A Concept for Increasing Trustworthiness in Deep Learning Perception for UAS Using Map Data

Rebecca Schmidt, Joachim Rüter, Sebastian Schirmer, Stefan Krause, Johann C. Dauer
German Aerospace Center (DLR), Institute of Flight Systems, Dept. Unmanned Aircraft, Braunschweig, Germany

Deep Learning (DL)-based perception models provide state-of-the-art results in semantic segmentation and object detection, allowing an unmanned aircraft system (UAS) to perceive and understand its environment. This capability is particularly crucial for applications like onboard risk-based planning, where a UAS dynamically adjusts its trajectory given critical changes that are detected during flight. However, fully trusting the output of a DL model without safeguards is inadvisable, as DL models are regarded as black boxes, lacking explainability and interpretability of their outputs. To enhance trust in the DL output, regulatory bodies recommend monitoring the model output during flight by an independent system. In this work, we propose a runtime monitoring concept for a DL-based UAS environment perception system that detects static and dynamic objects. We increase the trustworthiness of the system by computing the plausibility of the model output using trustworthy map data and known contextual relationships. For static objects such as buildings and streets, the plausibility is calculated against the map data. The plausibility of dynamic objects such as pedestrians or vehicles, which are not present in the map data, is calculated using an ontology derived from map data. After presenting the concept as well as its strengths and weaknesses, future avenues for deploying the concept are highlighted.

I. Introduction

The ability to perceive the environment is crucial for the automated operation of an unmanned aircraft system (UAS), e.g., for inspection [1], search and rescue [2], and cargo delivery missions [3]. During such missions, objects obstructing the pre-planned trajectory must be detected, and the trajectory must be adjusted to prevent safety incidents. Deep Learning (DL)-based semantic segmentation and object detection improved significantly over the last years [4, 5] and are promising techniques to improve the perception capabilities of a UAS by enabling a better understanding of the scene captured in aerial images. For this reason, several studies are already investigating the usage of DL-based perception techniques for a UAS [6–11]. Despite the fact that DL models for segmentation and object detection have become increasingly robust over the last few years [5, 12], the essential problem remaining for deployment is that their output cannot be considered trustworthy, e.g., due to a lack of predictability and explainability of their behaviour [13]. To address this problem, ongoing efforts are focused on defining guidelines and regulations for the use of DL-based models in aviation [14]. In [13], the European Union Aviation Safety Agency (EASA) introduces four trustworthy Artificial Intelligence (AI) building blocks, which are considered essential for reaching trustworthiness of DL models for aviation: AI trustworthiness analysis, learning assurance, AI explainability, and AI safety risk mitigation, illustrated in Fig. 1. These building blocks are designed to address seven shared ethical values defined by the European Commission (AI HLEG, [15]) in 2019. Within this paper, we address the AI safety risk mitigation approach, which considers the problem that a DL model might retain a certain level of intransparency despite efforts in learning assurance or increasing explainability and must still be considered as a black box. One risk mitigation strategy mentioned by the EASA is to monitor the output of such DL models through an independent system [16].

In this work, we propose a concept to increase the trustworthiness of a DL-based environment perception system for risk-based path planning by monitoring the model output during runtime. We assume the environment perception system consists of a semantic segmentation model well suited to detect and classify static objects, such as buildings and streets [8], and an object detection model better suited to detect small dynamic objects like cars [17] and pedestrians [9]. In order to increase the trustworthiness, we validate the outputs of the models against trustworthy map data and known contextual relationships. We assume that high-quality map data will be able to be classified as trustworthy for use in aviation. This assumption is valid, as existing regulations for the certification of aeronautical data in aviation have been published by the EASA [18]. Furthermore, a successful example of certifying such data is demonstrated by the digital elevation model "Lido Surface Data NEXTView" from Lufthansa Systems and Intermap Technologies [19]. Using

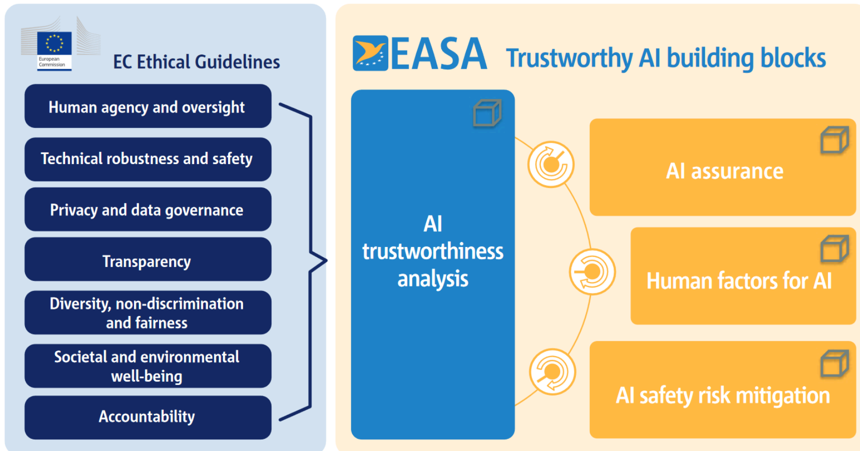


Fig. 1 EASA AI trustworthiness building blocks [16]

the trusted reference data, we derive the plausibility of both static objects in the scene, determined by the semantic segmentation model, and dynamic objects, assessed by the object detection model. Finally, the plausibility of both static and dynamic objects is provided to inform a risk-based path planning module. Section III outlines the example architecture for the environment perception system considered in this work. It is designed for risk-based path planning and supervised by the proposed monitoring concept, which is described in more detail in Section IV.

With this concept, we aim to address trustworthiness challenges in DL-based systems for aviation applications, working toward the safe use of DL-based perception methods for a UAS. The key contributions of this paper are as follows. First, we present two runtime monitors: one for DL-based semantic segmentation and another for DL-based object detection, both utilizing trustworthy map data, with the latter additionally incorporating a derived scene ontology. Second, we combine these monitors into a unified concept for plausibility validation of the entire environment perception system.

II. Related Work

Due to the high-dimensional input of the DL-based UAS perception system, the output cannot be validated through exhaustive testing, leaving a residual risk. As a result, safety risk mitigation strategies are crucial for ensuring a safe application. The primary goal of these strategies is to minimize unexpected behavior in the DL model by identifying and rejecting unsafe data encountered at runtime, thereby preventing the system from being impacted by erroneous outputs [16, 20, 21]. To provide a suitable risk mitigation approach, runtime monitors have been developed, which can serve as a key mechanism in runtime assurance standards, such as ASTM F3269 (Standard Practice for Methods to Safely Bound Flight Behavior of Unmanned Aircraft Systems Containing Complex Functions) [22].

Multiple data-based monitoring approaches have been proposed [23]. One group of approaches aims to detect out-of-distribution (OOD) data. The term OOD refers to data that differs significantly from the training set [20, 24] and therefore might lead to erroneous outputs. Several methods monitor intermediate outputs by observing neuron activation patterns in the models layers to identify OOD data [25, 26]. Other methods verify the models output confidence levels by applying uncertainty quantification methods to the final layer of the model [27–29]. Guérin et al. [20] argue that OOD detection might not be sufficient to design a monitor. They propose to rather focuses on identifying incorrect predictions of the DL model, called out-of-model-scope detection, than identifying inputs that differ from the training distribution. Furthermore, they note that in practice this idea differs from OOD only in the way the monitor is evaluated experimentally.

Another group of monitoring approaches aims to perform plausibility checks on the model input or output. Plausibility checks using expert knowledge are employed by Kang et al. [30], who apply model assertions, i.e., constraints on the model output, to detect erroneous results. A rapid change of an object class from one frame to another in a video might indicate such an erroneous output. The method demonstrates that model assertions can effectively detect when the model returns high confidences on wrong outputs. Similarly, Harper et al. [31] validate the behavior of the DL model by applying assertion checks to detected objects in a geographical context, utilizing geo-spatial information stored

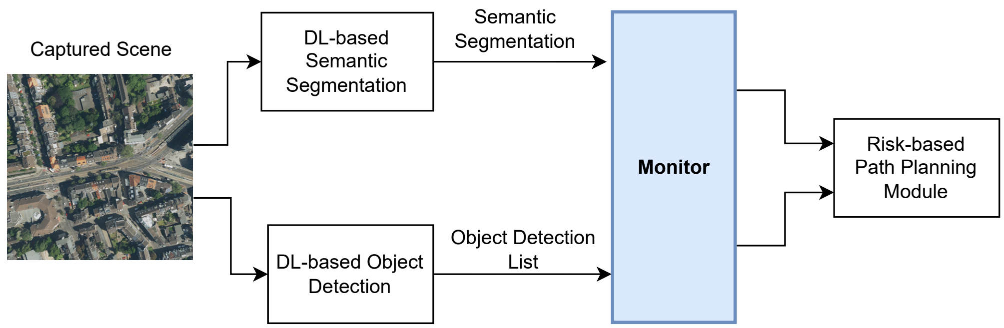


Fig. 2 Considered environment perception architecture. Perception outputs evaluated by the presented monitoring concept. Example scene image from [39]

in a Structured Query Language (SQL) database. Plausibility checks using external data sources, such as additional sensors, are considered by Zhou et al. [32]. In particular, they present a validation method for semantic segmentation in autonomous driving, using an additional Light Detection and Ranging (LiDAR) sensor to validate the segmentation output by comparing the classified areas from LiDAR and camera data. Similarly, Cheng et al. [33] introduce safety metrics to validate semantic segmentation against additional data, which can be provided by an additional sensor for runtime monitoring. Comparable approaches exist for object detection in autonomous driving [34, 35]. These studies monitor camera images by using a secondary perception system consisting of a LiDAR. However, the carrying of additional sensors for acquiring comparable scene data may not be feasible during UAS missions, as they can reduce the operation time due to increased weight and power consumption. Von Rueden et al. [36] validate DL-based semantic segmentation outputs by using a priori knowledge from OpenStreetMap (OSM) data for autonomous driving. The camera view is transformed into a bird's-eye view and compared to the map data by determining overlapping areas.

In this work, we use an external data source to perform plausibility checks on both the DL-based semantic segmentation and object detection models. Similar to [36], we use map data as trustworthy reference source and cross-check it against the model outputs. However, in comparison to that work, we address images from an aerial perspective and consider multiple semantic categories, including all those assessed by the segmentation and object detection models.

III. Considered Perception System Architecture

Onboard perception systems, comprising one or more camera sensors, are a common configuration for UAS environment perception due to their lightweight design and high-resolution capabilities [37]. These systems capture aerial images of the scene, such as the example shown in Fig. 2. The image data enables the identification of both static and dynamic objects within the scene. As described in Section I, reliably detecting these objects is crucial to prevent safety incidents. For a safe mission, static objects must be considered both during the pre-flight ground risk assessment, e.g., by using map data, and during the flight. Static objects can be subject to changes over a certain period of time, such as rivers altering their spatial appearance due to heavy rainfall or buildings undergoing reconstruction. These changes can cause discrepancies between map data used for pre-flight planning and the actual flight environment, which must be detected during the operation. Dynamic objects can only be partially accounted for during pre-flight planning. Whereas tools like population density maps [38] can provide partial insights, assessing the current scene during the flight is essential for mitigating potential risks. The image-based perception architecture shown in Fig. 2, represents an exemplary system designed to address these challenges. Static objects are identified through semantic segmentation, whereas dynamic objects are detected via object detection. The outputs from these DL models are monitored to ensure reliability before being used in the subsequent risk-based path planning module.

Although the shown architecture focuses on images captured from a downwards-facing perspective, the proposed monitoring concept is not restricted to such views, as perspective transformations can be applied to align the images with the map data. Additionally, we do not rely on specific DL model architectures for segmentation or object detection. Our monitoring concept is designed to consider arbitrary semantic classes that can be represented in map data or as contextual relationships.

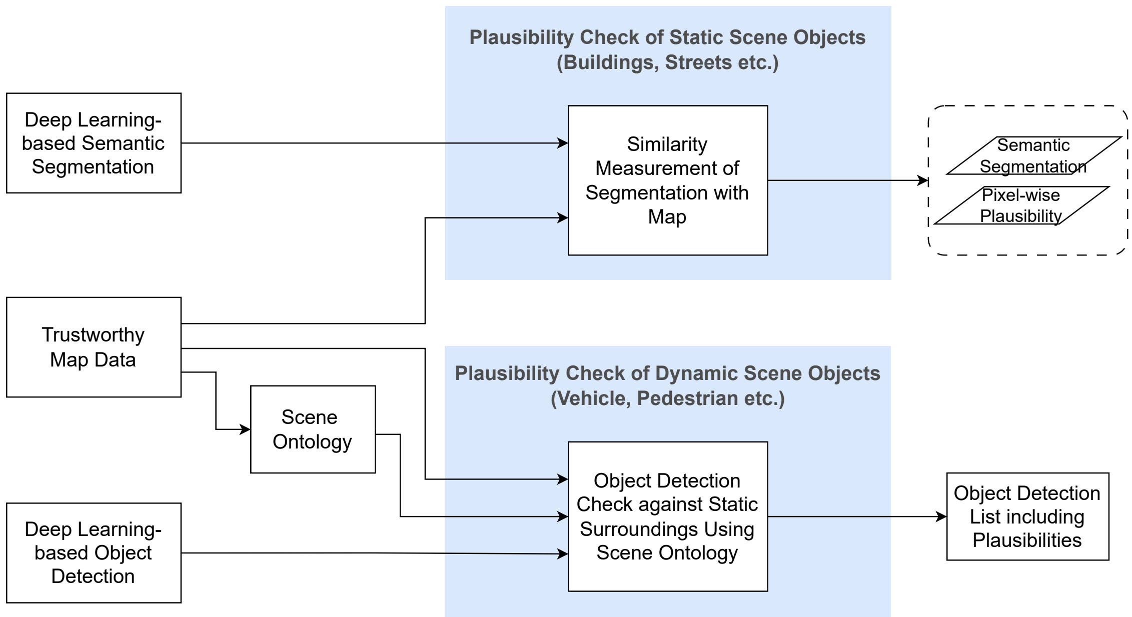


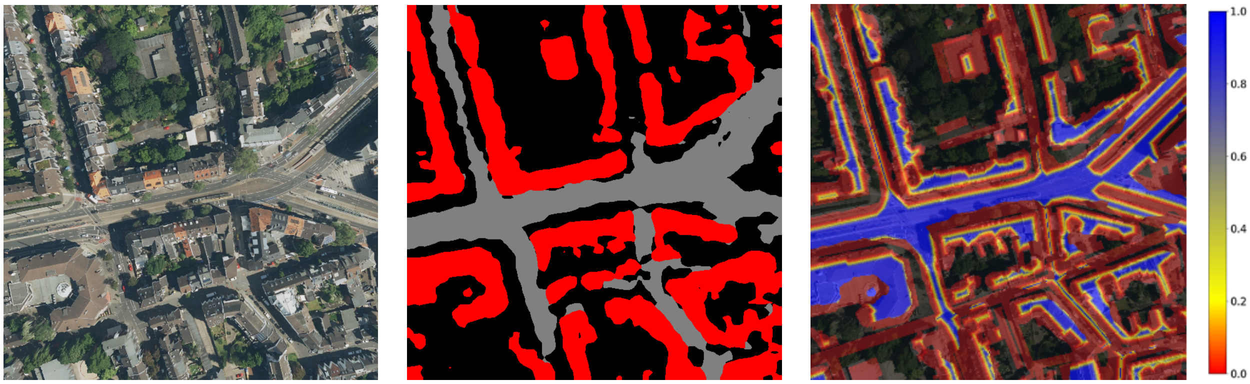
Fig. 3 Monitoring framework overview. Concept to increase the trustworthiness in a DL-based UAS perception system by monitoring the system output

IV. Monitoring Concept

Our monitoring concept aims to enhance the trustworthiness of an environment perception system by monitoring the outputs of the respective DL models. In this context, we define trustworthiness in alignment with the work of EASA [13, 16], which provides a concept for achieving trustworthiness. Our approach contributes to this goal by providing a risk mitigation strategy that assesses the plausibility of the DL model outputs through an architecture designed to cross-check outputs against trustworthy reference data. The plausibility checks are carried out within a framework consisting of two independent, parallelizable plausibility check modules, as shown in Fig. 3. The first component of the framework focuses on evaluating the plausibility of static objects, such as roads, buildings, and green spaces. This plausibility check is performed by comparing the segmentation output with the corresponding map data and measuring the scene similarity. The second component of the framework assesses the plausibility of dynamic objects, such as pedestrians or vehicles. Since dynamic objects are not present in map data, we use an ontology derived from the map data and known contextual relationships. The ontology is employed to assess the plausibility of the detected object in relation to its static surroundings within a defined area. In the following, the plausibility checks and the full output of the monitoring framework are described in more detail.

A. Plausibility Check of Static Scene Objects

Semantic segmentation provides access to both the spatial and semantic information of a scene captured in an image, making it particularly useful for detecting static objects. This process involves assigning each pixel in an image to a specific semantic category. The pixel-wise assignment is often visualized by overlaying the image with distinct colors, where each color represents a different category, as exemplary illustrated in Fig. 4b. The availability of pixel-wise semantic information enables a detailed, high-resolution plausibility check of the scene. In our approach, we use this information to cross-check it against trustworthy map data of the corresponding scene. This validation is carried out by comparing the semantic segmentation with the corresponding map section through a non-DL-based image matching process. Therefore, we assume access to prepared map data of the flight area. This map data includes only the semantic categories considered in the segmentation and is rendered using the same color scheme. We further assume that the captured scene includes a Global Navigation Satellite System (GNSS)-based geo-reference. However, such geo-reference is typically subject to some error, as GNSS accuracy can vary by up to 5 meters [40]. To account for this, we first search for the exact map section within a slightly larger area around the provided GNSS position. This is achieved using a template matching approach, such as [41], where the template slides across the search area and evaluates similarity at each position. The position with the highest similarity is selected as the matching result, and the



(a) (b) (c)

Fig. 4 Visual representation of the plausibility check for static objects. In (a), a scene image taken from UAS perspective is shown [39]. (b) illustrates an exemplary DL-based semantic segmentation of (a) with streets (gray), buildings (red) and background (black). In (c), an exemplary pixel-wise plausibility map with values between 0 (red) and 1 (blue) with no information (black) is provided

corresponding map section is subsequently used for a more detailed cross-check.

Although template matching provides a similarity measure between the map section and the segmentation, it is represented as a single value for the overlapping area, which fails to capture the desired level of detail. Furthermore, this value can only be interpreted in relation to other similarities determined by the template matching process, making it unsuitable for deriving a plausibility. Therefore, we propose a pixel-wise similarity measurement to obtain a more meaningful spatial and semantic similarity between the located map section and the semantic segmentation. This approach enables the detection of any deviations between the map and segmentation data. By comparing data at pixel-level, we eliminate the need to specify various types of deviations, such as noise or anomalies, that may arise between the data sources. For this comparison, we first split the segmentation image into the considered categories, generating a binary matrix for each category that indicates presence or absence. The map data is prepared in the same way to enable proper comparison between individual categories. We then perform the following procedure for each category and finally combine the binary results again. For each pixel in the binary matrix S_c, representing the semantic information of a specific category, we check whether the corresponding pixel in the binary map data M_c is assigned the same category. If the segmentation assigns a pixel to a different category than the map data, its plausibility is set to 0. Conversely, if the segmentation and map category match, we initially assign a plausibility value of 1. Afterwards, this value is multiplied by a weighting factor stored in a weight map W_c. This weighting factor accounts for the known limitations of semantic segmentation models, which tend to be less accurate at category boundaries, where misclassification is more frequent [42, 43]. Here, pre-determined weighting factors are used in W_c, assigning pixels at the center of an object a weight of 1, gradually decreasing to 0 near the objects contours. The determination of the plausibility map $Pl_{s,c}$ for static objects within a single category can be expressed by

$$Pl_{s,c} := (S_c \odot M_c) * W_c \tag{1}$$

where $\odot$ denotes the XNOR operator and * is an element-wise multiplication. An example of a plausibility map is illustrated in Fig. 4c. The derived plausibility can only be effectively interpreted in subsequent modules when combined with the corresponding semantic information. Therefore, the output of the static object plausibility check is a two-layer tensor consisting of the semantic segmentation from the DL model and the corresponding pixel-wise plausibility matrix.

B. Plausibility Check of Dynamic Scene Objects

As the detection of small and dynamic objects is challenging for the environment perception algorithms but crucial for a safe operation of the UAS, the trustworthiness of the detections is highly important. Current object detection models usually output a confidence score between 0 and 1 for each detection. However, this confidence cannot be fully trusted as DL models tend to be overconfident with their prediction compared to the observed accuracy [44]. Therefore, we aim to replace the confidence with a new value between 0 and 1 which is supposed to be more trustworthy. The new

(a) **(b)**

Fig. 5 Visual representation of the plausibility check for dynamic objects. In (a), the red box highlights a detected car in the image taken from UAS perspective [39]. In (b), the position of the detected car is projected onto the map data (red box) from OSM. The blue circle highlights the area in which present objects are considered for the plausibility calculation as described in Section IV.B

value should represent the plausibility of the object with respect to its surrounding objects. Since dynamic objects are not present in map data, we cannot directly apply the previously described approach for static objects. Our basic idea is to calculate a plausibility score for each detection based on the contextual relationships in its surroundings. This approach involves two steps: First, the detection is projected onto the trusted map data. Second, a plausibility score between 0 and 1 is determined by cross-checking with known contextual relationships of the object, which are represented as ontology. Figure 5 highlights the general concept.

The calculation of the plausibility is based on an ontology $\Omega = (E, C, R)$ that is derived from the map data before the flights are conducted. The ontology represents the set of relations R between relevant entities E in the map data under constraints C. The number of entities in the ontology is given by $|E|$. The outcome of a relation is the likelihood of occurrence, i.e., $R : E \times E \times C \longrightarrow [0, 1]$. For example, an ontology might contain the relation *likelihood of a street being within 3 meters of a street light* of 95 % and *likelihood of a street light being within 3 meters of a street* of 5 %. To create the ontology, we use map data of a pre-defined region in which the flights of the UAS should be conducted. Using the data of that region and pre-determined constraints, for each relevant object, we calculate the likelihood of each other relevant object being present in a pre-defined radius around the object. As this would only cover static objects present in the map and no dynamic objects, we then manually add the relations of dynamic objects of interest. For example, a person detected on top of a building could be added with a likelihood of 10 % whereas a person detected on the pavement may be more likely and therefore added with a likelihood of 80 %. Currently, these values are manually defined. In future work, they should be based on data to improve their validity. For example, they could be based on safety assessments derived from flight permits for the UAS.

Using this ontology, we propose to calculate the plausibility score as follows. The detection algorithm produces a set of detected, geo-referenced objects O_D. For each detected object $o \in O_D$, its plausibility with respect to entities $e \in E$ under constraints $c \in C$ is determined using the plausibility function Pl_Ω, defined as:

$$Pl_\Omega(o, c) = \frac{\sum_{e \in E} R(o, e, c)}{|\{e \in E \mid R(o, e, c) > 0\}|}. \tag{2}$$

For example, the constraint c can be defined to consider only those entities in Ω that are within an Euclidean distance of 3 m of the detected object o, i.e., $c := \|o - e\|_2 < 3$. Equation (2) presents a natural approach to compute the plausibility considering the surroundings of the detected object. Many other ways to calculate the plausibility using the ontology are possible and should be evaluated in the future. The output of this dynamic object plausibility check is an object list that additionally includes the plausibility score for each object.

V. Strengths and Weaknesses of the Concept

Compared to other monitoring approaches in the literature, this work introduces a concept designed to monitor a UAS perception system, which addresses both static and dynamic objects within the operational environment. Static and dynamic objects are monitored using a shared data source, consisting of a trusted map and an ontology derived from it. This data can either be pre-loaded onto the UAS or provided during operation via a communication link. Sharing map data across both monitors has the advantage of reducing the number of reference sources required to validate the DL output, which is particularly beneficial given the limited onboard memory of a UAS. However, scalability issues may emerge as map data can grow increasingly extensive in large operational areas. In addition to these general strengths and limitations, each part of the concept offers unique benefits and areas for further improvement, as described below.

The static object monitor applies a pixel-wise plausibility check to the segmentation output, allowing for detailed spatial analysis of the considered objects. However, the plausibility value assigned to each pixel is heavily influenced by the pre-defined weighting factor. Optimizing this weighting approach to better align with the characteristics of the semantic segmentation model could further improve the accuracy of the plausibility evaluation.

The dynamic object monitor benefits from the use of an ontology, which enhances the adaptability of the system. An ontology tailored to one region can also be applied to similar regions without extensive reconfiguration, making the approach suitable for dynamic applications across various geographic areas. However, the current method of manually adding dynamic objects to the ontology has some limitations. Although the likelihood values associated with objects in the ontology are reasonably assigned, they remain somewhat arbitrary and would benefit from a more systematic, data-driven approach to improve accuracy.

VI. Summary and Future Work

We introduce a map-based monitoring concept for a DL-based environment perception system, which consists of two models. The first model performs semantic segmentation of aerial images, whereas the second model detects objects within those images. Given that DL outputs are prone to errors and require trustworthiness validation, we use trusted map data and contextual relationships to assess the plausibility of the outputs. The cross-check between DL outputs and map data is twofold. For static scene objects detected by the semantic segmentation model, a similarity measurement evaluates how well the segmentation aligns with static objects in the map data, providing a pixel-wise plausibility. For dynamic scene objects detected by the object detection model, the plausibility is assessed based on an ontology derived from map data and manual augmentations of dynamic objects of interest.

We further discuss the strengths and limitations of this monitoring approach, highlighting that it covers two DL models for static and dynamic object identification, with both models being monitored by a shared data source within a single framework. We also address weaknesses, such as the pre-defined weighting factors in the static environment monitor and the manually assigned plausibility values within the ontology. To address these limitations, future work will involve a numerical evaluation of each monitoring component, i.e., the static and dynamic environment monitors. This evaluation will be used to refine the weighting approach for monitoring semantic segmentation outputs and to improve the plausibility metrics for object detection. A key question to explore through this evaluation is whether the proposed monitoring concept effectively enhances the trustworthiness of the UAS perception. Real-world flight tests may be conducted to answer this question, providing insights based on actual data and enabling an assessment of the runtime performance.

References

[1] S. Jordan, J. Moore, S. Hovet, J. Box, J. Perry, K. Kirsche, D. Lewis, and Z. T. H. Tse, "State-of-the-art technologies for uav inspections," *IET Radar, Sonar & Navigation*, vol. 12, no. 2, pp. 151–164, 2018.

[2] J. Scherer, S. Yahyanejad, S. Hayat, E. Yanmaz, T. Andre, A. Khan, V. Vukadinovic, C. Bettstetter, H. Hellwagner, and B. Rinner, "An autonomous multi-uav system for search and rescue," in *Proceedings of the first workshop on micro aerial vehicle networks, systems, and applications for civilian use*, 2015, pp. 33–38.

[3] T. Benarbia and K. Kyamakya, "A literature review of drone-based package delivery logistics systems and their implementation feasibility," *Sustainability*, vol. 14, no. 1, p. 360, 2021.

[4] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[5] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE access*, vol. 7, pp. 128 837–128 868, 2019.

[6] G. S., M. P. M.M., U. Verma, and R. M. Pai, "Semantic segmentation of uav aerial videos using convolutional neural networks," in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2019, pp. 21–27.

[7] S. Liu, J. Cheng, L. Liang, H. Bai, and W. Dang, "Light-weight semantic segmentation network for uav remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8287–8296, 2021.

[8] C. Hinniger and J. Rüter, "Synthetic training data for semantic segmentation of the environment from uav perspective," *Aerospace*, vol. 10, no. 7, 2023.

[9] J. Rüter, T. Maienschein, S. Schirmer, S. Schopferer, and C. Torens, "Filling the gaps: Using synthetic low-altitude aerial images to increase operational design domain coverage," *Sensors*, vol. 24, no. 4, 2024.

[10] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude uav datasets: A survey," *Image and Vision computing*, vol. 104, p. 104046, 2020.

[11] J. Rüter and R. Schmidt, "Using only synthetic images to train a drogue detector for aerial refueling," in *International Conference on Modelling and Simulation for Autonomous Systems*. Springer, 2023, pp. 392–413.

[12] P. de Jorge, R. Volpi, P. H. Torr, and G. Rogez, "Reliability in semantic segmentation: Are we on the right track?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7173–7182.

[13] European Union Aviation Safety Agency (EASA), "Artificial intelligence roadmap: A human-centric approach to ai in aviation," Tech. Rep., 2 2020.

[14] C. Torens, U. Durak, and J. C. Dauer, "Guidelines and regulatory framework for machine learning in aviation," in *AIAA Scitech 2022 Forum*, 2022, p. 1132.

[15] European Comission, "High-level expert group on artificial intelligence," 2024, accessed 20 September 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[16] European Union Aviation Safety Agency (EASA), "EASA concept paper: First usable guidance for level 1&2 machine learning applications," Tech. Rep., 2 2023.

[17] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Vehicle detection from uav imagery with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6047–6067, 2021.

[18] European Aviation Safety Agency (EASA), "Aeronautical database suppliers – dat provider's certificate," Tech. Rep., 2017, accessed 08 May 2024. [Online]. Available: https://www.easa.europa.eu/en/aeronautical-database-suppliers-DAT-providers

[19] Lufthansa Systems, "Lido surface data nextview (terrain)," 2020, accessed 15 May 2024. [Online]. Available: https://www.lhsystems.com/solutions/flight-operations-solutions/lido-data-solutions/lido-surface-data/lido-surface-data-0

[20] J. Guérin, K. Delmas, R. Ferreira, and J. Guiochet, "Out-of-distribution detection is not all you need," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 12, 2023, pp. 14 829–14 837.

[21] H. Bello, D. Geißler, L. Ray, S. Müller-Divéky, P. Müller, S. Kittrell, M. Liu, B. Zhou, and P. Lukowicz, "Towards certifiable ai in aviation: landscape, challenges, and opportunities," *arXiv preprint arXiv:2409.08666*, 2024.

[22] P. Nagarajan, S. K. Kannan, C. Torens, M. E. Vukas, and G. F. Wilber, "Astm f3269-an industry standard on run time assurance for aircraft systems," in *AIAA Scitech 2021 Forum*, 2021, p. 0525.

[23] R. S. Ferreira, J. Arlat, J. Guiochet, and H. Waeselynck, "Benchmarking safety monitors for image classifiers with machine learning," in *2021 IEEE 26th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2021, pp. 7–16.

[24] D. Siegismund, S. Heyse, and S. Steigele, "Uncertainty with deep learning: a practical view on out of distribution detection," in *2020 7th Swiss Conference on Data Science (SDS)*. IEEE, 2020, pp. 65–66.

[25] C.-H. Cheng, G. Nührenberg, and H. Yasuoka, "Runtime monitoring neuron activation patterns," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 300–303.

[26] T. A. Henzinger, A. Lukina, and C. Schilling, "Outside the box: Abstraction-based monitoring of neural networks," in *ECAI 2020*. IOS Press, 2020, pp. 2433–2440.

[27] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[28] M. Weiss and P. Tonella, "Fail-safe execution of deep learning based systems through uncertainty monitoring," in *2021 14th IEEE conference on software testing, verification and validation (ICST)*. IEEE, 2021, pp. 24–35.

[29] J. Guerin, K. Delmas, and J. Guiochet, “Evaluation of runtime monitoring for uav emergency landing,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9703–9709.

[30] D. Kang, D. Raghavan, P. Bailis, and M. Zaharia, “Model assertions for monitoring and improving ml models,” *Proceedings of Machine Learning and Systems*, vol. 2, pp. 481–496, 2020.

[31] C. Harper, G. Chance, A. Ghobrial, S. Alam, T. Pipe, and K. Eder, “Safety validation of autonomous vehicles using assertion checking,” *arXiv preprint arXiv:2111.04611*, 2021.

[32] W. Zhou, J. S. Berrio, S. Worrall, and E. Nebot, “Automated evaluation of semantic segmentation robustness for autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1951–1963, 2019.

[33] C.-H. Cheng, A. Knoll, and H.-C. Liao, “Safety metrics for semantic segmentation in autonomous driving,” in *2021 IEEE International Conference on Artificial Intelligence Testing (AITest)*. IEEE, 2021, pp. 57–64.

[34] C. Buerkle, F. Oboril, J. Burr, and K.-U. Scholl, “Safe perception-a hierarchical monitor approach,” in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 71–78.

[35] A. Vivekanandan, N. Maier, and J. M. Zöllner, “Plausibility verification for 3d object detectors using energy-based optimization,” in *European Conference on Computer Vision*. Springer, 2022, pp. 602–616.

[36] L. von Rueden, T. Wirtz, F. Hueger, J. D. Schneider, N. Piatkowski, and C. Bauckhage, “Street-map based validation of semantic segmentation in autonomous driving,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10 203–10 210.

[37] E. Balestrieri, P. Daponte, L. De Vito, and F. Lamonaca, “Sensors and measurements for unmanned systems: An overview,” *Sensors*, vol. 21, no. 4, p. 1518, 2021.

[38] S. Schopferer, J. Schlapbach, G. Strickert, L. González Combrink, and S. Fendel, “Minimum ground risk trajectory planning based on time-annotated population density maps,” in *AIAA SCITECH 2024 Forum*, 2024, p. 0233.

[39] M. Schleiss, F. Rouatbi, and D. Cremers, “Vpair–aerial visual place recognition and localization in large-scale outdoor environments,” *arXiv preprint arXiv:2205.11567*, 2022.

[40] W. Magiera, I. Vārna, I. Mitrofanovs, G. Silabrieds, A. Krawczyk, B. Skorupa, M. Apollo, and K. Maciuk, “Accuracy of code gnss receivers under various conditions,” *Remote Sensing*, vol. 14, no. 11, p. 2615, 2022.

[41] M. Schleiss, “Translating aerial images into street-map-like representations for visual self-localization of uavs,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 575–580, 2019.

[42] A. Bokhovkin and E. Burnaev, “Boundary loss for remote sensing imagery semantic segmentation,” in *International Symposium on Neural Networks*. Springer, 2019, pp. 388–401.

[43] E. Fernandez-Moral, R. Martins, D. Wolf, and P. Rives, “A new metric for evaluating semantic segmentation: leveraging global and contour accuracy,” in *2018 IEEE intelligent vehicles symposium (iv)*. IEEE, 2018, pp. 1051–1056.

[44] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.