, J., Bowden, R., Hadfield, S.: DeFeat-Net: General monocular depth via simultaneous unsupervised representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 14390–14401 (2020). <https://doi.org/10.1109/CVPR42600.2020.01441>

- [39] Shu, C., Yu, K., Duan, Z., Yang, K.: Feature-Metric Loss for Self-supervised Learning of Depth and Egomotion. In: European Conference on Computer Vision, pp. 572–588 (2020)
- [40] Chen, P.-Y., Liu, A.H., Liu, Y.-C., Wang, Y.-C.F.: Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2624–2632 (2019)
- [41] Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. arXiv preprint arXiv:2002.12319 (2020)
- [42] Jung, H., Park, E., Yoo, S.: Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12642–12652 (2021)
- [43] Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. Conference on Computer Vision and Pattern Recognition, 5667–5675 (2018) <https://doi.org/10.1109/CVPR.2018.00594>
- [44] Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.-M., Reid, I.: Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
- [45] Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1164–1174 (2021)
- [46] Rui, Jörg, S., Nan, C.D.Y., Wang: Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry. In: European Conference on Computer Vision, pp. 835–852 (2018)
- [47] Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., Lin, L.: Single View Stereo Matching. Conference on Computer Vision and Pattern Recognition, 155–163 (2018) <https://doi.org/10.1109/CVPR.2018.00024>
- [48] Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. Conference on Computer Vision and Pattern Recognition **2019-June**, 9791–9801 (2019) <https://doi.org/10.1109/CVPR.2019.01003>
- [49] Watson, J., Firman, M., Brostow, G., Turmukhambetov, D.: Self-supervised

- monocular depth hints. International Conference on Computer Vision **2019-October**, 2162–2171 (2019) <https://doi.org/10.1109/ICCV.2019.00225>
- [50] Gonzalez Bello, J.L., Kim, M.: PLADE-Net: Towards Pixel-Level Accuracy for Self-Supervised Single-View Depth Estimation with Neural Positional Encoding and Distilled Matting Loss. In: Conference on Computer Vision and Pattern Recognition, pp. 6847–6856 (2021). <https://doi.org/10.1109/CVPR46437.2021.00678>
- [51] Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: Monovit: Self-supervised monocular depth estimation with a vision transformer. International Conference on 3D Vision (2022)
- [52] Zhou, H., Greenwood, D., Taylor, S.: Self-Supervised Monocular Depth Estimation with Internal Feature Fusion. In: British Machine Vision Conference (2021)
- [53] Pillai, S., Ambrus, R., Gaidon, A.: SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation. In: International Conference on Robotics and Automation, pp. 9250–9256 (2019). <https://doi.org/10.1109/ICRA.2019.8793621>
- [54] Johnston, A., Carneiro, G.: Self-Supervised Monocular Trained Depth Estimation Using Self-Attention and Discrete Disparity Volume. In: Conference on Computer Vision and Pattern Recognition, pp. 4755–4764 (2020). <https://doi.org/10.1109/CVPR42600.2020.00481>
- [55] Yan, J., Zhao, H., Bu, P., Jin, Y.: Channel-Wise Attention-Based Network for Self-Supervised Monocular Depth Estimation. In: International Conference on 3D Vision, pp. 464–473 (2021)
- [56] Bhat, S.F., Alhashim, I., Wonka, P.: Localbins: Improving depth estimation by learning local distributions. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I, pp. 480–496 (2022). Springer
- [57] Gonzalez Bello, J.L., Kim, M.: Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12626–12637 (2020)
- [58] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. Conference on Computer Vision and Pattern Recognition, 4040–4048 (2016) <https://doi.org/10.1109/CVPR.2016.438>
- [59] Spencer, J., Qian, C.S., Russell, C., Hadfield, S., Graf, E., Adams, W., Schofield, A.J., Elder, J.H., Bowden, R., Cong, H., *et al.*: The monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Winter Conference on Applications

of Computer Vision, pp. 623–632 (2023)

- [60] Spencer, J., Qian, C.S., Trescakova, M., Russell, C., Hadfield, S., Graf, E., Adams, W., Schofield, A.J., Elder, J., Bowden, R., Others: The second monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2023)
- [61] Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity Invariant CNNs. International Conference on 3D Vision, 11–20 (2018) <https://doi.org/10.1109/3DV.2017.00012> arXiv:1708.06500
- [62] Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VI 12, pp. 611–625 (2012). Springer
- [63] Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
- [64] Nathan Silberman, P.K. Derek Hoiem, Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: ECCV (2012)
- [65] Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: Proc. of the International Conference on Intelligent Robot Systems (IROS) (2012)
- [66] Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2702–2719 (2019)
- [67] Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
- [68] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 (2020)
- [69] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nusenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11621–11631 (2020)
- [70] Schönberger, J.L., Frahm, J.-M.: Structure-from-Motion Revisited. In: Conference

- on Computer Vision and Pattern Recognition (CVPR) (2016)
- [71] Castellano, B.: PySceneDetect. GitHub (2014)
  - [72] Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free visual relocalization: Metric pose relative to a single image. In: ECCV (2022)
  - [73] Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2041–2050 (2018)
  - [74] Chen, Y., Schmid, C., Sminchisescu, C.: Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7063–7072 (2019)
  - [75] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
  - [76] Müller, S.G., Hutter, F.: Trivialaugument: Tuning-free yet state-of-the-art data augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 774–782 (2021)
  - [77] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
  - [78] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
  - [79] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
  - [80] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. Conference on Computer Vision and Pattern Recognition **2017-Janua**, 5987–5995 (2017) <https://doi.org/10.1109/CVPR.2017.634> arXiv:1611.05431
  - [81] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021). <https://openreview.net/forum?id=YicbFdNTTy>

- [82] Yuan, W., Zhang, Y., Wu, B., Zhu, S., Tan, P., Wang, M.Y., Chen, Q.: Stereo matching by self-supervision of multiscopic vision. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5702–5709 (2021). IEEE
- [83] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
- [84] Eigen, D., Fergus, R.: Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In: International Conference on Computer Vision, pp. 2650–2658 (2015). <https://doi.org/10.1109/ICCV.2015.304>
- [85] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
- [86] Wightman, R.: PyTorch Image Models. GitHub (2019). <https://doi.org/10.5281/zenodo.4414861>
- [87] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
- [88] Örnek, E.P., Mudgal, S., Wald, J., Wang, Y., Navab, N., Tombari, F.: From 2D to 3D: Re-thinking Benchmarking of Monocular Depth Prediction. arXiv preprint (2022)
- [89] Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(5), 824–840 (2009) <https://doi.org/10.1109/TPAMI.2008.132>
- [90] Toft, C., Turmukhambetov, D., Sattler, T., Kahl, F., Brostow, G.J.: Single-image depth prediction makes feature matching easier. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pp. 473–492 (2020). Springer
- [91] Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931 (2021)