RANSAC-Scaled Depth: A Dual-Teacher Framework for Metric Depth Annotation in Data-Scarce Scenarios

M. V. Lazukov [1,2], *A. V. Shoshin* [2], *P. V. Belyaev* [2], *E. A. Shvets* [2]
[1] *Moscow Institute of Physics and Technology, 117303, Russia, Kerchenskay 1 "A";*
[2] *NVI Solutions LLC, 115191, Moscow, Russia, Kholodilny Lane 3, block 1, building 3, office 3103*

Abstract

This paper addresses the problem of training metric monocular depth estimation models for specialized domains in the absence of labeled real-world data. We propose a hybrid pseudo-labeling method that combines the predictions of two models: a metric "teacher," trained on synthetic data to obtain the correct scale, and a foundational relative "teacher" for structurally accurate scene geometry and depth. The relative depth map is calibrated via a linear transformation, whose parameters are found using the outlier-robust RANSAC algorithm on a subset of "support" points. Experiments on the KITTI dataset show that the proposed approach improves the quality of the pseudo-labels, reducing the commonly used error metric AbsRel by 21.6 % compared to the baseline method. A compact "student" model trained on these labels demonstrated superiority over the baseline model, achieving a 23.8 % reduction in AbsRel and a 13.8 % reduction in RMSE log. The results confirm that the proposed method significantly improves domain adaptation from general purpose to the specific domain, allowing for the creation of high-precision metric models without the need to collect and annotate volumes of real data.

Keywords: monocular metric depth estimation, synthetic data, RANSAC, pseudo-labeling, domain adaptation.

Citation: Lazukov MV, Shoshin AV, Belyaev PV, Shvets EA, RANSAC-Scaled Depth: A Dual-Teacher Framework for Metric Depth Annotation in Data-Scarce Scenarios. Computer Optics 2025; 49(6): 1174-1181. DOI: 10.18287/COJ1810

Introduction

Monocular depth estimation is the task of reconstructing the three-dimensional structure of a scene from a single two-dimensional image. This problem can be approached in two alternative ways: estimating the absolute distance to each point, known as metric depth estimation, or determining the relative arrangement of objects, referred to as relative depth estimation [1].

Metric depth estimation is considered the more challenging formulation due to the difficulty of creating a model that is invariant to variations in camera parameters, prior scene knowledge, etc. [2, 3]. As a result, metric models perform well when trained on data from their target environment but generalize poorly to unseen domains. By contrast, relative-depth models depend less on camera specifics and scene priors. They can be trained on images captured under diverse conditions and across domains, which yields stronger cross-domain generalisation.

This training flexibility is possible because the fundamental difference between metric and relative depth can be resolved with a simple linear transformation. A metric depth map (d_{metric}) can be approximated from a relative depth map ($d_{relative}$) using the equation $d_{metric} \approx s \cdot d_{relative} + t$, where s and t are unknown scale and shift parameters. This calibration enables the seamless integration of datasets with varying scales and camera optics, facilitating the training of robust, foundational models for relative depth estimation.

Despite these advantages, metric monocular depth remains essential for practical applications in fields such as autonomous driving and robotics [4], or industrial problems, including estimating the load volume of a truck [5] and determining safe distances from high-voltage power lines [6, 7], etc.

In turn, traditional stereo vision systems are also widely adopted for depth perception in robotics and autonomous navigation [8]. Commercial implementations such as Intel RealSense and ZED cameras [9] simplify deployment through factory calibration and onboard processing, offering ready-to-use solutions for many applications. However, fundamental constraints remain: stereo systems require two cameras with a rigid baseline maintaining precise calibration, making them vulnerable to mechanical vibrations and environmental stress that cause calibration drift. Depth accuracy degrades quadratically with distance, and matching algorithms fail on texture-less, reflective, or transparent surfaces [1]. Active sensing devices such as Time-of-Flight (ToF) cameras and LiDAR offer alternatives but face their own limitations. ToF cameras degrade significantly in strong ambient light, limiting outdoor reliability [1], and struggle with reflective or transparent surfaces at lower spatial resolution. LiDAR sensors provide accurate long-range measurements but remain cost-prohibitive for mass adoption and produce sparse point clouds rather than dense, per-pixel depth maps.

Beyond these technical limitations, traditional depth sensing approaches present practical deployment challenges that are particularly relevant for industrial applications. The requirement for specialized hardware significantly increases system costs compared to single-camera solutions. Moreover, many real-world deployment scenarios involve existing camera infrastructure such as surveillance systems at industrial sites or mounted cameras on stationary equipment where retrofitting with stereo or active sensing systems would require substantial hardware modifications and complex calibration procedures. In such cases, a pure software solution that leverages existing monocular cameras becomes highly attractive. Given these constraints: hardware complexity, calibration requirements, environmental sensitivity, deployment costs, and infrastructure compatibility monocular depth estimation presents a compelling alternative, particularly for applications where deployment flexibility, system cost, power consumption, and the ability to utilize existing camera infrastructure are critical factors.

A key challenge is the need to fine-tune metric models on relevant data to achieve the performance required for industrial applications and large-scale deployment. The process of collecting and annotating this data is time-consuming, expensive, and complex, particularly in industrial settings where data is often proprietary. Furthermore, ground truth data acquired from sensors has inherent limitations: it is often sparse, struggles with transparent or reflective surfaces, and can be distorted by adverse weather conditions. This issue is particularly evident with LiDAR sensors, where environmental clutter from elements like dust or snow can corrupt the data, and there is a noted lack of datasets that capture such naturally occurring distortions [10].

Even datasets marketed as high-precision can contain significant outliers and inaccuracies in their ground truth annotations. A notable example is the DIODE dataset [11], which was created using a high-precision laser scanner with a reported precision of $\pm 1mm$. However, a closer inspection reveals significant artifacts. For instance, on surfaces that appear visually flat, such as building rooftops, the depth values can vary by more than five meters.

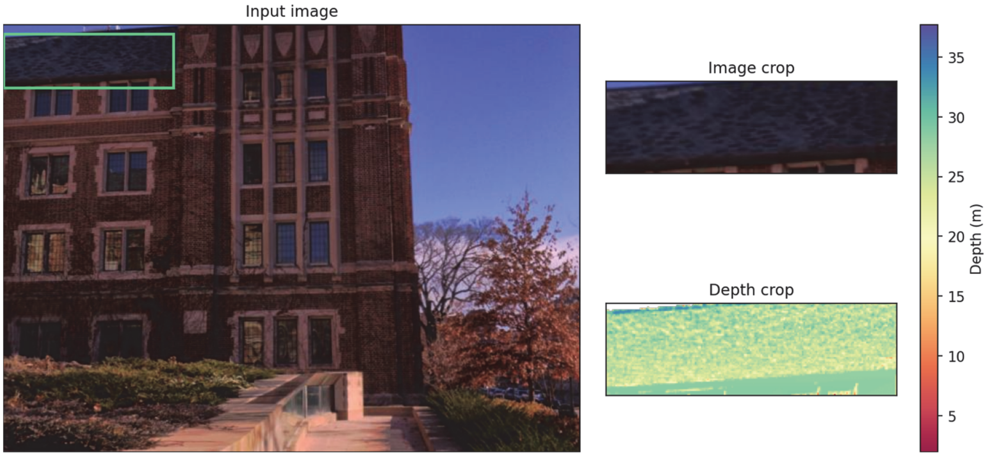

Fig. 1. Example of noisy region in the depth annotations in the DIODE dataset

The method presented in this work is designed to solve these problems. It enables the training of depth estimation models without relying on labeled real-world images, thereby avoiding the issues of noise and inaccuracies from measurement sensors.

1. Related work

To circumvent the fundamental challenge of data collection, two main research directions have emerged: (i) self-supervised learning from unlabelled videos or images [12], and (ii) the use of photorealistic synthetic datasets with perfect ground truth [13].

The first approach, self-supervised learning, aims to extract training signals directly from unlabeled real-world images. For example, Structure-from-Motion (SfM) [14] algorithm can analyze consecutive frames in a video stream to reconstruct a (typically sparse) scene structure and then it can be used to supervise a neural network. A similar idea can be applied to large image collections scraped from the Internet to generate training data. This allows for the automatic generation of depth maps, eliminating the need for expensive annotation.

A notable example is the MegaDepth dataset [15], which was generated by applying modern SfM techniques to a large collection of internet images. The primary advantage of this approach is its ability to use existing, readily available real-world images. However, it suffers from significant drawbacks: the resulting ground truth is often sparse, covering only static and well-textured areas of an image [14]. It can also be noisy and contain outliers in uniform or dynamic regions [15]. Most critically, the depth is estimated at an unknown metric scale, making it relative and requiring further calibration [16].

The second approach involves using synthetic datasets. These are typically created for a specific domain and consist of diverse, photorealistic images with pixel-perfect ground truth and known camera parameters. Synthetic data can model rare and complex scenarios, and the size of the dataset can be scaled as needed. The value and potential of this approach are underscored by the growing number of synthetic datasets [15 – 19] and their widespread adoption in monocular depth estimation research [20 – 23].

A primary limitation of this approach is the "sim-to-real gap"– the discrepancy between the data distributions of the real and simulated worlds that persists even with highly realistic rendering [26]. This gap can be minimized if a sufficient amount of unlabeled real-world data from the target domain is available. This is achieved through domain adaptation techniques, which transfer knowledge learned from a label-rich source domain (e.g., synthetic data) to a target domain where labeled data is scarce or nonexistent.

While there are many ways to classify these techniques, they can be conceptually divided into two main categories based on the level at which the domain distributions are aligned [27]: feature-level adaptation and prediction-level adaptation. Feature-level adaptation aims to train a neural network to extract internal representations from real and synthetic images that are statistically indistinguishable. The goal is to make the early layers of the model, which are responsible for recognizing basic patterns like edges and textures, perceive both data types similarly. Foundational works in this area proposed minimizing statistical discrepancies such as the Maximum Mean Discrepancy (MMD) [28] or aligning covariance matrices (e.g., Deep CORAL) [28]. This encourages the subsequent layers, which produce the final prediction, to become domain-invariant. While this approach can yield high-quality results, it is often more complex to implement and tune.

A more pragmatic and common alternative is prediction-level adaptation [29], often referred to as pseudo-labeling. Instead of intervening in the model's intermediate layer activations, this method operates on its final output. A "teacher" model, trained on synthetic data, generates predictions for a set of unlabeled real-world images [30]. These predictions, while imperfect, are then used as ground truth labels to train a second "student" model. Due to its conceptual simplicity, ease of implementation, and proven effectiveness [22, 28], pseudo-labeling is considered a powerful tool for bridging the domain gap [31].

Despite its popularity, pseudo-labeling has a fundamental limitation: the performance of the student model is typically capped by the performance of the teacher. If the teacher, trained on synthetic data, produces depth maps with artifacts or low level of detail, the student will inevitably inherit these flaws. Research shows that modern relative depth models, trained on vast amounts of real-world images, often produce maps with significantly higher detail and structural accuracy than even fine-tuned metric models [1, 20]. Metric models trained on sparse LiDAR data tend to generate smoothed and less-detailed maps, whereas relative depth models excel at reconstructing sharp boundaries and fine details [32]. Although these models do not predict absolute scale, they demonstrate an impressive ability to capture scene geometry and achieve strong zero-shot performance [33]. They possess a fundamentally better "understanding" of real-world geometry.

2. Proposed method

The primary objective of this work is to propose a method for training high-quality metric monocular depth estimation models for a specialized application domain. As previously mentioned, a key constraint is the absence of labeled real-world data from the target domain.

To address this task, we have the following components at our disposal:

A set of unlabeled real-world images: Images from the target domain for which metric depth needs to be predicted.

"Metric Teacher" Model: A neural network trained on a photorealistic synthetic dataset from the target domain. It can generate metrically correct depth maps, but its predictions may be smoothed, lack fine details, and contain various artifacts.

"Relative Teacher" Model: A neural network pre-trained on vast and diverse real-world image datasets. It generates structurally accurate and highly detailed relative depth maps.

As discussed in the literature review, the vast majority of existing domain adaptation approaches for metric depth estimation rely solely on a metric model to generate pseudo-labels. However, modern relative depth models capture more precise and detailed geometric information, such as fine textures and sharper object boundaries. We hypothesize that a predicted relative depth map, up to a linear transformation, is a better representation of the true scene geometry than a predicted metric map.

We propose a pseudo-labeling method that uses the predicted metric map as a source of scale information to calibrate the highly detailed relative depth map:

$$d_{pseudo} = s \cdot d_{relative} + t, (1)$$

where d_{pseudo} is the final metric pseudo-label, $d_{relative}$ is the relative depth map, s is the scale factor, and t is the shift factor.

The task of finding the scale and shift parameters (s, t) is framed as a linear regression problem. As a baseline, we consider the method of least squares (MSE), which provides an analytical solution but is sensitive to outliers. A more robust alternative is the RANdom Sample Consensus (RANSAC) algorithm [34]. It is widely used in image analysis tasks such as the automatic alignment of tomographic volumes [35], document image processing [34], and its modern implementations are being actively optimized to solve problems like 2D homography estimation [36]. RANSAC iteratively generates transformation hypotheses from random subsets of point pairs and selects the hypothesis that best explains the largest number of points (inliers) within a specified error threshold.

The robustness of the estimation depends not only on the chosen algorithm but also on the strategy for selecting the data points for calibration. We explore two pixel selection approaches: using all available points, or using a subset of "support" points formed by the extreme percentiles of the depth distribution. This second approach excludes the central, often smoother, part of the distribution. Specifically, we use the 10th and 90th percentiles of the metric depth map as thresholds:

$$T = \{ \, i \in \Omega | d_{metric}(i) \leq p_{10}(d_{metric}) \ \vee \ d_{metric}(i) \geq p_{90}(d_{metric}) \, \}, (2)$$

where, T is the resulting set of support points, i is a pixel from the set of all image pixels Ω, $d_{metric}(i)$ is the depth value of that pixel from the metric teacher's prediction, and $p_{10}(d_{metric})$ and $p_{90}(d_{metric})$ are the 10th and 90th percentile values of the entire metric depth map, respectively. This selection creates a mask of the nearest and farthest points, which are used for a more stable calibration.

Our support point selection strategy relies on the metric depth map to identify the nearest and farthest points in the scene, which provide robust geometric constraints for calibration. Because the metric teacher was fine-tuned on synthetic data from the target domain, its predictions provide a consistent, albeit smoothed, representation of the scene's absolute scale. While relative depth maps offer superior structural detail, they lack a meaningful scale; their extreme percentile values might highlight regions of high-frequency texture rather than the actual nearest and farthest scene points. Using the metrically-correct depth map ensures that our support points are genuinely representative of the scene's overall depth range, leading to a more stable and accurate scale and shift estimation.

We will evaluate the proposed methods in two ways: (i) by directly comparing the generated pseudo-labels against the ground truth, which is treated as a held-out set and is not used during training; and (ii) by comparing the performance of neural networks trained on the pseudo-labels generated by the different methods.

3. Experiments and results

3.1. Experimental Setup

The KITTI dataset [37] with the standard Eigen Split [38] was chosen as the real-world dataset for our experiments. This dataset, captured in an urban environment using LiDAR for ground truth generation, is a standard benchmark for evaluating monocular depth estimation methods. The training set contains 23,000 images, and the test set includes 652 labeled images.

Our study involves three neural network models. Their roles and specifications are described in Tab. 1.

Tab. 1. Models used in the experiments

Role	**Abbr.**	**Architecture**	**Description and Training Data**
Relative Teacher	$\mathrm{T}_{\mathrm{rel}}$	DepthAnything V2 Large	Foundational model trained on large-scale real-world datasets to produce structurally accurate relative depth maps [22]
Metric Teacher	$\mathrm{T}_{\mathrm{met}}$	DepthAnything V2 Large	Fine-tuned on the synthetic VKITTI2 dataset [37] to generate metrically correct predictions in the target domain [22]
Student Model	$\mathrm{S}_{\mathrm{met}}$	DepthAnything V2 Small	Compact model trained on the generated pseudo-labels to produce the final solution [22]

The quality of the predictions was assessed using a standard set of metrics for depth estimation:

Accuracy Metrics (δ_1, δ_2, δ_3): These measure the percentage of pixels for which the model's prediction is sufficiently close to the true value within a certain threshold. Higher is better (↑)

Error Metrics (AbsRel, RMSE, RMSE log): These measure the average discrepancy between predicted and true depth values across the entire image, in either absolute or logarithmic units. Lower is better (↓).

3.2. Direct Pseudo-Label Evaluation

The pseudo-label generation process begins with running both teachers on a source image. The calibration parameters are then determined to align the relative teacher's predictions. These parameters are calculated based on a set of pixels that can either include all image pixels or a subset of "support" points defined by the 10th and 90th percentiles of the metric depth distribution.

To evaluate effectiveness, five different configurations were established:

Baseline: Pseudo-labels are generated directly from the metric teacher's predictions without any calibration.

MSE: A hybrid method where calibration parameters for the relative depth map are determined using the least squares (MSE) method on all points.

Masked MSE: Similar to the above, but the regression is performed only on the subset of "support" points (10th and 90th depth percentiles).

RANSAC: A hybrid method that uses the outlier-robust RANSAC algorithm to find calibration parameters using all points.

Masked RANSAC: A combination of the RANSAC method with the "support" point selection strategy.

We configured our RANSAC implementation to perform a maximum of 1000 iterations. In each step, the algorithm constructed a hypothesis from a random sample of 50 points and classified points as inliers if their residual error was below a threshold of 0.1.

The first stage of experiments involved a direct quality assessment of the pseudo-labels generated by these methods by comparing them against the ground truth from the train and test KITTI Eigen Split. The results are presented in Table 2.

Tab. 2. Direct comparison of pseudo-label quality on the KITTI

Method	$\boldsymbol{\delta_1}$ ↑	$\boldsymbol{\delta_2}$ ↑	$\boldsymbol{\delta_3}$ ↑	**AbsRel↓**	**RMSE↓**	**RMSE log↓**
T_{met} (Baseline)	0.773	0.953	0.990	0.171	4.406	0.199
MSE	0.806	0.962	0.991	0.152	4.501	0.186
Masked MSE	0.836	0.969	0.993	0.134	4.047	0.168
RANSAC	0.809	0.965	0.992	0.153	4.308	0.183
Masked RANSAC	0.836	0.970	0.994	0.134	4.043	0.167

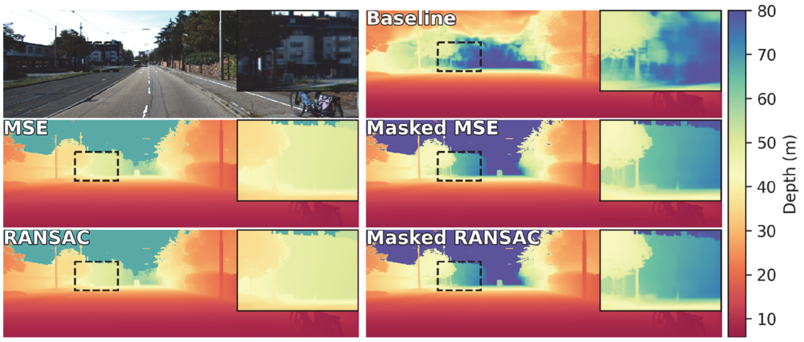

Fig. 2. Visual comparison of pseudo-label generation methods

The hybrid approaches, which use the structurally accurate relative depth maps from T_{rel}, show a significant advantage over the baseline method. The Masked RANSAC method achieved the best results across all metrics, reducing the Absolute Relative error from 0.171 to 0.134 and the RMSE by from 4.406 to 4.043 compared to the Baseline. A visual comparison of the pseudo-labels generated by the different methods is presented in Fig. 2.

3.3. Student Training on Pseudo-Labels

To provide a final validation of our method's effectiveness, a second experiment was conducted by training a compact student model (S_{met}) on the pseudo-labels generated by different methods. The goal was to assess how improvements in pseudo-label quality translate to the final performance of the target model. We compared two sets of training data: pseudo-labels from the baseline metric teacher ($\mathrm{T_{met}}$) and those from the best-performing Masked RANSAC method.

The $\mathrm{S_{met}}$ model was trained for 120 epochs with a batch size of 32 and an initial learning rate of $5 \cdot 10^{-6}$. Data augmentation consisted of random horizontal flips with a probability of 0.5. The input image resolution was 196×644 pixels. The evaluation results on the KITTI Eigen Split test set are shown in Tab. 3.

Tab. 3. Performance of the S_{met} trained on different pseudo-labels

Training Labels	$\boldsymbol{\delta_1}$ ↓	$\boldsymbol{\delta_2}$ ↑	$\boldsymbol{\delta_3}$ ↑	**AbsRel↓**	**RMSE↓**	**RMSE log↓**
$\mathrm{T_{met}}$ (Baseline)	0.815	0.959	0.989	0.151	4.032	0.181
Masked RANSAC	0.863	0.971	0.994	0.115	4.055	0.156

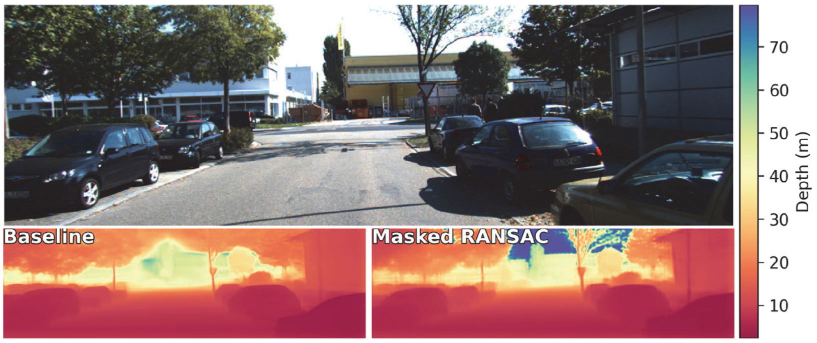

Fig. 3. Visual comparison of student models prediction

The results of this indirect comparison confirm the conclusions from the previous stage. The model trained on pseudo-labels generated by the Masked RANSAC method demonstrates a significant superiority over the model trained on the baseline labels from T_{met}. The most substantial improvements are seen in the AbsRel (a 23.8 % reduction from 0.151 to 0.115) and RMSE log (a 13.8 % reduction from 0.181 to 0.156) metrics.

The fact that the baseline model shows a marginally superior RMSE merits some explanation. Firstly, the difference is negligible (approximately 1%), whereas our method demonstrates significant improvements across alternative metrics. Secondly, the relatively higher RMSE can be explained by the high sensitivity of RMSE to large outliers in distant regions. Our method produces sharper and more detailed depth maps, which can occasionally lead to significant errors on a small number of pixels at complex object boundaries. The lower RMSE log values obtained corroborates this observation, as the logarithmic metric is less sensitive to such sparse, large-magnitude errors.

A visual comparison of the results from the models trained on the baseline and the best-performing pseudo-labels is shown in Fig. 3. This experiment proves that enhancing the quality and detail of pseudo-labels by combining metric and relative model predictions directly translates into higher accuracy for the final model.

3.4. Discussion

The experimental results validate our dual-teacher framework, confirming that hybrid pseudo-labels significantly improve model quality in the absence of annotated real-world data. The core of this success lies in robustly calibrating the structurally superior relative depth map using the metric teacher for scale information. Our analysis revealed that the most critical component of this process is our "support points" selection strategy. By focusing calibration on the nearest and farthest points in the depth distribution, this strategy leverages strong geometric constraints while avoiding noise from uninformative, uniform regions. The impact of this targeted selection proved more substantial than the choice between the RANSAC and MSE calibration methods. Despite this, RANSAC remains the theoretically preferred method due to its inherent robustness against outliers.

Conclusion

In this work, we proposed a method for training metric monocular depth estimation models that does not require labeled real-world data from the target domain. The approach aims to combine the fine-grained structure of relative depth with the absolute scale of metric depth by creating hybrid pseudo-labels.

We experimentally determine that the most effective calibration is achieved using the RANSAC algorithm applied to a subset of "support" points formed by the closes/farthest subsets of the depth distribution. A direct comparison on the KITTI dataset demonstrated that this method (Masked RANSAC) improves the quality of pseudo-labels compared to the baseline approach, reducing the AbsRel by 21.6 % (from 0.171 to 0.134) and the RMSE by 8.2 % (from 4.406 to 4.043).

Furthermore, a student model trained on the generated pseudo-labels significantly outperformed a model trained on the baseline labels, with improvements in AbsRel and RMSE log metrics of 23.8 % (from 0.151 to 0.115) and 13.8 % (from 0.181 to 0.156), respectively.

Our research proves that the proposed method is an effective solution to the domain adpatation of depth estimators. It enables the creation of domain-specific metric models with high level of detail for practical applications without the expensive process of collecting and annotating large volumes real-world data.

References

[1] Zhang J. Survey on Monocular Metric Depth Estimation. arXiv Preprint. 2025. Source: https://arxiv.org/abs/2501.11841. DOI: 10.48550/arXiv.2501.11841.

[2] Zhao Y, Bian H, Chen K, Ji P, Qu L, Lin S-Y, Yu W, Li H, Chen H, Shen J, Raj B, Xu M Metric from Human: Zero-shot Monocular Metric Depth Estimation via Test-time Adaptation. NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems 2024; 104724-104753. ISBN: 9798331314385

[3] Piccinelli L, Yang Y-H, Sakaridis C, Segu M, Li S, Van Gool L, Yu F. UniDepth: Universal monocular metric depth estimation. In: Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR); 2024. DOI: 10.1109/CVPR52733.2024.00963.

[4] Zhang Z, Zhang Y, Li Y, Wu L. Review of monocular depth estimation methods. Journal of Electronic Imaging 2025; 34(2):020901. DOI: 10.1117/1.JEI.34.2.020901.

[5] Chen J, Lu W, Yuan L, Wu Y, Xue F. Estimating construction waste truck payload volume using monocular vision. Resources, Conservation and Recycling 2022; 177:106013. DOI: 10.1016/j.resconrec.2021.106013.

[6] Yang T, Wei S, Fan L, Zhang L. Perspective transform-based depth estimation of monocular camera for electrocution threat determination of construction machinery. The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences 2024; XLVIII-4-2024:725-730. DOI: 10.5194/isprs-archives-XLVIII-4-2024-725-2024.

[7] Wang L, Wang B, Wang S, Ma F, Dong X, Yao L, Ma H, Mohamed MA. An effective method for sensing power safety distance based on monocular vision depth estimation. International Transactions on Electrical Energy Systems 2023; 2023(1):8480342. DOI: 10.1155/2023/8480342.

[8] Butt M, Nasir N, Rashid R. A review of perception sensors, techniques, and hardware architectures for autonomous low-altitude UAVs in non-cooperative local obstacle avoidance. Robotics and Autonomous Systems 2024; 173:104629. DOI: 10.1016/j.robot.2024.104629.

[9] Andriyanov N. Estimating Object Coordinates Using Convolutional Neural Networks and Intel Real Sense D415/D455 Depth Maps. International Conference on Information Technology and Nanotechnology (ITNT) 2022; 1-4. DOI: 10.1109/ITNT55410.2022.9848700

[10] Kokhan VL, Konyushenko ID, Bocharov DA, Seleznev IO, Nikolaev IP, Nikolaev DP. TSQ-2024: A categorized dataset of 2D LiDAR images of moving dump trucks in various environment conditions, in: Osten W, Nikolaev D, Debayle J . ICMV 2024; 2024. 13517: 1351709-1-1351709-6. DOI: 10.1117/12.3055203.

[11] Vasiljevic I, Kolkin N, Zhang S, Luo R, Wang H, Dai FZ, Daniele AF, Mostajabi M, Basart S, Walter MR, Shakhnarovich G. DIODE: A Dense Indoor and Outdoor Depth Dataset. arXiv Preprint. 2019. Source: https://arxiv.org/abs/1908.00463. DOI: 10.48550/arXiv.1908.00463

[12] Dong Q, Zhou Z, Qiu X, Zhang L. A survey on self-supervised monocular depth estimation based on deep neural networks. IEEE Transactions on Neural Networks and Learning Systems 2025; 1-21. DOI: 10.1109/TNNLS.2025.3552598.

[13] Zhihang S, He Z, Qiming M, Ming R, Mao Z, Pei H, Peng L, Hu J, Yao D, Zhang Y. Synthetic datasets for autonomous driving: A survey. IEEE Transactions on Intelligent Vehicles 2023; PP:1-19. DOI: 10.1109/TIV.2023.3331024.

[14] Izquierdo S, Civera J. SfM-TTR: Using structure from motion for test-time refinement of single-view depth networks. In: Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR); 2023. 21466-21476. DOI: 10.1109/CVPR52729.2023.02056.

[15] Li Z, Snavely N. MegaDepth: Learning single-view depth prediction from Internet photos. In: Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR); 2018. 2041-2050. DOI: 10.1109/CVPR.2018.00218.

[16] Liu X, Sinha A, Ishii M, Hager GD, Reiter A, Taylor RH, Unberath M. Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Transactions on Medical Imaging 2020; 39(5):1438-1447. DOI: 10.1109/TMI.2019.2950936.

[17] Roberts M, Ramapuram J, Ranjan A, Kumar A, Bautista MA, Paczan N, Webb R, Susskind JM. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: International Conference on Computer Vision (ICCV); 2021. Source: https://arxiv.org/pdf/2011.02523.

[18] Gaidon A, Wang Q, Cabon Y, Vig E. Virtual worlds as proxy for multi-object tracking analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. 4340-4349. DOI: 10.1109/CVPR.2016.470.

[19] Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. 3234-3243. DOI: 10.1109/CVPR.2016.352

[20] Karlo K. CARLA dataset for monocular depth estimation with varying camera parameters. Zenodo. 2023. DOI: 10.5281/zenodo.7899804

[21] McCormac J, Handa A, Leutenegger S, Davison AJ. SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation? In: IEEE International Conference on Computer Vision (ICCV); 2017. 2697-2706. DOI: 10.1109/ICCV.2017.292.

[22] Yang L, Kang B, Huang Z, Zhao Z, Xu X, Feng J, Zhao H. Depth Anything V2. arXiv Preprint. 2024. Source: https://arxiv.org/abs/2406.09414. DOI: 10.48550/arXiv.2406.09414

[23] Bochkovskii A, Delaunoy A, Germain H, Santos M, Zhou Y, Richter S-R, Koltun V. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. arXiv Preprint. 2024. Source: https://arxiv.org/abs/2410.02073. DOI: 10.48550/arXiv.2410.02073

[24] Li Z, Bhat SF, Wonka P. PatchRefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. Computer Vision - ECCV 2024; 2024. 15125:250-267. DOI: 10.1007/978-3-031-72855-6_15.

[25] Ke B, Obukhov A, Huang S, Metzger N, Daudt RC, Schindler K. Repurposing diffusion-based image generators for monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024. 9492-9502. DOI: 10.1109/CVPR52733.2024.00907.

[26] Man K, Chahl J. A Review of Synthetic Image Data and Its Use in Computer Vision. Journal of Imaging 2022;8(11):310. DOI: 10.3390/jimaging8110310.

[27] Toldo M, Maracani A, Michieli U, Zanuttigh P. Unsupervised domain adaptation in semantic segmentation: A review. Technologies 2020;8(2):35. DOI: 10.3390/technologies8020035.

[28] Wilson G, Cook DJ. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology 2020; 11(5):1–46. DOI: 10.1145/3400066.

[29] Chhabra S, Venkateswara H, Li B. Domain Adaptation Using Pseudo Labels. arXiv Preprint. 2024. Source: https://arxiv.org/abs/2402.06809. DOI: 10.48550/arXiv.2402.06809

[30] Yen Y-T, Lu C-N, Chiu W-C, Tsai Y-H. 3D-PL: Domain adaptive depth estimation with 3D-aware pseudo-labeling. Computer Vision - ECCV 2022. 710-728. DOI: 10.1007/978-3-031-19812-0_41.

[31] Kage P, Rothenberger JC, Andreadis P, Diochnos DI. A Review of Pseudo-Labeling for Computer Vision. arXiv Preprint. 2024. Source: https://arxiv.org/abs/2408.07221. DOI: 10.48550/arXiv.2408.07221

[32] Pham D-H, Do T, Nguyen P, Hua B-S, Nguyen K, Nguyen R. SharpDepth: Sharpening Metric Depth Predictions Using Diffusion Distillation. arXiv Preprint. 2024. Source: https://arxiv.org/abs/2411.18229. DOI: 10.48550/arXiv.2411.18229

[33] Marsal R, Chapoutot A, Xu P, Filliat D. A Simple yet Effective Test-Time Adaptation for Zero-Shot Monocular Metric Depth Estimation. arXiv Preprint. 2024. Source: https://arxiv.org/abs/2412.14103. DOI 10.48550/arXiv.2412.14103

[34] Skoryukina N, Arlazarov VV, Nikolaev DP. Fast method of ID documents location and type identification for mobile and server application. In: ICDAR 2019. 850–857. DOI: 10.1109/ICDAR.2019.00141.

[35] Bugai OA, Kulagin PA, Polevoy DV, Nikolaev DP. Orthotropic alignment for X-ray computed tomography images. In: Pang X, editor. Fifth Symposium on Pattern Recognition and Applications; 2025. 135400E-1–135400E-7. DOI: 10.1117/12.3056308.

[36] Rybakova EO, Trusov AV, Limonova EE, Skoryukina NS, Bulatov KB, Nikolaev DP. PESAC, the generalized framework for RANSAC-based methods on SIMD computing platforms. IEEE Access 2023;11:82151–82166. DOI: 10.1109/ACCESS.2023.3301777.

[37] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012. 3354-3361. DOI: 10.1109/CVPR.2012.6248074.

[38] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS); 2014. 2:2366-2374.

[39] Cabon Y, Murray N, Humenberger M. Virtual KITTI 2. arXiv Preprint. 2020. Source: https://arxiv.org/abs/2001.10773. DOI: 10.48550/arXiv.2001.10773

Author's information

Mikhail Vladimirovich Lazukov (b. 2000), completed Bachelor's degree (2021) from Tomsk Polytechnic University and Master's degree (2024) from Moscow Institute of Physics and Technology (MIPT). Currently he works as a Video Analytics Product *Development* Team Lead at NVI Research LLC and is a Ph.D. student at the MIPT. Research interests: computer vision, synthetic data, domain adaptation. E-mail: *mikhail.lazukov@nvi-research.com*

Alexey Valeryevich Shoshin (b. 1999), completed Bachelor's (2021) and Master's (2023) degrees from the Moscow Institute of Physics and Technology. Currently he works as a Machine Learning Engineer at NVI Research LLC. Research interests: 3D modeling, synthetic data, object detection. E-mail: *alexey.shoshin@nvi-research.com*

Philipp Vladimirovich Belyaev (b. 1984), completed higher education in Jurisprudence (2006) at the Academy of National Economy under the Government of the Russian Federation. In 2014, received an MBA with a specialization in International Oil and Gas Business from MGIMO University. Currently he works as Chief Executive Officer of the NVI Company Group. Specializes in video analytics, artificial intelligence, and industrial safety solutions for the energy sector; leads projects that enhance operational efficiency and safety standards using advanced AI technologies.
E-mail: *philipp.belyaev@nvi-solutions.com*

Evgeny Aleksandrovich Shvets (b. 1990) graduated from the Moscow Institute of Physics and Technology, Moscow, Russia, then received the Ph.D. degree in Technology from the Institute for Information Transmission Problems, Moscow in 2017. His Ph.D. thesis focused on distributed control of a robotic swarm for distributed area surveillance. His research interests include image processing, image registration and deep learning, including leverage of synthetic data for zero-shot training. Evgeny currently works as Chief AI Officer in NVI Research LLC.
E-mail: *evgeny.shvets@nvi-solutions.com*

Received August 20, 2025. The final version – October 23, 2025.