ROMANIAN ACADEMY

School of Advanced Studies of the Romanian Academy

„Simion Stoilow" Institute of Mathematics

PHD THESIS SUMMARY

Towards Unsupervised Multi-Task Consensus Learning in Space and Time for Aerial Scene Understanding

THESIS ADVISOR:

Prof. Dr. Marius Leordeanu

PHD STUDENT:

Eng. Alina-Elena Marcu

2024

Contents

ABSTRACT

This thesis advances aerial scene understanding through innovative deep-learning approaches, focusing on semantic segmentation, depth estimation, and multi-task learning with limited human supervision. It introduces a novel dual-stream convolutional neural network for object segmentation, achieving state-of-the-art performance in building and road segmentation. Our research proposes new datasets and methods for road detection, lightweight CNNs for real-time onboard UAV applications and unsupervised depth estimation using real drone flights. To reduce manual annotation, this thesis introduces SegProp, an automatic label propagation method, and introduces Ruralscapes, a large-scale dataset for real-world semantic segmentation. Our work culminates with the Dronescapes dataset, designed for multi-task learning in complex real-world scenes, and a semi-supervised learning approach using a graph of neural networks and a multi-task hypergraph model. The hypergraph framework leverages multiple representations for holistic scene understanding with very limited human supervision.

Our research enhances aerial scene understanding through multi-task learning and consensus-based approaches. It contributes to Computer Vision, Remote Sensing, and Robotics by developing novel architectures, by introducing datasets and methodologies for aerial imagery analysis. The research paves the way for autonomous drones capable of understanding complex real-world environments.

Keywords – *Aerial Scene Understanding, Unmanned Aerial Vehicles (UAVs), Aerial Robots, Remote Sensing, Semantic Segmentation, Local-Global, Unsupervised Metric Depth Estimation, Consensus Learning, Multi-task Hypergraphs, Multi-task Consensus, Deep Learning, Artificial Intelligence, Robotics*

Chapter 1

INTRODUCTION

The field of Artificial Intelligence (AI) has witnessed remarkable advancements that can be attributed to advancements in software algorithms, the availability of vast amounts of data, and the exponential growth in computational power, especially graphics processing units (GPUs) [1]. Robotics tackles the challenge of creating physical agents (robots) capable of interacting with the real world, bridging the gap between digital intelligence and tangible action and has been profoundly impacted by the rise of AI.

Peter Corke's succinct description captures the essence of what constitutes a *robot* [2] – *"a goal-oriented machine that can sense, plan, and act"*. This definition highlights the key components of robotic systems: perception, decision-making, and interaction with the physical world. Among these components, perception, particularly in the form of scene understanding, plays a crucial role in enabling robots to navigate and interact with their environment effectively, presenting its own set of unique challenges. The complexity of scene understanding stems from several factors. First, the visual world is inherently diverse and dynamic, requiring systems to generalize across a vast array of possible scenarios. Second, understanding a scene goes beyond mere object recognition; it involves comprehending spatiotemporal relationships, contextual cues, and even predicting intent (forecasting). Finally, scene understanding must operate in real-time

for many robotic applications, adding computational constraints to an already complex problem.

This thesis introduces methods for gaining a comprehensive understanding of the environment through multiple visual cues. This holistic approach involves tackling multiple tasks to obtain complementary information and diverse visual interpretations, all within the context of robotics and real-world applications. We consider both the three-dimensional spatial structure and the temporal evolution striving for a semantic understanding that enhances navigation capabilities. Adding another layer of complexity, we explore these concepts from an aerial perspective, which presents unique challenges and opportunities. To achieve our goals, we leverage the consensus between multiple information sources and combine data-driven learning methods. Recognizing the critical importance and difficulty of obtaining high-quality data we focus on developing approaches that maximize the use of sensors while minimizing human intervention, thereby addressing the complexities of the real world more efficiently.

Our work evolved from addressing individual tasks to developing comprehensive approaches for holistic scene understanding. It bridges the gap between supervised and unsupervised learning by exploiting the consensus that emerges between multiple visual cues using minimal supervision. Our focus is on improving scene understanding in the field of aerial robotics, as they present unique challenges that demand innovative solutions. The **main contributions** this thesis makes are summarized below:

Local - Global Multi-stage Consensus Learning – Our journey began with satellite imagery, where we introduced local-global multi-stage consensus learning for aerial segmentation. This approach demonstrated the power of combining local object appearance with global contextual information for two important tasks in Remote Sensing, building and road segmentation. We have also introduced the multi-stage multi-task neural networks capable of simultaneous segmentation and geolocalization. In a similar multi-stage manner, we also show improvement in road segmentation of varied widths and difficult remote environments, with state-of-the-art performance.

Lightweight CNNs for real-time UAV deployment – The progression from satellite to UAV-based research was a natural and necessary evolution, driven by the scarcity of

comprehensive UAV datasets and the unique challenges posed by this domain. Recognizing the computational constraints of UAV systems, we developed lightweight models suitable for onboard deployment, such as SafeUAVNets. These efficient architectures constitute the processing units for all of our subsequent contributions proving their adaptability and effectiveness for multiple tasks.

Hybrid (analytical and learned) approaches – We combined analytical and geometric approaches with data-driven deep learning techniques, as seen in our depth distillation method and UFODepth to cancel out each other's limitations, for efficient unsupervised metric depth estimation.

Learning with minimal human supervision – We developed novel semi-supervised learning techniques, including our SegProp method that leverages spatiotemporal consensus for automatic label propagation and the Neural Graph Consensus (NGC) model, to effectively leverage limited labeled data. These methods substantially reduce the need for extensive manual annotations, a critical advantage in aerial imagery analysis where labeled data is often scarce and costly.

Multi-task Consensus Learning – Our research also introduced innovative multi-task learning frameworks, evolving from the NGC model to the more complex model, the Multi-Task Consensus Hypergraph. These approaches exploit interdependencies between various scene interpretations and sensor modalities, leading to more robust and accurate scene understanding. By finding consensus among multiple tasks, we demonstrated improved performance across all tasks simultaneously with less than 2% labeled data.

Real-world UAV Benchmarks – To support and validate our research, we introduced new datasets such as *Ruralscapes* for semantic scene segmentation, *odometry datasets* for unsupervised metric depth estimation and *Dronescapes* for multi-task learning from real UAV flights. These contributions provide valuable resources for the research community and establish new standards for evaluating aerial scene understanding methods.

Chapter 2

LOCAL-GLOBAL MULTI-STAGE CONSENSUS LEARNING FOR AERIAL SEGMENTATION

2.1 Local-Global Consensus Learning

We study the importance of visual context in object segmentation in the context of top-down view satellite images, which are taken under poor lighting conditions and contain low-resolution objects, many times occluded. This domain also offers specific scientific challenges to Computer Vision. The local appearance of objects in aerial images is often degraded due to occlusions, illumination, shadows, and distance, leading to poor resolution. In such cases, contextual cues provide semantic insights that improve object recognition. We propose a dual-stream approach using deep convolutional neural

*This section is based on the paper – **Alina Marcu**, and Marius Leordeanu. "Object contra context: Dual local-global semantic segmentation in aerial images." In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. 2017. [3]*

networks that combines the *local* appearance of the object with *global* information retrieved from a larger scene. Thus, the object is seen both as a separate entity from the perspective of its own appearance, but also as a part of a larger scene which acts as its complement and implicitly contains information about it.

We studied the role of local and global information separately (Figure 2.1). A model trained only with local patches (*L-Seg*) is capable of accurately identifying the shape of the object. However, it has a high false positive rate. On the other side, a model trained exclusively with global patches (*G-Seg*) misses out on important details but responds confidently in areas of high residential density and manages to remove the misclassifications of the previous model.

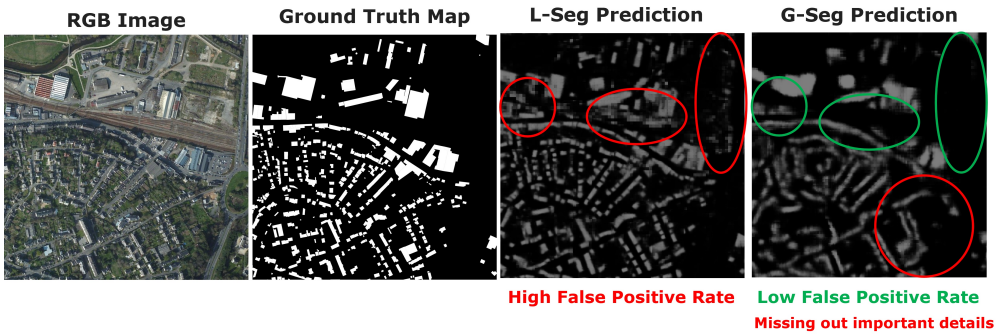


FIGURE 2.1: Models trained exclusively with local information identify object shapes but yield high false positives (*L-Seg*). Global models miss details yet excel in high-density areas, reducing misclassifications (*G-Seg*). Our approach leverages the complementary roles of local and global information.

Based on these observations we develop two novel dual-stream architectures, by first combining the L-Seg model and G-Seg model into a single local-global deep network, termed *LG-Seg* (Figure 2.2 (Left)). The two pathways process information in parallel, taking as input image patches of different sizes. The local stream focuses on capturing the fine-grained details of the object's appearance, while the global stream gathers a broader semantic context from the surrounding environment. Although complementary, both streams combined (through fully connected layers) reach consensus through an end-to-end learning procedure for improved segmentation performance. Unlike others, we intentionally chose two different types of networks with different image region sizes as input, to encourage learning different representations along the two pathways.

The LG-Seg model is a dual-stream combined architecture, an adapted VGG-Net [4] and an adapted AlexNet [5] joining into two last FC layers, which add up to $\approx$133M

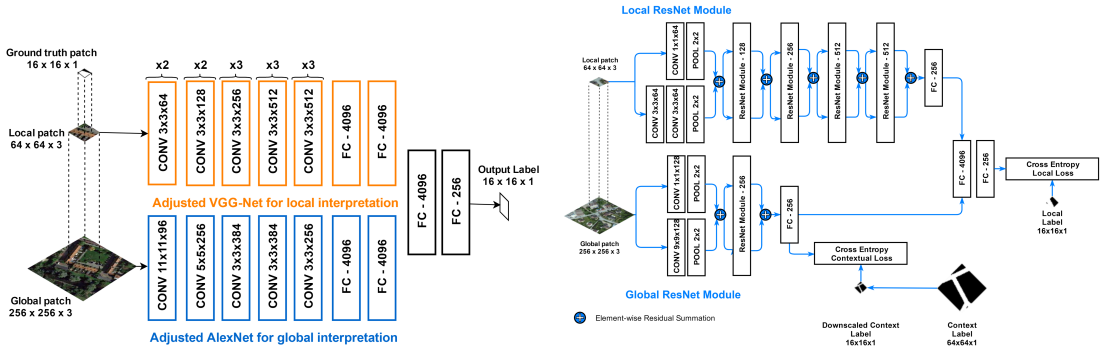


FIGURE 2.2: (**Left**) Our LG-Seg architecture (≈133M params) combines local and global information in a dual-stream CNN with different input region sizes, encouraging diverse representations along two pathways. (**Right**) Our LG-Seg-ResNet-IL uses residual blocks and an intermediary contextual loss on the global branch, enhancing contextual understanding with greater efficiency than LG-Seg and just ≈23M params.

parameters. Given the substantial computational demands of the LG-Seg model, we leverage residuals connections [6], since they are capable of bypassing extra levels of depth and thus simultaneously combining shallow and deep pathways into a unified multi-path neural network with filters of different sizes also acting in parallel, with a reduced number of parameters, termed *LG-Seg-ResNet*. On top of this architecture, we also introduce *LG-Seg-ResNet-IL* (Figure 2.2 (Right)) in which we added an extra intermediate loss for the global pathway to enforce learning of global context in the pathway receiving larger image patches. Thus we hope to improve the training time and also the quality of segmentation in places where context matters more.

TABLE 2.1: Quantitative evaluation of our models on different datasets. From left to right in order, we present results on the Massachusetts Buildings dataset and on our European Buildings and European Roads datasets.

Method	F-measure	Method	F-measure	Method	F-measure
Mnih et al. [7]	92.11	G-Seg	62.71		
Saito et al. [8]	92.30	L-Seg	82.66	LG-Seg	70.46
LG-Seg	94.23	LG-Seg	**84.20**	LG-Seg-ResNet	72.07
LG-Seg-ResNet-IL	**94.30**	LG-Seg-ResNet-IL	83.87	LG-Seg-ResNet-IL	**73.42**

Our work addresses the task of object segmentation, particularly, objects with regular structures, such as **buildings**, and those with variable and continuous shapes, such as **roads**. We also propose two new datasets, much larger than the most well-known dataset at that time, Massachusetts Buildings [7]. Our experimental analysis demonstrates the following. Firstly, we obtain state-of-the-art results on the public dataset with both proposed architectures (Table 2.1 (Left)). Secondly, the complexity of the proposed datasets is also reflected by the performance drop from 94% to 84% for building segmentation on the proposed European Buildings dataset (Table 2.1 (Middle)).

Lastly, we can also observe that in the case of objects with complex, continuous, and varied structures (results on proposed European Roads), context is truly important since the architecture with residual modules and explicitly modeled context obtains the best results for the road segmentation problem, with a bigger performance increase, of 3% compared to the initial LG-Seg model (Table 2.1 (Right)).

2.2 Multi-stage Multi-task Neural Networks

While our previous dual local-global networks demonstrated the importance of leveraging both local and contextual information for tasks such as building and road segmentation, they were limited to treating each task independently and required separate training for each dataset. To overcome these limitations we introduce a novel multi-stage multi-task (MSMT) neural network that tackles two crucial aerial tasks concurrently: semantic segmentation and vision-based geolocalization.

Our proposed architecture (Figure 2.3) employs a shared encoder branch (*UniEncoder*), followed by separate decoding branches (regression or segmentation) for each task (*LocDecoder*). This design choice is inspired by our previous work, which showed that processing different types of information through separate streams, especially when branches solve two different scopes, is beneficial. The multi-stage aspect of our network allows for progressive refinement of results. From a structural point of view, our architecture uses encoder-decoder modules (similar to U-Net [10] but using dilated convolutions of progressively increasing rates at the central bottleneck level) at each stage, having the same encoder structure re-used.

Our MSMT-Stage-1 model sets a new state-of-the-art benchmark on the Massachusetts Buildings dataset, significantly outperforming existing methods (Table 2.2 (Left) measures F-measure with a relaxed factor of 3 pixels [11]). Notably, our approach achieves

*This section is based on the paper – **Alina Marcu**, Dragos Costea, Emil Slusanschi, and Marius Leordeanu. "A multi-stage multi-task neural network for aerial scene interpretation and geolocalization." arXiv preprint arXiv:1804.01322 (2018). [9]*

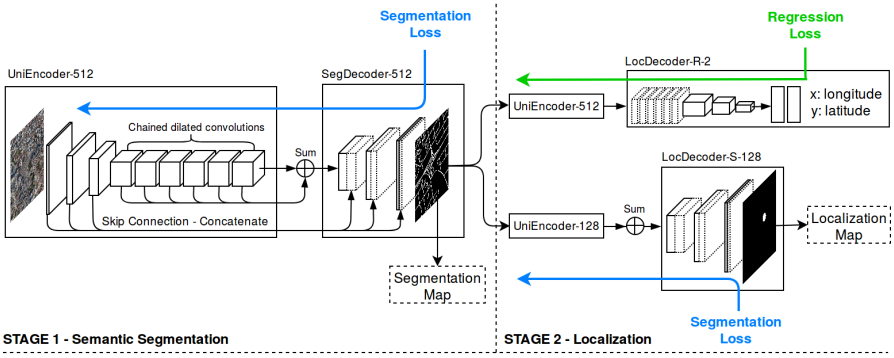


FIGURE 2.3: Our proposed multi-stage multi-task (MSMT) architecture for semantic segmentation and geolocalization in aerial images.

these results using only a single RGB image and class label, in contrast to methods like [12] that employ network ensembles. The same can be observed for the Inria dataset [13] on the testing dataset (Table 2.2 (Right) measuring IoU), surpassing other competitors (results verified on the official leaderboard).

TABLE 2.2: (**Left**) Comparison to state-of-the-art buildings segmentation methods on the Massachusetts Buildings dataset [7]. (**Right**) Comparison to state-of-the-art methods for the task of building segmentation from aerial images on the Inria dataset [13].

Method		F-measure
Deeplab	[14]	89.7
Mnih et al.	[7]	91.5
Saito et al.	[8]	92.3
U-Net	[10]	94.1
Saito et al.	[12]	94.3
Marcu et al.	[3]	94.3
Hamaguchi et al.	[15]	94.3
MSMT-Stage-1 (Ours)		**96.04**

Method		Austin	Chicago	Kitsap Co.	West Tyrol	Vienna	Overall
MLP [13]	IoU	61.20	61.30	51.50	57.95	72.13	64.67
	Acc.	94.20	90.43	98.92	96.66	91.87	94.42
Mask R-CNN [16]	IoU	65.63	48.07	54.38	70.84	64.40	59.53
	Acc.	94.09	85.56	97.32	**98.14**	87.40	92.49
SegNet MT-Loss[17]	IoU	**76.76**	67.06	**73.30**	66.91	76.68	73.00
	Acc.	93.21	**99.25**	97.84	91.71	**96.61**	95.73
MSMT-Stage-1 (Ours)	IoU	75.39	**67.93**	66.35	**74.07**	**77.12**	**73.31**
	Acc.	**95.99**	92.02	**99.24**	97.78	92.49	**96.06**

For the geolocalization task, we collected our own dataset from randomly sampled locations covering a city-wide area. As demonstrated by Costea et al. [18], roads serve as a unique footprint of an urban area, therefore we decided to use road segmentations (trained using MSMT-Stage-1). The second stage of our network takes the road segmentation output as input and learns to map it to a specific location using two branches. The regression branch outputs longitude and latitude coordinates, whereas the segmentation branch innovatively frames localization as a segmentation problem, generating a map with potential locations marked as dots. In our experiments, we proved that while the segmentation branch generally outperforms the regression one, it can sometimes produce multiple locations or fail to identify any (blank image). In such cases, we rely on the regression output. To increase localization precision as a final refinement step we use the Iterative Closest Point (ICP) algorithm [19], to align the road segmentation with the OpenStreetMap [20] roads from the predicted location. In Figure 2.4 we present,

from left to right, in order, the RGB input image, city-wide dot localization generated by MSMT-Stage-2-LocDecoder-S-128, predicted roads (green) from the RGB image on top of ground truth roads from the dot's location (white), before and after alignment.

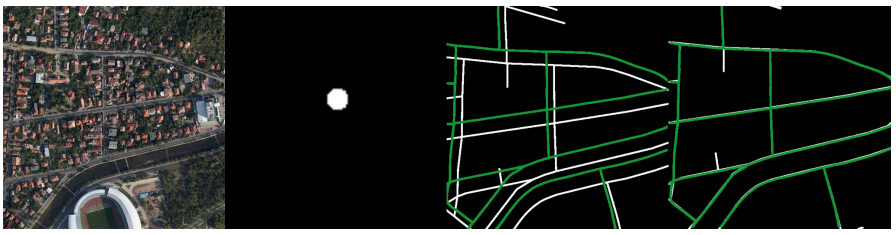


FIGURE 2.4: Qualitative results for geolocalization using MSMT-Stage-2-LocDecoder-S-128 branch, the road segmentation map using MSMT-Stage-1 network and the roads retrieved from OpenStreetMap, before and after alignment.

2.3 Multi-stage Ensemble of Deep Neural Networks

Building upon our Multi-Stage Multi-Task (MSMT) architecture detailed in the previous section, we address the challenging task of roadmap generation from satellite imagery. This section presents our top-performing approach in the DeepGlobe challenge [22], where we achieved a significant improvement of over 4% compared to the second-place contestant. An overview of our approach is presented in Figure 2.5.

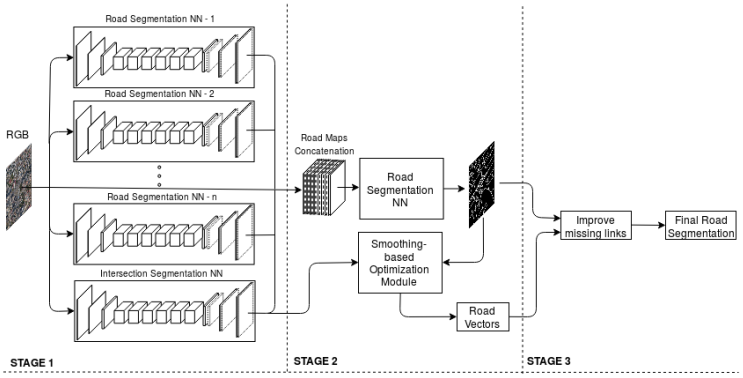


FIGURE 2.5: Overview of our three-stage method for improved roadmap generation.

Our first stage builds upon the MSMT-Stage-1 architecture (the same U-Net-like architecture with dilated convolutions at the central level) used to create an ensemble of

*This section is based on the paper – **Alina Marcu***, Dragos Costea*, Emil Slusanschi, and Marius Leordeanu. "Roadmap generation using a multi-stage ensemble of deep neural networks with smoothing-based optimization." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 220-224. 2018. [21]*

multiple such networks, each employing different dilation rates (and a different number of parameters). This approach enables the model to learn diverse aspects of the scene, enhancing the binary segmentation of roads. The varied dilation rates allow the network to capture features at multiple scales, which is crucial for accurately identifying roads of different widths and in various contexts. We train the following types of architectures (with progressive dilation rates in parentheses) for road segmentation and form two ensembles from them (Ensemble 1 - summation, Ensemble 2 - stage 2, learning): (1) **Max dilation 32** (1, 2, 4, 8, 16, 32), (2) **Max dilation 48** (1, 2, 4, 8, 16, 32, 48) and (3) **Max dilation 64** (1, 2, 4, 8, 16, 32, 48, 64). The second stage introduces a novel refinement process. We train a separate Max dilation 32 network to learn how to improve road segmentations in a strictly supervised manner. The key innovation here is the fusion of multiple road maps alongside the RGB image and also predicted intersection segmentation from a different trained network to improve road connections. This allows the model to learn a consensus among these various representations, resulting in more accurate and consistent road segmentation. This stage can be seen as an extension of the geolocalization refinement step in our MSMT architecture but is now applied specifically to road segmentation. The final stage incorporates previous work by Costea et al. [21]. This step applies an optimization procedure to the segmented roads, treating them as a graph. It aims to add missing links based on the inferred graph structure, further improving the overall segmentation quality. This stage complements our deep learning approach with classical computer vision techniques, potentially bridging gaps in the neural network's output.

We show consistent improvement on the DeepGlobe test set in both rounds of the competition using our architectures trained in semi-supervised iterations (with predictions on the testing data used in the second iteration) (Table 2.3 (Left)) with the best performing models being the learned ensemble as expected (Table 2.3 (Right)), surpassing the state-of-the-art on this benchmark by $\approx 4\%$.

TABLE 2.3: (**Left**) (Round 1) – Roads segmentation results on the official evaluation set of 1,243 images for which no ground truth data was provided during the competition to avoid cheating and overfitting. (**Right**) (Round 2) – Roads segmentation results on the official testing set of 1,101 images. Results were reported after adding the missing links. The results were provided by the submission site for both rounds. We report percentages.

Model	Iteration	IoU Evaluation
Max dilation 32	1	59.24
	2	59.75
Max dilation 48	1	60.39
	2	**60.58**

Model	IoU Evaluation
Baseline [22]	54.5
Ensemble 1	57.88
Ensemble 2	**58.62**

2.4 Lightweight CNNs for Bird's Eye View Scene Understanding

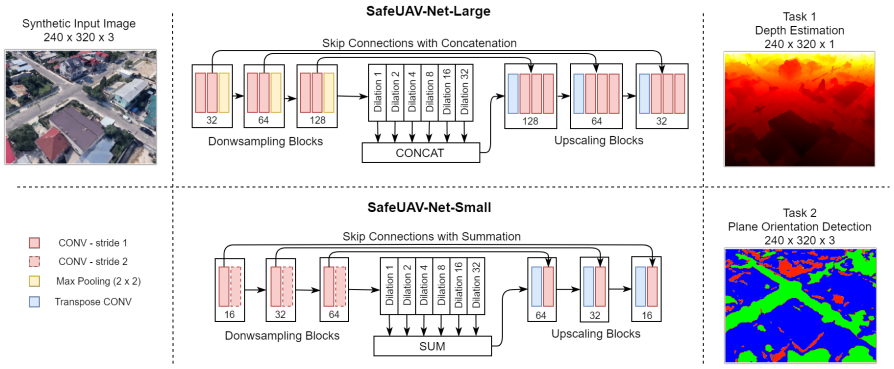


FIGURE 2.6: Our proposed SafeUAVNets for both onboard and offboard processing are trained for depth estimation and plane orientation prediction.

So far we have focused primarily on analyzing satellite imagery, which provides a top-down view of the Earth's surface. While satellites offer global coverage from very high altitudes, unmanned aerial vehicles (UAVs) have 6 degrees of freedom, capable of capturing both top-down views and perspective views at much lower altitudes compared to satellites. This flexibility allows UAVs to adjust their altitude for optimal resolution and coverage, collect data at varying scales, from broad overviews to detailed close-ups and provide real-time, on-demand imagery for critical applications. This section presents

*This section is based on the paper – **Alina Marcu**, Dragos Costea, Vlad Licaret, Mihai Pirvu, Emil Slusanschi, and Marius Leordeanu. "SafeUAV: Learning to estimate depth and safe landing areas for UAVs from synthetic data." In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0-0. 2018. [23]*

one of the first efforts in developing efficient solutions that use visual perspective cues for semantic and structural scene understanding from 3D reconstructions of real environments whilst also running onboard resource-constrained devices. We built upon our previous architectures and developed two new variants of the MSMT-Stage-1 (or Max dilation 32) neural network. The first, named *SafeUAVNet-Large* below, runs at $\approx$ 35 FPS on Nvidia's Jetson TX2. The second, named *SafeUAVNet-Small* below, is a simplified version of the first and runs at $\approx$ 130 FPS on an embedded device for an image resolution of $240 \times 320 \times 3$. Both architectures can be depicted in Figure 2.6.

We extract data from Google Earth [24] and reconstruct the scenes in 3D. We obtain rendered RGB images and dense depth maps, with pixel-level values in meters, from the perspective of a drone flying at low altitudes and at a 45-degree angle. The constructed dataset offers a high degree of realism. Based on the surface normals, a set of semantic annotations is constructed to delimit safe areas from unsafe ones, which can be further used in a drone's safe landing procedure, which we model and solve as a semantic segmentation task. The quantitative evaluation demonstrated the efficiency of the two architectures compared to well-known dense prediction architectures, both for the task of segmenting safe, unsafe, or oblique (other) areas in the scene (Table 2.4 (Left)), as well as for estimating depth in meters from an image (Table 2.4 (Right)).

We also examined the performance of models trained on synthetic data when applied to real-world scenarios. Despite visual similarities, we found significant performance differences between synthetic and real inputs. This highlights the need for training on real-world data to ensure model effectiveness in similar challenging scenarios.

TABLE 2.4: Quantitative evaluation of our SafeUAVNets for (**Left**) The task of safe landing detection, modeled as a segmentation task of horizontal, vertical and oblique planes within a scene compared to well-known segmentation NNs. Results are reported in percentages with bolded values being the best. (**Right**) The task of depth estimation from aerial images. Lower numbers are better.

Model	mAcc.	mPrec.	mRec.	mIoU
U-Net [10]	72.9	56.0	50.5	35.6
DeepLabv3+ [25]	84.0	75.3	73.9	59.7
SafeUAVNet-Small	82.3	72.8	69.3	55.1
SafeUAVNet-Large	**84.6**	**76.1**	**74.8**	**60.7**

Model	RMSE	Meters
U-Net [10]	0.041	9.63
DeepLabv3+ [25]	0.034	8.49
SafeUAVNet-Small	0.031	7.22
SafeUAVNet-Large	**0.026**	**6.09**

Chapter 3

CONSENSUS-BASED UNSUPERVISED METRIC DEPTH ESTIMATION

This chapter addresses a critical challenge in autonomous aerial systems: accurate and efficient metric depth estimation from monocular imagery. On top of that, compared to the previous methods proposed in this thesis, we make a huge step towards efficiently leveraging the temporal cues in video and derive metric depth in a completely unsupervised manner. Traditional approaches to this problem have relied on either purely geometric methods, which can be precise but lack robustness, or learning-based approaches, which often struggle with generalization and metric scale. The work presented in this chapter tackles these challenges through a series of novel contributions that synergistically combine analytical methods with deep learning approaches. To address the scarcity of relevant real-world data, we introduce two significant datasets – a 20-minute continuous UAV flight dataset covering two European mountain town resorts and an extended 33-minute dataset encompassing a variety of scenes, including urban, rural, and the Danube Delta. These datasets provide crucial benchmarks for evaluating metric depth estimation methods in realistic scenarios.

3.1 Depth Distillation: Finding Consensus between Kinematics, Optical Flow and Deep Learning

We consider the scenario of a flight within a given perimeter, with a focus on unsupervised learning to estimate depth within the larger context of aerial scene understanding. We introduce a hybrid model that combines an analytical, vision-with-odometry approach with deep unsupervised learning techniques. While the analytical path is mathematically precise, it lacks robustness in the presence of noise and numerically fails in focus of expansion areas. On the other hand, the unsupervised data-driven approach is robust and smooth over time, but not as accurate and more importantly, not metric. We employ a "Teacher-Student" paradigm to distill the knowledge [27] of the "Teacher" into a more compact "Student". By forming an ensemble from the two pathways (Teacher) and distilling them into a single net (Student) we manage to improve both accuracy and lower the computational requirements.

This method distills the consensus between scene geometry, camera pose, kinematics, and RGB imagery into a single neural network, achieving both computational efficiency and high accuracy. Importantly, this approach is designed for deployment on embedded devices, making it suitable for real-world UAV applications, since it builds on top of the SafeUAVNets architectures used in the distillation process for depth estimation. We present the overview of our approach in Figure 3.1 (Left) in which we combine several complementary pathways for accurate metric depth estimation. Along one path, we estimate consistent non-metric depth in an unsupervised way (D_{Unsup}). Along a different path, we use odometry and optical flow to estimate exact, metric depth ($D_{OdoFlow}$). $D_{OdoFlow}$ is used to scale D_{Unsup} and make it metric. The two form an ensemble Teacher used to distill a Student model for metric depth estimation. Along a third path, depth is reconstructed with Structure-from-Motion software [28] (D_{SfM}), which is made metric

*This section is based on the paper – Mihai Pirvu, Victor Robu, Vlad Licaret, Dragos Costea, **Alina Marcu**, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu. "Depth distillation: unsupervised metric depth estimation for UAVs by finding consensus between kinematics, optical flow and deep learning." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3215-3223. 2021. [26]*

by aligning its predicted trajectory with the metric trajectory from GPS. D_{SfM}, computed offline, plays the role of reference metric or ground truth (even though is not perfect) and is used for evaluation only.

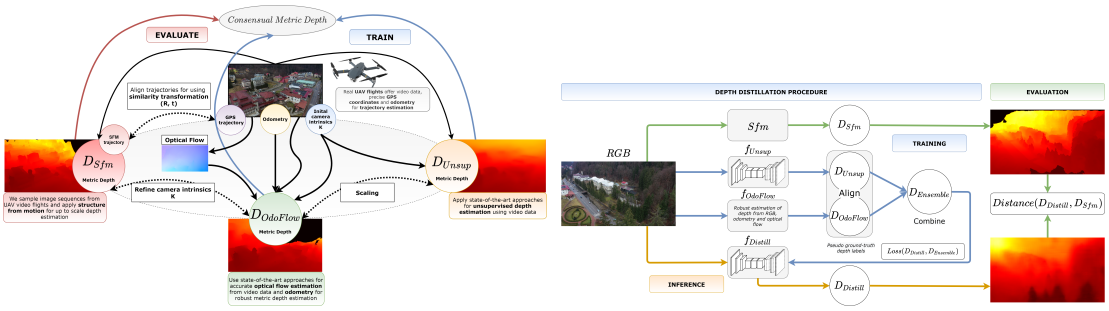


FIGURE 3.1: (**Left**) Overview of our approach, combining several complementary pathways for accurate metric depth estimation. (**Right**) Our depth distillation procedure for accurate metric depth estimation.

We introduce a novel dataset of two continuous UAV flights in two mountain town resorts, **Slanic** and **Herculane**. We divide Slanic in two non-overlapping subsets, one used for *training* the depth-distillation neural network and the other for *testing* purposes, measuring how well a UAV would perform in the same scene on a new flight. Herculane is used only for testing, to estimate the generalization capabilities of the proposed solutions, in different novel scenes, not seen during training.

In Figure 3.1 (Right) we showcase our depth distillation procedure for accurate metric depth estimation. We combine two label-free methods (unsupervised and analytical) into a single result and evaluate it against the SfM reconstruction. In our experimental analysis (Table 3.1), we proved that the distilled Student can significantly improve over the Teacher on the test video from the first scene while remaining competitive in the second scene, unseen during training.

TABLE 3.1: Absolute and relative errors on the areas where both $D_{OdoFlow}$ and D_{Unsup} are valid. We observe that these areas yield more stable overall results. Herculane has higher errors mainly due to the higher height of the dataset. The distilled students (SafeUAVNets) have the best results on the Slanic test set. Lower numbers are better and best are **bolded**.

	Slanic		**Herculane**	
	Metric (m)	Relative (%)	Metric (m)	Relative (%)
D_{Unsup}	21.06	15.31	31.61	16.60
$D_{OdoFlow}$	19.56	14.39	**24.97**	**12.72**
$D_{Ensemble}$	19.03	13.81	27.10	13.79
SafeUAVNet-Large	16.66	13.41	37.43	22.41
SafeUAVNet-Small	**16.11**	**12.90**	37.42	22.95

3.2 UFODepth: Unsupervised Learning with Flow-based Odometry Optimization for Metric Depth Estimation

Building on the depth distillation concept, we present *UFODepth*, an unsupervised learning framework for metric depth estimation that incorporates a novel flow-based odometry optimization procedure. This method significantly improves the accuracy of analytical depth estimation and demonstrates superior generalization capabilities across diverse scenes using video data and odometry from real-world UAV flights.

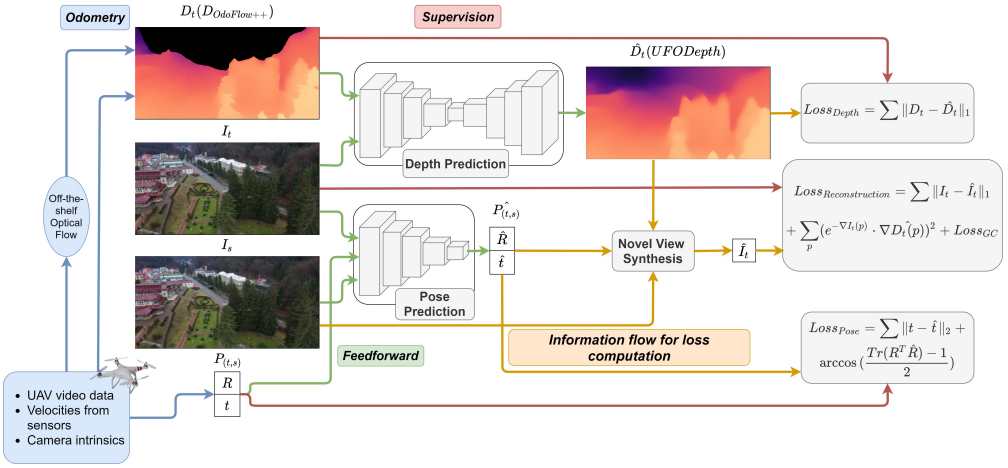


FIGURE 3.2: Overview of our approach (UFODepth). We combine three types of losses with an improved mathematical formulation for depth from optical flow, which results in a more robust depth estimation across multiple and varied scenes while being able to maintain real-time inference.

We correct the noisy odometric measurements by optimizing the alignment between the derotated optical flow and the projected linear speed in the image. Then, we detail an analytical depth estimation method based on optical flow and corrected camera velocities ($D_{OdoFlow++}$). Subsequently, the improved depth and camera velocities obtained analytically are used, as additional cost terms, for training our novel unsupervised learning architecture for metric depth estimation (*UFODepth* presented in Figure 3.2). We

This section is based on the paper – Vlad Licaret, Victor Robu*, **Alina Marcu***, Dragos Costea, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu."UFODepth: Unsupervised learning with flow-based odometry optimization for metric depth estimation." In 2022 International Conference on Robotics and Automation (ICRA), pp. 6526-6532. IEEE, 2022. [29]*

extensively experiment with a UAV dataset, collected from the real world, which we significantly extend by adding completely novel scenes. The dataset extension consists of approximately 33 minutes of additional flight time with odometry and GPS information. It covers a variety of scenes in Eastern Europe such as urban, rural and the Danube Delta (Figure 3.3).

FIGURE 3.3: Dataset samples from all video sequences in the extended dataset. Sample frames (top) along with estimated trajectories overlaid on top of SfM reconstruction (bottom). We show samples from previous work Slanic, Herculane [26], alongside the newly introduced scenes - Oveselu, Olanesti, Chilia.

We outperform by significant margins different kinds of state-of-the-art approaches, ranging from analytical and unsupervised solutions to transformer-based architectures that require heavy computation and pre-training (Table 3.2). Even though small details are lost, our method provides a globally consistent and accurate depth. One can assume that the small receptive field (for DPT) and several visually pleasing features of BMD (such as gradient-based scaling and blending for each patch) yield locally accurate structure and foreground/background separation. However, they are not as suitable for practical depth estimation tasks such as obstacle avoidance due to their intensive computational requirements. Our resulting algorithm could be deployed on embedded devices, being a good candidate for practical robotics use cases, such as obstacle avoidance and safe landing for UAVs.

TABLE 3.2: Mean absolute and relative errors on entire map against D_{SfM} ground truth depth. Metric is represented in meters (m) and Relative as a percentage (%). Since we report the errors, both metrics mean that lower numbers are better. Methods that do not provide metric estimations are scaled towards $D_{OdoFlow++}$ for a fair comparison. Overall denotes the average over all scenes.

Method	Slanic		Chilia		Olanesti		Herculane		Oveselu		Overall	
	Metric	Relative	Metric	Relative	Metric	Relative	Metric	Relative	Metric	Relative	Metric	Relative
D_{Unsup} [30]	25.00	15.4	44.52	23.4	25.85	18.4	34.40	16.0	31.10	22.3	32.17	19.1
$D_{Ensemble}$ [26]	24.83	14.9	37.93	18.4	22.46	**15.2**	34.28	16.6	33.15	22.1	30.53	17.44
SafeUAVNet-Small [26]	26.34	16.7	46.11	23.7	26.76	19.5	41.30	19.8	32.76	23.8	34.65	20.7
DPT [31]	34.33	22.8	**23.87**	**13.9**	26.36	20.1	30.48	14.8	28.57	26.8	28.72	19.7
BMD [32]	42.09	33.1	46.83	36.5	33.75	31.4	80.44	44.1	38.83	41.4	48.4	37.3
PackNet [33]	34.36	21.4	43.82	25.4	31.34	22.9	42.64	20.1	33.41	25.2	37.11	23.0
UFODepth-RGB (ours)	**21.52**	**14.4**	49.90	27.6	25.28	18.5	32.52	16.2	30.80	23.0	32.0	19.9
UFODepth (ours)	22.36	14.9	33.56	17.0	**21.98**	15.4	**26.45**	**13.0**	**26.73**	**19.4**	**26.21**	**15.9**

Chapter 4

TEMPORAL CONSENSUS FOR SEMANTIC SCENE SEGMENTATION

In the context of video semantic scene segmentation, it is impractical to manually label each frame independently, especially considering that there are relatively few changes from one frame to another in modern captured videos that have a high frame rate. Therefore, the ability to perform automatic annotation of the entire scene would be extremely valuable. Towards this direction, we introduce *SegProp*, a novel iterative flow-based method for video semantic segmentation - densely annotating each pixel of every frame. SegProp leverages sparsely annotated frames, alongside a large range of motion through optical flow chains and propagates semantic labels to intermediary frames lacking annotations. This method is grounded in spectral clustering, exploiting spatial and temporal coherence to enhance label propagation. By doing so, SegProp significantly reduces the annotation burden in video without sacrificing the quality of the generated labels.

*This chapter is based on the paper – **Alina Marcu**, Vlad Licaret, Dragos Costea, and Marius Leordeanu. "Semantics through time: Semi-supervised segmentation of aerial videos with iterative label propagation." In Proceedings of the Asian Conference on Computer Vision. 2020. [34]*

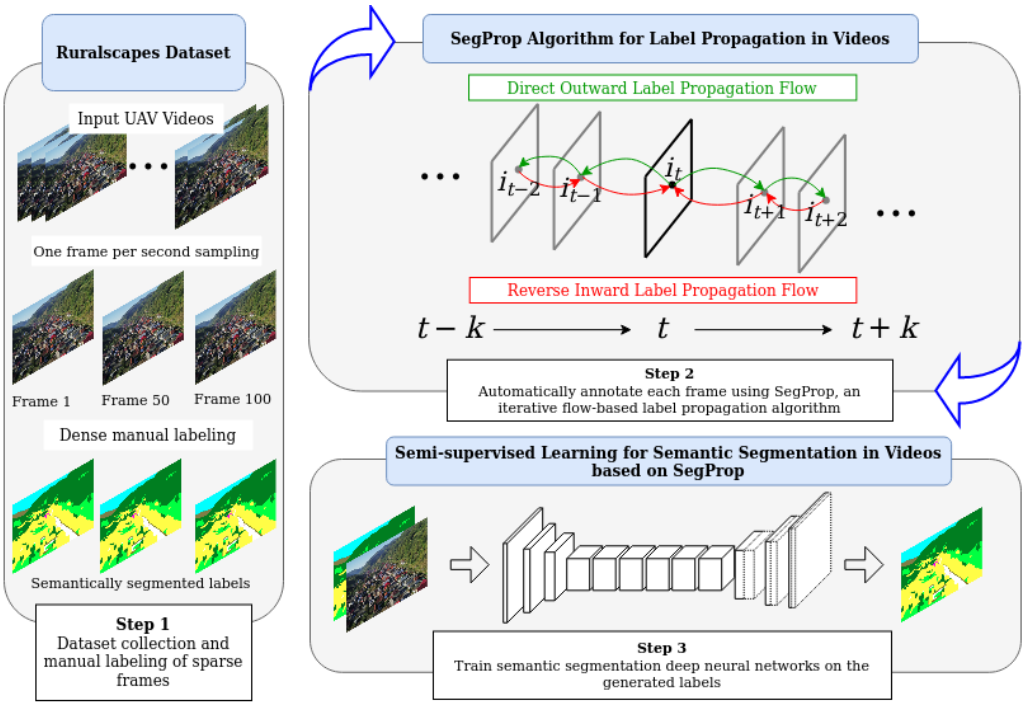


FIGURE 4.1: **SegProp** – our novel method for automatic propagation of semantic labels in the context of semi-supervised segmentation in aerial videos. Each step depicts the contributions we make in terms of (1) the Ruralscapes dataset, (2) SegProp label propagation method and (3) show efficiency in semi-supervised learning scenarios.

One of the motivations behind developing SegProp is the scarcity of comprehensive video aerial datasets. Existing datasets often lack the resolution and dense annotations necessary for training robust segmentation models. To fill this gap, we introduce *Ruralscapes*, a new dataset comprising high-resolution (4K) images with manually annotated dense labels every 50 frames, being the largest publicly available dataset for the task of video scene segmentation from aerial flights at the time of publication.

We present our main contributions in Figure 4.1. In **Step 1** we sample the UAV videos, at regular intervals (e.g. one frame, every second). The resulting frames are then manually labeled (*Ruralscapes*). In **Step 2** we automatically propagate labels to the remaining unlabeled frames using our *SegProp* algorithm - based on class voting, at the pixel level, according to inward and outward label propagation flows between the current frame and an annotated frame. The propagation flows could be based on optical flow (default), homography transformation, or another propagation method, as shown in our experiments (Table 4.1). SegProp propagates iteratively the segmentation class voting until convergence, improving performance over iterations. In **Step 3** we mix all the generated annotations with the ground truth manual labels to train state-of-the-art

deep CNNs for semantic segmentation and significantly improve performance in unseen videos (generalization to novel scenarios).

TABLE 4.1: Automatic label propagation results. We present Zhu et al. [35] (which has only 1 iteration) vs. SegProp starting either from [35] or the initial anchor frames from our algorithm with consistent improvements over multiple iterations.

Methods \ Iterations		1	2	3	4	5	6	7	+ Filt.
Zhu et al. [35]	mF1	.846	-	-	-	-	-	-	-
	mIOU	.747	-	-	-	-	-	-	-
SegProp from [35]	mF1	.846	.874	.877	.885	.888	.891	.893	.896
	mIOU	.747	.785	.790	.801	.805	.810	.813	.818
SegProp	mF1	.884	.894	.896	**.897**	**.897**	**.897**	**.897**	**.903**
	mIOU	.801	.817	.819	**.821**	**.821**	**.821**	**.821**	**.829**

SegProp demonstrates remarkable performance in automatically annotating the remaining 98% of frames in the Ruralscapes dataset, achieving an accuracy exceeding 90% in F-measure. This accuracy significantly surpasses that of existing state-of-the-art label propagation methods (Zhu et al. [35]). Moreover, SegProp's modularity allows for the integration of other methods within its iterative label propagation loop, resulting in a further boost in performance over starting or baseline labels (Table 4.1).

In addition to its label propagation capabilities, SegProp is tested in a semi-supervised learning setting (Table 4.2). Here, we train several well-known semantic segmentation architectures on the frames automatically labeled by SegProp and evaluate their performance on novel, unseen videos. The results consistently show a substantial improvement over models trained in a purely supervised manner. This highlights SegProp's potential to enhance the efficiency of training CNNs for semantic segmentation with limited annotated data.

TABLE 4.2: Quantitative results after training the neural networks on the generated labels, including the labeled ones. We report mean F-measure over all videos from the testing set, for each class: (1) - land, (2) - forest, (3) - residential, (4) - haystack, (5) - road, (6) - church, (7) - car, (8) - water, (9) - sky, (10) - hill, (11) - person, (12) - fence and the average over all classes. The best results for each class and each trained model, are bolded. Results clearly show a significant performance boost over the baseline, when training with SegProp (SP). We mark with ✗the purely supervised case and with ✓the one in which we augment the labels with SP and train the NN with the mix.

Methods	SP	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	All
Unet	✗	.681	.497	.834	.000	.000	.000	.000	.000	**.967**	.000	.000	.000	.248
[10]	✓	**.757**	**.544**	**.838**	.000	**.556**	**.672**	.000	.000	.900	**.454**	.000	.000	**.393**
DeepLab	✗	.500	.416	.745	.000	.220	.073	.000	.000	.909	.242	.000	.000	.259
v3+ [25]	✓	**.570**	**.452**	**.776**	**.022**	**.369**	**.122**	**.007**	.000	**.926**	**.272**	**.004**	**.043**	**.297**
SafeUAV	✗	.713	.475	.757	.000	.371	.640	.000	.000	.953	**.260**	.000	.003	.348
Net [23]	✓	**.783**	**.488**	**.836**	**.364**	**.552**	**.748**	**.031**	**.428**	**.973**	.176	**.481**	**.610**	**.515**

Chapter 5

MULTI-TASK CONSENSUS GRAPHS AND HYPERGRAPHS FOR AERIAL SCENE UNDERSTANDING

Our efforts so far have focused on improving one task at a time, but our ultimate goal is to solve multiple simultaneously. Our motivation stems from the understanding that the real world is not limited to being viewed through an RGB lens; instead, we have various methods to interpret what we see, and with multiple sensors available, it makes sense to leverage them to enhance our perception and understanding. Our world is interconnected, with some tasks closely linked, and we aim to utilize similar information from different sources. This led us to the following key questions: How can we efficiently leverage these interconnections? How can we use multiple visual cues to predict other more complex ones that cannot be easily derived? Can we develop a framework for integrating various data sources and representations, to create a holistic understanding of the scene, without the need for additional human effort? In this chapter, we present our efforts to address these open questions.

5.1 Multi-task Scene Understanding by Neural Graph Consensus

We address the challenging problem of learning multiple visual interpretations of the world by finding consensus in a graph of neural networks. As our answer to the aforementioned open questions, we introduce Neural Graph Consensus (NGC) which integrates various interpretations of dynamic scenes into a unified neural graph, encompassing 3D structure, pose, motion, and semantic segmentation of objects and activities across space and time. Each graph node is a scene interpretation layer, while each edge is a deep network that transforms one layer at one node into another from a different node. Within this graph, multiple pathways converging on a node act as collective teachers through consensual agreements, guiding individual edge networks to the same node. This self-supervised training approach proves to be effective even in the context of unsupervised learning, with unlabeled data.

We employ an initialization procedure for the graph by training each edge independently in a supervised manner. Afterwards, the edges are trained using pseudo-ground truth generated from the consensus among multiple paths that reach a particular node. We employ the well-known "Teacher-Student" paradigm for continuous learning. The paths directed at a node function as an ensemble "Teacher" for each edge, providing high-confidence supervisory signals when there is strong consensus. The process is repeated over several iterations, in which each edge becomes a "Student" and also part of a different ensemble "Teacher" for training other students. By optimizing the consensus among different paths, the graph achieves consistency and robustness across various interpretations and iterations, even in the absence of labeled data.

Experimental analysis is conducted on a synthetic, but realistic, dataset that closely replicates real UAV flights extracted from the CARLA simulator [37] (Figure 5.1). This

*This section is based on the paper – Marius Leordeanu, Mihai Pirvu, Dragos Costea, **Alina Marcu**, Emil Slusanschi, and Rahul Sukthankar. "Semi-supervised learning for multi-task scene understanding by neural graph consensus." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, pp. 1882-1892. 2021. [36]*

simulator produces high-quality, multi-representational data, enabling comprehensive validation of our method. The paths simulate a drone trajectory, with small random variations in all angles. While the training path is a grid-based, traditional surveying flight, the test path aims to capture as many viewpoints as possible, to increase complexity.

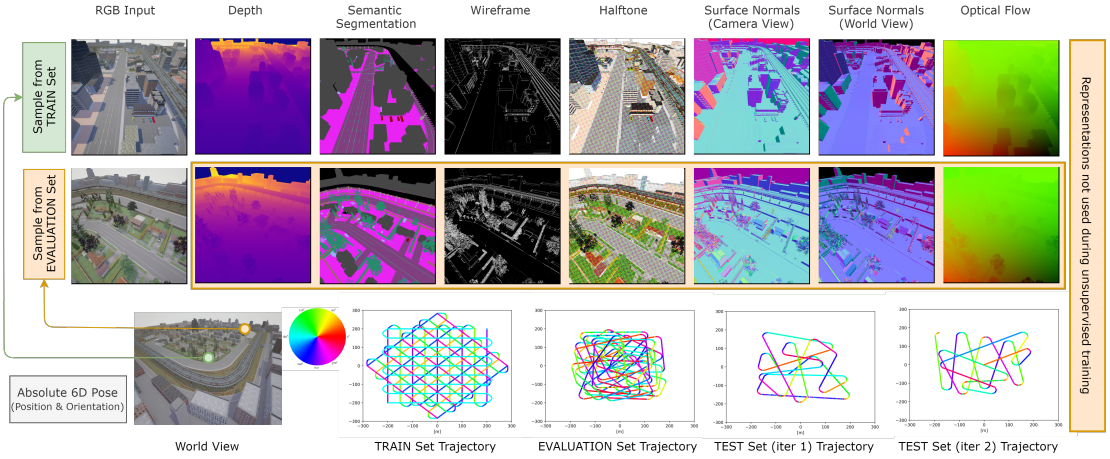


FIGURE 5.1: Samples from the synthetic dataset (train, first row, evaluation, second row) we collected using the CARLA virtual environment.

Our experimental analysis (Table 5.1), based on a pre-configured graph structure, which means we applied a sub-optimal approach to determine the best set of edges for each task, demonstrated the effectiveness of the method in improving results for six complex tasks over two learning iterations, not just at the NGC mean ensemble level, but also at the level of a single edge with $\approx$ 1.1M parameters [23].

TABLE 5.1: Quantitative results for our proposed NGC on 6 representations, over 2 iterations of unsupervised learning. We show the best results over NGC ensemble teachers (**bolded**) and single edge students (**bolded**). Note the consistent iterative improvements.

Representation	Evaluation Metric	Iteration 0	Iteration 1		Iteration 2	
		EdgeNet	NGC	Distil. EdgeNet	NGC	Distil. EdgeNet
Depth	L1 ↓ (meters)	4.98	3.48	4.28	**3.29**	**3.95**
	Pixels ↑ (%)	-	79.30	60.66	**79.69**	**61.90**
Surface Normals (C)	L1 ↓ (degrees)	8.48	7.79	8.28	**7.45**	**7.67**
	Pixels ↑ (%)	-	74.18	53.59	**74.61**	**53.94**
Surface Normals (W)	L1 ↓ (degrees)	11.88	8.82	10.75	**8.52**	**8.67**
	Pixels ↑ (%)	-	79.95	57.88	**81.12**	**61.14**
Semantic Segmentation	Accuracy ↑ (%)	90.01	91.81	90.19	**92.45**	**92.83**
	mIOU ↑ (%)	48.40	49.78	49.80	**52.58**	**51.59**
	Pixels ↑ (%)	-	79.46	69.62	**81.49**	**71.95**
Wireframe	Accuracy ↑ (%)	96.17	96.55	96.54	**96.61**	**96.55**
	Pixels ↑ (%)	-	77.71	72.57	**78.02**	**73.46**
Position	L2 ↓ (meters)	25.75	15.53	20.02	**12.07**	**15.55**
Orientation	L1 ↓ (degrees)	3.84	2.50	3.39	**2.20**	**3.00**

5.2 Multi-task Hypergraphs for Aerial Understanding

Building on the NGC concept, we introduce a novel hypergraph structure for multi-task learning and show its effectiveness in more challenging scenarios with even less human supervision than before. We apply our model in two distinct domains (Figure 5.2): 1) complex real-world scenes captured in the Dronescapes dataset, which we introduce, a large real-world video collection recorded with UAVs, and 2) the NASA NEO dataset [40], a comprehensive Earth Observation dataset spanning 22 years. The Dronescapes dataset, with its multiple representations, is ideal for multi-task learning, while the NASA NEO dataset presents challenges such as missing data and temporal distribution shifts.

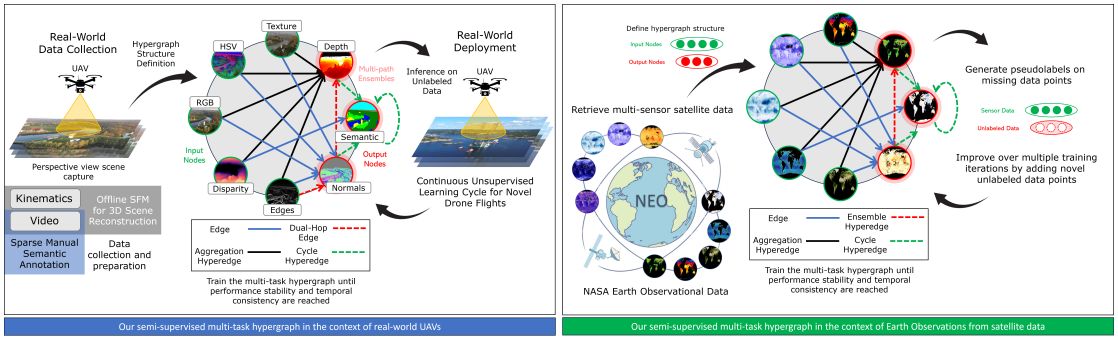


FIGURE 5.2: Overview of our semi-supervised multi-task hypergraph for multi-domain applications using real-world UAVs flights or Earth Observations.

This model goes beyond the pairwise relationships in NGC, incorporating higher-order connections, through multiple types of hyperedges (Figure 5.3) that capture more complex interdependencies. Similar to NGC, in our hypergraph, each node is an interpretation layer of the scene. The basic processing units of the hypergraph are **direct neural links (DNL)** which represent the RGB→Task edge. The neural links that connect an input node to an output node are simple edges (E), while the others, that connect multiple nodes to an output node form complex hyperedges. Each hyperedge is modeled

*This section is based on the paper – **Alina Marcu**, Mihai Pirvu, Dragos Costea, Emanuela Haller, Emil Slusanschi, Ahmed Nabil Belbachir, Rahul Sukthankar, and Marius Leordeanu. "Self-supervised hypergraphs for learning multiple world interpretations." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 983-992. 2023. [38] and a small part on*
*Mihai Pirvu, **Alina Marcu**, Maria Alexandra Dobrescu, Ahmed Nabil Belbachir, and Marius Leordeanu. "Multi-Task Hypergraphs for Semi-supervised Learning using Earth Observations." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3404-3414. 2023. [39]*

using a lightweight neural network [23] (SafeUAVNet-Small). Previous works use only edges, while we introduce different types of hyperedges to capture more complex relations between different layers. Given input nodes: A, B, C, and outputs D and E, we form four types of hyperedges of different complexities: pairwise edges (E), dual-hop edges (DH-E), ensemble hyperedges (EH), aggregation hyperedges (AH) and cycle hyperedges (CH). We have demonstrated on multiple tasks that these hyperedges far exceed the performance of simple edges (Figure 5.4 (Left)).

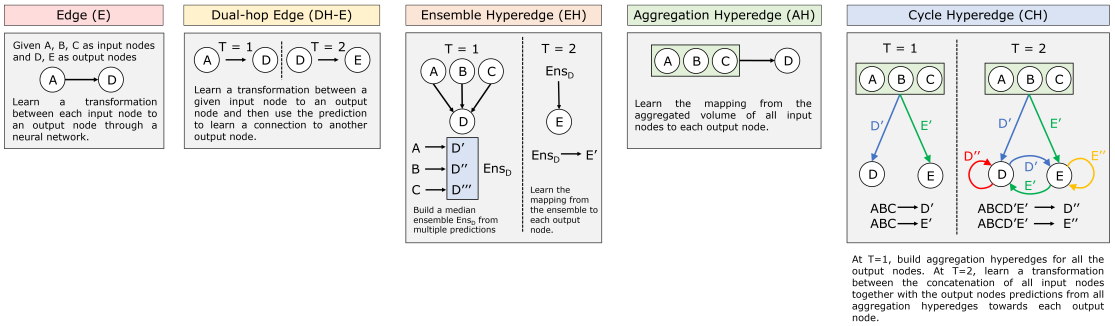


FIGURE 5.3: Types of edges and hyperedges in the hypergraph.

	Type	Train Unlabeled (iter 2)			Train Unlabeled (iter 3)		
		(1)	(2)	(3)	(1)	(2)	(3)
Edges	E: rgb	42.85	5.04	10.37	32.79	21.66	12.40
	E: hsv	41.70	4.69	10.54	33.51	19.90	12.48
	E: softedges	32.47	6.26	11.56	27.28	18.61	13.53
	E: softseg	30.71	5.97	11.14	24.68	22.70	12.76
	E: ufo	20.77	7.19	11.69	16.93	17.55	12.89
	DH-E: sseg	-	6.25	11.39	-	19.00	12.93
	DH-E: depth	29.24	-	12.22	24.11	-	13.79
	DH-E: norm	30.56	6.17	-	26.35	21.15	-
	mean	32.61	5.94	11.27	26.52	**20.08**	12.97
Hyperedges	AH	41.80	5.33	10.37	33.63	23.96	12.24
	AH-ufo	41.96	5.16	10.78	33.82	21.10	12.72
	CH	44.63	4.93	10.32	36.92	20.36	12.23
	mean	**42.80**	**5.14**	**10.49**	**34.79**	21.81	**12.40**

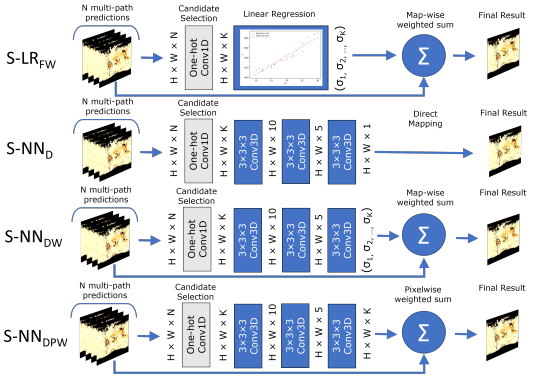


FIGURE 5.4: (**Left**) Evaluation of edges and hyperedges, on the Dronescapes dataset (Figure 5.5 (Left)), for multiple tasks: 1 - semantic segmentation (sseg); 2 - depth estimation (depth); 3 - surface normals (norm). We report mean IoU (% - higher values are better (↑) for the task of semantic segmentation and L1 error * 100 (lower is better (↓)) for depth and normals estimation. Bolded results highlight the mean performance gain of training hyperedges over edges. The meaning of the training sets are in Figure 5.5 (Right). (**Right**) Proposed learned ensembles architectures.

Multiple paths can reach the same node to form ensembles from which we obtain robust pseudolabels and leverage the power of semi-supervised learning in the hypergraph over multiple iterations by adding novel unlabeled data. Different from previous ensemble formation methods (aggregation by mean, median or distance-based) we introduce four types of ensembles, all with an initial learnable candidate selection model, which keeps only the relevant candidates before combining them: S-LR$_{FW}$ learns one fixed weight

per candidate. S-NN$_{DW}$ dynamically outputs a weight per candidate, depending on the input, while S-NN$_{DPW}$ dynamically outputs a weight for each pixel of each candidate. Instead of linearly combining the candidates S-NN$_D$ learns a direct non-linear mapping from candidates to output (Figure 5.4 (Right)). All are learned end-to-end.

TABLE 5.2: Comparison to previous multi-task graph-based methods – (**Left**) On the Dronescapes dataset, we show considerable improvements by adding the proposed hyper-edges (denoted with HE in the table) on top of existing work that uses only edges within their graph structure. (**Right**) On the NEO dataset, we evaluate on the Test Set of different ensemble models, for each output node (1) - *AOD*, (2) - *CM*, (3) - *FIRE*, (4) - *LAI*, (5) - $LSTD_{AN}$, (6) - $LSTN_{AN}$, (7) - *WV*. The best numbers are bolded, while the second best are underlined.

Method	IoU(↑)			
	Barsana	Comana	Norway	Mean
NGC [36] (Mean)	41.53	40.75	27.38	36.55
NGC (Mean) + HE	42.61	42.17	27.96	37.58
NGC [36] (Median)	39.25	37.41	27.01	34.56
NGC (Median) + HE	44.34	38.99	22.63	35.32
CShift [41] (Mean)	43.91	42.13	29.68	38.57
CShift (Mean) + HE	44.71	43.88	30.09	39.56
CShift [41] (Median)	43.30	40.62	29.51	37.81
CShift (Median) + HE	46.27	43.67	29.09	39.68
S-LR$_{FW}$	46.51	**45.59**	**30.17**	**40.76**
S-NN$_D$	45.53	42.92	28.37	38.94
S-NN$_{DW}$	45.48	43.25	26.36	38.36
S-NN$_{DPW}$	**48.21**	44.85	28.94	40.67

Ens. Type	(1)	(2)	(3)	(4)	(5)	(6)	(7)	ARPI (↑)
NGC [36]	0.44	12.97	12.15	5.44	-50.2	-39.0	2.21	-8.01
CShift [41]	1.86	13.07	8.66	6.64	-43.5	-28.0	2.65	-5.52
S-Mean	3.31	<u>14.71</u>	**12.42**	8.97	-9.50	-2.48	5.93	4.77
S-LR$_{FW}$	4.85	11.15	8.62	8.45	0.21	<u>4.90</u>	5.84	6.29
S-NN$_D$	**6.67**	13.20	10.56	**9.68**	0.69	3.57	**7.86**	<u>7.46</u>
S-NN$_{DW}$	4.79	**14.80**	<u>12.39</u>	9.50	0.25	**4.91**	6.03	**7.52**
S-NN$_{DPW}$	5.74	6.51	11.21	8.26	<u>1.33</u>	5.62	4.98	6.24

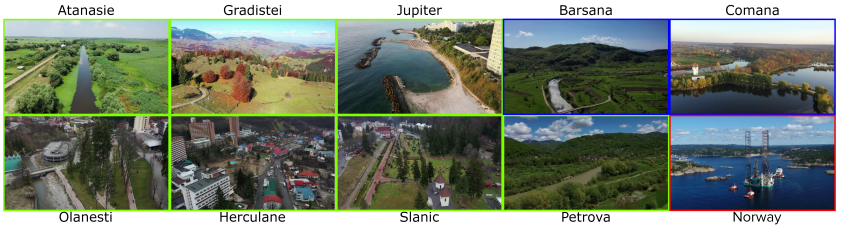
We evaluate the performance of our learned ensembles on our proposed Dronescapes dataset, for the task of semantic segmentation (Table 5.2 (Left)) and on all the output nodes from the NEO dataset (Table 5.2 (Right)). On Dronescapes, we bring a performance boost by adding hyperedges and also by allowing the ensembles to learn. Both NGC [36] and CShift [41] models use only edges and relatively simple non-parametric ensemble models at nodes (NGC - simple average and CShift - non-parametric pixel-wise kernel weighted average). Our experiments show that learning parametric ensemble models, even a simple linear one, improves significantly (above 2% on average) over previously published work. Performance is reported on the Train Unlabeled (iter 3), which includes only the test scenes. We show a similar improvement pattern on the NEO dataset where all of the learned ensembles with selection offer a positive relative improvement compared to previous methods without selection (NGC and CShift).

TABLE 5.3: Iterative learning performance on the Dronescapes dataset, at the level of the direct neural links for each task. The evaluation was done on the test scenes (averaged).

	Type	Semantic		Depth		Normals	
		IoU (↑)	Cons. (↑)	L1 (↓)	Cons. (↑)	L1 (↓)	Cons. (↑)
DNL	supervised	25.04	88.85	-	-	-	-
	(iteration 1)	32.79	94.04	21.66	5.89	12.40	98.32
	(iteration 2)	37.26	95.72	17.34	7.06	11.93	98.87
	(iteration 3)	**40.31**	**98.13**	**16.64**	**30.26**	**11.71**	**99.30**

We also show that leveraging pseudolabels generated from ensemble predictions over multiple learning iterations continuously improves model performance, even at the simplest edge (DNL) (Table 5.3) with improvements not only in terms of accuracy over multiple tasks but also at the level of temporal consistency.

We introduce *Dronescapes*, a new, large-scale UAV video dataset with diverse scenes and automatically generated annotations for multiple tasks (Figure 5.5 (Left)). All video sequences include GPS information, linear and angular velocities, and absolute camera angles (except for the out-of-distribution scene of Norway). The total video length is about 50 minutes, with 4K frames and odometry provided at 10 Hz. We collect a total of 10 widely varied scenes that we split into 7 training and 3 test scenes (dataset split details in Figure 5.5 (Right)). There is a large variation in spatial distributions of classes among the different Dronescapes scenes, which range from rural (Atanasie, Gradistei, Petrova, Barsana, Comana), to urban (Olanesti, Herculane, Slanic) and seaside (Jupiter, Norway), while also being geographically far apart. This dataset provides a challenging testbed and to the best of our knowledge, is the first one for evaluating multi-task learning approaches in real-world scenarios with real UAV flights.



Scene Name	Train Labeled	Train Unlabeled (iter 1)	Train Unlabeled (iter 2)	Train Unlabeled (iter 3)
Atanasie	76	4501 (76)	4500 (75)	-
Gradistei	18	726 (18)	484 (12)	-
Herculane	12	484 (12)	363 (9)	-
Jupiter	21	847 (21)	605 (15)	-
Olanesti	18	726 (18)	484 (12)	-
Petrova	12	484 (12)	363 (9)	-
Slanic	76	4501 (76)	4500 (75)	-
Barsana	-	-	-	1452 (36)
Comana	-	-	-	1210 (30)
Norway	-	-	-	2941 (50)
TOTAL	233	12269 (233)	11299 (207)	5603 (116)

FIGURE 5.5: (**Left**) Sample frames from each of the 10 scenes from our Dronescapes dataset. The scenes framed with **green** borders represent training scenes for which we have access to a small fraction of manual annotations during training. The others depict unseen, test scenes with semantic distributions that are closer to the training set (in **blue**) or out-of-distribution (**red**). (**Right**) Dronescapes dataset split.

These contributions collectively address the challenges of learning comprehensive scene representations from aerial perspectives with minimal supervision. By leveraging the power of hypergraphs, ensemble learning, and iterative semi-supervised training, this work pushes the boundaries of what is possible in multi-task aerial scene understanding. The methods developed here have broad implications for various applications, from autonomous UAV navigation to long-term environmental monitoring and climate change studies.

Chapter 6

CONCLUSIONS

This thesis addresses key challenges in aerial scene understanding, offering both practical solutions and theoretical contributions applicable to various robotics fields, specifically from the aerial perspective. We acknowledge that the complexities of aerial scene understanding are far from fully solved and propose several directions for future research. These include expanding from 2D to 3D world representation, leveraging synthetic data and domain adaptation techniques, integrating a wider range of sensor modalities, improving semantic understanding through open-set learning, advancing cross-domain adaptation and generalization, exploring adaptive and continual learning, and exploiting advanced architectures such as transformer-based models.

We also address ethical considerations, particularly regarding safety and privacy in UAV data collection. We adhered to drone flight regulations in Romania, prioritizing safety by conducting flights in sparsely populated areas and at higher altitudes. Privacy concerns were mitigated by obtaining consent where possible and informing individuals about the research purpose of the flights. We conclude with reflections on the current AI landscape, emphasizing the responsibility of AI experts to shape public perception and understanding of AI's trajectory, stressing the importance of maintaining AI safety, conducting interdisciplinary research, and fostering informed discussions to build trust and reach a global consensus on AI as a tool for human progress and empowerment.

Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[2] Peter Corke. *Robotics, Vision and Control: fundamental algorithms in Python*, volume 146. Springer Nature, 2023.

[3] Alina Marcu and Marius Leordeanu. Object contra context: Dual local-global semantic segmentation in aerial images. *AAAI Workshops*, 2017.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Volodymyr Mnih. *Machine learning for aerial image labeling*. PhD thesis, University of Toronto (Canada), 2013.

[8] Shunta Saito and Yoshimitsu Aoki. Building and road detection from large aerial imagery. In *IS&T/SPIE Electronic Imaging*, pages 94050K–94050K. International Society for Optics and Photonics, 2015.

[9] Alina Marcu, Dragos Costea, Emil Slusanschi, and Marius Leordeanu. A multi-stage multi-task neural network for aerial scene interpretation and geolocalization. *arXiv preprint arXiv:1804.01322*, 2018.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[11] Christian Wiedemann, Christian Heipke, Helmut Mayer, and Olivier Jamet. Empirical evaluation of automatically extracted road axes. *Empirical Evaluation Techniques in Computer Vision*, pages 172–187, 1998.

[12] Shunta Saito, Takayoshi Yamashita, and Yoshimitsu Aoki. Multiple object extraction from aerial imagery with convolutional neural networks. *Electronic Imaging*, 2016(10):1–9, 2016.

[13] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 2017.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[15] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. *arXiv preprint arXiv:1709.00179*, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[17] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. *arXiv preprint arXiv:1709.05932*, 2017.

[18] Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. In *BMVC*, 2016.

[19] P.J. Besl and N.D. McKay. A method for registration of 3D shapes. *PAMI*, 14(2): 239–256, 1992.

[20] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017. Accessed: July, 2024.

[21] Dragos Costea, Alina Marcu, Emil Slusanschi, and Marius Leordeanu. Roadmap generation using a multi-stage ensemble of deep neural networks with smoothing-based optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[22] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[23] Alina Marcu, Dragos Costea, Vlad Licaret, Mihai Pîrvu, Emil Slusanschi, and Marius Leordeanu. Safeuav: Learning to estimate depth and safe landing areas for uavs from synthetic data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[24] Google. Google Earth, 2018. URL `https://www.google.com/earth/`. Available at `https://www.google.com/earth/`, version 7.3.0.

[25] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[26] Mihai Pirvu, Victor Robu, Vlad Licaret, Dragos Costea, Alina Marcu, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu. Depth distillation: unsupervised metric depth estimation for uavs by finding consensus between kinematics, optical flow and deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3215–3223, 2021.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL `http://arxiv.org/abs/1503.02531`.

[28] AliceVision. Meshroom: A 3D reconstruction software., 2018. URL `https://github.com/alicevision/meshroom`.

[29] Vlad Licăret, Victor Robu, Alina Marcu, Dragoş Costea, Emil Sluşanschi, Rahul Sukthankar, and Marius Leordeanu. Ufo depth: Unsupervised learning with flow-based odometry optimization for metric depth estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6526–6532. IEEE, 2022.

[30] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems*, pages 35–45, 2019.

[31] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

[32] S. Mahdi, H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proc. CVPR*, 2021.

[33] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] Alina Marcu, Vlad Licaret, Dragos Costea, and Marius Leordeanu. Semantics through time: Semi-supervised segmentation of aerial videos with iterative label

propagation. In *Asian Conference on Computer Vision (ACCV), 2020*, pages 2881–2890, 2020.

[35] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.

[36] Marius Leordeanu, Mihai Cristian Pîrvu, Dragos Costea, Alina E Marcu, Emil Slusanschi, and Rahul Sukthankar. Semi-supervised learning for multi-task scene understanding by neural graph consensus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1882–1892, 2021.

[37] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

[38] Alina Marcu, Mihai Pirvu, Dragos Costea, Emanuela Haller, Emil Slusanschi, Ahmed Nabil Belbachir, Rahul Sukthankar, and Marius Leordeanu. Self-supervised hypergraphs for learning multiple world interpretations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 983–992, October 2023.

[39] Mihai Pirvu, Alina Marcu, Maria Alexandra Dobrescu, Ahmed Nabil Belbachir, and Marius Leordeanu. Multi-task hypergraphs for semi-supervised learning using earth observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3404–3414, October 2023.

[40] Earth Observations NASA. Nasa earth observations dataset, 2020. URL `https://neo.gsfc.nasa.gov/`. [Online; accessed: 10.06.2024].

[41] Emanuela Haller, Elena Burceanu, and Marius Leordeanu. Self-supervised learning in multi-task graphs through iterative consensus shift. *BMVC*, 2021.