

Revisiting CNNs for Monocular Depth Estimation: A Synthetic-to-Real Perspective

Luís Gustavo Lorgus Decker  [Institute of Computing, University of Campinas | luisgustavo.decker@gmail.com]

Jose Luis Flores Campana  [Institute of Computing, University of Campinas | jluisflores357@gmail.com]

Marcos Roberto e Souza  [Institute of Computing, University of Campinas | marcoosrs@gmail.com]

Helena de Almeida Maia  [Institute of Computing, University of Campinas | helena.maia@ic.unicamp.br]

Helio Pedrini   [Institute of Computing, University of Campinas | helio@ic.unicamp.br]

 Institute of Computing, University of Campinas, Campinas, SP, 13083-852, Brazil.

Received: 10 April 2025 • Accepted: 16 September 2026 • Published: 05 October 2026

Abstract. Monocular depth estimation is crucial for autonomous systems, augmented reality, and scene understanding. While Vision Transformers (ViTs) excel at capturing global context, their need for large amounts of data makes them impractical for smaller datasets, leaving convolutional neural networks (CNNs) as a viable alternative. This study evaluates CNN-based architectures for monocular depth estimation, proposing a fully convolutional model optimized for efficiency and generalization. Trained solely on synthetic datasets with precise depth annotations, our model avoids real-world data limitations. We assess its performance on synthetic validation sets and real-world datasets without fine-tuning, examining its capacity for domain generalization. Results show that CNNs, with optimized architectures and training, achieve strong depth estimation and enable synthetic-to-real transfer, with an AbsREL of 0.077 on KITTI and 0.099 on NYU Depth. These findings challenge the assumption that real-world data is essential for generalization.

Keywords: Depth Estimation, Convolutional Neural Networks, Synthetic Datasets, Real-World Data

1 Introduction

The ability to perceive the three-dimensional structure of the world from visual input is a cornerstone of intelligent interaction, underpinning critical technologies from autonomous navigation and robotic manipulation to immersive augmented and virtual reality experiences.

Monocular Depth Estimation (MDE) [Decker *et al.*, 2024; Zeng *et al.*, 2024; Spencer *et al.*, 2024], the task of inferring depth information from a single two-dimensional (2D) image, represents a fundamental challenge in computer vision. Lacking explicit geometric cues available in stereo or multi-view settings, MDE must rely on learned interpretations of monocular cues such as perspective, texture gradients, object sizes, and occlusion relationships, making it an inherently ill-posed problem.

Historically, MDE research saw significant progress with the advent of deep learning. Convolutional Neural Networks (CNNs) [Laina *et al.*, 2016] became the dominant paradigm, demonstrating remarkable capabilities in learning complex mappings from image pixels to depth values. More recently, Vision Transformers (ViTs) [Dosovitskiy *et al.*, 2020] have emerged, setting new performance benchmarks in MDE [Ranftl *et al.*, 2021; Yang *et al.*, 2024b] by leveraging their strength in capturing long-range dependencies and global scene context. However, this performance often comes with a need of a large quantity of training data [Oquab *et al.*, 2023], while CNNs show comparable performance when trained on smaller, task-specific datasets [Liu *et al.*, 2022].

Training robust depth estimation models typically requires vast amounts of diverse data with accurate, dense depth an-

notations. Acquiring such ground truth for real-world scenes using sensors such as LiDAR or structured light is expensive, time-consuming, and often results in sparse or noisy labels, particularly for challenging surfaces or distant regions [Geiger *et al.*, 2013; Silberman *et al.*, 2012]. This has spurred significant interest in using synthetic data [Ros *et al.*, 2016; Wrenninge and Unger, 2018; Roberts *et al.*, 2021]. Generated using computer graphics engines, synthetic datasets offer pixel-perfect, dense depth maps, controllable scene variations, and scalability at a lower cost. Moreover, the persistent “domain gap” between synthetic and real-world imagery poses a major hurdle; models trained purely on synthetic data often struggle to generalize to the complexities and nuances of real scenes. Addressing this gap typically involves domain adaptation techniques or fine-tuning on real data, negating some benefits of synthetic training.

This paper explores a compelling alternative: can modern, efficiently designed CNNs, trained exclusively on diverse synthetic datasets, achieve effective zero-shot generalization to real-world depth estimation tasks? We hypothesize that by leveraging large-scale synthetic data, advanced CNN architectures, and appropriate training strategies (specifically, scale-and-shift invariant loss functions suitable for relative depth estimation [Ranftl *et al.*, 2020]), it is possible to learn domain-invariant features sufficient for robust performance on unseen real-world data without any fine-tuning.

Therefore, our primary objective is to develop and evaluate a fully convolutional MDE model trained solely on a curated mix of synthetic datasets. We focus on relative depth estimation, mitigating issues arising from inconsistent absolute scales across datasets. The model’s zero-shot transfer capabil-

ity is rigorously assessed on standard real-world benchmarks (KITTI and NYU Depth V2). By demonstrating competitive performance in this challenging setting, this work aims to highlight the continued relevance of optimized CNNs and underscore the potential of synthetic data to significantly reduce reliance on annotated real-world data for practical depth estimation systems.

This study builds upon our previous work [Decker *et al.*, 2024], introducing several key advancements. Specifically, we propose an enhanced convolutional decoder that incorporates a more robust upscaling method and refines the skip-connection mechanism by employing concatenation instead of summation. Moreover, this research integrates a more extensive and diverse dataset during the training phase, improving the model’s generalizability. We also transition from direct depth estimation to disparity estimation, aligning with recent advances in the field, and implement the median-based robust scale and shift predictor from [Ranftl *et al.*, 2020] instead of the least-squares-based approach used in our earlier work.

Our main contributions in this work are:

- **A Novel Neural Network Architecture:** This work introduces a cutting-edge encoder-decoder convolutional neural network (CNN) architecture designed for monocular depth estimation. The proposed model incorporates a ConvNext [Liu *et al.*, 2022] as the backbone image encoder to capture rich visual representations. This is complemented by a modernized U-Net-inspired decoder [Ronneberger *et al.*, 2015], which integrates learnable upscaling functions and multi-scale feature aggregation, facilitating precise depth predictions.
- **Innovative Training Protocol for Cross-Domain Generalization:** A robust training strategy is proposed to effectively train the neural network on diverse synthetic datasets. This protocol enhances the model’s capability to learn representative features, enabling seamless generalization from synthetic to real-world data. Remarkably, this approach achieves outstanding performance in monocular depth estimation on real datasets without requiring real data during the training phase.

2 Related Work

Monocular depth estimation is a task within the broader field of depth estimation, where the model is constrained to using only a single image of the scene to generate a depth map. This task presents significant challenges due to the inherent ambiguity of inferring depth from a single viewpoint. Uncommon perspectives or unusual object arrangements within the scene can lead to inaccurate depth estimations [Masoumian *et al.*, 2022].

The work proposed by Eigen *et al.* [2014], a cornerstone for deep learning-based monocular depth estimation, presents a method for predicting depth from a single image by employing two deep network stacks: one for coarse global prediction and another for local refinement. The approach integrates both global and local information to address the inherent ambiguity and scale uncertainty in depth estimation from a single image. The method leverages large raw datasets for training,

accurately matching detailed depth boundaries without the need for superpixelation. The paper also introduces a scale-invariant error to measure depth relations rather than scale, enhancing the overall depth prediction accuracy.

Following the work developed by Eigen *et al.* [2014], several approaches have been proposed to improve prediction accuracy [Li *et al.*, 2019; Fu *et al.*, 2018; Liu *et al.*, 2015; Roy and Todorovic, 2016; Laina *et al.*, 2016]. However, these methods share a common limitation: they are typically trained on a single dataset, which restricts the diversity and quantity of data available for training, potentially limiting generalization performance.

To address the limitations of training on single data sets and to create larger, more diverse training sets, Ranftl *et al.* [2020] introduced a method for improving monocular depth estimation by combining multiple data sets with different characteristics and biases. They proposed a robust training loss function that is invariant to variations in depth range and scale, and advocated principled multi-objective learning to effectively integrate data from different sources. Extensive experiments demonstrated that combining data from complementary sources significantly improves monocular depth estimation performance.

Building upon the cost function employed in the MiDaS model [Ranftl *et al.*, 2020], Yang *et al.* [2024a] further expanded the available training data by proposing a two-stage teacher-student training protocol. Initially, a teacher model is trained on a large, mixed corpus of annotated depth estimation datasets. This trained teacher model is then used to generate depth pseudo-labels for a substantial collection of unlabeled internet images. These pseudo-labeled images are subsequently combined with the original annotated dataset to train a student network. A comparative study demonstrates the benefits of incorporating these pseudo-labeled images into the training pipeline.

Another approach to addressing the scarcity of large and diverse annotated depth estimation datasets is the use of synthetic data. A pioneering effort in this direction was conducted by Atapour-Abarghouei and Breckon [2018], who presented a monocular depth estimation method leveraging synthetic data and domain adaptation through image style transfer. Recognizing the challenge of domain bias, which impedes the transfer of models trained on synthetic data to real-world scenarios, the authors employed style transfer and adversarial training. This approach enabled the prediction of accurate, pixel-level depth from single real-world color images, despite being trained on a large corpus of synthetic environment data.

Lopez-Rodriguez and Mikolajczyk [2023] also leveraged synthetic data through domain adaptation. The authors introduced a method that leverages semantic predictions and low-level edge features to bridge the domain gap between a fully annotated source dataset and an unlabeled target dataset. By enforcing consistency between the primary depth estimation model and a secondary model trained with semantic segmentation and edge maps, and further incorporating instance height priors, the approach demonstrates consistent improvements over state-of-the-art domain adaptation methods on standard benchmarks. These results underscore the effectiveness of semantic consistency in enhancing depth estimation performance in scenarios where direct depth annotations are

unavailable.

As an advancement of the previously proposed method, Yang *et al.* [2024b] modified their two-step teacher-student training protocol to use labeled synthetic datasets only during teacher model training. This choice of combining synthetic datasets only instead of real-world datasets or a combination of the two was made given the nature of synthetic datasets having pixel-perfect annotations, while real-world datasets have blurry or incomplete annotations due to the nature of the sensors used to label the data.

This study critically examines the recent paradigm shift toward Vision Transformers and the integration of real-world data in monocular depth estimation. Contrary to the prevailing trend, it demonstrates that Convolutional Neural Networks (CNNs), when carefully architected and rigorously optimized, can rival the performance of state-of-the-art transformer-based models.

The proposed method, trained exclusively on synthetic datasets, highlights the capacity of CNNs to learn robust and transferable feature representations. These representations enable effective zero-shot domain adaptation, yielding strong generalization to real-world scenarios without the need for any domain-specific fine-tuning.

The findings reaffirm the relevance of CNN-based architectures and synthetic data pipelines in contemporary depth estimation research, offering a compelling alternative to transformer-centric approaches.

3 Datasets

In this section, we present the datasets used in our experiments. The datasets can be divided into some categories, according to their characteristics and their use in this work:

- **Synthetic or Real:** Monocular depth estimation datasets are categorized as synthetic or real. Synthetic datasets use computer graphics to render virtual scenes, offering precise ground-truth depth and experimental control. Real-world datasets pair actual photographs with sensor-acquired depth maps (e.g., LiDAR, stereo), providing authentic environmental complexity but introducing noise and occlusions.
- **Indoor or Outdoor scenes:** Depth estimation datasets are categorized by scene content into indoor and outdoor sets. Indoor datasets, capturing enclosed environments, feature limited depth variations but high object diversity and spatial complexity. Conversely, outdoor datasets encompass open environments, characterized by significantly larger depth ranges and variability in scale and perspective. Combining both controlled indoor complexity and dynamic outdoor variability is essential for training models that generalize effectively.

Table 1 shows an overview of the datasets used in this work and their categorization according to the categories mentioned previously, while Table 2 presents a summary of the data splits and dataset after we performed our dataset splitting and data filtering, as described on each dataset subsection.

The following sections detail the datasets used, providing an overview, their training/evaluation roles, and data-splitting

Table 1. Datasets used in our experiments.

Dataset	Scene Type	Type
SYNTHIA		
Synscapes	Outdoor	Synthetic
Virtual KITTI 2		
Mid-Air		
Hypersim	Indoor	
TartanAir	Mixed	
KITTI	Outdoor	Real
NYU Depth V2	Indoor	

Table 2. Dataset split sizes and total image counts, as we use in our experiments. Details on the splitting criteria we used for each dataset and sample filtering conditions are provided in the respective dataset subsection.

	Train	Validation	Test	Total
SYNTHIA	77,325	19,150	0	96,475
Synscapes	15,000	5,000	0	20,000
Virtual KITTI 2	9,500	3,390	0	12,890
Mid-Air	282,386	141,290	0	423,676
Hypersim	39,449	4,289	0	43,738
TartanAir	159,336	60,394	0	219,730
KITTI	0	0	652	652
NYU Depth V2	0	0	654	654

criteria. Necessary preprocessing, filtering, or unique dataset characteristics are also detailed.

3.1 Synthetic Datasets

This section describes the main characteristics of the synthetic datasets used in our experiments on the training protocol of our model.

3.1.1 SYNTHIA

The SYNTHIA dataset [Ros *et al.*, 2016] simulates urban driving environments. Although it provides precise annotations, it exhibits lower visual realism compared to modern engines, and its depth labels do not treat transparent surfaces (such as car windows) as solid barriers, instead annotating them with the depth of background objects (see Figure 1). We utilize this dataset for training and validation, partitioning the original training split into 80% for active training and 20% for validation.

3.1.2 Synscapes

The Synscapes dataset [Wrenninge and Unger, 2018] features high-fidelity urban street scenes rendered with photorealistic camera effects, such as motion blur and varying exposure, as shown in Figure 2. In our experiments, Synscapes is used for both training and evaluation. Since no official splits exist, we employ a geographic split: scenes “01”, “02”, and “06” are used for training, while scene “18” serves as the validation set.

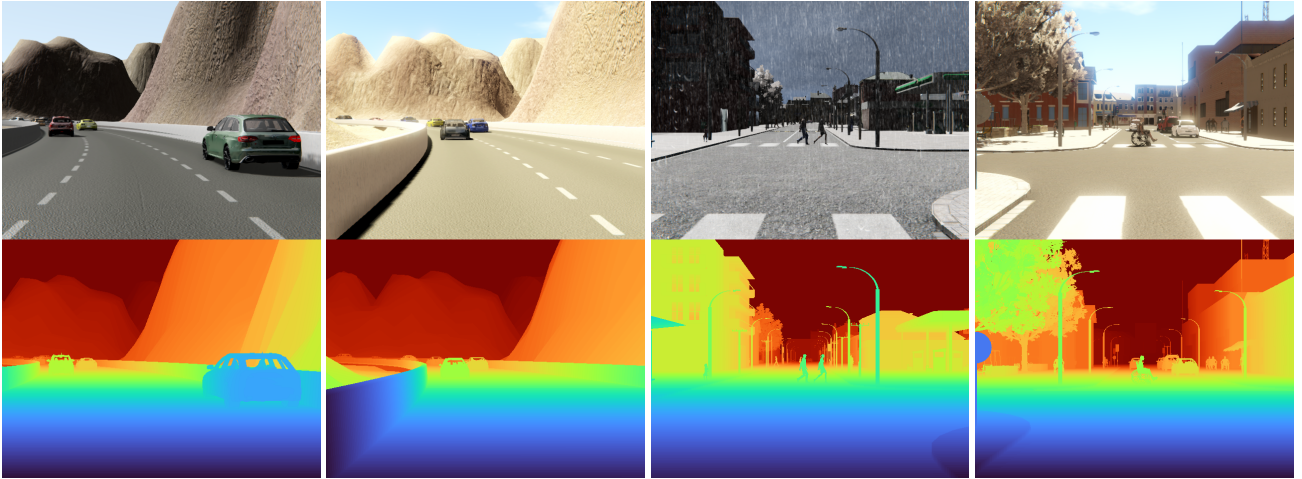


Figure 1. Synthetic RGB images (top row) and their ground-truth depth maps (bottom row) from the SYNTHIA dataset highlight the challenges associated with this dataset. The images often lack realism, and the depth maps suffer from incomplete annotations, particularly for transparent surfaces, like the car windows in the first example.

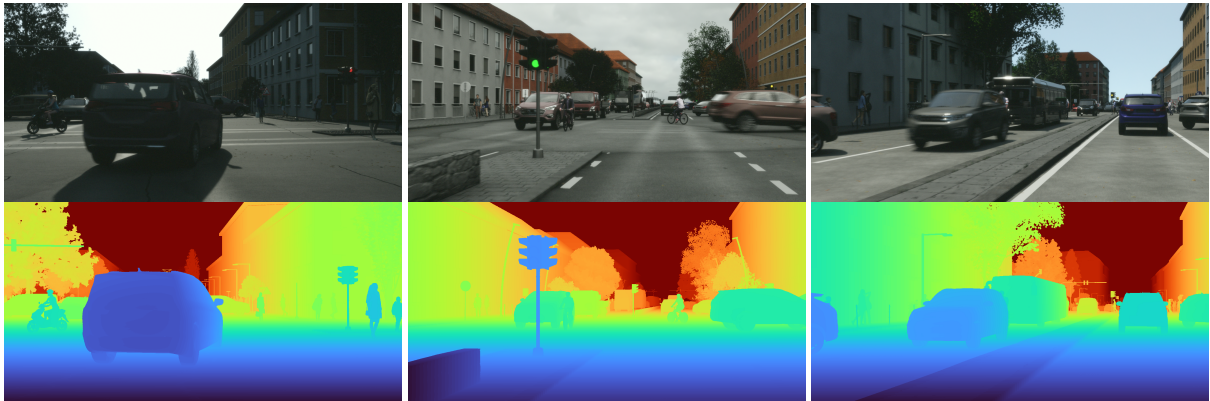


Figure 2. Synthetic RGB images (top row) and their corresponding ground-truth depth maps (bottom row) from the Synscapes dataset, showcasing the realism achievable through synthetic data generation. This dataset also simulates camera artifacts, such as the underexposure in the first image and the blurring of the moving objects in the other two. The trees in all the depth maps further highlight the richness of detail in the annotations.

3.1.3 Virtual KITTI 2

Virtual KITTI 2 [Cabon *et al.*, 2020] is a synthetic version of the real-world KITTI tracking benchmark that adds diverse weather conditions (e.g., rain, fog) and camera viewpoint variations. Although its photorealistic quality is limited compared to real images or high-fidelity synthetic datasets (see Figure 3), it provides dense, pixel-accurate ground-truth annotations even in heavily occluded regions. We incorporate this dataset into our training and validation pipeline by splitting the original training set into 80% for training and 20% for validation.

3.1.4 Hypersim

Hypersim [Roberts *et al.*, 2021] is an indoor synthetic dataset containing diverse, artist-designed scenes with complex geometries and surfaces made of reflective and transparent materials. Some samples of this dataset can be seen in Figure 4. Prior to training, we filtered out fully black frames resulting from rendering errors as well as frames containing invalid depth pixels. This dataset is utilized to train and validate our models on challenging indoor structures, allocating 90% of the original training set for training and 10% for validation.

3.1.5 Mid-Air

The Mid-Air dataset [Fonder and Van Droogenbroeck, 2019] captures low-altitude quadcopter flight trajectories under varying climate conditions. It introduces challenges such as aerial perspectives and obscured distant objects. We utilize Mid-Air to train our model in long distance outdoor scenes, like the ones depicted in Figure 5, splitting the official training data to allocate two-thirds for training and one-third for validation.

3.1.6 TartanAir

TartanAir [Wang *et al.*, 2020] is a massive synthetic dataset covering more than 30 diverse indoor and outdoor environments, including urban, rural, indoor, and nature settings, all presented under varying lighting conditions, weather scenarios, and with dynamic objects. We leverage TartanAir to improve overall model generalization, splitting two thirds of the original training scenes for training and one third for evaluation.

It is important to note that some pixels within the depth annotations of the TartanAir dataset are marked as invalid. These invalid pixels may result from occlusions, reflections, or other challenges within the simulated environments. During the training process, these invalid pixels are ignored to

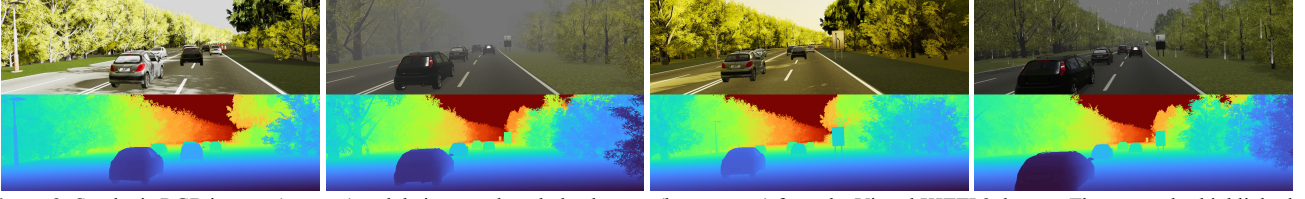


Figure 3. Synthetic RGB images (top row) and their ground-truth depth maps (bottom row) from the Virtual KITTI 2 dataset. These samples highlight the dataset’s ability to provide realistic simulations of various weather conditions, paired with accurate depth information. The second image also displays one of the challenges of this dataset: The depth annotation provides labels for every pixel on the image, even for those objects occluded by the fog.

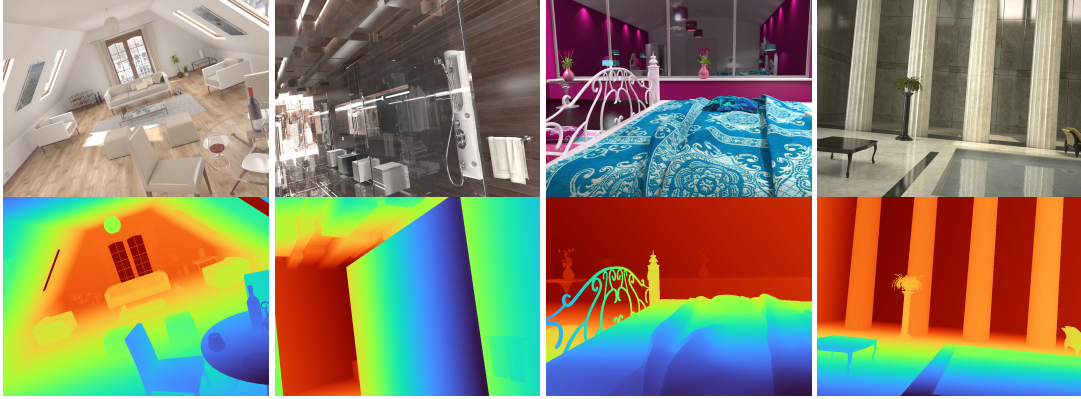


Figure 4. Synthetic RGB images (top row) and their corresponding ground-truth depth maps (bottom row) generated using the Hypersim dataset. This dataset offers a significant advantage by providing examples of transparent and reflective surfaces, like the glass plane in the second image or the mirror in the third one, which are often challenging to handle in real-world scenarios.

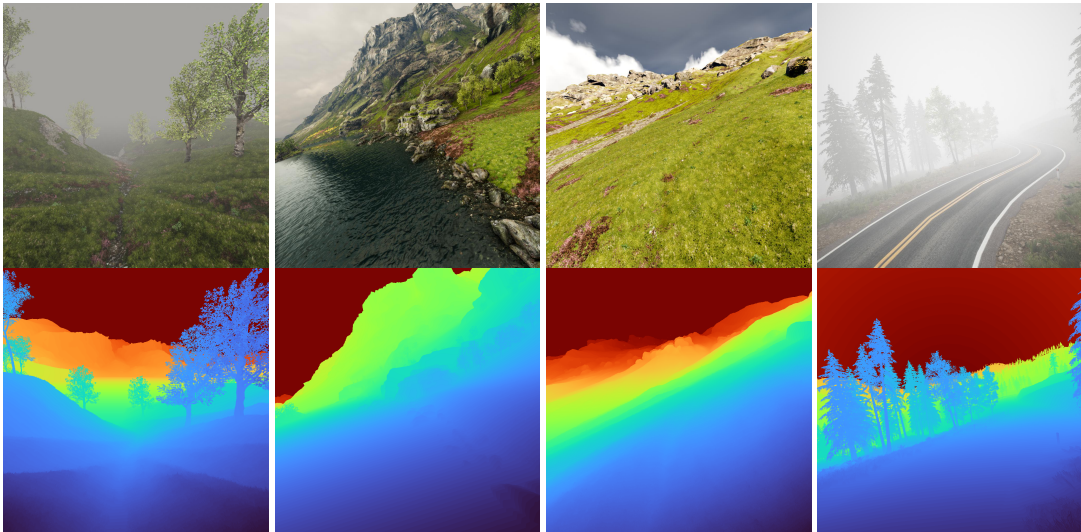


Figure 5. Synthetic RGB images (top row) and their corresponding ground-truth depth maps (bottom row) generated using the Mid-Air dataset. This dataset offers challenging scenarios of low-altitude drone flights in diverse weather conditions, providing valuable data for training and evaluating computer vision models. The first and last images show some annotated pixels occluded by fog, which presents a challenge.

ensure that the depth estimation model is not influenced by potentially misleading data. Figure 6 shows samples of this dataset with some corrupted pixels painted in black.

3.2 Real Datasets

This section describes the main characteristics of the real datasets used in our experiments to evaluate the performance of the model.

3.2.1 KITTI

The KITTI dataset [Geiger *et al.*, 2013] consists of real-world outdoor driving sequences captured in Germany. The ground-

truth depth maps, as seen in Figure 7, are derived from a rotating 3D LiDAR sensor, resulting in sparse, incomplete annotations that do not cover upper image regions. To benchmark the real-world outdoor performance of the trained model, we used the test split proposed by Eigen *et al.* [2014].

3.2.2 NYU Depth V2

The NYU Depth V2 dataset [Silberman *et al.*, 2012] represents a standard indoor benchmark captured using a Microsoft Kinect sensor. Its depth maps are characterized by typical real-world sensor limitations, including noise on flat walls and missing details around reflective or transparent surfaces (see Figure 8). We utilize the officially annotated pairs to

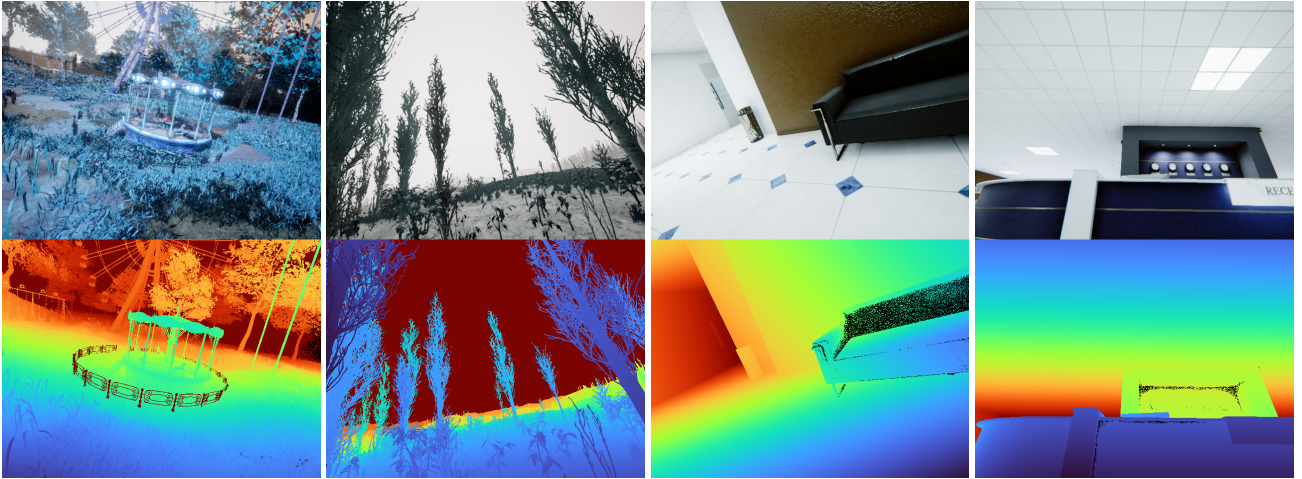


Figure 6. Synthetic RGB images (top row) and their corresponding ground-truth depth maps (bottom row) from the TartanAir dataset. In the first two images, there are samples of the fine annotated depth estimation maps present in this dataset, with the following two images demonstrating some of the annotation issues (the dark pixels on the ground-truth images).

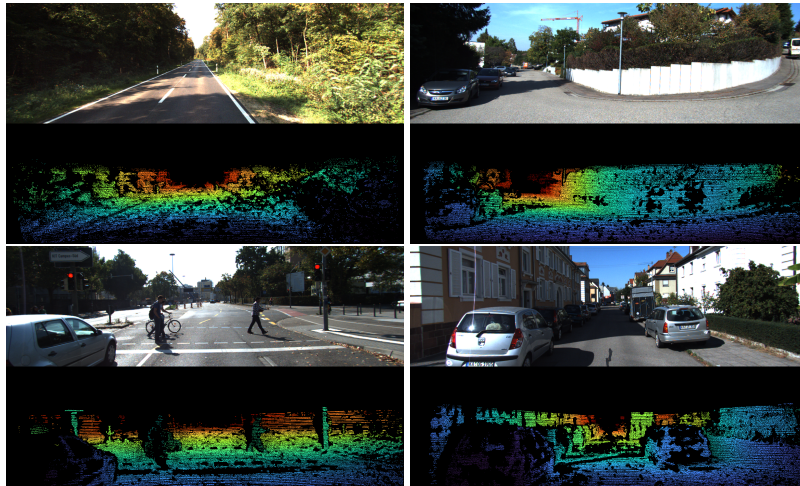


Figure 7. RGB images (top row) and their corresponding ground-truth depth maps (bottom row) from the KITTI dataset. While the RGB images provide rich visual information, the depth maps, generated using LiDAR, suffer from sparsity and limited detail, particularly for distant objects and complex scenes.

evaluate real-world indoor performance. We used the official test split to evaluate our model in a real-world indoor environment.

4 Proposed Method

In this section, we present the proposed monocular depth estimation method, a fully convolutional deep neural network trained only on synthetic images. A high-level design of the architecture can be seen in Figure 9.

4.1 Overview

This work introduces a novel monocular depth estimation network based on a fully convolutional encoder-decoder architecture. The network strategically combines a ConvNeXt backbone, chosen for its competitive performance against transformer-based models and its advantageous data efficiency, with a custom-designed, U-Net-inspired decoder. This combination aims to optimize both robust feature extraction and precise depth map reconstruction.

Formally, the proposed method can be represented as a

learnable function F composed of two distinct components: an encoder E , implemented as a convolutional backbone network, and a convolutional decoder D . Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the image height and width, respectively, the encoder E extracts a set of multi-scale feature vectors $E(I) = \{e_1, e_2, e_3, e_4\}$. These feature vectors are derived from successive layers of the encoder, with e_4 representing the final, most abstract latent representation of the input image.

The decoder D then processes the feature vectors generated by the encoder to produce the final estimated depth map $d \in \mathbb{R}^{H \times W}$. This is achieved through a series of upsampling operations applied sequentially to e_4 . After each upsampling step, the resulting feature maps are concatenated with the corresponding feature vectors from the encoder in reverse order. This process of progressive upsampling and feature fusion allows the decoder to refine the depth map, incorporating both coarse and fine-grained information from the encoded feature representations.

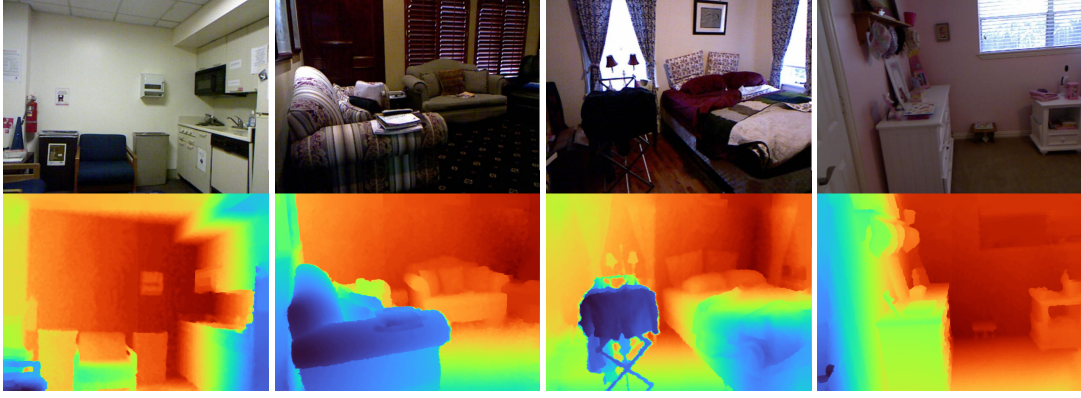


Figure 8. RGB images (top row) and their corresponding ground-truth depth maps (bottom row) from the NYU V2 dataset. Despite the dense annotations, those samples demonstrate the inaccurate nature of real-world depth annotations, such as the jagged edges and prevalent noise, as can be seen on walls.

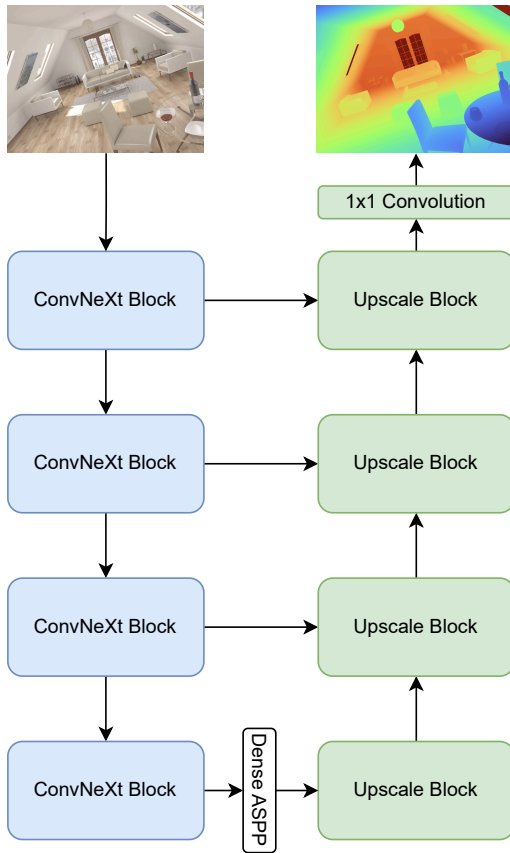


Figure 9. Diagram of the proposed deep neural network architecture. The input image goes through a series of ConvNeXt blocks, which extract features while reducing image size. The resulting encoded features are then processed by a Dense ASPP layer to capture multi-scale relations of features. The decoder then recovers the original image size while simultaneously predicting depth information. Finally, the upscaled features are combined to produce the final depth estimation. The partial downscaled feature maps are summed to the equivalent upscaled ones, forming a residual connection.

4.2 Architecture

The selection of ConvNeXt as the encoder backbone is motivated by its demonstrated efficacy in visual representation learning. Liu *et al.* [2022] established that ConvNeXt achieves performance comparable to or superior to transformer architectures while exhibiting higher data efficiency, a crucial factor given our dependence on synthetic training data. Furthermore, utilizing a pre-trained ConvNeXt backbone significantly accelerates the training process, capitalizing on established feature representations and reducing convergence

time.

Following the work of Yang *et al.* [2018], a Dense Atrous Spatial Pyramid Pooling (Dense ASPP) layer was included between the deeper output of the encoder and its corresponding upscale block. This allows the model to better capture multi-scale relationships between the encoder features. This layer first applies a 1×1 convolution to reduce the number of features to half the input. Then, it performs four convolutions with different dilation rates: 1, 2, 4 and 8, each one with the concatenation of the results of the previous ones. The final results are then concatenated and rearranged to the same feature count as the input by 1×1 convolutions. Figure 10 shows a simplified diagram of this layer.

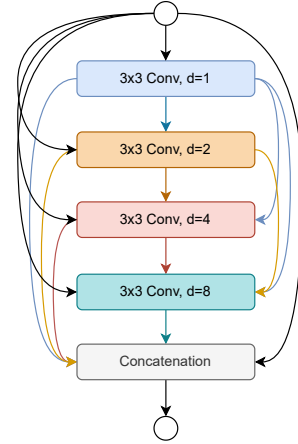


Figure 10. Simplified diagram of a Dense ASPP layer. The features are processed by a sequence of dilated convolutions, with expanding dilation rates d . An arrow from a convolution to another represents the concatenation of the output of one to the input of another.

The decoder, inspired by the U-Net architecture [Ronneberger *et al.*, 2015], is designed to facilitate precise upsampling and effective feature fusion for depth-map generation by progressively reconstructing the depth map from the encoded feature representation. This design allows for the integration of multiscale features extracted by the ConvNeXt backbone, enabling the network to capture both global context and fine-grained details essential for accurate depth estimation. To enhance the quality of upsampling, we replace traditional bilinear interpolation with pixel shuffle [Shi *et al.*, 2016], a learnable upscaling method that improves spatial resolution and preserves structural information. The decoder consists

of four sequential upsampling modules, each meticulously crafted to refine the feature maps and recover spatial resolution. This convolutional architecture aims to demonstrate the feasibility of robust monocular depth estimation on real-world datasets using models trained exclusively on synthetic data, thereby addressing the domain-gap challenge. Figure 11 illustrates the structure of an upsampling module.

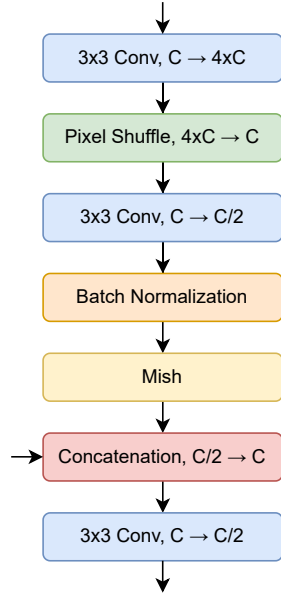


Figure 11. Diagram of the proposed upsampling module. The channels of the input feature map are quadrupled, then a pixel shuffle layer rearranges those extra channels to double the image dimension. The number of channels of the resulting feature map is then halved, a normalization and an activation layer are applied, and then the feature map extracted from the encoder is concatenated. A final convolution merges those features.

Each upsampling module consists of the following sequence: (1) a 3×3 convolutional layer to increase the channel dimensionality, preparing the data for pixel shuffle; (2) a pixel shuffle layer, which efficiently doubles the spatial resolution by rearranging channel information, offering a learned and more detailed upscaling compared to bilinear interpolation which can introduce blurring; (3) a 3×3 convolutional layer to reduce the channel count for subsequent processing; (4) a batch normalization layer to stabilize training; and (5) a Mish activation function [Misra, 2019] to introduce non-linearity. The ability of a pixel shuffle to learn the optimal arrangement of pixels for upscaling results in sharper and more accurate depth maps, particularly crucial for preserving fine details.

To further enhance depth map accuracy, skip connections are incorporated to integrate high-resolution features from the encoder stages. Specifically, the output of each upsampling module is concatenated with the corresponding encoder feature map, followed by a 3×3 convolution to reduce the channel dimension by half, effectively fusing multiscale information. This integration enables the decoder to take advantage of both the global context and fine-grained details, leading to more precise depth estimations.

Following the final upsampling module, where the feature map matches the input image’s spatial dimensions, a 1×1 convolutional layer is employed. This allows the model to aggregate the learned features into a single depth value, maintaining the fine-grained details learned through the process.

Finally, a ReLU activation function [Agarap, 2018] is applied to the output depth map. This ensures that all predicted depth values are non-negative, aligning with the physical constraint that depth or disparity values cannot be negative in our context. The use of a final ReLU layer is essential to enforce physically plausible predictions.

4.3 Training Protocol

Diverse synthetic datasets were used to train our model, and two real-world datasets commonly used in the literature, KITTI and NYU Depth V2, were used to evaluate the capabilities of our model to generalize to real data without any further training. Table 3 shows the datasets used in the training and evaluation.

Table 3. Usage of datasets on training and evaluation steps.

Dataset	Training	Evaluation
SYNTHIA	✓	✗
Synscapes	✓	✗
Virtual KITTI 2	✓	✗
Mid-Air	✓	✗
Hypersim	✓	✗
TartanAir	✓	✗
KITTI	✗	✓
NYU Depth V2	✗	✓

During the training of our model, a series of transformations originally defined on the training procedure of the pre-trained encoder are applied to the input image. Initially, the values are rescaled to $[0.0, 1.0]$. The normalization is then performed using the mean of $[0.485, 0.456, 0.406]$ and the standard deviation of $[0.229, 0.224, 0.225]$. These transformations are applied to preserve the expected input distribution for RGB images, given the pre-training of the encoder with these data.

4.3.1 Loss Function

Training machine learning models for depth estimation, especially when integrating multiple datasets in a concatenated manner, presents significant challenges. These challenges stem from the inherent diversity of the datasets, including variations in depth range, distribution shifts, and potential ambiguities between data sources. To address these difficulties, Ranftl *et al.* [2020] introduced a scale- and shift-invariant loss function. This innovative approach allows relative depth estimates to be learned from a wide range of data sets, thereby mitigating the effects of absolute depth value discrepancies.

The downside of this loss function is that it can only supervise the estimation of relative depth maps. These maps do not represent the absolute distance from a pixel to the corresponding object but instead maintain depth values that are consistent relative to each other within the image. Since the magnitude of these estimates has no absolute significance, the depth map can be normalized to any range without losing information. Figure 12 illustrates an example of three equivalent relative depth maps.

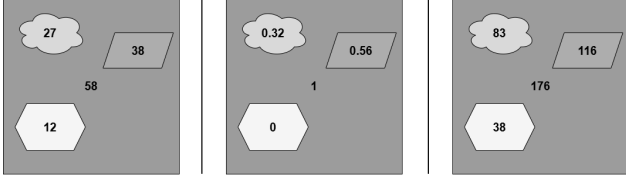


Figure 12. Illustration of relatively equivalent depth maps. The central depth map is generated by applying a min-max normalization to either the left or right depth map, scaling the values to the range $[0, 1]$.

This loss function is defined in two elements: The residual loss and the gradient loss. The residual loss is the component that ensures the depth values are being learned by the proposed network. However, a simple comparison of values would not be enough to address the challenges of training with multiple datasets, with some normalization of the predicted values needed beforehand. To that end, the residual loss aligns both the prediction and the ground truth, taking into account the scale and shift factors. The estimation of the shift $t(\cdot)$ and the scale $s(\cdot)$ for a depth map d is derived as follows:

$$t(d) = \text{median}(d), \quad s(d) = \frac{1}{M} \sum_{i=1}^M |d - t(d)|, \quad (1)$$

where M is the number of pixels in d . Then, with the estimation of those factors, both ground truth and prediction are aligned to have zero scale and shift:

$$\hat{d} = \frac{d - t(d)}{s(d)}. \quad (2)$$

The error of the aligned prediction $\hat{d}^*$ with the aligned ground truth $\hat{d}$ is then computed using a simple MSE error:

$$\mathcal{L}_{\text{residual}}(\hat{d}^*, \hat{d}) = \frac{1}{M} \sum_{i=1}^M (\hat{d}_i^* - \hat{d}_i)^2. \quad (3)$$

To avoid the influence of outliers on the training process, the larger residuals 20% are ignored in every image.

The gradient loss is a second element of this loss function with the task of encouraging the network to preserve sharp edges on object borders. Since there are several parts of depth maps that are composed of smooth depth gradients, like walls and floors, sharp depth changes are often ignored by the model. Based on the gradient matching term proposed by Li and Snavely [2018], the gradient loss is defined as

$$\mathcal{L}_{\text{gradient}}(\hat{d}^*, \hat{d}) = \frac{1}{M} \sum_{k=1}^K \sum_{i=1}^M (|\nabla_x R_i^k| + |\nabla_y R_i^k|), \quad (4)$$

where $R_i = \hat{d}_i^* - \hat{d}_i$ and R^k denotes the disparity difference at the scale k of $K = 4$ scale levels, with the image size being halved at each scale level.

The final loss function used to train our models is

$$\mathcal{L}(\hat{d}^*, \hat{d}) = \mathcal{L}_{\text{residual}}(\hat{d}^*, \hat{d}) + \alpha \mathcal{L}_{\text{gradient}}(\hat{d}^*, \hat{d}), \quad (5)$$

where α is a weighting parameter, empirically set to 1.

5 Implementation Details

Given that a substantial part of our network architecture is the ConvNeXt-based encoder, we followed some design decisions from the ConvNeXt authors. We trained our model for

300 epochs on a setup where we concatenated all the training datasets into a single larger virtual dataset.

To minimize the loss function over the training epochs, we choose the AdamW [Loshchilov, 2017] Optimizer. By incorporating decoupled weight decay, AdamW effectively mitigates the vanishing gradient problem, enabling the network to learn meaningful representations even in deep architectures. Additionally, its adaptive learning rate scheme allows for efficient convergence, particularly in the presence of noisy gradients. Furthermore, AdamW’s regularization properties promote generalization and prevent overfitting, leading to more accurate and reliable models.

We also used a Cosine Annealing with Warm Restarts [Loshchilov and Hutter, 2016] as a learning rate scheduler, with a single restart at 100 epochs. Cosine annealing is a learning rate scheduling technique characterized by an initial, relatively high learning rate that is progressively reduced to a minimum value, followed by a rapid increase. This periodic resetting of the learning rate emulates a restart of the training process. When the optimization process resumes from the previously attained, advantageous weights, this is termed a “warm restart”. We set up the initial learning rate as $1e^{-3}$.

As data augmentation techniques, we applied a random number of 90° rotations to the input and label images alongside a random adjustment of contrast and brightness in the input image.

Given that the main task of the backbone of our proposed method is to provide a robust representation of the input image, we started our training process with a set of pre-trained parameters for the encoder. While this pre-training approach speeds up convergence and improves performance, it introduces a limitation: the pre-trained network was designed to process square images.

To adapt the different aspect ratios of our datasets to our model, we center-crop each input image and its corresponding depth map to a square, using the smaller image dimension as the side length and cropping the larger dimension equally from both sides. This cropping strategy ensures compatibility with the pre-trained encoder while mitigating potential performance degradation due to information loss. It also standardizes the input format, simplifying the training process when combining datasets with different image aspect ratios.

A significant challenge in outdoor depth estimation is the handling of pixels representing the sky. Since depth annotations are typically unavailable for such regions, naive approaches, such as ignoring these pixels, can negatively impact model performance. In particular, objects adjacent to the sky may suffer from reduced accuracy, and the model may generate incorrect depth predictions for parts of the image adjacent to the sky. To address these issues, we adopt the strategy proposed by Birkl *et al.* [2023]. Using the object segmentation annotations available in our datasets, we assign sky pixels a depth value corresponding to the most distant object in the scene.

6 Experimental Results

In this section, we present an analysis of the results of our proposed model in two real-world datasets: KITTI [Geiger *et al.*, 2013], composed of street scenes from a roaming car, and NYU Depth V2 [Silberman *et al.*, 2012], composed of indoor scenes.

6.1 Evaluation Methodology

To evaluate our models, we applied the same transformations used during training to the input data, excluding the data augmentation techniques. Additionally, before computing the evaluation metrics, we rescaled the predictions to match the distribution of the ground truth, as we are dealing with relative depth. This rescaling follows the scale and shift estimation procedure proposed by Ranftl *et al.* [2020]: Given a prediction d^* and a ground truth d , we first invert the ground truth to obtain the disparity representation of the depth map:

$$d' = \frac{1}{d}. \quad (6)$$

An estimate of the scale and shift, denoted by s and t respectively, is computed between d^* and d' using the closed-form least-squares solution:

$$(s, t)^\top = \left(\sum_{i=1}^M \vec{d}_i^* \vec{d}_i'^\top \right)^{-1} \left(\sum_{i=1}^M \vec{d}_i^* d'_i \right), \quad (7)$$

where d_i^* and d'_i are individual pixels of the predicted and ground truth disparity maps, and $\vec{d}_i^* = (d_i^*, 1)^\top$. The aligned prediction d^{**} is then obtained as:

$$d^{**} = d^* s + t. \quad (8)$$

To compute the evaluation metrics in depth space, the aligned disparity prediction is inverted, yielding the estimated depth map d^{*-1} , which is then compared to the original ground-truth depth map d .

Given the scale- and shift-invariant properties of the proposed network’s loss function, model evaluation was conducted using metrics independent of the absolute depth in the ground truth. This approach focuses solely on the relative depth information within each image, quantified as the ratio between a pixel’s depth and the image’s minimum and maximum depth. This ensures that the evaluation process remains agnostic to any inherent scaling or shifting factors in the data.

We assess the performance of our depth estimation model using two primary metrics: the delta score and the absolute relative error (AbsREL). The delta score quantifies the percentage of pixels where the relative error between the predicted and ground-truth depth is less than a specified threshold. It is defined as:

$$\delta_i(d^{*-1}, d) = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left[\max \left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i} \right) < 1.25^i \right], \quad (9)$$

where N is the number of pixels in the depth map and $\mathbb{I}$ is the indicator function, which is equal to 1 if the condition is true and 0 otherwise. In simpler terms, the delta score

measures the proportion of pixels where the ratio between the predicted and ground-truth depths (in either direction) is within a threshold of 1.25^i . Higher delta scores indicate better model performance, as they correspond to a larger percentage of accurate pixel predictions.

The AbsREL metric, on the other hand, measures the average absolute difference between the predicted and ground-truth depth, normalized by the ground-truth depth. It is defined as:

$$\text{AbsREL}(d, d^*) = \frac{1}{N} \sum_{i=1}^N \left| \frac{d_i - d_i^*}{d_i} \right|. \quad (10)$$

Lower AbsREL values reflect better performance, as they indicate a smaller average relative difference between the predicted and ground-truth depths.

By employing both the delta score and AbsREL, we obtain a comprehensive evaluation of our model’s capability to generate accurate depth maps. The delta score provides insight into the model’s ability to predict accurate depths across different ranges, while AbsREL offers a complementary measure of the overall average error.

6.2 Quantitative Analysis

To contextualize the performance of our method within the broader literature, we compared the extracted metrics with those reported by other state-of-the-art relative depth estimation methods: Depth Anything V2 [Yang *et al.*, 2024b], DPT [Ranftl *et al.*, 2021], and MiDaS [Ranftl *et al.*, 2020]. These methods were selected for comparison because they employ the same loss function as the proposed method, producing relative disparity maps. Furthermore, Depth Anything V2 incorporates synthetic data into its training protocol, while MiDaS employs a convolutional network architecture similar to ours. In contrast, DPT and Depth Anything V2 are based on Vision Transformers. Tables 4 and 5 summarize our results on the KITTI and NYU datasets, respectively.

Table 4. Depth estimation performance comparison on the KITTI dataset. MiDaS does not report AbsRel values.

Method	AbsRel ↓	Delta1 ↑
Ours	0.075	0.914
Depth Anything V2	0.080	0.939
DPT	0.127	0.850
MiDaS	-	0.761

Table 5. Depth estimation performance comparison on the NYU dataset. MiDaS does not report AbsRel values.

Method	AbsRel ↓	Delta1 ↑
Ours	0.104	0.885
Depth Anything V2	0.046	0.979
DPT	0.048	0.980
MiDaS	-	0.904

The quantitative results demonstrate the promising performance of our model. On the KITTI dataset, our model

achieves competitive results compared to the Depth Anything V2 model, despite it being larger and using a large quantity of real data during training, while outperforming both DPT and the convolutional-based MiDaS. However, in the NYU dataset, the Transformer-based approach yields better results, with MiDaS performing similarly to our model. A key point to emphasize is that our method relies entirely on synthetic data, whereas the other methods are trained using real images, either annotated or unannotated.

Our strong performance on the KITTI dataset can be attributed to the significant presence of synthetic vehicle-mounted camera data in our training set, particularly from the Virtual KITTI dataset – a synthetic counterpart to KITTI. This highlights the capability of our model to effectively transfer knowledge from a synthetic environment to a real-world setting. In contrast, our model underperforms on the NYU dataset due to the lack of a sufficiently similar synthetic dataset, limiting its ability to execute a refined zero-shot synthetic-to-real domain transfer.

6.3 Qualitative Analysis

To further assess the quality of the predictions of our model, this section presents a qualitative analysis of our results.

Figure 13 presents representative depth predictions generated by our proposed monocular depth estimation model on the KITTI autonomous driving dataset. A qualitative evaluation of these results reveals several strengths of our approach, especially taking into account that no real data was used in the training process. The model demonstrates an impressive ability to capture fine details, such as lamp heads, foliage, and pedestrians, with coherent depth predictions across these complex scene elements. Moreover, the model maintains robust performance despite the presence of abrupt color transitions on planar surfaces – arising from lane markings, material changes (e.g., asphalt to grass), or varying illumination conditions due to reflections and shadows – which indicates a strong understanding of the underlying scene geometry.

Further analysis highlights the model’s resilience in low-contrast settings, like on the edge of the images in Figure 13, a common characteristic of the KITTI dataset owing to the limited dynamic range of the capturing camera. Even in such challenging conditions, the proposed model produces geometrically precise depth estimations, reinforcing its applicability in real-world scenarios. Additionally, the model correctly predicts depth for transparent surfaces, such as car windows, thereby avoiding the typical error of incorrectly assigning depth to objects visible through these surfaces.

However, the qualitative assessment also identifies areas for improvement. The model occasionally fails to accurately predict the depth of extremely thin structures, such as wires and slender poles. This shortcoming is likely attributable to the reduced resolution of the input images, which may lead to undersampling and subsequent loss of critical features necessary for detecting these fine objects. Similarly, dark objects and overexposed regions, such as the sky, continue to pose challenges, resulting in less reliable depth estimations in these areas.

Overall, while the proposed model exhibits considerable strengths in capturing detailed and geometrically consistent

depth predictions in a variety of challenging conditions, further refinements in input resolution and feature extraction strategies may be necessary to address its limitations in handling fine and low-contrast structures.

Figure 14 presents representative predictions generated by our model in the NYU V2 dataset. Unlike autonomous driving datasets, indoor environments are characterized by a more limited depth range but exhibit a high degree of variability in object types, arrangements, and textures. This variability introduces a significant challenge for monocular depth estimation. Nevertheless, our model demonstrates competitive performance, particularly considering the substantial domain shift from synthetic training data to real-world indoor scenes.

Despite the complexity of indoor scenes, our method exhibits a strong understanding of indoor geometry. Large objects such as beds, tables, and wardrobes are consistently identified with well-defined edges and precise relative depth predictions. The model also handles textured flat surfaces without confusion and demonstrates robustness to visual artifacts such as reflections and shadows, further highlighting its capacity to generalize to real-world scenarios.

However, the model shows limitations in predicting the depth of small, detailed objects. This suggests that the model’s geometric understanding, acquired during training, does not fully extend to fine structures absent from the training data. Additionally, large indoor spaces with extended depth ranges pose a challenge, likely due to the limited depth diversity in the training datasets, which were primarily composed of smaller-scale indoor environments.

Overall, our qualitative analysis indicates that the model effectively adapts to domain shifts between synthetic and real-world datasets when the training data includes synthetic datasets closely resembling the target domain. This is evident in the case of the KITTI dataset, where training included Virtual KITTI, Synscapes, and SYNTHIA – synthetic datasets with scenes highly similar to real driving scenarios. In contrast, for NYU V2, the model was trained on Hypersim and some scenes from TartanAir, which, despite being indoor datasets, lack the specific structural and depth characteristics of NYU V2. This suggests that the absence of a more closely related synthetic indoor dataset limits the model’s capacity to generalize to complex indoor environments.

7 Conclusions

This research addressed the critical challenge of monocular depth estimation, focusing on achieving a balance between accuracy and computational efficiency, a key requirement for deployment in real-world applications like autonomous systems and augmented reality. While Vision Transformers (ViTs) have demonstrated strong performance by capturing global context, their significant computational demands often render them impractical. This work revisited the potential of Convolutional Neural Network (CNN) architectures, specifically tailoring a fully convolutional model for this task.

The primary contributions of this study are:

- **Efficient CNN Design for Depth Estimation:** We proposed and validated a specific CNN architecture optimized for monocular depth estimation, demonstrating

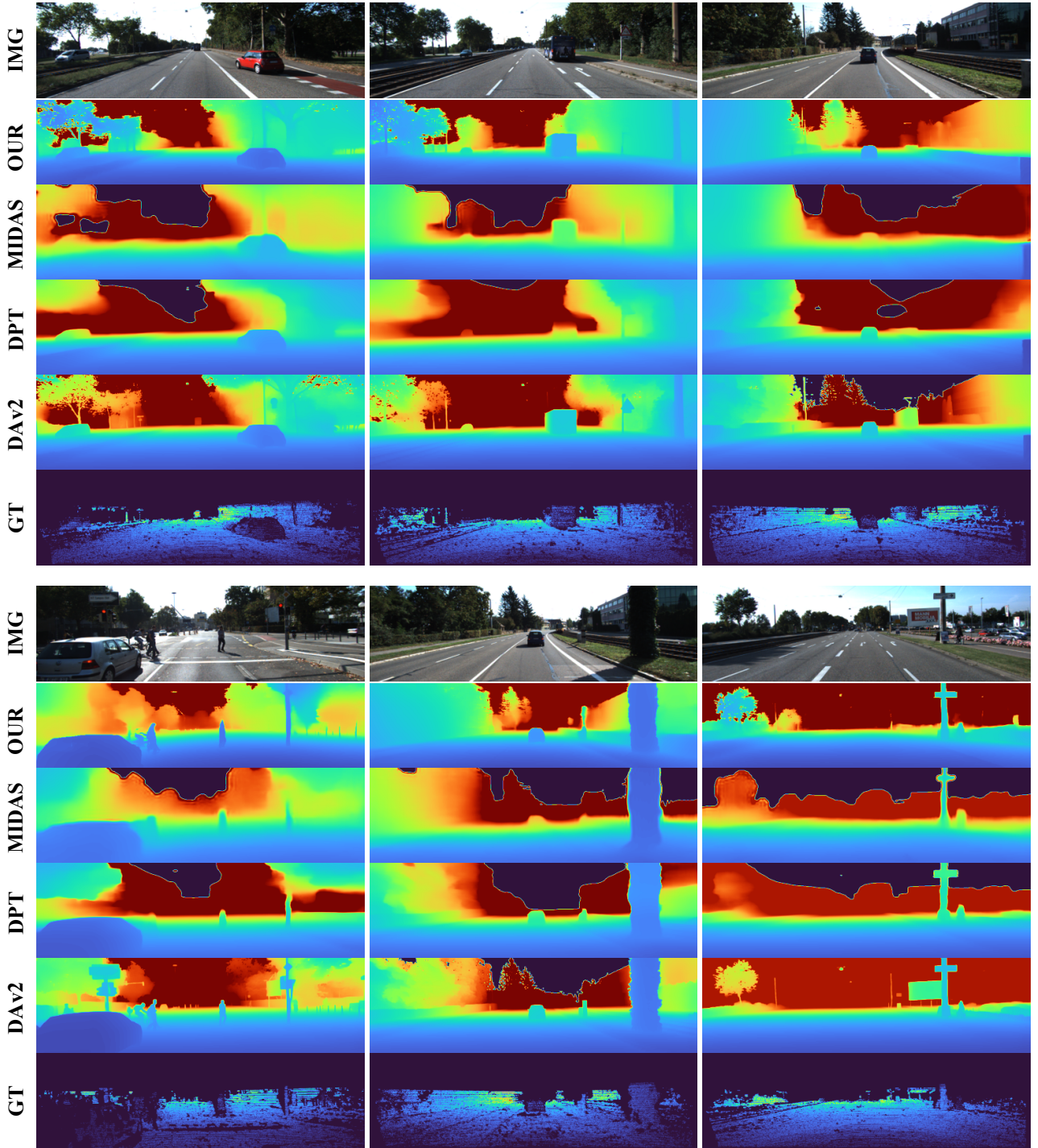


Figure 13. Samples of KITTI dataset and their prediction on the evaluated networks. From top to bottom: Input Image (IMG), our model prediction (OUR), MiDaS V1 prediction (MIDAS), DPT prediction (DPT), Depth Anything V2 prediction (DAV2) and the ground truth (GT). The network outputs, after the scale and shift alignment described in our evaluation methodology, are clipped to the maximum depth of the dataset (10 meters).

that carefully designed convolutional networks can still offer competitive performance against more computationally intensive models, particularly when resource constraints are a factor.

- **Effective Synthetic-to-Real Generalization:** We demonstrated the efficacy of training exclusively on synthetic datasets, leveraging their perfectly annotated depth information. Our findings show that the proposed model, despite being trained solely on synthetic data (like Synscapes, inferred from the references), achieves notable generalization performance on unseen real-world sce-

narios, mitigating the need for expensive and often imperfect real-world depth annotations.

- **Benchmarking Against Modern Approaches:** By evaluating our model and comparing its performance (implicitly, as detailed results are assumed to be in the paper body) against the backdrop of contemporary methods, including potentially ViT-based approaches like Depth Anything (referenced in the paper), this work provides valuable insights into the current trade-offs between CNNs and ViTs for monocular depth estimation, reaffirming the continued relevance of optimized CNNs.

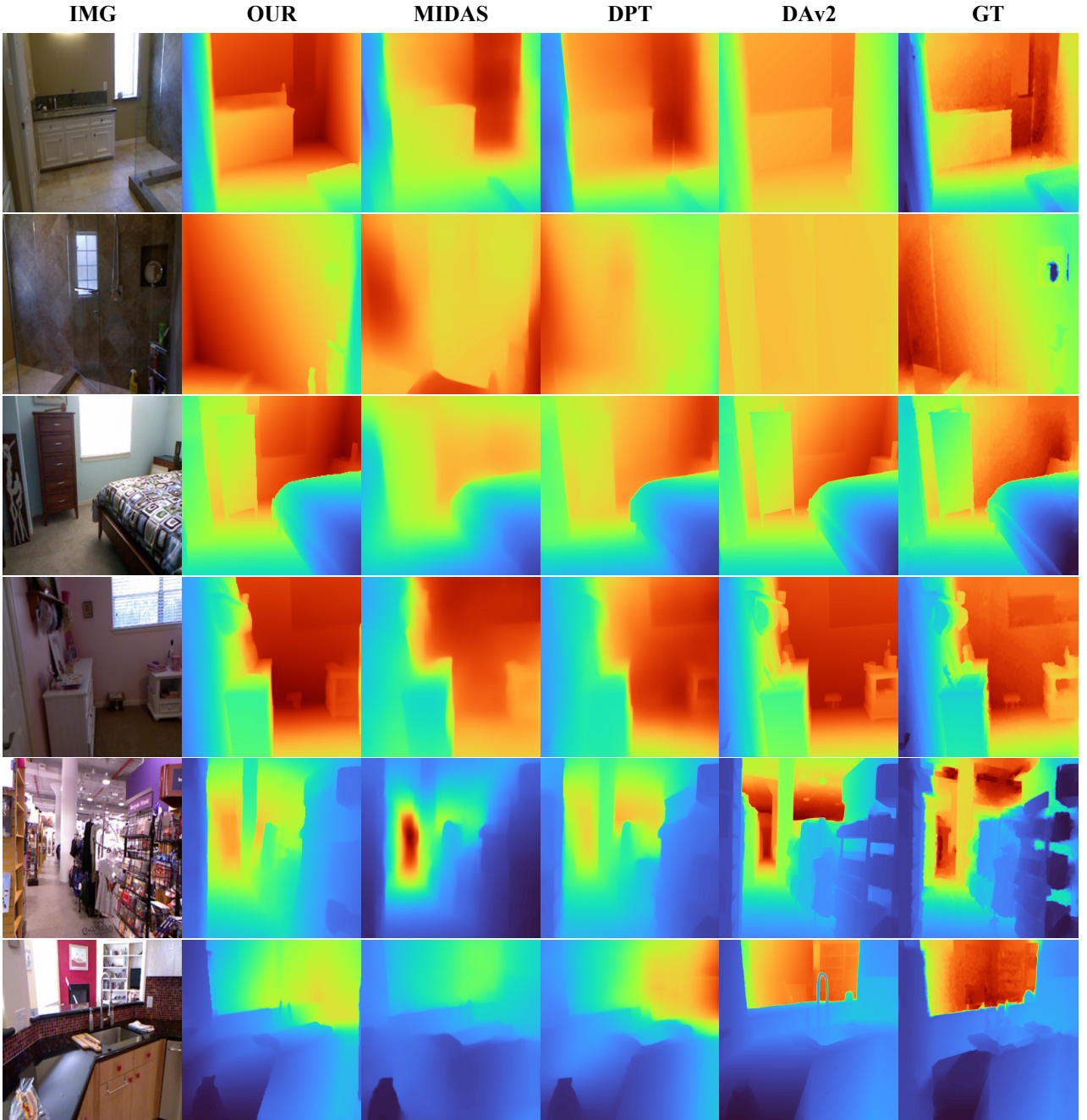


Figure 14. Samples of NYU dataset and their prediction on the evaluated networks. From left to right: Input Image (IMG), our model prediction (OUR), MiDaS V1 prediction (MIDAS), DPT prediction (DPT), Depth Anything V2 prediction (DAv2) and the ground truth (GT). The network outputs, after the scale and shift alignment described in our evaluation methodology, are clipped to the maximum depth of the dataset (10 meters).

In summary, this paper contributes a computationally efficient CNN-based solution for monocular depth estimation that successfully bridges the synthetic-to-real domain gap. By achieving robust generalization from synthetic data alone, this work offers a practical pathway for developing depth perception systems where real-world training data is scarce or unavailable, while highlighting the enduring value of optimized convolutional architectures in the current deep learning landscape. Future work could explore hybrid approaches or further investigate techniques to minimize the residual domain gap between synthetic and real data.

Declarations

Authors' Contributions

LGLD contributed to the conception of this study, the implementation of codes, the conduction of experiments, the evaluation of results, and the preparation of the manuscript. JLFC, MRS, HAM, and HP contributed by editing and reviewing the manuscript, as well as supervising the research work. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank the Artificial Intelligence (Recod.ai) Laboratory for the support with computational resources.

Funding

This research received no funding.

Availability of data and materials

Data can be made available upon request.

References

- Agarap, A. (2018). Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375*, pages 1–7. DOI: 10.48550/arXiv.1803.08375.
- Atapour-Abarghouei, A. and Breckon, T. P. (2018). Real-Time Monocular Depth Estimation Using Synthetic Data with Domain Adaptation via Image Style Transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2800–2810. DOI: 10.1109/CVPR.2018.00296.
- Birkel, R., Wofk, D., and Müller, M. (2023). MiDaS v3. 1 – A Model Zoo for Robust Monocular Relative Depth Estimation. *arXiv preprint arXiv:2307.14460*, pages 1–14. DOI: 10.48550/arXiv.2307.14460.
- Cabon, Y., Murray, N., and Humenberger, M. (2020). Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, pages 1–11. DOI: 10.48550/arXiv.2001.10773.
- Decker, L. G. L., Campana, J. L. F., Souza, M. R., Almeida Maia, H., and Pedrini, H. (2024). Zero-Shot Synth-to-Real Depth Estimation: From Synthetic Street Scenes to Real-World Data. In *37th SIBGRAP Conference on Graphics, Patterns and Images*, pages 1–6. IEEE. DOI: 10.1109/SIBGRAP62404.2024.10716265.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2020). An Image is Worth 16×16 Words. *arXiv preprint arXiv:2010.11929*, 7. DOI: 10.48550/arXiv.2010.11929.
- Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Advances in Neural Information Processing Systems*, 27:2366–2374. Available at: https://proceedings.neurips.cc/paper_files/paper/2014/file/91c56ce4a249fae5419b90cba831e303-Paper.pdf.
- Fonder, M. and Van Droogenbroeck, M. (2019). Mid-Air: A Multi-Modal Dataset for Extremely Low Altitude Drone Flights. In *Computer Vision and Pattern Recognition Workshops*, pages 1–10. DOI: 10.1109/CVPRW.2019.00081.
- Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011. DOI: 10.1109/CVPR.2018.00214.
- Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision Meets Robotics: The KITTI Dataset. *The International Journal of Robotics Research*, 32(11):1231–1237. DOI: 10.1177/0278364913491297.
- Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). Deeper Depth Prediction with Fully Convolutional Residual Networks. In *Fourth International Conference on 3D Vision*, pages 239–248. IEEE. DOI: 10.1109/3DV.2016.32.
- Li, R., Xian, K., Shen, C., Cao, Z., Lu, H., and Hang, L. (2019). Deep Attention-based Classification Network for Robust Depth Prediction. In *14th Asian Conference on Computer Vision*, pages 663–678, Perth, Australia. Springer. DOI: 10.1007/978-3-030-20870-7_41.
- Li, Z. and Snavely, N. (2018). MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050. DOI: 10.1109/CVPR.2018.00218.
- Liu, F., Shen, C., and Lin, G. (2015). Deep Convolutional Neural Fields for Depth Estimation from a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170. DOI: 10.1109/CVPR.2015.7299152.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A ConvNet for the 2020s. In *Computer Vision and Pattern Recognition*, pages 11976–11986. DOI: 10.1109/CVPR52688.2022.01167.
- Lopez-Rodriguez, A. and Mikolajczyk, K. (2023). DESC: Domain Adaptation for Depth Estimation via Semantic Consistency. *International Journal of Computer Vision*, 131(3):752–771. DOI: 10.1007/s11263-022-01718-1.
- Loshchilov, I. (2017). Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*. DOI: 10.48550/arXiv.1711.05101.
- Loshchilov, I. and Hutter, F. (2016). SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint arXiv:1608.03983*. DOI: 10.48550/arXiv.1608.03983.
- Masoumian, A., Rashwan, H. A., Cristiano, J., Asif, M. S., and Puig, D. (2022). Monocular Depth Estimation Using Deep Learning: A Review. *Sensors*, 22(14):5353. DOI: 10.3390/s22145353.
- Misra, D. (2019). Mish: A Self Regularized Non-Monotonic Activation Function. *arXiv preprint arXiv:1908.08681*. DOI: 10.5244/C.34.191.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*. DOI: 10.48550/arXiv.2304.07193.
- Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision Transformers for Dense Prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 12179–12188. DOI: 10.1109/ICCV48922.2021.01196.
- Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637. DOI: 10.1109/TPAMI.2020.3019967.
- Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. (2021).

- Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *International Conference on Computer Vision*, pages 10912–10922. DOI: 10.1109/ICCV48922.2021.01073.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, Munich, Germany. Springer. DOI: 10.1007/978-3-319-24574-4_28.
- Ros, G., Sellart, J., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243. DOI: 10.1109/CVPR.2016.352.
- Roy, A. and Todorovic, S. (2016). Monocular Depth Estimation using Neural Regression Forest. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514. DOI: 10.1109/CVPR.2016.594.
- Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution using an Efficient Sub-pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883. DOI: 10.1109/CVPR.2016.207.
- Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision*, pages 746–760. Springer. DOI: 10.1007/978-3-642-33715-4_54.
- Spencer, J., Tosi, F., Poggi, M., Arora, R. S., Russell, C., Hadfield, S., Bowden, R., Zhou, G., Li, Z., and Rao, Q. (2024). The Third Monocular Depth Estimation Challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–14. DOI: 10.1109/CVPRW63382.2024.00005.
- Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., and Scherer, S. (2020). TartanAir: A Dataset to Push the Limits of Visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE. DOI: 10.1109/IROS45743.2020.9341801.
- Wrenninge, M. and Unger, J. (2018). Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2742–2749. Available at: <https://www.alphaxiv.org/pdf/1810.08705>.
- Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. (2024a). Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381. DOI: 10.1109/CVPR52733.2024.00987.
- Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. (2024b). Depth Anything V2. *38th Conference on Neural Information Processing Systems*, pages 1–37. DOI: 10.52202/079017-0688.
- Yang, M., Yu, K., Zhang, C., Li, Z., and Yang, K. (2018). DenseASPP for Semantic Segmentation in Street Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692. DOI: 10.1109/CVPR.2018.00388.
- Zeng, Z., Wang, D., Yang, F., Park, H., Soatto, S., Lao, D., and Wong, A. (2024). WorDepth: Variational Language Prior for Monocular Depth Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9708–9719. DOI: 10.1109/CVPR52733.2024.00927.