Enhancing Unsupervised Domain Adaptation for Monocular Depth Estimation: Research Plan

Arya Araban

October 2023

1 Motivation

Monocular depth estimation, the task of predicting depth from a single image, is a fundamental problem in computer vision. It finds applications in various domains, including autonomous navigation to ensure safe movement of vehicles and robots, 3D reconstruction for generating three-dimensional models, and augmented reality for realistic interactions with virtual objects. Monocular depth estimation also plays a vital role in robotics, enabling tasks like object recognition and manipulation, as well as assisting visually impaired individuals in healthcare settings.

Despite significant advancements in this field, most existing methods heavily rely on supervised learning and necessitate large-scale annotated datasets. The main challenge lies in the cost associated with obtaining accurate depth maps for real-world images, which typically requires the use of specialized and often expensive LiDAR sensors or stereo rigs. For example, the popular KITTI dataset [1] only contains 80,000 annotated frames, which is insufficient to train high-capacity models without overfitting.

As a consequence, these limitations often lead to restricted generalization capabilities, where models trained on one dataset may struggle to perform effectively on new target domains [2]. Unsupervised approaches that utilize structure from motion cues, such as estimating scene geometry from camera motion in a video sequence, circumvent the need for expensive labeling [3]. However, they encounter challenges when confronted with domain shifts, which occur when the distribution of the training data deviates from that of the test data.

Unsupervised domain adaptation (UDA) provides a promising solution by leveraging labeled data from source domains to improve target performance without annotation [4]. UDA is well-suited for this task as it directly addresses the domain shift problem. In contrast, other semi-supervised techniques still require some target labels. UDA can align source and target distributions through either adversarial training [5], feature normalization [6], or self-supervised pretraining [7].

Applying UDA to monocular depth estimation presents unique challenges. One of the key issues is maintaining content consistency during the adaptation process, ensuring that the semantic structures within the data are preserved even as the model adjusts to new input. This is crucial as these structures often hold valuable information for depth estimation.

Another significant challenge is adapting to complex target domains that feature varied lighting conditions or reflections that can degrade monocular cues. For instance, a model trained on clear daytime imagery may perform poorly when confronted with rainy or nighttime scenes. These conditions present a stark shift in lighting and can introduce elements like glare or shadows that the model may not be equipped to handle.

Addressing these challenges requires innovative approaches that can effectively handle domain shifts while preserving content consistency. While some prior works have explored UDA for monocular depth estimation [8, 9], performance remains limited due to the aforementioned challenges.

This research aims to enhance UDA techniques for monocular depth estimation by adapting synthetic models to real-world target domains while maintaining content consistency, facilitating the development of depth estimation networks with broad generalization capabilities across various environments

and illumination conditions, and minimizing labeling efforts. The research focuses on street view imagery due to the prevalence of these scenes in existing benchmark datasets for monocular depth estimation [10, 11], but the proposed approach has high potential to generalize to other domains and environments with available or easily obtainable synthetic data.

2 Research Question

Can domain randomization be combined with generative adversarial domain adaptation networks to improve unsupervised adaptation for monocular depth estimation from synthetic to real environments?

UDA techniques have demonstrated promise in adapting models trained on synthetic data to real-world target domains for monocular depth estimation. However, as previously mentioned, significant challenges remain in bridging the domain gap between synthetic source domains and complex real-world target environments. This research proposes a novel approach that combines domain randomization and adversarial domain adaptation to improve model transferability for monocular depth estimation.

Domain randomization refers to a technique where synthetic training data is randomized by applying a variety of transformations such as lighting changes, color shifts, and added noise [12]. This exposes models to diverse variations of the source domain and improves generalization to real-world target distributions. However, domain randomization alone is often insufficient to match real-world complexity due to its limited capacity to replicate intricate environmental factors and scene-specific nuances [13].

On the other hand, adversarial adaptation methods, specifically generative adversarial networks (GANs) have achieved impressive results for UDA across various tasks, including depth estimation [14, 15]. GANs consist of two components: a generator, responsible for producing images that closely resemble samples from the target domain while maintaining their semantic content, and a discriminator, which aims to distinguish between real and generated images. This framework enables models to adapt to target distributions without sacrificing crucial task-specific information. However, when dealing with intricate shifts, such as transitioning from synthetic to real data, GANs face challenges in generating truly realistic target-style images, which can potentially compromise details [16].

Combining these approaches provides an intriguing path to address their individual limitations. The key insight is that domain randomization can act as a "warm start", exposing the model to basic environmental variations and simplifying the adaptation problem. The GAN can then focus on handling the nuanced residual shift between the randomized source and real-world target.

This unified approach is well-suited for monocular depth estimation, where preserving spatial structures and layouts is critical during adaptation [17]. Domain randomization can allow the model to learn robustness to basic variations like lighting and texture changes. The GAN then adjusts the fine details to match real-world image statistics without distorting critical geometric cues. The expected outcome is improved content consistency during adaptation and better generalization.

While the idea of combining domain randomization and adversarial adaptation methods has been explored in other fields [18], Investigating this technique for monocular depth estimation remains an open challenge, despite the potential benefits. Monocular depth estimation relies heavily on spatial relationships and precise depth boundaries, which could be compromised by improper adaptation techniques [17]. Developing an approach that can adapt models to handle complex real-world shifts while still maintaining geometric consistency would be a valuable contribution.

3 Investigative Method

To investigate the proposed approach of combining domain randomization and GANs for UDA in monocular depth estimation, this study will utilize a mixed-method experimental design. A large-scale synthetic dataset will be generated to serve as the source domain, while real-world street view images will comprise the target domain.

The foundation of the synthetic source dataset will consist of photorealistic urban driving simulations generated via the CARLA platform [19]. This open-source simulator produces high-fidelity synthetic data and facilitates precise control over environmental conditions and scene contents. Static snapshots from the driving simulator will be extracted as source images. To implement domain randomization [12], the synthetic source images will be augmented with a wide array of random transformations including variations in lighting, weather, textures, camera angles, and field of view. Physics-based simulations will introduce realistic reflections, shadows, and lighting effects. By exposing the model to diverse synthetic variations during training, domain randomization aims to improve adaptation and generalization capabilities.

Real-world street view images will be collected from public datasets, including KITTI [1], Cityscapes [11], and ETH3D [20] to comprise the complex target domain. These datasets provide urban driving imagery captured under varied real-world conditions, featuring illuminations changes, weather patterns, seasonal differences, and unstructured environments. Images will be extracted from designated train and test splits to enable rigorous evaluation. The diversity of conditions represented across these target datasets will facilitate assessment of the model's ability to generalize.

The proposed approach will utilize a convolutional neural network (CNN) architecture derived from AUTODEPTH [21], which is a high-performing network used for monocular depth estimation. The chosen architecture effectively maintains critical structural and geometric cues for depth estimation. It incorporates Hourglasses, which are specialized image block modules that extract multi-scale features which retain both local details and global depth structure. The model will be first trained on only the randomized source dataset in a fully-supervised manner to learn basic depth estimation skills. Then, the UDA process will commence by continuing training on the source while also exposing the model to unlabeled real-world images in an adversarial fashion via the proposed combined framework.

Specifically, the pre-trained model will be integrated into a cyclic GAN framework similar to CycleGAN [22], which excels at learning cross-domain mappings. The key modification is the introduction of domain randomization on the source side, which allows the generator to learn a robust representation of the input data that can "fool" the discriminator. The generator will receive randomized source images and attempt to translate them towards realistic target-domain styles while preserving content critical for depth estimation. Meanwhile, the discriminator will ensure translated outputs are indistinguishable from real target samples, thereby forcing the generator to produce high-quality translations that are difficult to distinguish from real-world data. This adversarial process aims to close the domain gap, and the generator will ultimately be used to produce high-quality depth estimates for real-world scenarios. Critically, randomization provides a strong initialization before the nuanced adaptation to complex real-world domains begins, allowing the GAN to focus efforts on bridging the residual domain shift without distorting spatial relationships.

The proposed approach may potentially encounter challenges in generating highly realistic target-style images, especially under complex illumination changes. To address this, monitoring of the GAN's outputs throughout training is planned, ensuring that the translation ability is well-assessed. Additionally, maintaining content consistency during adaptation is crucial. Strategies will be employed to ensure that spatial relationships are preserved during the adversarial adaptation process, with further details about the evaluation of these outputs being discussed in the next section.

4 Analysis method

The performance of the proposed approach will be thoroughly evaluated both quantitatively and qualitatively to assess its ability to improve UDA for monocular depth estimation.

Quantitative analysis will rely on standard metrics used for evaluating depth estimation, including absolute relative difference (Abs Rel), squared relative difference (Sq Rel), root mean squared error (RMSE), and accuracy under threshold ($\delta < 1.25$) [23]. These metrics will be computed by comparing the predicted depth maps on the target real-world test sets [1, 11, 20] to the available ground truth depths. Lower error and higher accuracy indicates better performance. The values on the target test set will be compared to multiple baselines:

- Source only model - trained on just synthetic data, no adaptation

- Domain randomization only - adapted with randomized source data
- GAN adaptation only - adapted without domain randomization
- State-of-the-art unsupervised adaptation techniques [24]

Statistical significance will be evaluated using paired t-tests. Outperforming these alternatives would demonstrate the value of the proposed unified approach.

In addition to quantitative analysis, a qualitative assessment will be conducted by visually inspecting the generated depth maps. To facilitate this evaluation, two experts that specialize in depth estimation will be recruited. They will rate the perceptual quality on a Likert scale considering various factors such as depth boundaries, spatial layout preservation, robustness in challenging cases, and realism. Higher scores on the Likert scale indicate better adherence to these desired traits. The inter-rater reliability will also be measured to ensure consistent judgment. The qualitative feedback provided by the experts will offer valuable insights into the strengths of the proposed approach as well as areas that require refinement.

Furthermore, visualizations will be created to examine the intermediate outputs of the GAN model during training, allowing for an assessment by the recruited experts. These visualizations will focus on evaluating whether the source images appear realistic after randomization, whether the GAN successfully generates high-fidelity target-style translations, and whether the adaptation process effectively preserves spatial relationships. Through continuous monitoring of these characteristics during the training process, it will be possible to determine if the GAN is behaving as intended. In case adjustments to the model parameters are necessary to enhance the quality of the translations, the expert feedback will guide the refinement process.

To evaluate content consistency, structural similarity index measures (SSIM) will be computed to compare the input and translated images. Higher similarity scores indicate minimal distortion to critical elements such as object boundaries and scene layout. Additionally, a pre-trained state-of-the-art semantic segmentation model, CCNet [25], will be utilized to measure changes in pixel-level class assignments before and after adaptation. The preservation of semantic maps indicates the effectiveness of content retention.

Finally, the generalizability of the proposed approach will be tested by evaluating its performance on entirely new target domains with different characteristics. Specifically, a model adapted to clear daytime street scenes from Cityscapes [11] will be applied to rainy street images from RainCityscapes [26]. To assess performance, Quantitative depth metrics and qualitative expert ratings will be computed on these datasets and compared to alternative methods, including source-only, domain randomization-only, and GAN-only approaches.

By employing this comprehensive methodology that encompasses quantitative and qualitative analyses, the performance, outputs, and generalizability of the proposed unified domain adaptation approach can be rigorously assessed. Thorough comparisons with alternative methods will elucidate the specific benefits derived from combining domain randomization and GANs. Detailed analysis will also reveal areas for improvement in the technique. Overall, this experimental design will provide valuable insights into enhancing UDA for the challenging task of monocular depth estimation across divergent synthetic and real-world domains.

5 Contribution to Knowledge

This research aims to advance the state-of-the-art in UDA for monocular depth estimation by developing a novel approach that unifies domain randomization and adversarial learning. As discussed in the motivation, monocular depth estimation has immense value for various applications, but obtaining labeled data is costly and models often fail to generalize across domains with different characteristics. Previous works have explored the use of GANs for UDA in depth estimation [15], However, directly aligning complex synthetic and real-world image distributions remains challenging for GANs alone due to the vast domain shift that can distort important depth cues. This study hypothesizes that by incorporating domain randomization, the nuanced residual adaptation task presented to the GAN is simplified compared to traditional approaches.

If successful, this work would be the first demonstration of leveraging both domain randomization and adversarial adaptation in a cyclic framework for depth estimation from monocular cues. The key innovation lies in using randomization as a "warm start" to simplify the complex synthetic-to-real shift, allowing the GAN to focus efforts on bridging the residual gap without compromising spatial relationships critical for the task. Both the quantitative and qualitative analyses outlined aim to rigorously validate that this unified approach leads to improved performance and preservation of geometric cues compared to alternatives.

Furthermore, the proposed technique has the potential to minimize the need for costly manual labeling of depth data, facilitating the development of models that can generalize robustly across diverse real-world domains. This could expand the applicability of monocular depth estimation to new environments where capturing expensive ground truth is infeasible, such as underwater [27] or aerial scanning [28]. The approach may also inspire novel UDA solutions for other vision tasks that rely heavily on spatial semantics like semantic segmentation or object detection.

Beyond academic contributions, this research envisions real-world impact by making monocular depth estimation more accessible. The improved generalizability and reduced labeling requirements would enable broader deployment of models to benefit key applications. For instance, augmented reality systems could leverage models adapted via this method to integrate virtual objects seamlessly into any environment, without expensive calibration for each new condition. Similarly, autonomous navigation systems would gain enhanced environmental perception and safety across varying operational domains [29].

While the focus of this proposal is on synthetic-to-real transfer in the context of monocular depth estimation, the concepts could be extended to other scenarios where domain shifts occur, like cross-season or cross-city adaptation. It outlines an exciting new direction that could push the boundaries of UDA and beyond. The combination of domain randomization and GANs could greatly enhance the applicability of this technique, reducing the need for costly data collection, and enabling safer autonomous systems, more immersive extended reality experiences, and breakthroughs in other vision tasks. The proposal also anticipates positive outcomes from both quantitative and qualitative analyses, which would underscore the merits of this approach in advancing the field. Future work may include exploring optimal randomization strategies to simplify the domain mapping problem.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, June 2012.

[2] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022.

[3] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 1180–1189, PMLR, July 2015.

[5] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] J. Xu, L. Xiao, and A. M. López, "Self-Supervised Domain Adaptation for Computer Vision Tasks," *IEEE Access*, vol. 7, pp. 156694–156706, 2019.

[8] A. Tonioni, M. Poggi, S. Mattoccia, and L. D. Stefano, "Unsupervised domain adaptation for depth prediction from images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2396–2409, 2020.

[9] S. Saha, A. Obukhov, D. P. Paudel, M. Kanakis, Y. Chen, S. Georgoulis, and L. Van Gool, "Learning to relate depth and semantics for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8197–8207, June 2021.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017.

[13] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[14] A. R. Sanabria, F. Zambonelli, and J. Ye, "Unsupervised domain adaptation in activity recognition: A gan-based approach," *IEEE Access*, vol. 9, pp. 19421–19438, 2021.

[15] A. Pilzer, D. Xu, M. Puscas, E. Ricci, and N. Sebe, "Unsupervised adversarial depth estimation using cycled generative networks," in *2018 International Conference on 3D Vision (3DV)*, pp. 587–595, 2018.

[16] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] J. Tobin, L. Biewald, R. Duan, M. Andrychowicz, A. Handa, V. Kumar, B. McGrew, A. Ray, J. Schneider, P. Welinder, W. Zaremba, and P. Abbeel, "Domain randomization and generative models for robotic grasping," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3482–3489, 2018.

[19] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning* (S. Levine, V. Vanhoucke, and K. Goldberg, eds.), vol. 78 of *Proceedings of Machine Learning Research*, pp. 1–16, PMLR, 13–15 Nov 2017.

[20] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] S. Kumari, R. R. Jha, A. Bhavsar, and A. Nigam, "Autodepth: Single image depth map estimation via residual cnn encoder-decoder and stacked hourglass," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 340–344, 2019.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[24] F. Khan, S. Salahuddin, and H. Javidnia, "Deep learning-based monocular depth estimation methods—a state-of-the-art review," *Sensors*, vol. 20, no. 8, 2020.

[25] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "Shelfnet for fast semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[26] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8022–8031, 2019.

[27] X. Ye, Z. Li, B. Sun, Z. Wang, R. Xu, H. Li, and X. Fan, "Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3995–4008, 2020.

[28] Y. Zhang, Q. Yu, K. H. Low, and C. Lv, "A self-supervised monocular depth estimation approach based on uav aerial images," in *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, pp. 1–8, 2022.

[29] I. M. Elzayat, M. Ahmed Saad, M. M. Mostafa, R. Mahmoud Hassan, H. A. El Munim, M. Ghoneima, M. S. Darweesh, and H. Mostafa, "Real-time car detection-based depth estimation using mono camera," in *2018 30th International Conference on Microelectronics (ICM)*, pp. 248–251, 2018.