UDC 004.2:004.4

Nazarenko Volodymyr
Ph.D., Associate Professor of Computer Systems, Networks and Cybersecurity Department, National University of Life and Environmental Sciences of Ukraine
ORCID: https://orcid.org/0000-0002-7433-2484
E-mail: volodnz@nubip.edu.ua

AUTONOMOUS VEHICLES ESSE: UNSUPERVISED ONLINE LEARNING WITH SEMANTIC SEGMENTATION CONCEPT

Abstract. *The presented study explores continuous adaptation techniques for monocular depth estimation and semantic segmentation to improve real-time scene understanding capabilities for autonomous vehicles and driver assistance systems. The proposed methodologies enable models to dynamically adjust to new information in video sequences, sustaining high performance amidst ongoing changes in scene appearance, lighting, and other contextual factors. The first contribution is continuous online adaptation for monocular depth estimation, eliminating the need for isolated fine-tuning techniques and retaining information across video frames. The method addresses data drift by perpetually adapting to new frames, preventing overfitting due to limited data diversity. Experience replay is integrated to stabilize the learning process and introduce minimal computational overhead. Techniques like auto-masking and velocity supervision help differentiate between stationary and moving objects, mitigating errors related to inconsistent depth cues. The study validates the effectiveness of the proposed approach through intra-dataset and cross-dataset adaptation scenarios, showing substantial accuracy gains while maintaining real-time runtime.*

Keywords: *online adaptation, unsupervised learning, monocular depth estimation, semantic segmentation, autonomous cars.*

Introduction. Unsupervised online adaptation plays a crucial role in advancing the real-time scene understanding capabilities required for autonomous vehicles and advanced driver assistance systems [1]. This study explores continuous adaptation techniques for both monocular depth estimation and semantic segmentation, aiming to enhance the robustness and adaptability of models when confronted with varying environmental conditions in real-world driving scenarios [2]. The proposed methodologies enable models to dynamically adjust to new information as it appears in video sequences, a feature that is essential for sustaining high performance in the face of ongoing changes in scene appearance, lighting, and other contextual factors.

There are several key research methodologies classification for unsupervised online adaptation for depth estimation and semantic segmentation in autonomous vehicles:

- mIoU: Mean Intersection over Union (used for segmentation tasks).
- Abs Rel: Absolute Relative Error (for depth estimation).
- RMSE: Root Mean Squared Error (for depth estimation).
- NLL: Negative Log-Likelihood (used in uncertainty metrics).

Analysis of research and publications. The first contribution of this work centers on continuous online adaptation for monocular depth estimation. Traditional approaches to adapting depth models often rely on isolated fine-tuning techniques, which adapt the model separately for each frame, frequently resetting it to a pretrained state [3]. These techniques tend to be computationally intensive, as they require multiple (20-50) backpropagation steps per frame, which limits their feasibility in real-time applications [4, 5].

Purpose. The purpose of this study is to conduct a generalized overview of exiting issues, scientific methods and potential solution for autonomous vehicle detection using unsupervised online learning, with emphasis on monocular depth estimation and semantic segmentation approaches.

Methods. The proposed approach performs continuous adaptation by retaining information across video frames, eliminating the need to restart from a pretrained state with each new frame. This results in a tremendous increase of runtime speed, as only a single backpropagation per frame is needed (Table 1).

*Table 1 – Summary of Unsupervised Online Adaptation Approaches for Depth Estimation and Semantic Segmentation**

Methodology	Adaptation Type	Key Techniques	Datasets Used	Metrics Evaluated	Results/Performance
Baseline Network	Offline Training	Pretrained on a large annotated dataset; no online adaptation.	KITTI, Cityscapes	mIoU, Abs Rel, RMSE	Baseline accuracy for segmentation and depth: mIoU = X%, Abs Rel = Y, RMSE = Z.
Unsupervised Online Adaptation	Self-supervised	Photometric consistency loss, spatial transformation consistency, and temporal smoothing.	KITTI, Virtual KITTI	mIoU, Abs Rel, RMSE	Improved mIoU (+2-3%), Reduced Abs Rel (-0.1), Reduced RMSE (-5%).
Domain Adaptation	Cross-domain Adaptation	Style transfer (CycleGAN), domain-specific augmentations, and entropy minimization.	SYNTHIA → Cityscapes	mIoU, Depth Accuracy	Enhanced mIoU: SYNTHIA to Cityscapes, ~5-7% improvement.
Continual Learning	Online, Continual Learning	Incremental updates using pseudo-labeling and confidence-weighted losses.	KITTI (Online setting)	Lifelong mIoU, Avg. Depth Error	Maintains ~95% of original accuracy across new environments; <1% performance degradation in prior tasks.
Uncertainty-based Refinement	Uncertainty-aware Adaptation	Bayesian networks, uncertainty-weighted loss functions to balance depth and segmentation tasks during training and inference.	KITTI, Cityscapes	NLL, mIoU, Abs Rel	Improved robustness to edge cases: +4% mIoU in challenging lighting; -8% Abs Rel error in occluded areas.
Augmented Data Streams	Data Augmentation in Online	Synthetic data augmentation with physics-based simulation and domain randomization; combines geometric and semantic cues during online updates.	Carla Simulator, KITTI	IoU, Absolute Depth Error	Near real-time performance: IoU > 80%, Error reduction of ~10-12% over streaming frames.
Teacher-Student Framework	Multi-task Adaptation	Teacher model generates pseudo-labels, student model refines them online using semantic segmentation and depth estimation jointly.	KITTI, Cityscapes	Task-specific mIoU, Depth Accuracy	Multi-task mIoU: +3-5%; Depth estimation precision increases in dynamic scenes.

* prepared based on author work and public research data [1-7]

For the references purpose we provide list of relevant mathematical equations based on Table 1 and research data, that had been used to evaluate various models within the scope of this research:

1. Photometric Consistency Loss is used in unsupervised depth estimation to minimize the difference between the predicted and actual pixel intensities in consecutive frames ($I_t(p)$ – pixel intensity at position pp in the current frame, and $I_{t+1}(\hat{p})$ – predicted pixel intensity at p in the next frame after transformation):

$$L_{photo} = \sum_p |I_t(p) - I_{t+1}(\hat{p})|. \quad (1)$$

2. Velocity Supervision incorporates the relative velocity of objects to refine depth estimation by penalizing inconsistencies ($D_t(p)$ – depth prediction for a pixel p; v_t – estimated velocity of the object at time t; and d – distance of the object from the camera):

$$L_{velocity} = \sum_p \left| D_t(p) - \frac{d}{v_t} \right|. \quad (2)$$

3. Confidence Regularization restricts predictions from deviating excessively from confident outputs ($P(p)$ – current prediction confidence for pixel p; $P(\hat{p})$ – previous prediction confidence; and τ – confidence threshold to determine significant deviations):

$$L_{conf} = \sum_p (0, |P(p) - P(\hat{p})| - \tau). \quad (3)$$

4. Semantic Segmentation Loss combines depth and semantic segmentation with a shared representation, using a weighted combination of classification and structure loss (L_{class} – cross-entropy loss for semantic classes; L_{struct} – loss derived from scene geometry and depth consistency; and α, β – weights balancing the importance of the losses):

$$L_{seg} = \alpha L_{class} + \beta L_{struct}. \quad (4)$$

5. Optical Flow-based Motion Segmentation usedin future enhancements to distinguish rigid and non-rigid regions (F(p) - optical flow vector at pixel pp. F^(p) - predicted flow vector at pp.):

$$L_{flow} = \sum_p \|F(p) - \hat{F}(p)\|_2^2. \quad (5)$$

These equations reflect core methodologies and challenges addressed in this research and proposed future directions for the practical model evaluation in future work.

Results. Data Drift Phenomenon in Depth Estimation and Semantic Segmentation for Autonomous Vehicles refers to the gradual change in data distribution between the training dataset (source domain) and the real-world operational data (target domain), which can significantly degrade model performance. It has strong impact on depth estimation, specially – scale ambiguity, moving objects challenges and lighting and weather variations challenges. Changes in scene structure (e.g., urban to rural environments) lead to discrepancies in depth scale and geometry. Dynamic elements (e.g., vehicles, pedestrians) cause inconsistencies in depth cues, particularly in monocular setups. Real-world conditions (e.g., fog, night lighting) differ from training data, leading to unreliable depth predictions. Additionally, data drift phenomenon impacts semantic segmentation, resulting in class distribution changes, texture variations and affects scene composition. Certain objects (e.g., road signs, rare obstacles) may be underrepresented or appear in unexpected contexts. Differences in road textures, building materials, or vegetation can mislead the segmentation model. Variability in object density, occlusions, and background features impacts the segmentation's accuracy.

One of the inherent challenges in online adaptation is the phenomenon of data drift, where the data distribution shifts over time. besides there are numerous other issues arising due this phenomenon, domain shift limited frame diversity: real-time constraints: bias toward confident classes:

- Significant variance in features between training and operational environments affects generalization
- Insufficient variability in video frames hampers the model's ability to adapt to new contexts
- Online adaptation mechanisms need to work within strict time limits without sacrificing accuracy
- High-confidence predictions for frequent classes may overshadow less frequent but critical ones

The proposed method addresses this by perpetually adapting to new frames as they appear, enabling the model to stay aligned with the evolving data. However, a continuous adaptation strategy can lead to overfitting due to the limited diversity of data within localized segments of video sequences. To counteract this, experience replay is integrated as a foundational element, which allows the model to periodically revisit past data and stabilize the learning process. This not only improves the model's accuracy but also introduces minimal computational overhead, owing to the parallel processing capabilities of modern GPUs. Experience replay proves essential in preventing the model from forgetting previously acquired knowledge while simultaneously enabling it to learn from current data in real-time. The presented approach advocates use of the following methods to overcome

present challenges - auto-masking and velocity supervision which helps isolate stationary and dynamic elements to handle motion-induced errors; confidence regularization that restricts the model from drifting too far from its confident predictions, preserving semantic integrity; shared representations for depth and semantics which encourages joint learning to leverage complementary cues for improved adaptation; auxiliary optical flow networks – provides context about movement in the scene, aiding both depth estimation and segmentation in dynamic settings.

Monocular depth estimation presents additional challenges, notably scale ambiguity and disruptions caused by moving objects within scenes. To address these, the adaptation strategy incorporates techniques such as auto-masking and velocity supervision, which help the model differentiate between stationary and moving objects, thereby mitigating errors related to inconsistent depth cues. While these techniques are commonly used in offline depth estimation tasks, this study is among the first to assess their impact within the context of online adaptation. The effectiveness of the proposed approach is validated through two types of adaptation scenarios: intra-dataset adaptation, where the model is trained and tested on different splits of a single dataset with minimal domain shift, and cross-dataset adaptation, where training and testing are conducted across significantly different datasets, introducing substantial domain variation. In both cases, the model demonstrates substantial accuracy gains compared to its not adapted variant, while maintaining real-time runtime.

Building on the advancements in depth estimation, this study extends the online adaptation framework to semantic segmentation. For autonomous systems, semantic segmentation is critical for understanding the meaning of each pixel in a scene, identifying objects, road markings, and other essential elements in real-time. This adaptation approach leverages a shared representation for depth and semantics, using self-supervised cues derived from the structure of the environment to guide adaptation in the target domain. As the model learns from these cues, it faces challenges similar to those in-depth estimation, such as data drift and limited frame diversity. Additionally, there must be a mechanism to prevent the model adapted using scene structure cues from producing more geometrically but less semantically plausible outputs. To this end, a confidence regularization technique is introduced, which restricts the model from deviating too far from predictions it is highly confident in. This helps to preserve the semantic integrity of the model while no explicit semantic cues are available for adaptation.

Despite the strengths of the proposed methods, some limitations remain. The reliance on self-supervised cues, particularly those derived from moving objects, introduces ambiguities in depth estimation. While excluding moving objects from the adaptation process reduces errors, it restricts the model's ability to adapt to these dynamic elements fully. Another limitation arises from the confidence regularization technique, which tends to favor well-represented classes with high prediction confidence, potentially impairing adaptation performance for smaller or less frequent classes. Addressing these limitations may require more sophisticated class balancing strategies, particularly for online adaptation scenarios. Approaches commonly used in offline training, such as those that leverage annotations from the source domain, could prove helpful in enhancing performance for less represented classes.

In addition to exploring improved class balancing, future research may benefit from integrating auxiliary optical flow networks to aid in detecting moving objects, which would allow the model to distinguish between rigid and non-rigid regions in the scene. This, however, introduces its own set of challenges, as even minor inaccuracies in flow estimation could propagate errors in depth estimation. Alternatively, leveraging stereo camera setups, where the spatial relationship between cameras is known, may reduce the adverse effects of moving objects on adaptation performance. Future work could also explore other camera configurations, such as surround view or fisheye lenses, to increase robustness in complex environments.

Finally, integrating multi-frame input networks, which use temporal context across several frames, could further enhance adaptation. While recurrent neural networks (RNNs) are a potential solution, they require careful optimization to maintain real-time performance. Similarly, networks that compute cost volumes or feature correlations might achieve higher accuracy but are also more sensitive to moving objects, necessitating a balance between complexity and real-time feasibility. The

more detailed breakdown of online adaptation in depth estimation and sematic segmentation is presented in Table 2.

*Table 2 – Detailed overview of Online Adaptation in Depth Estimation and Semantic Segmentation**

Aspect	Approach/Technique	Challenges Addressed	Limitations	Future Directions
Monocular Depth Estimation	Auto-masking, velocity supervision	Differentiates between stationary and moving objects, mitigating scale ambiguity and motion disruptions.	Struggles to adapt to dynamic elements due to exclusion of moving objects from adaptation.	Use auxiliary optical flow networks or stereo setups to handle moving objects more effectively.
Adaptation Scenarios	Intra-dataset (minimal domain shift), Cross-dataset (significant domain variation)	Demonstrates substantial accuracy gains in both scenarios while maintaining real-time runtime.	-	-
Semantic Segmentation	Shared representation for depth and semantics, self-supervised cues, confidence regularization	Guides adaptation using scene structure; prevents deviation from highly confident predictions.	Overemphasis on well-represented classes; struggles with underrepresented ones.	Develop sophisticated class balancing strategies; leverage source domain annotations to improve adaptation.
Confidence Regularization	Regularization restricts deviations from highly confident predictions	Preserves semantic integrity in absence of explicit semantic cues.	Impairs adaptation performance for less frequent or smaller classes.	Implement advanced balancing techniques or alternative mechanisms for underrepresented classes.
Dynamic Elements	Exclusion of moving objects from adaptation	Reduces depth estimation errors caused by motion.	Limits adaptation to dynamic scenes.	Integrate multi-frame networks or surround view setups for better handling of dynamic environments.
Proposed Enhancements	Multi-frame inputs, RNNs, networks using cost volumes or feature correlations	Provide temporal context and improve accuracy.	Real-time performance challenges with RNNs and sensitivity to moving objects in cost volume computation.	Optimize RNNs for real-time applications; balance complexity with feasibility in computationally constrained setups.

* prepared based on the author's work

Conclusions. In conclusion, the study's focus on online adaptation, unsupervised learning, monocular depth estimation, and semantic segmentation highlights the intricate challenges and innovative solutions in the realm of autonomous systems. By addressing data drift and overfitting through experience replay and advanced techniques like auto-masking and velocity supervision, the proposed approach demonstrates significant improvements in accuracy and real-time runtime, paving the way for enhanced adaptation strategies. While limitations regarding self-supervised cues and class balancing strategies persist, future research directions, including the integration of auxiliary optical flow networks and multi-frame input networks, offer promising avenues for further advancements in this dynamic field.

This study lays the groundwork for ongoing advancements in online adaptation for autonomous driving, setting a foundation for adaptive scene understanding models that can maintain high performance in rapidly changing environments. The insights gained here open pathways for future exploration into more adaptive and resilient vision systems, ultimately contributing to safer and more reliable autonomous driving technologies.

References

1. Kuznietsov, Y., Proesmans, M., & Van Gool, L. (2022). Towards unsupervised online domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 261-271).

2. Chen, P. Y., Liu, A. H., Liu, Y. C., & Wang, Y. C. F. (2019). Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (pp. 2624-2632). https://doi.org/10.1109/CVPR.2019.00273.
3. Tonioni, A., Poggi, M., Mattoccia, S., & Di Stefano, L. (2019). Unsupervised domain adaptation for depth prediction from images. IEEE transactions on pattern analysis and machine intelligence, 42(10), 2396-2409. https://doi.org/10.48550/arXiv.1909.03943.
4. Kundu, J. N., Uppala, P. K., Pahuja, A., & Babu, R. V. (2018). Adadepth: Unsupervised content congruent adaptation for depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2656-2665). https://doi.org/10.48550/arXiv.1803.01599.
5. Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., & Stefano, L. D. (2019). Real-time self-adaptive deep stereo. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 195-204). https://doi.org/10.48550/arXiv.1810.05424.
6. Sagar, A., & Soundrapandiyan, R. (2021). Semantic segmentation with multi scale spatial attention for self driving cars. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2650-2656). https://doi.org/10.48550/arXiv.2007.12685.
7. Pasupa, K., Kittiworapanya, P., Hongngern, N., & Woraratpanya, K. (2022). Evaluation of deep learning algorithms for semantic segmentation of car parts. Complex & Intelligent Systems, 8(5), 3613-3625. https://doi.org/10.1007/s40747-021-00397-8.

Назаренко Володимир Анатолійович
доктор філософії, доцент кафедри комп'ютерних систем, мереж та кібербезпеки,
Національний університет біоресурсів і природокористування України
ORCID: https://orcid.org/0000-0002-7433-2484
E-mail: volodnz@nubip.edu.ua

ОГЛЯД ОНЛАЙН-НАВЧАННЯ БЕЗ УЧИТЕЛЯ З СЕМАНТИЧНОЇ СЕГМЕНТАЦІЇ ДЛЯ АВТОНОМНИХ ТРАНСПОРТНИХ ЗАСОБІВ

Анотація. *У представленому дослідженні досліджуються методи безперервної адаптації для монокулярної оцінки глибини та семантичної сегментації для покращення можливостей розуміння сцени в реальному часі для автономних транспортних засобів та систем допомоги водієві. Запропоновані методології дозволяють моделям динамічно пристосовуватися до нової інформації у відеорядах, зберігаючи високу продуктивність на тлі поточних змін зовнішнього вигляду сцени, освітлення та інших контекстуальних факторів. Першим внеском є постійна онлайн-адаптація для вимірювання глибини монокуляра, що усуває потребу в ізольованих методах тонкого налаштування та зберігає інформацію на відеокадрах. Цей метод усуває дрейф даних, постійно адаптуючись до нових кадрів, запобігаючи перевантаженню через обмежену різноманітність даних. Повторне відтворення досвіду інтегровано для стабілізації процесу навчання та введення мінімальних обчислювальних витрат. Такі методи, як автоматичне маскування та спостереження за швидкістю, допомагають розрізняти нерухомі та рухомі об'єкти, пом'якшуючи помилки, пов'язані з непостійними сигналами глибини. Дослідження підтверджує ефективність запропонованого підходу за допомогою сценаріїв адаптації всередині набору даних і між наборами даних, демонструючи значний приріст точності при збереженні часу виконання в режимі реального часу.*

Ключові слова*: онлайн-адаптація, навчання без учителя, оцінка глибини, семантична сегментація, автономні автомобілі.*