Self-supervised Hypergraphs for Learning Multiple World Interpretations

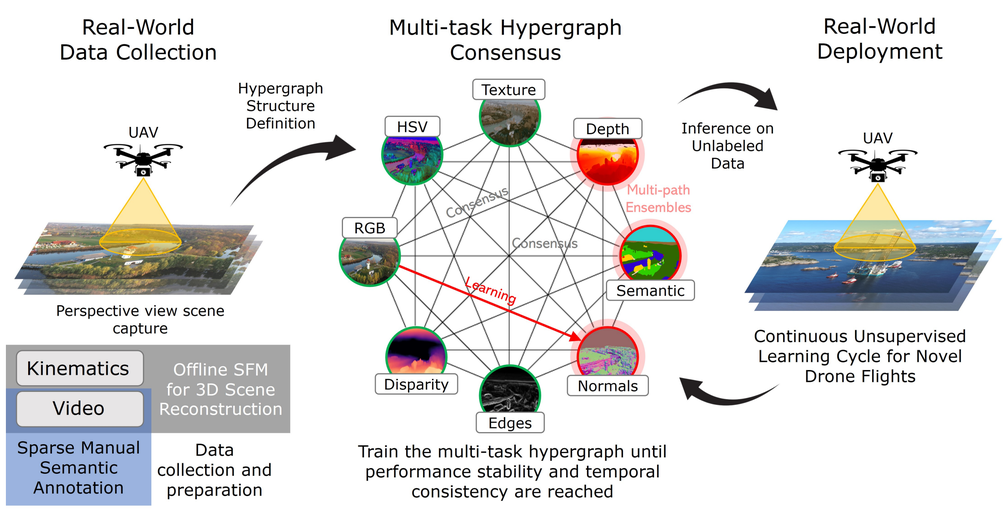
Alina Marcu[1,2] Mihai Pîrvu[1,2] Dragoş Costea[1,2] Emanuela Haller[3] Emil Sluşanschi[2] Ahmed Nabil Belbachir[4]
Rahul Sukthankar[5] **Marius Leordeanu**[1,2]

[1]Institute of Mathematics of the Romanian Academy [2]University Politehnica of Bucharest [3]Bitdefender [4]NORCE [5]Google Research

Contributions

Scope: Our research focuses on developing real-time embedded solutions using different kinds of sensing capabilities for drones to understand and function in the real world with little to no human supervision. Toward this end we make the following contributions:

- We propose a novel method for learning multiple scene representations given a small labeled set, by exploiting the relationships between such representations through a multi-task hypergraph.
- We introduce Dronescapes, a novel large-scale aerial video dataset with odometry data and semantically annotated frames.



Multi-task Hypergraph Consensus

- **Graph Structure.** Nodes represent different visual representations of a scene. We model relations between nodes through NNs (edges/hyperedges). Input nodes represent easily accessible information from sensors (e.g. RGB), whilst output nodes are complex tasks (e.g. semantic segmentation).
- **Learned Ensembles.** Multiple pathways within a graph/hypergraph can point towards a given node, meaning that we have several candidate predictions that we can use to form ensembles.
- **Iterative Learning.** We use ensembles at the level of each output node to produce pseudolabels, on novel unlabeled data, used to distill (retrain) the connections for the next learning iteration, until multi-task consensus.
- Hyperedges are modeled by lightweight U-Nets, with the same structure of 1.1M parameters, but independently learned.

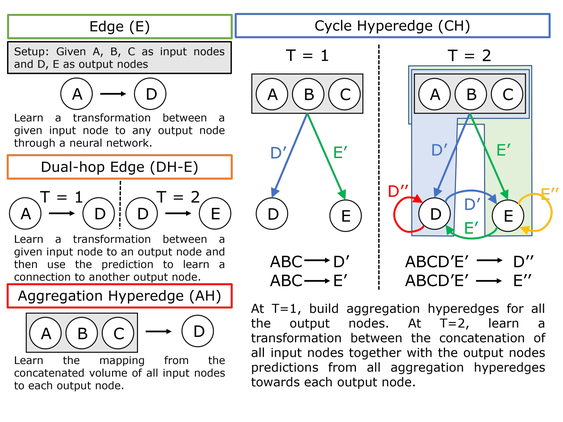


Figure 1. Hyperedges Types

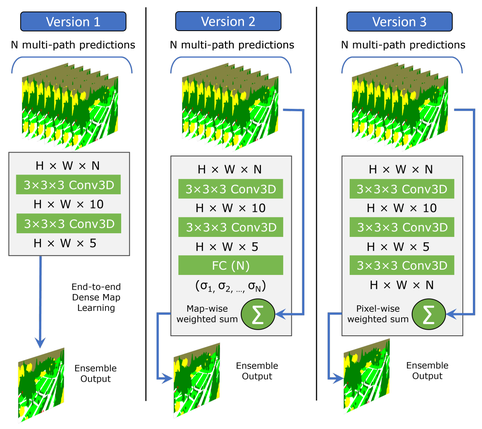


Figure 2. Proposed Learned Ensembles

Dronescapes Dataset

- Real-world drone data suitable for multi-task learning of semantic segmentation, depth estimation and surface normals estimation.
- 4K drone flight videos (30 FPS) with kinematics (10Hz) from 10 varied scenes (urban, rural, seaside), geographically far apart (25k+ video frames).
- Less than 2% are manually annotated frames with 8 semantic classes - *land*, *forest*, *residential*, *road*, *little-objects*, *water*, *sky*, *hill*.
- Auto-annotated metric depth maps by processing the structure-from-motion and telemetry data and Blender to extract camera normals (surface normals w.r.t camera).

Acknowledgements: This work was funded in part by UEFISCDI, under Projects EEA-RO-2018-0496 and PN-III-P4-ID-PCE-2020-2819, and by a Google Research Gift. We want to express our sincere gratitude towards Aurelian Marcu and The Center for Advanced Laser Technologies (CETAL) for providing access to GPU resources.

Experimental Analysis

Study 1. Impact of Hyperedges and Learned Ensembles

- **Hypergraph Definition.** – Input nodes (given): RGB, HSV, SoftEdges [3], SoftSegmentation [3], Unsup. Metric Depth Map [5]. Output nodes (predicted): Semantic Segmentation, Depth, Camera Normals.
- **Performance evaluation.** We report mean IoU (% - higher is better (↑)) for semantic segmentation and L1 error * 100 (lower is better (↓)) for depth and normals estimation.
- **Temporal consistency.** We use optical flow to establish temporal chains that connect corresponding pixels across frames. Pixel-wise consistency is the percentage of votes received by the winning class along neighboring pixels for semantic segmentation and the variance of predictions along the same chain for depth and camera normals.

	Type	Overall (test scenes)		
		SSeg	Depth	Normals
Edges	E: rgb	32.79	21.66	12.40
	E: hsv	33.51	19.90	12.48
	E: softedges	27.28	18.61	13.53
	E: softseg	24.68	22.70	12.76
	E: ufo	16.93	17.55	12.89
	DH-E: sseg	-	19.00	12.93
	DH-E: depth	24.11	-	13.79
	DH-E: norm	26.35	21.15	-
	mean	26.52	**20.08**	12.97
Hyperedges	AH	33.63	23.96	12.24
	AH-ufo	33.82	21.10	12.72
	CH	36.92	20.36	12.23
	mean	**34.79**	21.81	**12.40**

Table 1 Impact of Hyperedge Complexity

Method	Overall IoU(↑) (test scenes)
NGC [4] (Mean)	36.55
NGC (Mean) + HE	37.58
NGC [4] (Median)	34.56
NGC (Median) + HE	35.32
CShift [2] (Mean)	38.57
CShift (Mean) + HE	39.56
CShift [2] (Median)	37.81
CShift (Median) + HE	39.68
Logistic Regression (Ours)	**40.76**
NN (v1) (Ours)	38.94
NN (v2) (Ours)	38.36
NN (v3) (Ours)	40.67

Table 2 Comparison to previous multi-task graph-based methods. Numbers are reported for semantic segmentation.

Study 2. Impact of Iterative Self-supervised Learning

- We consistently show improvement over multiple self-supervised learning iterations, by progressively adding novel scenes (information) and applying our hypergraph consensus procedure. We show considerable improvement at the level of RGB→*Task* edge on the test scenes for multiple tasks. As a baseline, we compare to the purely supervised learning scenario.

	Semantic		Depth		Normals	
	IoU (↑)	Cons. (↑)	L1 (↓)	Cons. (↑)	L1 (↓)	Cons. (↑)
RGB-sup.	25.04	88.85	-	-	-	-
RGB-iter1	32.79	94.04	21.66	5.89	12.40	98.32
RGB-iter2	37.26	95.72	17.34	7.06	11.93	98.87
RGB-iter3	**40.31**	**98.13**	**16.64**	**30.26**	**11.71**	**99.30**

Study 3. Adapting to Novel Scenes

- Experiments show the advantages of using our hypergraph procedure in multiple learning phases and using an off-the-shelf VisTransformer-based method [1] as the supervisory signal for the task of semantic segmentation.

Method	Phase 1	Phase 2	Phase 3	Overall (test scenes) IoU (↑)	Cons. (↑)
Mask2Former [1]	-	-	-	52.77	96.99
RGB → SSeg (1)	✗	✓	✓	52.88	98.95
RGB → SSeg (2)	✓	✓	✗	53.57	98.78
RGB → SSeg (3)	✓	✓	✓	**54.26**	**99.06**

- **Phase 1** - training the edge in two iterations of learning (before seeing any data from the test scenes).
- **Phase 2** - distilling the edge (fine-tuning after Phase 1 or from scratch when Phase 1 is missing) on the output of Mask2Former on the test scenes.
- **Phase 3** - training the edge during one iteration of self-supervised hypergraph learning, only on unlabeled data from test scenes, after Phase 2.

References

[1] Cheng et al., *Masked-attention mask transformer for universal image segmentation*, CVPR (2022).
[2] Haller et al., *Self-supervised learning in multi-task graphs through iterative consensus shift*, BMVC (2021).
[3] Leordeanu et al., *Generalized boundaries from multiple image interpretations*, T-PAMI (2014).
[4] Leordeanu et al., *Semi-supervised learning for multi-task scene understanding by neural graph consensus*, AAAI (2021).
[5] Licăret et al., *Ufodepth: Unsupervised learning with flow-based odometry optimization for metric depth estimation*, ICRA (2022).

1st Workshop on Representation Learning with Very Limited Images, ICCV 2023