Dense Depth Distillation with Out-of-Distribution Simulated Images

Junjie Hu[a], Chenyou Fan[b], Mete Ozay[c], Hualie Jiang[d] and Tin Lun Lam[d,a,*]

[a] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[b] South China Normal University, China
[c] METU, Turkey
[d] The School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen, China

ARTICLE INFO

Keywords:
Monocular depth estimation
knowledge distillation
data-free KD
dense distillation.

ABSTRACT

We study data-free knowledge distillation (KD) for monocular depth estimation (MDE), which learns a lightweight model for real-world depth perception tasks by compressing it from a trained teacher model while lacking training data in the target domain. Owing to the essential difference between image classification and dense regression, previous methods of data-free KD are not applicable to MDE. To strengthen its applicability in real-world tasks, in this paper, we propose to apply KD with out-of-distribution simulated images. The major challenges to be resolved are i) lacking prior information about scene configurations of real-world training data and ii) domain shift between simulated and real-world images. To cope with these difficulties, we propose a tailored framework for depth distillation. The framework generates new training samples for embracing a multitude of possible object arrangements in the target domain and utilizes a transformation network to efficiently adapt them to the feature statistics preserved in the teacher model. Through extensive experiments on various depth estimation models and two different datasets, we show that our method outperforms the baseline KD by a good margin and even achieves slightly better performance with as few as 1/6 of training images, demonstrating a clear superiority.

1. Introduction

As a cost-effective alternative solution to depth sensors, monocular depth estimation (MDE) predicts scene depth from only RGB images and has wide applications in various tasks, such as scene understanding [27], autonomous driving [49], 3D reconstruction [14], and augmented reality [9]. In recent years, the accuracy of MDE methods has been significantly boosted and dominated by deep learning based approaches [13, 21, 28], where the advances are attributed to modeling and estimating depth by complex nonlinear functions using large-scale deep neural networks.

On the other hand, many practical applications, *e.g.*, robot navigation, demand a lightweight model due to the hardware limitations and requirement for computationally efficient inference. In these cases, we can either perform model compression on a well-trained large network [54] or apply supervised learning to directly train a compact network [38]. These solutions assume that the original training data of the target domain is known and can be freely accessed. However, since data privacy and security are invariably a severe concern in the real world, the training data is routinely unknown in practice, especially for industrial applications. A potential solution under this practical constraint is to distill preserved knowledge from a well-trained and publicly available model without accessing the original training data. The task is called data-free knowledge distillation (KD) [35] and has been shown to be effective for image classification.

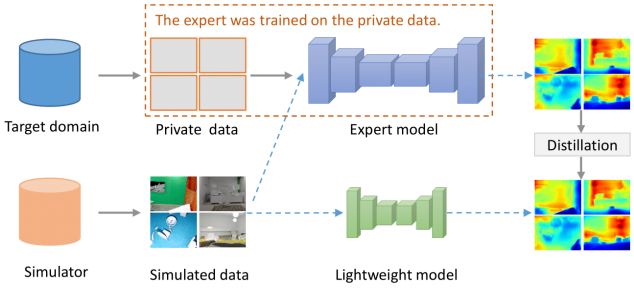


Figure 1: A visualization of the problem of data-free depth distillation. We propose to use simulated images as an alternative solution to the challenges of applying knowledge distillation for monocular depth estimation when original training data is not available.

Most existing methods of data-free KD proposed to synthesize training images from random noise [12, 57]. Specifically, assuming that y is a target object attribute, it is an element that inherently exists in the last layer of a classifier and is easily pre-specified, such that we can enforce a classifier to produce the desired output by gradually optimizing its input data. We refer to this property as the inherent constraint of classification. Unfortunately, in the case of MDE, the output is a high-resolution two-dimensional map with interrelated objects, not a score for a category; the inherent constraint does not hold for MDE, making most existing data-free approaches incompatible. More formal analyses are given in Sec. 3.

Given the above challenges, in this paper, we propose to leverage out-of-distribution (OOD) images as an alternative solution for applying KD. For the MDE task, intuitively, we consider three critical elements for choosing the alternative

*Corresponding author
hujunjie@cuhk.edu.cn (J. Hu); fanchenyou@gmail.com (C. Fan); meteozay@gmail.com (M. Ozay); hl.jiang@siat.ac.cn (H. Jiang); tllam@cuhk.edu.cn (T.L. Lam)
ORCID(s):

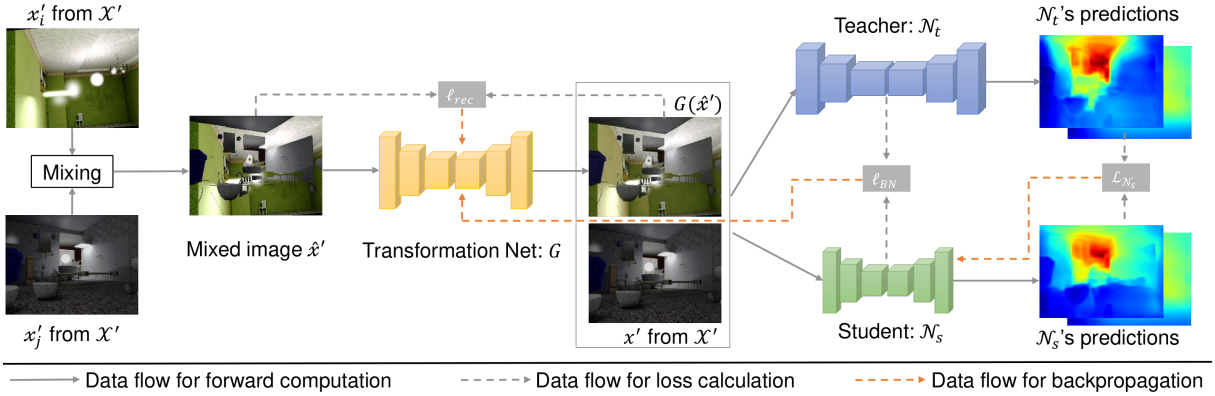


Figure 2: A flowchart of the proposed approach for distilling a trained model in the real world with simulated images. We firstly mix two images x'_i and x'_j sampled from the simulated dataset $\mathcal{X}'$ to generate a new sample $\hat{x}'$, and use a transformation network to fit $\hat{x}'$ to the feature distribution provided by the trained teacher model. The distillation is applied from the teacher $\mathcal{N}_t$ to the target student $\mathcal{N}_s$ with the new input $G(\hat{x}')$ and the original simulated data x' where the latter ensures a lower bound performance. The transformation and the student network are trained jointly.

set: i) the similarity in scene structures between the original scenarios and the OOD set, ii) the number of training images of the OOD set, and iii) domain gap between the original domain and the OOD domain. We analyze the effect of the above three factors on the accuracy of KD through quantitative experiments. Unsurprisingly, high scene similarity, sufficient data, and a small domain gap contribute to better accuracy. Another valuable observation is that the teacher still estimates meaningful depth maps correctly representing relative depths among objects, even from simulated images. It reveals that DNNs may utilize some geometric cues [23, 20, 8], or can learn some domain-invariant features [4] for inferring depths rather than the straightforward fitting. Therefore, it is still possible to perform KD even though the predicted depth maps are completely wrong in scales.

The effective yet impractical solution is to collect a dataset similar to the original training data. In reality, data collection is always costly and time-consuming. Besides, due to the lack of prior information about scene structures of the original training data, we have no sufficient clues to guide this data collection process. For these reasons, we prefer using synthetic images collected from simulators, as visualized in Figure 1. In this way, we can handily collect enough images to ensure the multitude of data for distillation. However, we need to overcome the significant domain gap between simulated and real-world data.

In general, the difficulties of depth distillation utilizing simulated data are two-fold. The first is the unknown scene/object configurations in the target domain. The second is the unavoidable domain discrepancy between the simulated and original (real-world) training sets. Our distillation framework is composed of two sub-branches. The first branch applies the plain KD using initial simulated images to ensure a lower-bound performance. The second branch expands our training set by generating additional training samples to tackle the above challenges. Specifically, we first generate new images that aim to encompass a diverse array of scene configurations in the target domain by applying random object-wise mixing between two simulated images. Then, we propose to regularize the mixed images to fit the target domain by tackling an efficient image-to-feature adaption problem with a transformation network. Figure 2 shows the diagram of the proposed distillation framework where we learn the transformation network G and the target student network $\mathcal{N}_s$ simultaneously.

To the best of our knowledge, we are the first to distill knowledge for MDE in data-free scenarios. We extensively evaluate the proposed method on different depth estimation models and two indoor datasets, including NYU-v2 and ScanNet. In all datasets, our approach demonstrates the best performance. It outperforms the baseline KD by a good margin, *e.g.*, gaining 0.05 and 0.08 average improvements in RMSE (meters) on NYU-v2 and ScanNet, respectively, and shows slightly better performance with as few as $1/6$ of the image dataset.

In summary, our contributions include:

1. The first attempt performing data-free KD for monocular depth estimation using OOD simulated images with quantitative studies to ascertain the essential requirements for selecting the OOD data.

2. A specialized framework for efficient depth distillation by learning image-to-feature adaption. We propose to use a transformation network for fast distillation.

3. We validate the proposed method for different depth estimation models on multiple datasets. As a result, our method outperforms the baseline methods by a good margin.

The rest of the paper is organized as follows. In Section. 2, we introduce the related works. In Section. 3, we give formal analyses regarding the difficulties of applying data-free KD for MDE. Section. 4 presents our method in detail. Section 5 shows detailed experimental settings and

results to verify the effectiveness of our method. Section. 6 concludes the paper.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation (MDE) aims to predict scene depths from only a single image. Deep learning-based approaches have dominated recent progress [26, 36, 13, 25, 28, 59, 58, 62, 17] in which the advanced performances are attributed to modeling and estimating depth using large and complex networks with data-driven supervised/unsupervised learning.

On the other hand, deploying MDE algorithms into real-world applications often faces practical challenges, such as limited hardware resources and inefficient computation. Therefore, an emerging requirement of MDE is to develop lightweight models to meet the above demands. This problem has been specifically considered in previous studies [38, 42, 54, 18] where several different lightweight networks have been designed.

However, lightweight networks inevitably degrade their MDE performance due to the trade-off between model complexity and accuracy. Hence, it remains an open question: how can the model complexity be reduced while maintaining high accuracy? One potential solution to this problem is KD, which transfers the knowledge from a cumbersome teacher network to a compact student network with decent accuracy improvement. However, KD requires the original training dataset for implementation. Currently, there are no existing solutions in data-free scenarios for MDE.

2.2. Knowledge Distillation

Knowledge distillation (KD) [16] was initially introduced in image classification, where either the soft label or the one-hot label predicted by the teacher is used to supervise student training. Existing methods can be generally categorized into two groups depending on whether they can access the original training set: 1) standard data-aware KD and 2) data-free KD.

The effectiveness of data-aware KD has been demonstrated for various vision tasks, such as image classification [16], semantic segmentation [19, 34], object detection [1], and depth estimation [45, 53], *etc*. In addition to the conventional setup, researchers have proposed to improve KD via distilling intermediate features [24, 34], distilling from multiple teachers [50, 33], employing an additional assistant network [40], and adversarial distillation [5, 47].

For data-free KD, researchers resorted to synthesizing training sets from random noise [35, 57, 12, 11, 60] or employed other large-scale data from different domains [2, 55, 10, 41]. However, existing methods are only effective for classification tasks and cannot be applied to MDE. In this paper, we propose the first method of data-free distillation for MDE. Our method leverages data from simulated environments to distill a model trained on a real-world dataset.

We would like to emphasize the distinctions between knowledge distillation (KD) and domain adaptation (DA), as there may be some potential misunderstandings. Primarily, Domain Adaptation (DA) entails the transfer of a model from its original domain (source domain) to a novel and distinct domain (target domain). In contrast, Knowledge Distillation (KD) pertains to a single domain context, concentrating on upholding the model's accuracy within the original domain. The scenario shifts when training images from the original domain remain undisclosed and inaccessible, leading KD and DA to manifest as specific instances of data-free KD and source data-free DA [29, 56], correspondingly. Nevertheless, it remains noteworthy that DA necessitates access to data from the target domain.

3. Preliminary Analyses

Our target is to perform data-free KD for MDE tasks, facilitating the development of a lightweight MDE model. We commence by delineating the challenges inherent in this task through meticulous analysis and elucidate why existing methods are unsuitable for MDE. Subsequently, we endeavor to leverage OOD data, supported by quantitative analyses, to discern the critical requisites for effectively incorporating OOD data into the KD framework.

3.1. Difficulty of Data-Free Depth Distillation

Suppose that $\mathcal{N}_t$ is a model trained using data from the target domain $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ where $\mathcal{X}$ and $\mathcal{Y}$ denote input data (*i.e.*, image) and label space, respectively. For any $x \in \mathcal{X}$, its corresponding label is estimated by $y = \mathcal{N}_t(x)$.

KD aims at learning a smaller network $\mathcal{N}_s$ with the supervision from $\mathcal{N}_t$. Usually, $\mathcal{N}_t$ is called the teacher network and $\mathcal{N}_s$ is called the student network, respectively. Then, the learning is formulated as:

$$\min_{\mathcal{N}_s} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \lambda\mathcal{H}\left(\mathcal{N}_t(x), \mathcal{N}_s(x)\right) + (1-\lambda)\mathcal{H}\left(y, \mathcal{N}_s(x)\right), \quad (1)$$

where $\mathcal{H}$ is a loss function, $\lambda > 0$ is a weighting coefficient and usually is a relatively large number, *e.g.*, 0.9, for giving more weights to the teacher predictions than ground truths. In practice, the second term of Eq. (1) is sometimes discarded. In these cases, Eq. (1) is simplified using $\lambda = 1$ by

$$\min_{\mathcal{N}_s} \sum_{x\in\mathcal{X}} \mathcal{H}\left(\mathcal{N}_t(x), \mathcal{N}_s(x)\right). \quad (2)$$

As shown above, the standard KD requires accessing the original training data sampled from $\mathcal{X}$. Contrarily, data-free KD attempts to train the student model without being aware of $\mathcal{X}$. It is formulated by

$$\min_{\mathcal{N}_s} \sum_{x'\in\mathcal{X}'} \mathcal{H}\left(\mathcal{N}_t(x'), \mathcal{N}_s(x')\right), \quad (3)$$

where $\mathcal{X}'$ is a proxy to $\mathcal{X}$ and can be either i) a set of images synthesized from $\mathcal{N}_t$, or ii) other alternative OOD datasets. Then, Eq. (3) can be solved by searching for the optimal $\mathcal{X}'$.

For image classification, the success is attributed to an inherent constraint for identifying $\mathcal{X}'$. As y denotes an object

Table 1
Accuracy of the student model employed on the NYU-v2 test set. The student model is trained via knowledge distillation with different OOD data. Except for (g), all datasets have approximately 50K images.

	Dataset	Scene	Domain	Data	δ_1
(a)	NYU-v2 [48]	indoor scene	real world	50K	**0.808**
(b)	ImageNet [7]	single object	real wrold	50K	0.685
(c)	Random noises	-	-	50K	0.194
(d)	ScanNet [6]	indoor scene	real world	50K	0.787
(e)	KITTI [51]	outdoor scene	real world	50K	0.705
(f)	SceneNet [39]	indoor scene	simulation	50K	0.712
(g)	SceneNet [39]	indoor scene	simulation	300K	0.742

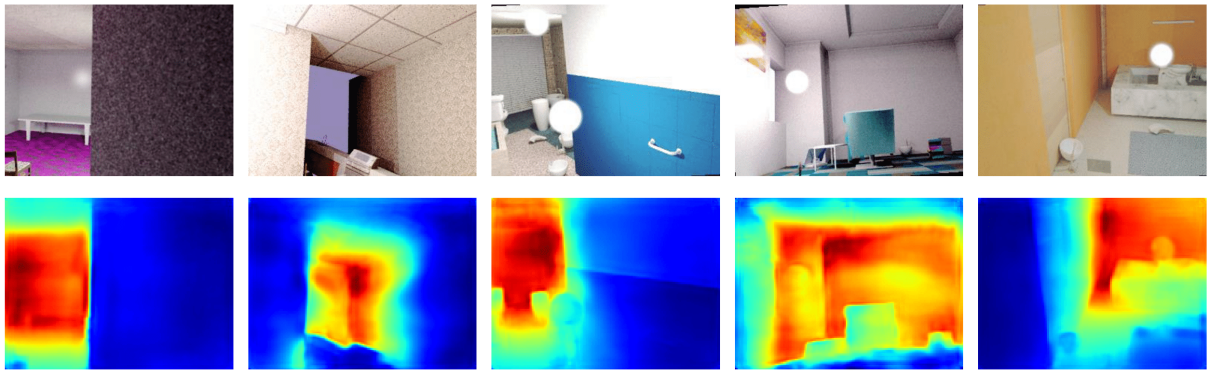


Figure 3: Visualization of the simulated images and the depth maps estimated by the teacher model.

category, it corresponds to an index of the SoftMax outputs from the last fully convolutional layer of the model and thus provides prior information about the desired model output. Then, $\mathcal{X}'$ is constructed by

$$\arg\min_{x'} \sum_{x' \in \mathcal{X}'} \mathcal{H}(\mathcal{N}_t(x'), y) + \mathcal{R}(x'), \tag{4}$$

where $\mathcal{R}$ denotes regularization terms.

Remark 1. *The first term of Eq. (4) is an inherently strong constraint of image classification that enforces the output consistency such that two posterior probabilities satisfy* $\mathcal{P}(y|x) \approx \mathcal{P}(y|x')$.

We can specify any category corresponding to an actual label of $\mathcal{Y}$ and generate sufficient images from random noises. Besides, in some works, this inherent constraint is used to transform the OOD data to the target distribution [10] or identify the most relevant data with low entropy from a large-scale dataset to the distribution of the target domain for efficient KD [2].

Based on the above discussions, existing approaches of data-free KD on image classification cannot be deployed into our depth estimation task, as predicted depth maps are 2D high-resolution maps with interrelated objects but not score values, invalidating the strong constraint used in Eq. (4). Since directly synthesizing data from noise is intractable, we have to seek alternative data to perform KD.

3.2. Depth Distillation with OOD Data

According to Remark 1, data-free KD for MDE can be intractable. A plausible way is to use some OOD data if we can decipher the essential requirements for $\mathcal{X}'$. Here, we consider that three factors are essential for selecting $\mathcal{X}'$: i) scene structure similarity to $\mathcal{X}$, ii) data-scale for performing KD, and iii) domain gap between $\mathcal{X}'$ and $\mathcal{X}$.

We conducted preliminary experiments to analyze how a depth estimation model reacts to different types of OOD data. We employ a model trained on the NYU-v2 [48] dataset as the teacher model and apply KD using different OOD datasets with the same number of randomly sampled images.

We observed that the depth range of the depth maps predicted from different types of OOD data still fits into the target domain. However, this constraint is insufficient to ensure KD, as shown in Table 1, where random Gaussian noises led to the lowest performance even though they yield a similar depth range to other types of OOD data.

Then, we analyze the effect of the above three factors by comparing scene similarity, data scale, and data domain. Except for (g), all datasets have 50,000 (50K) images. By closely comparing (d) and (e), (d) and (f), and (f) and (g), with (a), not surprisingly, we observe that high scene similarity, small domain gaps in images, and large-scale training datasets are beneficial for the performance boost.

However, it is challenging to satisfy all these three conditions simultaneously. Considering the difficulties of data collection in real-world applications, we propose to apply data-free KD for MDE with simulated images. Figure 3 shows depth maps estimated from simulated images. It is observed that the inferred depth maps are perceptually correct despite being wrong in absolute depth scales, providing a strong prior of data-free distillation for MDE with simulated data. We posit that this property is pivotal to the success of depth distillation utilizing simulated data.

4. Framework for Dense Depth Distillation with Simulated Data

Our distillation framework consists of three models: the fixed teacher, the target student, and a data transformation model, where the latter two models are jointly trained. We present the learning objectives of the transformation model and the student model as follows.

4.1. Training the Transformation Model

We have shown that the trained model would estimate reasonable depth maps with correct relative distances among objects from the OOD simulated data. However, due to the domain gap, there is a significant discrepancy between $\mathcal{X}$ and $\mathcal{X}'$. Thus, we wish to mitigate this domain gap and accordingly improve KD.

Since we have no clues about the original training data, such as identities or representations of objects, we naturally consider applying data augmentation to maximally encompass the configurations of scenarios in the target domain. Inspired by ClassMix [43], we randomly change half of objects between two simulated images to obtain a new image with the help of semantic maps collected from the simulator. More formally, for two images x'_i and x'_j where $x'_i \in \mathcal{X}'$, $x'_j \in \mathcal{X}'$, we generate a new mixed image $\hat{x}'$ by

$$\hat{x}' = m \odot x'_i + (1 - m) \odot x'_j, \tag{5}$$

where m is a binary mask obtained from the semantic map of x'_i, and randomly selects half of the classes observed in x'_i.

We leverage the running average statistics captured inside neural networks as DeepInversion [57] to regularize $\hat{x}'$. Specifically, assuming that feature statistics follow the Gaussian distribution and can be defined by mean μ and variance σ^2, then, $\hat{x}'$ is optimized through the following loss

$$\ell_{BN} = \sum_{l\in[L]} \|u_l(\hat{x}') - \overline{u}_l\|_2 + \sum_{l\in[L]} \|\sigma_l^2(\hat{x}') - \overline{\sigma}_l^2\|_2, \tag{6}$$

where $u_l(\hat{x}')$ and $\sigma_l^2(\hat{x}')$ are the batch-wise mean and variance of feature maps of the l-th convolutional layer of $\mathcal{N}_t$, respectively. $\overline{u}_l$ and $\overline{\sigma}_l^2$ are the running mean and variance of the l-th BN layer of $\mathcal{N}_t$, respectively. Eq.(6) allows regularizing $\hat{x}'$ to fit the feature distribution provided by the teacher model. However, this optimization requires thousands of iterations[1] for a single batch and is highly time consuming. To tackle this problem, we propose to turn the optimization process for estimating $\hat{x}'$ into a representation learning problem by training an additional model G for data transformation. Then, Eq.(6) can be rewritten by

$$\ell_{BN} = \sum_{l\in[L]} \|u_l(G(\hat{x}')) - \overline{u}_l\|_2 + \sum_{l\in[L]} \|\sigma_l^2(G(\hat{x}')) - \overline{\sigma}_l^2\|_2. \tag{7}$$

It is essential to ensure the fidelity of the original scenes to keep the geometry structure. Thus, we adopt an image reconstruction loss. Finally, the loss function for training the transformation network is defined by

$$\mathcal{L}_G = \sum_{\hat{x}'\in\mathcal{X}'} \left(\alpha\ell_{BN} + \beta\ell_{rec}\right), \tag{8}$$

where $\ell_{rec} = \|\hat{x}' - G(\hat{x}')\|_1$ is the reconstruction error that penalizes the ℓ_1 norm of image difference, and α and β are weighting coefficients.

Algorithm 1 Depth Distillation Algorithm.

Input: $\mathcal{X}'$: OOD images collected from a simulator; $\mathcal{N}_t$: The teacher model trained on a target domain; α, β: Weighting coefficients used for defining loss in training G; T: Number of iterations;

Output: $\mathcal{N}_s$: The student model; G: The transformation model.

```
Freeze N_t;
Initialize N_s and G;
for j = 1 to T do
    Set gradients of N_s and G to 0;
    Select a batch x' from X';
    Let x'_i = x' and x'_j = random_shuffle(x');
    Generate mixed images x̂' by Eq. (5);
    ▷ % Updating the student model %
    Calculate N_t(x'), N_s(x'), N_t(G(x̂')), N_s(G(x̂'));
    Calculate the depth loss by Eq. (9);
    Update N_s;
    ▷ % Updating the transformation network %
    Calculate the loss for training G by Eq. (8);
    Update G;
end for
```

4.2. Training the Student Model

We formally describe the distillation framework to enable data-free student training. We amalgamate the prediction loss from both the teacher and student models using the initial simulated images x' without amalgamation and the transformed images $G(\hat{x}')$ derived from the mixed images $\hat{x}'$. The former adopts plain distillation to ensure a lower bound performance, and the latter contributes to further performance improvement.

The optimization objective of depth distillation from the teacher model to the student model is defined by

[1] 3000 iterations are used for DeepInversion.

Table 2
Details of the RGBD datasets used in the experiments.

	Dataset	Training scenarios / images	Test scenarios / images
Target domain	NYU-v2	249 / 50688	215 / 614
	ScanNet	1513 / 50473	100 / 17607
Simulated data	SceneNet $\mathcal{X}'_1$	1000 / 50K	-
	SceneNet $\mathcal{X}'_2$	1000 / 300K	-

$$\mathcal{L}_{\mathcal{N}_s} = \sum_{x' \in \mathcal{X}'} \mathcal{H}\left(\mathcal{N}_t(x'), \mathcal{N}_s(x')\right) + \sum_{\hat{x}' \in \hat{\mathcal{X}}'} \mathcal{H}\left(\mathcal{N}_t(G(\hat{x}')), \mathcal{N}_s(G(\hat{x}'))\right), \tag{9}$$

where $\mathcal{H}$ is a function used for measuring the depth errors. We employ the function proposed in [21] that penalizes losses of depth, gradient, and normal.

To facilitate efficient learning, the transformation model and the student model are trained jointly by solving the following optimization problem

$$\min_{G,\mathcal{N}_s} (\mathcal{L}_G + \mathcal{L}_{\mathcal{N}_s}). \tag{10}$$

The details of our method are given in Algorithm 1 where random_shuffle denotes the operation of randomizing images.

5. Experimental Analyses

5.1. Experimental Settings

5.1.1. Implementation Details

Our learning framework includes three models: (1) a teacher model $\mathcal{N}_t$ trained on a given target domain and is fixed while training a student model; (2) a student model $\mathcal{N}_s$, which we aim to train; and (3) a transformation network G which will also be optimized during training. We train $\mathcal{N}_s$ and G for 20 epochs using the Adam optimizer [30]. The learning rate begins at 0.0001 and is halved every five epochs. The hyper-parameters α and β controlling the data transformation are set to 0.001 for all experiments throughout the paper. We trained models with a batch size of 8 in all the experiments and developed the code-base using PyTorch [44].

5.1.2. Datasets

NYU-v2 [48] The NYU-v2 dataset is the benchmark most commonly used for depth estimation. The depth range is 0 to 10 meters. It is captured by Microsoft Kinect with an original resolution of 640 × 480, and contains 464 indoor scenes. Among them, 249 scenes are chosen for training, and 215 scenes are used for testing. We use the pre-processed data by Hu et al. [21, 20] with approximately 50,000 unique pairs of an image and a depth map with a resolution of 640×480. Following most previous studies, we resize the images to 320×240 pixels and then crop their central parts of 304×228 pixels as inputs. For testing, we use the official small subset of 654 RGBD pairs.

ScanNet [6] ScanNet is a large-scale RGBD dataset that contains 2.5 million RGBD images. The depth range is 0 to 6 meters. We randomly and uniformly select a subset of approximately 50,000 samples from the training splits of 1513 scenes for training and evaluate the models on the test set of another 100 scenes with 17K RGB pairs. We apply the same image pre-processing methods, such as image resizing and cropping, as utilized on the NYU-v2 dataset.

SceneNet [39] SceneNet is a large-scale synthesized dataset that contains 5 Million RGBD indoor images from over 15,000 synthetic trajectories. Each trajectory has 300 rendered frames. The original image resolution is 320×240. Thus, we only apply the center crop to yield an image resolution of 304×228.

We sample two subsets from 1000 indoor scenes of the official validation set. The two subsets have 50,000 and 300,000 images, respectively, and are denoted by $\mathcal{X}'_1$ and $\mathcal{X}'_2$ in the following texts. Detailed information on the datasets used in the experiments is given in Table 2.

5.1.3. Models of the Teacher and the Student

We choose multiple combinations of the teacher and student models to extensively evaluate our models and methods. For the first combination, we let the teacher and student models have the same lightweight depth estimation network (LDEN) proposed in [18] built on ResNet-34 [15] to investigate the performance without model compression. For the second combination, we use the above ResNet-34 based (LDEN) as the teacher model and the MobileNet-v2 [46] based network as the student model in [18]. For the next two combinations, the teacher models are implemented using a ResNet-50 [15] based encoder-decoder network (EDN) [26] and multi-branch feature fusion network (FFN) [21], respectively. Networks of the student models are modified from the teacher networks by replacing ResNet-50 with ResNet-18. For the last combination, the teacher model is a SeNet-154 [22] based structure-aware residual pyramid network ((SARPN) [3]. Similarly, the student model is derived from the teacher model by replacing the backbone with a smaller ResNet-34.

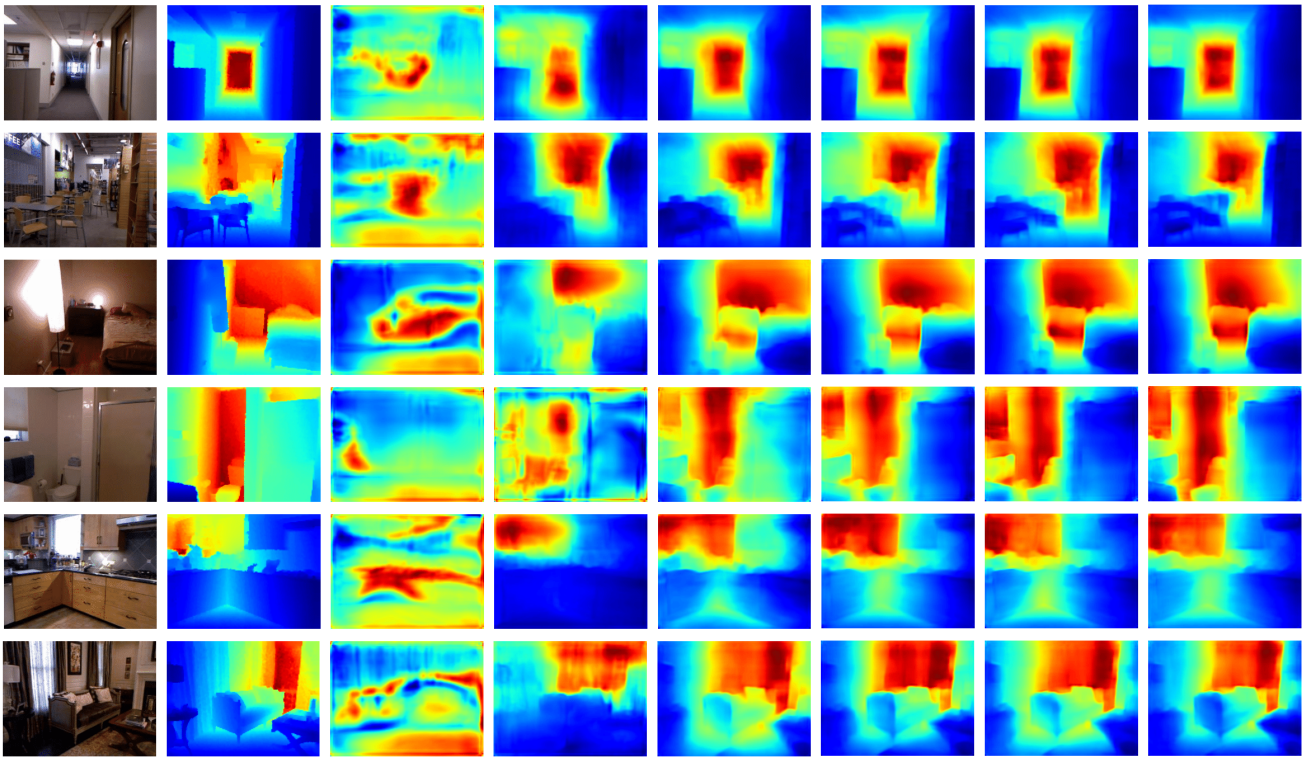
To implement the network of the transformation model, we use the dilated convolution [61] based encoder-decoder network modified from the saliency prediction network [23, 20] by adding symmetric skip connections between the encoder and the decoder.

5.1.4. Baselines

As discussed in Sec. 3, most of the previous data-free KD methods cannot be applied to depth regression tasks. Thus,

Table 3
Quantitative results on the NYU-v2 dataset. We use two popular measures including the root mean squared error (RMSE) and thresholding accuracy δ_1.

Model		LDEN [18]		LDEN [18]		EDN [26]		FFN [21]		SARPN [3]	
Teacher (Backbone) → Student (Backbone)		ResNet-34 → ResNet-34		ResNet-34 → MobileNet-v2		ResNet-50 → ResNet-18		ResNet-50 → ResNet-18		SeNet-154 → ResNet-34	
Parameter Reduction		None		21.9M → 1.7M		63.6M → 13.7M		67.6M → 14.9M		258.4M → 38.7M	
Method	Data	RMSE	δ_1	RMSE	δ_1	RMSE	δ_1	RMSE	δ_1	RMSE	δ_1
Teacher	NYU-v2	0.481	0.829	0.481	0.829	0.497	0.824	0.465	0.843	0.418	0.878
Student		0.481	0.829	0.518	0.802	0.522	0.805	0.494	0.826	0.459	0.843
Random noises	None	1.673	0.193	1.702	0.194	1.935	0.102	1.934	0.112	1.953	0.107
DFAD		0.958	0.402	1.090	0.329	1.004	0.382	1.163	0.338	1.208	0.278
KD-OOD	SceneNet $\mathcal{X}_1'$	0.596	0.753	0.648	0.712	0.729	0.660	0.660	0.710	0.665	0.695
Ours		**0.555**	**0.774**	**0.600**	**0.742**	**0.676**	**0.701**	**0.639**	**0.722**	**0.581**	**0.759**
KD-OOD	SceneNet $\mathcal{X}_2'$	0.590	0.761	0.611	0.742	0.705	0.676	0.663	0.713	0.605	0.738
Ours		**0.537**	**0.789**	**0.558**	**0.778**	**0.648**	**0.726**	**0.584**	**0.760**	**0.569**	**0.776**



(a) RGB images. (b) Ground truth. (c) Random noise. (d) DFAD. (e) KD-OOD with $\mathcal{X}_1'$. (f) Our results with $\mathcal{X}_1'$. (g) KD-OOD with $\mathcal{X}_2'$. (h) Our results with $\mathcal{X}_2'$.

Figure 4: Qualitative comparison of depth maps predicted by different methods on the NYU-v2 test set.

we choose DFAD [12] as a baseline since this method does not apply the inherent constraint for synthesizing images. Overall, we consider the following methods as baselines for comparison.

Teacher: The teacher model trained on the target dataset.

Student: The student model trained on the target dataset.

KD-OOD: For the sake of comparison, we take KD [16] using the OOD simulated data as the strong baseline of our method. It is the first loss term of Eq. (9).

Random noise: The student model is trained via KD with random Gaussian noise. It is also a baseline commonly used for image classification.

DFAD: The student model is trained with data-free adversarial distillation [12] that synthesizes images from random noise with adversarial training.

5.2. Quantitative Comparisons

5.2.1. NYU-v2 Dataset

We first thoroughly evaluate the proposed method on the NYU-v2 dataset. We measure depth maps using the root mean squared error (RMSE) and the thresholding δ_1 accuracy. Table 3 shows the quantitative results of different methods for various teacher-student combinations where the performance of the student (trained in supervised learning) exhibits an upper bound that we aim to reach. As seen, distillation with random noise yields the lowest performance, although they are shown to be effective for some toy datasets, *e.g.*, MNIST [32] and CIFAR-10 [31], for image classification. Moreover, DFAD has also failed on the task.

Compared to the above methods, KD-OOD demonstrates much better results, showing the advancement of our route that utilizes OOD simulated images. In the case of using the smaller set $\mathcal{X}'_1$, it provides 0.165 mean increase in RMSE (meters) and 0.115 decrease in δ_1 over the five different model combinations. Most importantly, the proposed method outperforms all baselines and attains consistent performance improvement for all different teacher-student combinations. It yielded 0.115 and 0.081 performance degradation in RMSE (meters) and δ_1. Compared to KD-OOD, it achieves 0.05 meters and 4.0% mean improvement in RMSE and δ_1, respectively.

We then analyze the effect of utilizing the larger set $\mathcal{X}'_2$. As a result, we found a performance boost for both KD-OOD and our method in all experiments when using a larger scale set. Our method consistently outperforms KD-OOD by 0.06 meters and 4.9% in RMSE and δ_1, respectively. Besides, our method using $\mathcal{X}'_1$ even outperforms KD-OOD using $\mathcal{X}'_2$.

Another observation is that the first two teacher-student model combinations outperform the latter three. The results agree well with previous studies [52], which verified that the performance of the student model degrades when the gap in model capacity between them is significant. This problem can be well handled by using an additional assistant model [40], distilling intermediate features [34], multiple teacher models [50], and the ensemble of distributions [37]. Since it is a common challenge, we leave it as future work.

Figure 4 visualizes a qualitative comparison of different methods. It is seen that random noises produce meaningless predictions, and DFAD estimates coarse depth maps. A closer observation of maps predicted by KD-OOD and our method shows that our proposed method can more accurately estimate depth in local regions. Overall, the quantitative and qualitative results verified the effectiveness of our approach.

5.2.2. ScanNet Dataset

To fully evaluate our method, we also test methods using the ScanNet dataset. We use the teacher and student models proposed in [18]. The results are given in Table 4. The final results are highly consistent with those obtained using NYU-v2. Both random noises and DFAD show extremely low accuracy. The proposed method outperforms KD-OOD even using the smaller set. We obtained 10.5% and 7.5% improvement in δ_1 and, 0.089 meters and 0.065 meters improvement in RMSE for $\mathcal{X}'_1$ and $\mathcal{X}'_2$, respectively.

Table 4
The results provided by the models on the ScanNet dataset.

Method	Data	RMSE	δ_1
Teacher Model	ScanNet	0.333	0.790
Student Model		0.357	0.764
Random noise	None	1.265	0.079
DFAD		0.725	0.368
KD-OOD	SceneNet $\mathcal{X}'_1$	0.542	0.541
Ours		**0.453**	**0.646**
KD-OOD	SceneNet $\mathcal{X}'_2$	0.477	0.618
Ours		**0.412**	**0.693**

5.3. Analyses for the Transformation Model

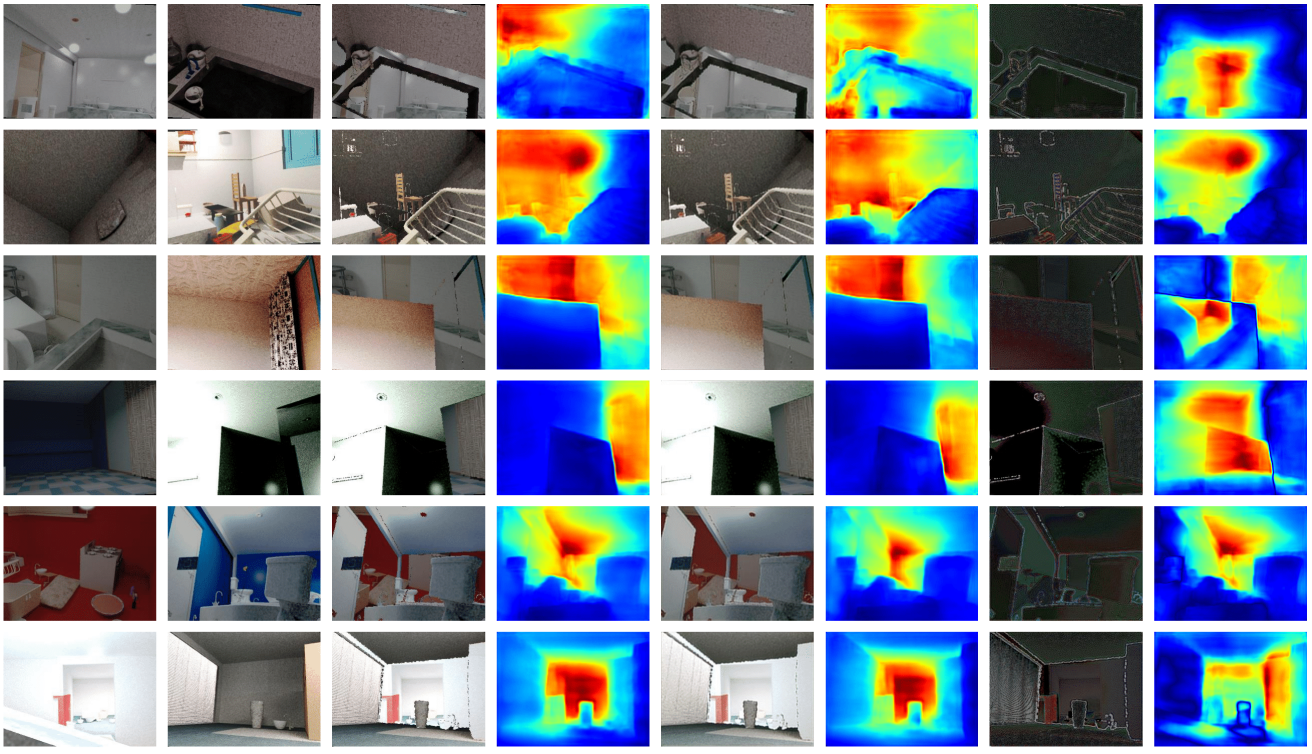
Figure 5 shows some examples of the input and output images of the transformation network as well as their corresponding predictions. In the figure, x'_i and x'_j denote two images randomly selected from the simulated set, and $\hat{x}'$ is the image generated by applying object-wise mixing between x'_i and x'_j. $G(\hat{x}')$ denotes the transformed image, *i.e.*, the output of G. We marked some regions of images in Figure 5 (c) and (d) with red boxes for better visualization. By visually comparing $\hat{x}'$ and $G(\hat{x}')$, we observe that G tends to reduce noises and alleviate artifacts around object boundaries such that G can produce more realistic images, mitigating domain gap from the real-world scenarios. It can be validated by $||\hat{x}' - G(\hat{x}')||_1$ where $||\cdot||_1$ is the ℓ_1 norm (Figure 5. (g)) where differences at object boundaries are highlighted. Furthermore, Figure 5. (d) and (f) show the predicted depth maps for $\hat{x}'$ and $G(\hat{x}')$, respectively. They demonstrate a clear difference, as observed in Figure 5. (h). We quantify these differences by evaluating the whole set $\mathcal{X}'_1$. As a result, the ℓ_1-norm of the image and depth difference is 0.156 and 0.227, respectively.

5.4. Ablation Studies

We conduct several ablation studies to analyze our approach and provide additional results on the NYU-v2 dataset. Table 5 gives the results. Specifically, we perform several experiments as follows:

Without using ℓ_{rec}: In our original method, we impose the reconstruction consistency between $\hat{x}'$ and $G(\hat{x}')$ to suppress undesirable noises while training the transformation model. We relax this constraint and observe that the RMSE and δ_1 dropped to 0.612 and 0.735, respectively.

Without using G: We also test the performance while removing the transformation model in the pipeline. We directly perform distillation using x' and mixed images $\hat{x}'$. As a result, the RMSE and δ_1 dropped to 0.635 and 0.722, respectively.



(a) x'_i. (b) x'_j. (c) $\hat{x}'$. (d) $\mathcal{N}_t(\hat{x}')$. (e) $G(\hat{x}')$. (f) $\mathcal{N}_t(G(\hat{x}'))$. (g) d_I. (h) d_D.

Figure 5: Visual comparisons of images and depth maps where (a) and (b) are original images from the simulated set, (c) and (d) are mixed images and estimated depth maps, (e) and (f) denote transformed images of (c) and estimated depth maps, (g) $d_I \triangleq ||\hat{x}' - G(\hat{x}')||_1$ and (h) $d_D \triangleq ||\mathcal{N}_t(\hat{x}') - \mathcal{N}_t(G(\hat{x}'))||_1$ are used to compute image discrepancy and depth discrepancy, respectively.

Table 5
Results for ablation studies.

	RMSE	δ_1
Original	**0.600**	0.742
Without using ℓ_{rec}	0.612	0.735
Without using G	0.635	0.722
Without using image mixing	0.635	0.724

Without using image mixing: We evaluate the effect without utilizing object-wise image mixing. We feed the images x' to G and apply distillation with both x' and $G(x')$. We find that the RMSE and δ_1 dropped to 0.635 and 0.724, respectively.

6. Conclusion

We have studied knowledge distillation for monocular depth estimation in data-free scenarios. By first thoroughly analyzing the challenges of the task, we showed that a promising approach to address the challenges is to utilize out-of-distribution (OOD) images as an alternative solution. We then empirically and quantitatively verified that i) a high degree of scene similarity, ii) the large-scale size of datasets, and iii) the small magnitude of domain gap contribute to the performance boost of depth distillation methods through detailed experimental analyses with different OOD data. Given the difficulty of data collection in practice, we proposed to utilize simulated images to strengthen the applicability of KD. We further presented a novel framework to perform data-free depth distillation with simulated data. As a practical solution to the task, we have evaluated the effectiveness of the proposed distillation framework on various depth estimation models and two real-world benchmark datasets. We consider that our framework and results can further inspire future explorations, shedding light on this unexplored problem.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (62073274, 62106156) and the funding AC01202101103 from the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

References

[1] Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M., 2017. Learning efficient object detection models with knowledge distillation, in: Advances in neural information processing systems, pp. 742–751.

[2] Chen, H., Guo, T., Xu, C., Li, W., Xu, C., Xu, C., Wang, Y., 2021a. Learning student networks in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6424–6433.

[3] Chen, X., Chen, X., Zha, Z.J., 2019. Structure-aware residual pyramid network for monocular depth estimation, in: International Joint Conferences on Artificial Intelligence, pp. 694–700.

[4] Chen, X., Wang, Y., Chen, X., Zeng, W., 2021b. S2r-depthnet: Learning a generalizable depth-specific structural representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3033–3042.

[5] Chung, I., Park, S., Kim, J., Kwak, N., 2020. Feature-map-level online adversarial knowledge distillation, in: International Conference on Machine Learning, pp. 2006–2015.

[6] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5828–5839.

[7] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255.

[8] van Dijk, T., de Croon, G.C., 2019. How do neural networks see depth in single images?, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2183–2191.

[9] Du, R., Turner, E., Dzitsiuk, M., Prasso, L., Duarte, I., Dourgarian, J., Afonso, J., Pascoal, J., Gladstone, J., Cruces, N., Izadi, S., Kowdle, A., Tsotsos, K., Kim, D., 2020. DepthLab: Real-Time 3D Interaction With Depth Maps for Mobile Augmented Reality, in: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, pp. 829–843.

[10] Fang, G., Bao, Y., Song, J., Wang, X., Xie, D., Shen, C., Song, M., 2021. Mosaicking to distill: Knowledge distillation from out-of-domain data, in: Advances in Neural Information Processing Systems, pp. 11920–11932.

[11] Fang, G., Mo, K., Wang, X., Song, J., Bei, S., Zhang, H., Song, M., 2022. Up to 100x faster data-free knowledge distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 6597–6604.

[12] Fang, G., Song, J., Shen, C., Wang, X., Chen, D., Chen, D., Song, M., 2019. Data-free adversarial distillation. arXiv preprint arXiv:1912.11006 .

[13] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2002–2011.

[14] Guo, X., Hu, J., Chen, J., Deng, F., Lam, T.L., 2021. Semantic histogram based graph matching for real-time multi-robot global localization in large scale environment. IEEE Robotics and Automation Letters 6, 8349–8356.

[15] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

[16] Hinton, G.E., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2.

[17] Hu, J., Bao, C., Ozay, M., Fan, C., Gao, Q., Liu, H., Lam, T.L., 2023a. Deep depth completion from extremely sparse data: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 8244–8264.

[18] Hu, J., Fan, C., Jiang, H., Guo, X., Gao, Y., Lu, X., Lam, T.L., 2023b. Boosting lightweight depth estimation via knowledge distillation, in: International Conference on Knowledge Science, Engineering and Management, pp. 27–39.

[19] Hu, J., Fan, C., Ozay, M., Feng, H., Gao, Y., Lam, T.L., 2022. Progressive self-distillation for ground-to-aerial perception knowledge transfer. arXiv preprint arXiv:2208.13404 .

[20] Hu, J., Okatani, T., 2019. Analysis of deep networks for monocular depth estimation through adversarial attacks with proposal of a defense method. arXiv preprint arXiv:1911.08790 .

[21] Hu, J., Ozay, M., Zhang, Y., Okatani, T., 2019a. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1043–1051.

[22] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141.

[23] Hu, J., Zhang, Y., Okatani, T., 2019b. Visualization of convolutional neural networks for monocular depth estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3869–3878.

[24] Huang, Z., Wang, N., 2017. Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 .

[25] Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E., Heikkilä, J., 2020. Guiding monocular depth estimation using depth-attention volume, in: European Conference on Computer Vision (ECCV), pp. 581–597.

[26] Iro, L., Christian, R., Vasileios, B., Federico, T., Nassir, N., 2016. Deeper depth prediction with fully convolutional residual networks, in: International Conference on 3D Vision (3DV), pp. 239–248.

[27] Jaritz, M., Charette, R.D., Wirbel, E., Perrotton, X., Nashashibi, F., 2018. Sparse and dense data with cnns: Depth completion and semantic segmentation, in: International Conference on 3D Vision (3DV), pp. 52–60.

[28] Jiang, H., Ding, L., Hu, J., Huang, R., 2021. Plnet: Plane and line priors for unsupervised indoor depth estimation, in: International Conference on 3D Vision (3DV), pp. 741–750.

[29] Kim, Y., Cho, D., Han, K., Panda, P., Hong, S., 2021. Domain adaptation without source data. IEEE Transactions on Artificial Intelligence 2, 508–518.

[30] Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR).

[31] Krizhevsky, A., 2009. Learning multiple layers of features from tiny images , 32–33URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[32] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

[33] Liu, I.J., Peng, J., Schwing, A.G., 2019a. Knowledge flow: Improve upon your teachers, in: International Conference on Learning Representations (ICLR).

[34] Liu, Y., Shun, C., Wang, J., Shen, C., 2019b. Structured knowledge distillation for semantic segmentation., in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2604–2613.

[35] Lopes, R.G., Fenu, S., Starner, T., 2017. Data-free knowledge distillation for deep neural networks. arXiv preprint arXiv:1710.07535 .

[36] Ma, F., Karaman, S., 2018. Sparse-to-dense: Depth prediction from sparse depth samples and a single image, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 1–8.

[37] Malinin, A., Mlodozeniec, B., Gales, M.J.F., 2019. Ensemble distribution distillation. arXiv preprint arXiv:1905.00076 .

[38] Mancini, M., Costante, G., Valigi, P., Ciarfuglia, T.A., 2016. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks, in: IEEE International Conference on Intelligent Robots and Systems (IROS), pp. 4296–4303.

[39] McCormac, J., Handa, A., Leutenegger, S., Davison, A.J., 2016. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. arXiv preprint arXiv:1612.05079 .

[40] Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H., 2020. Improved knowledge distillation via teacher assistant, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 5191–5198.

[41] Nayak, G.K., Mopuri, K.R., Chakraborty, A., 2021. Effectiveness of arbitrary transfer sets for data-free knowledge distillation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1430–1438.

[42] Nekrasov, V., Dharmasiri, T., Spek, A., Drummond, T., Shen, C., Reid, I., 2019. Real-time joint semantic segmentation and depth estimation using asymmetric annotations, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 7101–7107.

[43] Olsson, V., Tranheden, W., Pinto, J., Svensson, L., 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1368–1377.

[44] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, pp. 8024–8035.

[45] Pilzer, A., Lathuilière, S., Sebe, N., Ricci, E., 2019. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9760–9769.

[46] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520.

[47] Shen, Z., He, Z., Xue, X., 2019. Meal: Multi-model ensemble via adversarial learning, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 4886–4893.

[48] Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision (ECCV), pp. 746–760.

[49] Song, Z., Lu, J., Yao, Y., Zhang, J., 2022. Self-supervised depth completion from direct visual-lidar odometry in autonomous driving. IEEE Transactions on Intelligent Transportation Systems 23, 11654–11665.

[50] Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in Neural Information Processing Systems, pp. 1195–1204.

[51] Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A., 2017. Sparsity invariant cnns, in: International Conference on 3D Vision (3DV), pp. 11–20.

[52] Wang, L., Yoon, K.J., 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3048–3068.

[53] Wang, Y., Li, X., Shi, M., Xian, K., Cao, Z., 2021. Knowledge distillation for fast and accurate monocular depth estimation on mobile devices, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2457–2465.

[54] Wofk, D., Ma, F., Yang, T.J., Karaman, S., Sze, V., 2019. Fastdepth: Fast monocular depth estimation on embedded systems, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 6101–6108.

[55] Xu, Y., Wang, Y., Chen, H., Han, K., Xu, C., Tao, D., Xu, C., 2019. Positive-unlabeled compression on the cloud, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 2561–2570.

[56] Yang, S., van de Weijer, J., Herranz, L., Jui, S., et al., 2021. Exploiting the intrinsic neighborhood structure for source-free domain adaptation, in: Advances in neural information processing systems, pp. 29393–29405.

[57] Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J., 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8715–8724.

[58] Yin, W., Liu, Y., Shen, C., 2021. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 7282–7295.

[59] Yin, W., Liu, Y., Shen, C., Yan, Y., 2019. Enforcing geometric constraints of virtual normal for depth prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5684–5693.

[60] Yoo, J., Cho, M., Kim, T., Kang, U., 2019. Knowledge extraction with no observable data, in: Advances in Neural Information Processing Systems, pp. 2701–2710.

[61] Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 472–480.

[62] Zhao, C., Yen, G.G., Sun, Q., Zhang, C., Tang, Y., 2021. Masked gan for unsupervised depth and pose prediction with scale consistency. IEEE Transactions on Neural Networks and Learning Systems 32, 5392–5403.