Virtually Enriched NYU Depth V2 Dataset for Monocular Depth Estimation: Do We Need Artificial Augmentation?

Dmitry Ignatov[1]*, Andrey Ignatov[2], Radu Timofte[1]
[1] Computer Vision Lab, CAIDAS & IFI, University of Würzburg, Germany
[2] Computer Vision Lab, ETH Zürich, Switzerland

Abstract

We present ANYU, a new virtually augmented version of the NYU depth v2 dataset, designed for monocular depth estimation. In contrast to the well-known approach where full 3D scenes of a virtual world are utilized to generate artificial datasets, ANYU was created by incorporating RGB-D representations of virtual reality objects into the original NYU depth v2 images. We specifically did not match each generated virtual object with an appropriate texture and a suitable location within the real-world image. Instead, an assignment of texture, location, lighting, and other rendering parameters was randomized to maximize a diversity of the training data, and to show that it is randomness that can improve the generalizing ability of a dataset. By conducting extensive experiments with our virtually modified dataset and validating on the original NYU depth v2 and iBims-1 benchmarks, we show that ANYU improves the monocular depth estimation performance and generalization of deep neural networks with considerably different architectures, especially for the current state-of-the-art VPD model. To the best of our knowledge, this is the first work that augments a real-world dataset with randomly generated virtual 3D objects for monocular depth estimation. We make our ANYU dataset publicly available in two training configurations with 10% and 100% additional synthetically enriched RGB-D pairs of training images, respectively, for efficient training and empirical exploration of virtual augmentation at https://github.com/ABrain-One/ANYU.

1. Introduction

This paper focuses on depth estimation from single images, a technique widely used today for 3D image synthesis in augmented reality. The concept of depth estimation refers to the determination of the distance from camera to points in a three-dimensional scene based on the analysis of two-dimensional image. The challenging process of deriving such information is key for 3D scene reconstruction and augmented reality generation, robotics and autonomous driving, essential for perception, navigation, and planning. Herewith, monocular depth estimation based on deep neural networks has demonstrated a high ability for depth prediction from RGB images [1, 9, 24, 31] and exhibits numerous practical applications, including the highly demanded area of mobile devices [11, 12].

To train deep depth prediction models for indoor scenes, one often uses RGB-D datasets such as NYU Depth V2 [20] (NYU-v2). The importance of the NYU-v2 is illustrated by the following statistics: according to the resource "Papers with Code"[1], the number of papers released in 2019 – 2022 that are using indoor datasets was 439, 242, 216, and 14 for NYU-v2, Matterport3D [5], SUN RGB-D [26] (a mixture of NYU-v2 with other datasets), and iBims-1 dataset [15] specifically designed for validation, respectively.

Despite its popularity, NYU-v2, introduced more than 10 years ago, has such drawbacks as the inaccuracy of distances seen in depth maps and limited training data diversity typical for most real-world datasets. These flaws can be partially fixed by using RGB-D images with the accurate values of depth maps from virtual reality. In contrast to the limited nature of real-world datasets, virtual reality can provide an unlimited number of 3D images with precise depth values, where objects can be captured from all possible angles and in limitless spatial combinations. In addition to the well-known approach of creating virtual datasets using complete virtual world scenes, in this paper we propose to combine 3D images from virtual reality and RGB-D data from a real-world dataset to obtain a more accurate and diverse training set of RGB-D images.

2. Related Work

Depth estimation is one of the most important tasks in the domain of scene understanding, which is often based on the training of deep models and is continuously improv-

*Corresponding author: dmitri.ignatov@gmail.com

[1]Dataset statistics: https://paperswithcode.com/task/monocular-depth-estimation

ing with new architectures of deep convolutional neural networks and new training sets of data [5, 7, 14, 15, 18, 19, 29]. One strong trend for increasing performance in computer vision tasks is related to the use of synthetic data for training of deep convolution models [2, 3, 8, 10, 21, 25]. In particular, synthetically generated RGB-D datasets of varying complexity are utilized, ranging from simple ones, such as non-natural stacking rectangles rendering on top of each other presented by Courtois *et al*. [6], to more advanced datasets created for indoor [10, 17, 27, 28] and outdoor [8, 23] scenes, where a synthetic world with high quality rendering of virtual reality details is presented.

The potential effectiveness of virtual-world supervision in convolutional neural network training for monocular depth estimation was recently shown by Gurram *et al*. [9]. New synthetic RGB-D datasets provided by Roberts *et al*. [22] and Zhang *et al*. [30] yield better results in some indoor depth prediction-related tasks. More sophisticated approaches, such as training on a synthetic dataset and fine-tuning on real 3D data, can improve the performance of monocular depth estimation tasks [7, 25] and enhance the occluding contours location accuracy as shown by Ramamonjisoa and Lepetit [21].

Inspired by the ideas behind the synthetic Flying Chairs dataset [13], where 2D representations of virtual objects were successfully embedded in real-world RGB images to solve the optical flow estimation problem, in this work, we modify real RGB-D data from the NYU-v2 dataset by enriching it with randomly generated 3D virtual reality objects in order to train more accurate monocular depth estimation models.

3. Methodology of Virtual Enrichment of NYU Depth V2 Dataset

While previous works use either only real or artificial scenes to train deep neural networks for monocular depth estimation, we consider a different approach and augment real-world data for this task with 3D objects taken from virtual reality. To obtain a higher variety of training images with more accurate depth values that can be potentially used to achieve a higher depth estimation performance, we artificially enhance real-life NYU-v2 images using virtual reality. Specifically, a new virtually augmented NYU-v2 dataset (ANYU) was generated that extends the original NYU-v2 dataset with 10% and 100% synthetically modified RGB-D data.

We intentionally did not try to match each generated virtual object with the appropriate texture and place it in a suitable location. Instead, we used as much randomness as possible in generating virtual objects and choosing the place to embed them in real images to show that this way of maximizing the diversity of training data can improve the generalization ability of a dataset. The exact data generation procedure is described below.

Input. Virtual scenes are created using 84 3D objects of various sizes, ranging from small vases to large cabinets, and 308 specially selected diverse seamless textures that are publicly available on the Internet. For artificial modification, we randomly select RGB images and the corresponding depth maps from the entire set of 24231 pairs of RGB-D images in the NYU-v2 dataset, preserving their resolution of 640×480 pixels.

Virtual 3D objects generation. Virtual 3D scene rendering is performed with a popular *Unity*[2] game engine [4] that is frequently used for realistic high-quality 3D object generation. For every virtual scene, we arbitrarily select number of generated objects (up to 9), parameters for their lighting, shadows, and types of virtual surfaces reflection as follows. The "Directional", "Point" or "Spot" light types are used equally likely. With a probability of 20%, we choose a colored light where the brightness of red, green, and blue colors is random. From 4 to 6 light sources of the same type, with a random location behind the camera, are generated. The "Soft" shadows are added with a 50% probability to a virtual object at randomly selected values of the "shadowNormalBias" and "shadowBias" parameters. The "Standard" or "Diffuse" surface reflection types are utilized for every surface separately.

Seamless textures are randomly assigned to each surface of 3D objects. Affine transformation parameters are arbitrarily sampled for rendering of each virtual 3D object, its size in every dimension is scaled with a random multiplier from 0.9 to 1.1, and the colors of each texture are arbitrarily shifted in RGB space to further increase the diversity of the generated data. For each RGB-D recording, a virtual camera changes its position in the space of synthetic world to produce 3D images of virtual objects, which are subsequently utilized for augmentation of the NYU-v2 real-world RGB-D data. All the above-mentioned rendering and transformation parameters affect only virtual objects and do not directly influence the real-world scenes of the NYU-v2 dataset, which are modified only by the incorporation of virtual 3D images.

Augmentation and Culling. We randomly select RGB-D images from the NYU-v2 training pairs and on the basis of each depth map insert RGB pixels into a Unity3D scene, calculating the share of virtual objects which remains above the surface of the original real-world image. Data pairs, where virtual objects occupy from 10% to 50% of the RGB image area, are used for subsequent post-processing.

Post-processing. To maintain NYU-v2 color distribution in the virtualized RGB images, each color is normalized by the mean and standard deviation of distribution of its brightness in the original NUY-v2 training set.

Output. The resulting RGB-D data are stored in the file system using the existing NYU-v2 categorization and scene naming conventions. Some examples of generated

[2] https://unity.com/

RGB-D pairs of images are shown in Fig. 1. As one can see, depth maps of virtual objects are free from random distortions, show smooth depth variation, increase visual and depth variability of images, and in such way can provide a better quality of depth estimation.

4. Experiments

In the following sections, we empirically demonstrate the effectiveness of the augmented ANYU dataset for training of depth estimation models. For this, we train the latest state-of-the-art diffusion VPD architecture proposed by Zhao *et al*. [31] and transformer-based depth prediction PixelFormer neural network provided by Agarwal and Arora [1] on the original and augmented NYU-v2 datasets. Both models trained on our ANYU dataset yield performance improvements on the NYU-v2 and iBims-1 benchmarks described below.

4.1. Datasets

NYU depth v2 dataset (NYU-v2) provided by Silberman *et al*. [20] contains 24231 training and 645 test images of resolution 640 $\times$ 480 pixels, and covers 464 indoor scenes. It is widely used in various research involving a fine-grained depth estimation due to the indoors origin of the captured scenes.

Independent benchmark images and matched scans version 1 (iBims-1) is employed in our experiments for cross-dataset validation. This dataset provided by Koch *et al*. [15] consists of 100 high-quality RGB-D images and is specifically designed for validation of monocular depth estimation in different indoor scenarios. Compared to the NYU-v2, it provides lower noise levels, sharper depth transitions, fewer occlusions, and higher depth ranges. As iBims-1 has higher quality depth maps closer to real depth values compared to the NYU-v2, it can be used for more precise evaluation of depth prediction models.

4.2. Training and Testing

All experiments were performed on six NVIDIA H100 GPUs with 80GB of RAM. We used the official PyTorch implementation of the VPD model[3] provided by the authors [31] with a pre-trained image encoder, denoising UNet, and depth estimation decoder. The training architecture without text prompts is utilized. To reproduce the state-of-the-art results, we employed the same training and testing settings as in the original paper, except for fewer epochs and a smaller batch size, which appears to be more suitable for the virtually increased dataset. That is why by default, VPD model was trained during 18 epochs with a batch size of 12 (6 GPUs $\times$ batch size of 2), and demonstrates the same state-of-the-art depth estimation results (Fig. 2, 0% of augmented images) as in the original paper [31].

For the second transformer-based PixelFormer model with skip attention, we also utilized its official PyTorch implementation[4] provided by Agarwal and Arora [1], preserving all training and testing parameters from this paper. In addition to the original NYU-v2 test set, we created the virtualized one according to Section 3: artificially modified 2048 test RGB-D images are generated from the original 645 test pairs by enriching them with different virtual objects.

4.3. Evaluation Metrics

Following the standard evaluation protocol from prior works [1, 16, 31], for both NYU-v2 and iBims-1 datasets we report the root mean squared error (RMSE), accuracy metrics ($\delta_i < 1.25^i$ for $i \in 1, 2, 3$), absolute relative error (REL), and absolute error of log depths ($\log_{10}$).

5. Results and Discussion

The models are trained by running their open-source implementations on the NYU-v2 and ANYU datasets. We evaluate the depth estimation performance with the original, virtually enriched NYU-v2 test sets, and iBims-1 validation benchmark. Subsequent sections summarize the results of experiments conducted to answer the following questions:

- How does the accuracy on the NYU-v2 original and augmented *test sets* change with the increasing percentage of artificially modified training data.
- What is the *optimal share* of augmented training RGB-D image pairs.
- How *cross-dataset generalization performance* of the models changes when the augmented NYU-v2 dataset is used for training.

The results of these experiments are summarized in Fig. 2, 3, 6, 7, and Tab. 1, 2. Sample visual results obtained with the proposed solution are shown in Fig. 4, a comparison between the results produced with and without data augmentation is illustrated in Fig. 5. As one can see, the accuracy of depth estimation increased not only due to improved depth prediction for big 3D surfaces, but also due to better rendering of contours or even drawing of new details, which provides essentially better visual quality of 3D maps, though leads only to a slight improvement in the depth estimation metrics. Further samples of the training and test ANYU images are available in the supplementary material.

5.1. Monocular Depth Estimation on the ANYU Dataset

An exhaustive quantitative comparison of typical depth estimation metrics obtained on the ANYU dataset with dif-

[3]PyTorch implementation of VPD: https://github.com/wl-zhao/VPD

[4]PyTorch implementation of PixelFormer: https://github.com/ashutosh1807/PixelFormer

Figure 1. Examples of virtually augmented NYU-v2 RGB-D training pairs. Columns 1 and 3 show augmented RGB images, columns 2 and 4 — the corresponding depth maps.

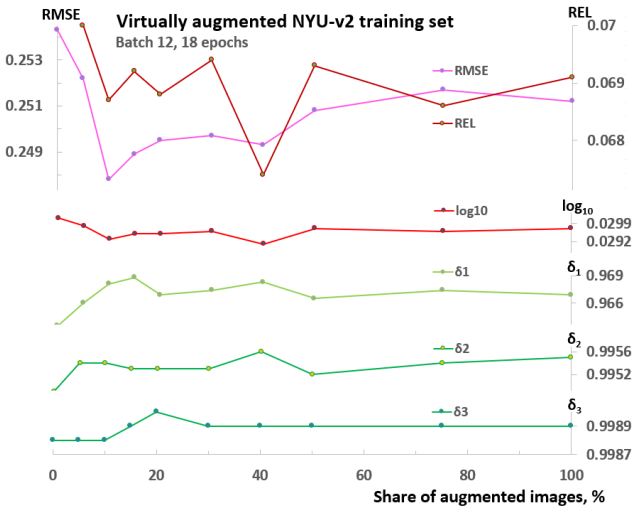

Figure 2. Performance breakdown of the VPD model [31] trained on the NYU-v2 dataset expanded up to a factor of 2 (100%) with virtualized RGB-D training images. All commonly used error metrics (RMSE↓, REL↓, $\log_{10}$↓) and performance metrics (δ_1↑, δ_2↑, δ_3↑) of the depth estimation show improvement over the results obtained on the original NYU-v2 dataset (0% of augmented images, abscissa axis).

ferent proportions of the original and virtually augmented training data is presented in Fig. 2, 3, 6, 7 and Tab. 1. More specifically, the results of the VPD neural network trained on RGB-D image sets with different amounts of the original and augmented training data are reported to compare their depth prediction performance on the NYU-v2 original and virtualized test sets. We quantify the augmentation-induced performance gains of computational depth prediction models in three aspects: by varying the proportion of augmented images, by reducing the NYU-v2 training set, and by altering the training conditions.

Full NYU-v2 dataset. Depth prediction metrics for the VPD model trained on the complete NYU-v2 dataset with additional 0% – 100% virtually augmented images are summarized in statistical plots in Fig. 2. The values of the RMSE (left axis), REL, $\log_{10}$, and δ_1 – δ_3 (right axes) are obtained on the original NYU-v2 test set. For comparison, RMSE for depth prediction on virtually altered NYU-v2 test set is presented in Fig. 3. As one can see, the artificial augmentation persistently effects the depth prediction for both the original Fig. 2 and the virtually modified Fig. 3 test sets.

VPD model trained on the augmented NYU-v2 dataset (5% – 100%) shows lower (better) values of the RMSE, REL and $\log_{10}$, and higher (better) levels of δ_1 – δ_3, compared to the values of these metrics on the NYU-v2 without virtual augmentation (0%). This means that supplementing virtually modified images yield a better depth estimation in all cases (5% – 100%) according to all commonly used metrics. The lowest RMSE value is achieved when using

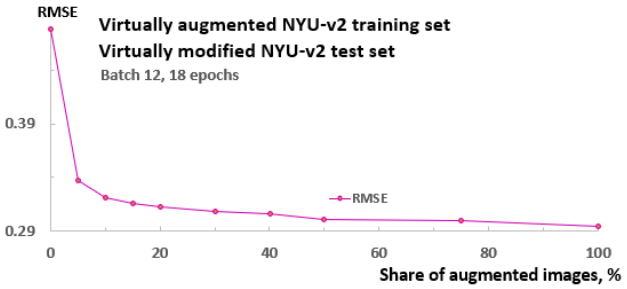

Figure 3. Performance of the VPD model [31] tested on virtually modified NYU-v2 test set after training on the NYU-v2 dataset expanded up to a factor of 2 with virtualized RGB-D images.

10% of augmented training images, while the best values for other standard metrics are obtained at 10% to 40% augmentation rate (Fig. 2). Further increase in the proportion of synthetically modified training images shows a tendency to degrade the depth prediction accuracy, which, however, remains better than on the original NYU-v2 dataset without augmentation. A probable reason for this effect is an incomplete correspondence between the quality of rendering surfaces / details of virtual objects compared to real ones. This assumption is confirmed by the fact that on the virtually modified test set augmenting the training data up to 100% leads to a continuous natural improvement in depth prediction, as shown by the constant decrease in RMSE that can be observed in Fig. 3.

In order to explore more deeply the regularities of this virtual augmentation process, we will further analyze the cases where dataset virtual enhancement yields performance gains, and in the following experiments compare the RMSE of depth prediction when reducing the size of the NYU-v2 training set.

Reduction of the NYU-v2 training set. As one can see in Fig. 6, shrinking from 100% to 1% of the original NYU-v2 training images naturally leads to an increase in the RMSE score, while expanding the quantity of the remaining training images through their virtually enriched clones results in a substantially smaller RMSE degradation. When the number of training images is reduced down to 1% from the initial size of the NYU-v2 training set, the RMSE rises approximately by a factor of 2, while expanding the dataset by virtually modifying the remaining training data improves the depth estimation accuracy by roughly 25% as shown by the corresponding RMSE scores. These experiments show that, as expected, the importance of augmentation is growing with the reduction of the original training data, however, the virtually modified images cannot maintain the same quality of depth prediction as the original ones, mainly due to a lack of image background diversity.

Varying training conditions on 1% and 5% of the NYU-v2 images. Furthermore, it is important to verify that changing the terms of VPD training, such as training pa-

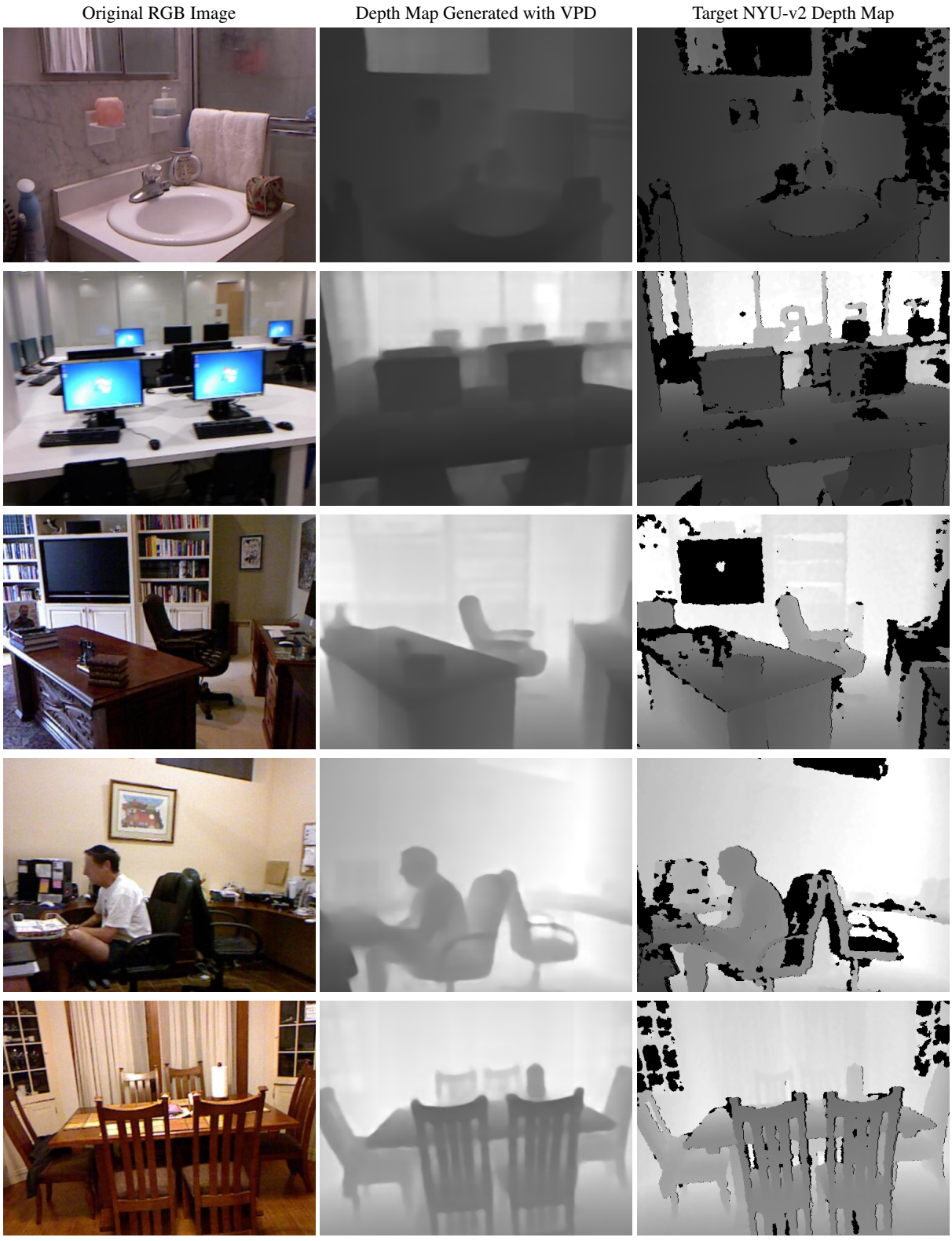

Figure 4. Sample visual results obtained with the VPD model [31] using the proposed augmented NYU-v2 dataset (ANYU).

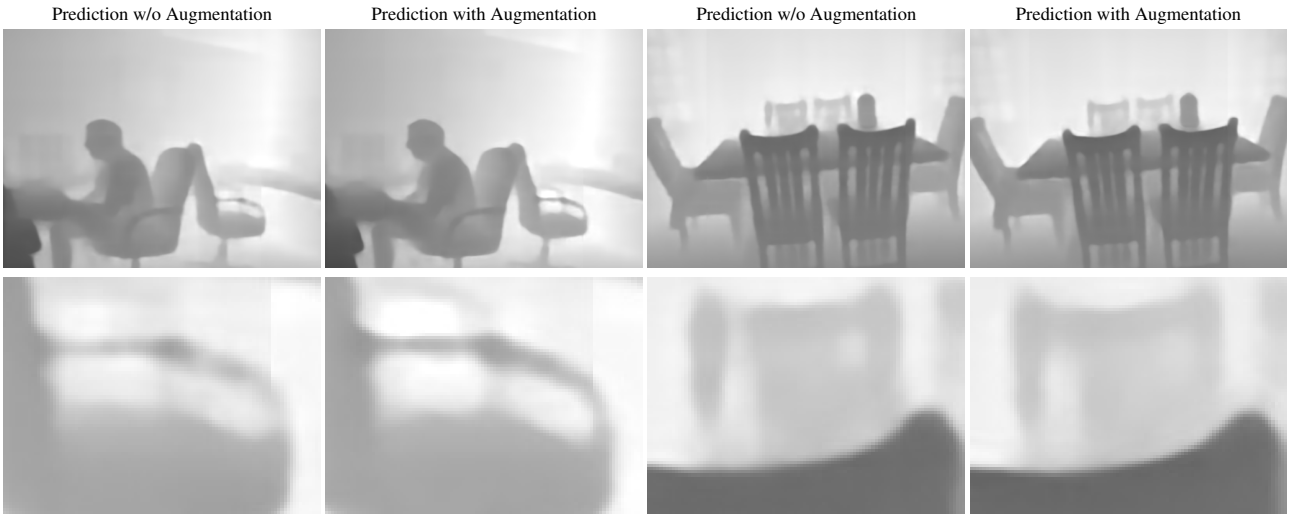

Figure 5. Sample visual results and the corresponding crops obtained with the VPD model [31] trained on the original and augmented NYU-v2 datasets. One can observe clearer, better-drawn objects and their contours when data augmentation is used.

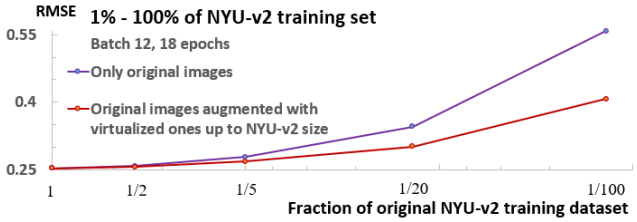

Figure 6. Performance of the VPD model [31] trained on 1% – 100% of the NYU-v2 training set. In the second series of experiments (red graph), the training set is virtually extended to the original NYU-v2 size with the proposed augmentations.

rameters or a portion of supplemented images, does not impact the fact that virtual augmentation reduces the RMSE of depth prediction. Therefore, in subsequent experiments we vary the parameters of these modalities to ensure that the observed patterns of augmentation-induced improvement in depth estimation are preserved across all cases examined.

As one can see from Fig. 7, the maximum reduction in the RMSE is observed when about 400% of additional virtually modified RGB-D image pairs are generated from a selected fractions (1% or 5%) of the NYU-v2 training data. This observation affirms the intuitive understanding that increasing diversity is more important for a small number of the original images than for the entire dataset, and the smaller the image quantity, the more valuable it is to increase their diversity. For all combinations of training parameters and fractions of additional images (from 1/1 to 20/1), the RMSE value is lower than in the case when only the original images are utilized. Thus, the augmentation-induced improvement in depth prediction is preserved in all

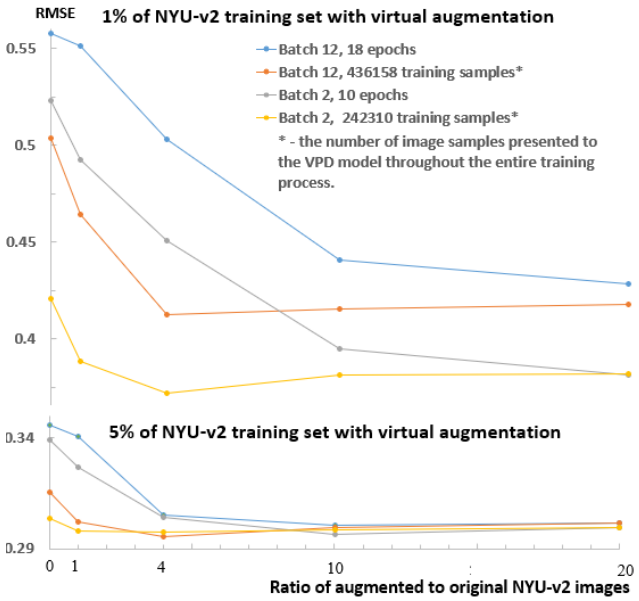

Figure 7. Performance of the VPD model [31] trained on a small portion of the NYU-v2 dataset depending on the level of augmentation: 1% or 5% of the original NYU-v2 training set is expanded by up to a factor of 20 with virtualized RGB-D training images. The artificially modified dataset is utilized for training of the VPD model with a batch size of 2 or 12. A fixed number of epochs or a constant quantity of total training samples (determining the number of epochs) presented to the VPD model is used.

cases studied, and the positive impact of artificially modified images increases as the original image count decreases. Further details and experimental results are presented in the supplementary material.

Table 1. Monocular depth estimation accuracy for deep models trained on the NYU depth v2 (NYU-v2) and augmented NYU-v2 (ANYU) datasets. ANYU yields new state-of-the-art results for the VPD model on the NYU-v2 test set and improves the performance of the PixelFormer on commonly accepted metrics.

Method	Dataset	RMSE↓	δ_1↑	δ_2↑	δ_3↑	REL↓	$\log_{10}$↓
Pixel-Former	NYU-v2 [1]	0.322	0.929	0.991	0.998	0.090	0.039
	ANYU	**0.320**	**0.930**	**0.999**	**0.999**	**0.090**	**0.038**
VPD	NYU-v2 [31]	0.254	0.964	0.995	0.999	0.069	0.030
	ANYU	**0.248**	**0.968**	**0.995**	**0.999**	**0.068**	**0.029**

Table 2. Cross-dataset validation on iBims-1 test set. Monocular depth estimation is performed by the models trained on the NYU depth v2 (NYU-v2) and augmented NYU-v2 (ANYU) datasets.

Method	Training Dataset	RMSE↓	δ_1↑	δ_2↑	δ_3↑	REL↓	$\log_{10}$↓
Liu *et al*. [16]	NYU-v2	2.665	0.192	0.601	0.876	0.329	0.184
Pixel-Former [1]	NYU-v2	1.595	0.165	**0.553**	0.907	0.335	0.184
	ANYU	**1.539**	**0.210**	0.549	**0.921**	**0.327**	**0.178**
VPD [31]	NYU-v2	1.493	0.196	0.649	**0.940**	0.313	0.169
	ANYU	**1.365**	**0.253**	**0.677**	0.939	**0.294**	**0.158**

At the end of this series of experiments, we compare the impact of virtual augmentation on the performance of two fundamentally different depth prediction models.

Depth prediction models with considerably different architectures. Tab. 1 summarizes and provides qualitative comparisons of the depth prediction metrics obtained with the VPD and PixelFormer models on the original NYU-v2 (ground truth) and ANYU datasets. As one can see, along with the VPD model from a novel family of generative diffusion neural networks, the well-established transformer-based PixelFormer deep neural network also exhibits improvement for all typical depth prediction metrics, confirming that the ANYU can be used for enhancement of monocular depth estimation on different types of depth prediction deep models.

The qualitative results presented in this section indicate that, in all considered scenarios, virtual augmentation leads to better performance of monocular depth estimation models compared to the training on the pure original NYU-v2. Therefore, the diversity of data in the NYU-v2 training set is not sufficient enough, and increasing it with artificial 3D images improves the results of depth estimation.

5.2. Cross-Dataset Validation

For a decent dataset, it is important that the models trained on it can show improvement while validating on another dataset acquired from different sources, which is called cross-dataset validation. Such approach provides an assessment of generalization of prediction models, e.g., by their training on the NYU-v2 and validating on the iBims-1 dataset, as done for example by Liu *et al*. [16]. The iBims-1 dataset contains different indoor scenarios and has higher quality depth maps closer to real depth values compared to the NYU-v2. Therefore, a validation on the iBims-1 dataset could verify the model efficiency for different data distributions between training and testing sets.

We employ the same PixelFormer and VPD networks in this experiment, utilizing models trained on the NYU-v2 by their authors, and models trained in previous experiments on the ANYU (full NYU-v2 with 10% of virtually augmented training images). These models are used without fine-tuning to generate the iBims-1 depth maps from RGB images. The common evaluation metrics for them are summarized in Tab. 2. Compared to the results [16] for the depth prediction model trained on the NYU-v2 and validated on the iBims-1, the PixelForme and VPD neural networks show essential improvements in depth estimation accuracy. The virtual augmentation of the NYU-v2 increases their performance across most metrics, acknowledging on the iBims-1 validation set the generalizing ability of these ANYU-trained depth prediction models.

6. Conclusion

In this paper, we introduced a new virtually augmented NYU depth v2 dataset, named ANYU, where artificially modified RGB-D training image pairs are enhanced with 3D objects from a virtual world. ANYU is provided in two training configurations with 10% and 100% of additional virtually enriched training images, respectively, for training new depth estimation models and for empirical exploration of the virtual augmentation. When generating ANYU, we deliberately did not match each virtual object with an appropriate texture and a suitable location within the real-world image. Instead, an assignment of texture, location, lighting, and other rendering parameters was randomized to maximize the diversity of training data, and to show that this randomness can improve the generalizability of a dataset. Considering common depth prediction metrics and validating on the NYU-v2 and iBims-1 benchmarks, we demonstrated that training depth estimation models on our ANYU data provides better generalization ability and improves the performance of indoor depth estimation. By training models with significantly different architectures on the ANYU, we improve the accuracy of the transformer-based PixelFormer model, and with the diffusion VPD neural network, achieve a new state-of-the-art result in monocular depth estimation.

Acknowledgments

This work was partially supported by the Alexander von Humboldt Foundation.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5861–5870, 2023. 1, 3, 8

[2] Shaojie Bai, Zhengyang Geng, Yash Savani, and J. Zico Kolter. Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 620–630, 2022. 2

[3] Fausto Tapia Benavides, Andrey Ignatov, and Radu Timofte. Phonedepth: A dataset for monocular depth estimation on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3049–3056, 2022. 2

[4] Ismail Buyuksalih, Serdar Bayburt, Gurcan Buyuksalih, AP Baskaraca, Hairi Karim, and Alias Abdul Rahman. 3d modelling and visualization based on the unity game engine–advantages and challenges. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4: 161–166, 2017. 2

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments, 2017. 1, 2

[6] A. Courtois, J. Morel, and P. Arias. Investigating neural architectures by synthetic dataset design. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4886–4895, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2

[7] Thibaud Ehret. Monocular Depth Estimation: a Review of the 2022 State of the Art. *Image Processing On Line*, 13: 38–56, 2023. 2

[8] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[9] Akhil Gurram, Ahmet Faruk Tuna, Fengyi Shen, Onay Urfalioglu, and Antonio M. López. Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12738–12751, 2022. 1, 2

[10] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J. Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1524–1531, 2014. 2

[11] Andrey Ignatov, Grigory Malivenko, David Plowman, Samarth Shukla, and Radu Timofte. Fast and accurate single-image depth estimation on mobile devices, mobile ai 2021 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2545–2557, 2021. 1

[12] Andrey Ignatov et al. Efficient single-image depth estimation on mobile devices, mobile ai & aim 2022 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 71–91, 2023. 1

[13] E. Ilg, T. Saikia, M. Keuper, and T. Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[14] Faisal Khan, Saqib Salahuddin, and Hossein Javidnia. Deep learning-based monocular depth estimation methods—a state-of-the-art review. *Sensors*, 20(8), 2020. 2

[15] Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Computer Vision and Image Understanding (CVIU)*, 191:102877, 2020. 1, 2, 3

[16] Peng Liu, Zonghua Zhang, Zhaozong Meng, and Nan Gao. Monocular depth estimation with joint attention feature distillation and wavelet-based loss function. *Sensors*, 21(1), 2021. 3, 8

[17] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[18] Alican Mertan, Damien Jade Duff, and Gozde Unal. Single image depth estimation: An overview. *Digital Signal Processing*, 123:103441, 2022. 2

[19] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 2

[20] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 3

[21] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 2

[22] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, pages 10912–10922, 2021. 2

[23] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[24] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J. Fleet. Monocular depth estimation using diffusion models, 2023. 1

[25] Roghani Seyed, Erfan Seyed, and Emre Koyuncu. Fine-tuning monocular depth-estimator artificial neural networks trained on synthetic rgb-d data sets for real scenes. In *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1105–1110, 2020. 2

[26] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1

[27] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[28] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment, 2018. 2

[29] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2023. 2

[30] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[31] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception, 2023. 1, 3, 5, 6, 7, 8