University of Genova

DRIM

Ph.D. Program of National Interest in Robotics and Intelligent Machines

Administrative Headquarters: Università di Genova

Real-Time Monocular Scene Analysis for UAV in Outdoor Environments

by

Yara AlaaEldin Abdelmottaleb

Thesis submitted for the degree of *Doctor of Philosophy* (38° cycle)

December 2025

Francesca Odone	Supervisor
Antonio Sgorbissa	Head of the PhD program

Thesis Jury:

Raffaella Lanzarotti, *Università degli Studi di Milano*	External examiner
Alessandra Sciutti, *IIT*	External examiner
Nicoletta Noceti, *Università degli Studi di Genova*	Internal examiner

Dibris

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

To all the children of Gaza.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Yara AlaaEldin Abdelmottaleb
February 2026

Acknowledgements

In the name of Allah, the merciful. I would like to thank my parents for their continuous support during my study. I would like to give a special thanking to my supervisor prof. Odone for her support, guidance, and advising throughout the three years of my PhD. I would like also to thank the PhDRIM board for all their efforts to assist us in all the steps from the beginning to the end. I thank all my family members and my friends in Egypt and in Italy for their encouragements and standing by my side throughout the journey. In the end, I thank Allah for giving me the health and well-being necessary to complete this thesis.

Abstract

Understanding the geometric and semantic properties of the scene is crucial in autonomous navigation and particularly challenging in the case of Unmanned Aerial Vehicle (UAV). Such information may be obtained by estimating depth and semantic segmentation maps of the surrounding environment, and for practicality, the procedure must be performed as fast as possible. In this thesis, we leverage monocular cameras on aerial robots to predict depth and semantic maps in low-altitude unstructured environments. We propose a joint deep-learning architecture, named Co-SemDepth, that can perform the two tasks accurately and rapidly, and validate its effectiveness on a variety of datasets.

The training of neural networks requires an abundance of annotated data, and in the UAV field, the availability of such data is limited due to the specificity of the domain and the burden of the annotation process. Simulation engines allow us to collect annotated data automatically with minimal effort. We introduce a new synthetic dataset in this thesis, *TopAir*[1] that contains images captured with a nadir view in outdoor environments at different altitudes, helping to fill the gap of the scarcity of annotated datasets in the aerial field.

While using synthetic data for the training is convenient, it raises issues when shifting to the real domain for testing. We conduct an extensive analytical study to assess the effect of several factors on the synthetic-to-real generalization in depth estimation and semantic segmentation. Co-SemDepth and TaskPrompter models are used for comparison in this study. The results reveal a superior generalization performance for Co-SemDepth in depth estimation and for TaskPrompter in semantic segmentation. Also, our analysis allows us to determine which training datasets lead to a better generalization for depth estimation and semantic segmentation. Using few-shot learning generally improved the generalization outcomes, and a visualization of the 3D semantic maps using the predictions is presented.

Moreover, to help attenuate the gap between the synthetic and real domains, image style transfer techniques are explored on aerial images to convert from the synthetic style to the

[1]Dataset is publicly available:https://huggingface.co/datasets/yaraalaa0/TopAir

realistic style. Cycle-GAN and Diffusion models are employed. The results reveal that diffusion models are better in the synthetic to real style transfer.

In the end, we focus on the marine domain and address its challenges. Co-SemDepth is trained on a collected synthetic marine data, called *MidSea*, and tested on both synthetic and real data. In addition, self-supervised approaches are tried to enhance the results and cope with the limited available annotated data. The results reveal good generalization performance of Co-SemDepth trained from scratch when tested on real data from the SMD dataset, which contains simple marine scenarios, while further enhancement is needed on the MIT Sea Grant dataset, which contains more challenging scenarios.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Table of contents

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

List of figures

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

List of tables

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 1

Introduction

1.1 Scene Analysis for UAV

The applications of unmanned aerial vehicles, also known as UAVs, are rapidly expanding across various fields, including environmental exploration, national security, package delivery, firefighting, and many more. The tasks included vary depending on the mission. Such tasks can include object detection, tracking, classification, depth estimation, and semantic segmentation. We focus on depth estimation and semantic segmentation. As it commonly happens in autonomous navigation, sensors are adopted to estimate scene depth and semantic information. Unlike ground autonomous vehicles, many types of UAVs, including drones, have limited computational capability and allowed carried weight. Thus, not all types of sensors can be mounted on them. For instance, sensors like LiDAR and RADAR that are usually adopted for estimating the depth of near or far objects cannot be adopted for lightweight UAVs as they are heavy and power-consuming. Also, while LiDAR point clouds contain accurate depth information, it is difficult to extract rich semantic information from them. To associate such depth points to their semantic meaning, an additional step of calibration between LiDAR and RGB cameras has to be done to estimate their relative transformation [7] and associate the points in the point cloud to their corresponding pixels in the image frame. However, such calibration is never fully accurate, and this leads to errors in the semantic association. Other depth sensors like stereo cameras are not common and may not be appropriate for UAVs since the small baseline distance between their two internal cameras, compared to the large distance between the stereo camera and the scene, produces inaccurate depth estimates [8]. Therefore, there is a particular need to achieve depth estimation for UAV applications using only monocular cameras, as they are cheap, light, and

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

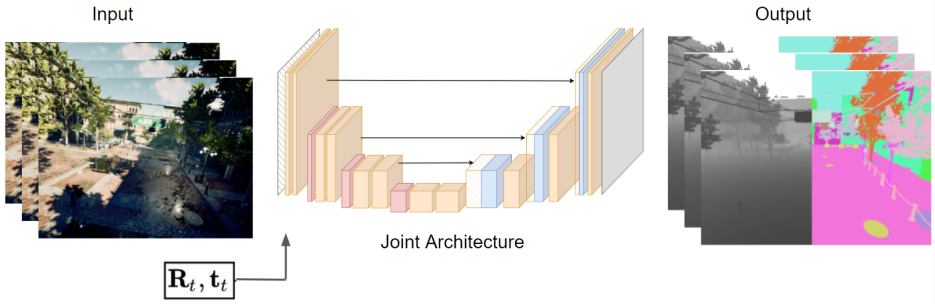


Figure 1.1 Our objective of designing a joint architecture for estimating both depth and semantic segmentation maps given as input video frames and camera pose data.

small in size. Current state-of-the-art methods use deep learning architectures for Monocular Depth Estimation (MDE) [9; 10; 11].

Extracting semantic information from the perceived scene can be achieved using multiple computer vision techniques: object detection, object classification, and semantic segmentation. In object detection and classification the goal is to predict the class of the objects appearing in the image and their corresponding bounding boxes in 2D pixel space or in 3D space. In semantic segmentation the goal is to assign a class label to every pixel in the image belonging to a predefined set of object categories. There are no known sensors that can achieve semantic segmentation directly. Instead, this task has to be realized by the semantic analysis of the RGB frames received from monocular video cameras. Current state-of-the-art methods [12; 13; 14; 15; 16] are using deep neural networks to learn the semantic features of the input image and decode the required output.

In this thesis, we leverage monocular cameras on aerial vehicles for obtaining two types of necessary information for scene understanding, *depth estimation* and *semantic segmentation*:

- In monocular depth estimation (MDE), the goal is to predict the depth of each pixel in each RGB frame captured by a video camera. Such depth expresses the distance (in meters) of the points in the world, projected in the pixels, with respect to the camera frame.

- In semantic segmentation, the goal, instead, is to predict the semantic class of each pixel in the input RGB frames. This semantic class belongs to a set of predefined semantic classes of interest.

The two are complementary since depth estimation expresses the geometric properties of the scene while semantic segmentation expresses the semantic properties.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

To address both challenges at once, we propose a joint deep architecture based on U-Net for achieving the two tasks accurately and in real-time, see Figure 1.1 for a broad overview. Using a *joint architecture* helps in saving computational time compared to performing each task separately, as well as saving GPU memory by having fewer model parameters. Also, joint-learning can help in sharing learned features between the two tasks, and this can, in turn, benefit both of them.

As for the architecture design, while currently there is a trend in using foundation models for achieving these tasks (see, for instance, DepthAnything [17], SegmentAnything [18] and Any-to-Any [19]), such models are more concerned with zero-shot prediction which is not our primary goal, and they are huge in size and not convenient for hardware deployment where memory and speed are critical factors. With current consumer hardware we can rely on at most 16 GB of primary memory and 16 GB of storage. To show these limitations, Table 1.1 reports the memory and power specifications of commonly used microcontrollers and single-board computers (SBCs) in robotic hardware.

Table 1.1 Specifications of commonly used micro-controllers and single board computers in robotics

Board	GPU?	Memory	Storage	Power
Arduino Uno [20]	No	32KB	-	5V
Arduino Portenta [21]	No	2MB	16MB	5V
Rapberry Pi [22]	No	8GB	8GB	15W
Jetson Nano [23]	Yes	4GB	16GB	5W-10W
Jetson TX2 [23]	Yes	4/8GB	16/32GB	7.5W-15W
Jetson Xavier NX [23]	Yes	8/16GB	16GB	10W-20W
Jetson AGX Xavier [23]	Yes	32/64GB	32GB	10W-30W

1.2 Collection of Synthetic Datasets

For the training of neural networks, different types of training are possible: supervised [24; 25], unsupervised [26; 27], or semi-supervised [28; 29]. Supervised training learns from input data (in our case, RGB images) and their labels. The goal of the network is to learn the mapping function from the input data to their corresponding labels to be able to predict the correct output for unseen input. Unsupervised training learns from unlabelled data. The network tries in this case to learn the pattern and the features from the data without any labels. Semi-supervised learning benefits from both labelled and unlabelled data. We hereby focus our attention on supervised learning to train our proposed architecture.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

The training of a supervised deep network requires an abundance of data that should contain RGB frames along with their corresponding annotation, in our case, both depth maps and semantic segmentation maps. In general, and especially in the UAV field, annotated real datasets are limited and small in size due to the huge effort required for carrying out the annotation process. For example, labeling a single semantic urban image in Cityscapes dataset [30] can take up to 60 minutes [31]. In addition, not all types and variants of the outdoor environments are represented in the currently available real datasets. As a consequence, it is convenient to use simulation engines like AirSim [32] and CARLA [33] for the collection of automatically annotated synthetic datasets in a variety of environments, and changing daytime and weather conditions to compensate for the lack in the available real datasets. Such synthetic data can be collected with large amounts and almost zero effort in the annotation process, and then, this data can be used for the training of supervised networks. We collect a synthetic aerial dataset *TopAir* using the AirSim simulator that is captured from a nadir (top) view in different environments and at various altitudes. Our dataset is annotated with depth and semantic segmentation maps, and it contains camera pose information.

1.3 Synthetic-to-Real

While the usage of synthetic datasets for training neural networks is convenient, this brings to the surface the issue of *synthetic-to-real domain generalization* where the neural network is trained on only or mostly synthetic data, and it is expected to perform well when tested on real data. The difference in the appearance of objects between the synthetic and the real world generally causes a drop in the neural network performance on real data. For example, in Figure 1.2, some of the variations between synthetic and real-world images are demonstrated. While there is a recent advancement in the 3D graphics of simulation engines, it can be observed that the lighting, texture, and semantic detail of the synthetic scenes still lag behind the real world. Most of the works in the literature addressing this problem were developed in the automotive field due to the latest trend in autonomous driving [34; 35; 36; 37], leaving the aerial field under-explored [38; 39]. Motivated by this, we investigate the problem of synthetic-to-real domain generalization in UAV monocular depth estimation and semantic segmentation by conducting an extensive set of experiments adopting a variety of synthetic and real aerial datasets that are publicly available and using joint deep architectures.

In particular, we evaluate the following factors:

- How changing the synthetic dataset used for training changes the model performance on the real data.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

(a) Synthetic images

(b) Real images

Figure 1.2 Sample images captured in simulation (a) and real-world (b) outdoor environments. It can be observed the differences in color, texture, lighting, and semantic detail between the two domains.

- How changing the model architecture can affect the synthetic-to-real generalization performance.
- Whether adding a small number of real data to the training would enhance the synthetic-to-real generalization output.

The analysis we carry out involves our joint architecture Co-SemDepth (a light, small network, more convenient for hardware deployment) as well as TaskPrompter (a big transformer-based network, more suitable for offline testing) [3]. This allows us to assess the impact of architectural properties in the reported analysis. In addition, we try out some of the image style transfer techniques to attenuate the gap between the synthetic and real images.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

1.4 Application to Marine Environments

An additional objective of our thesis is to apply the developed pipeline to hazardous outdoor scenarios; we consider, in particular, the marine environment. Such an environment is specifically addressed due to the additional challenges related to the sea (like water reflections and waves) and the limited datasets available in the maritime domain. There are many applications for Unmanned Surface Vehicles (USV) in the sea, including rescue, border surveillance, environmental monitoring, as well as maintenance of offshore systems [40].

First, synthetic marine data was collected using UnrealEngine5 [41], and we call the resulting dataset *MidSea*. Then, we apply the proposed Co-SemDepth architecture on MidSea and analyze the results obtained both quantitatively and qualitatively. In addition, we evaluate the synthetic-to-real performance by training the model on synthetic data and testing it on real, publicly available data. The analysis allows us to discuss the specifics of the marine environments and the potential and limits of our proposed solutions.

1.5 Problem Statement & Assumptions

In this section, the specifications of the problem, system assumptions, and the objectives are summarized.

Problem Statement: We are considering the problem of a UAV flying in outdoor unstructured environments (for instance, desert, forest, wild nature, and sea) and performing scene analysis of the video frames received from a monocular camera mounted on it. We focus our attention on depth estimation and semantic segmentation with the aim of carrying out these tasks in real-time. In most of the applications, the scene analysis has to be carried out in real-time due to the criticality of instantaneous decisions in UAVs. To address the scene analysis using deep neural networks, a sufficient amount of data has to be available for the training of the networks. However, a limited amount of data is available in the aerial field, and it is hard to find annotated real datasets. By using simulation engines, a huge amount of data can be collected and automatically and precisely annotated. Nevertheless, the synthetic-to-real domain gap problem remains to be addressed. The content, colors and style of collected synthetic images are different from real ones. We try to address this gap by varying the data used for training the network and by varying the architecture of the network. Also, image style transfer techniques are explored. One type of outdoor environments, the marine, has very limited data (synthetic and real), and it has challenges like transparent water,

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

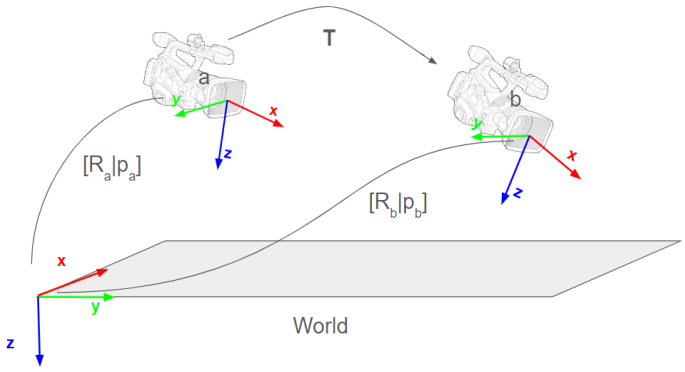


Figure 1.3 Illustration of the camera moving from one frame a to the next b

estimating the shoreline, and waves and sea dynamics that need attention. For this reason, we dedicate a separate chapter for it.

System Assumptions: In this thesis, we are considering a monocular camera is rigidly attached to a UAV. The camera intrinsic parameters are assumed to be known and constant. The intrinsic parameters describe the internal geometrical and optical characteristics of a camera, specifically used for mapping 3D world points to 2D image pixels. They include focal length f, the principal point c, and the sensor skew. The camera frame rate is relatively high, resulting in overlapping regions between consecutive frames. The UAV moves freely (6 Degrees of Freedom) in space and records the video frames as well as the camera position (using an IMU sensor for example) at each time step. As illustrated in Figure 1.3, using the camera position $p = [x, y, z]$ and orientation $R = [q_w, q_x, q_y, q_z]$ at each time step we can compute the motion transformation matrix ${}^{a}T_{b}$ from one frame a to the next b, as follows:

$$t = R_a^{-1}(p_b - p_a) \tag{1.1}$$

$$r_{3\times 3} = R_a^{-1} R_b \tag{1.2}$$

$${}^{a}T_{b} = \begin{bmatrix} r & t \\ 0 & 1 \end{bmatrix} \tag{1.3}$$

Objectives: Our first objective is to design a network, denoted by a function F, that takes at each time step the current frame I_t, previous n frames $I_{seq} = [I_{t-1}, I_{t-2}, ..., I_{t-n}]$ and camera motion transformations $T_{seq} = [T_{t-1}, T_{t-2}, ..., T_{t-n}]$ and outputs an estimated depth map $\hat{D}_t$ and semantic segmentation map $\hat{S}_t$ corresponding to the current frame:

$$(\hat{D}_t, \hat{S}_t) = F(I_t, I_{seq}, T_{seq}). \tag{1.4}$$

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Our second objective is to collect synthetic annotated data from a nadir view due to the scarcity of this type of data in the aerial field.

Our third objective is to train the network, developed in the first stage, on only or mostly synthetic data $Data_{synth}$ and make it generalize well to unseen real data $Data_{real}$ at test time.

The final objective is to focus on the marine environments by applying the developed pipeline on both synthetic and real marine data.

1.6 Contributions

In summary, the main contributions of this work are the following:

- We propose a lightweight joint architecture, Co-SemDepth, for performing monocular depth estimation and semantic segmentation on aerial data [42].
- We provide a benchmark for our method and other state-of-the-art methods in semantic segmentation and depth estimation, and highlight the advantages of using our joint architecture with regard to speed and memory efficiency.
- We introduce *TopAir*: an aerial synthetic dataset collected with a nadir view in various outdoor unstructured environments with annotations of depth, semantic segmentation, and camera location.
- We conduct a comparative analysis of the synthetic-to-real generalization between different synthetic data used for training and using different architectures [43].
- We demonstrate the positive effect of adding a small percentage of real data (few-shot learning) to the training of the networks on their synthetic-to-real generalization.
- We explore image style transfer techniques (Cycle-GAN and Diffusion models) to convert from the synthetic to the realistic style of input images
- We conduct an analysis on the application of Co-SemDepth in the marine environments

1.7 Thesis Structure

The rest of the thesis is structured as follows. In Chapter 2, the reviewed research papers related to the field are presented and analyzed. In Chapter 3, the Co-SemDepth joint architecture is proposed and the experiments done for validating it are presented. In Chapter 4,

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

we show the procedure followed to collect the TopAir dataset. In addition, the experiments carried out on the synthetic-to-real domain shift are analyzed. In Chapter 5, image style transfer techniques are explored for the sake of possible improvements. In Chapter 6, the approaches followed in application to the marine domain are discussed, and the experiments done in the maritime domain are presented. Finally in Chapter 7, the conclusions of this PhD thesis are presented with a focus on the achieved results.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 2

Related Work

Summary

In this chapter, the research found in the literature addressing the monocular depth estimation, semantic segmentation, and joint architectures are presented. The problem of monocular depth estimation in deep learning can be tackled with supervised methods (require large amount of annotated data) or self-supervised (supervision signals are drawn from stereo image pairs or video sequences). Semantic Segmentation is mostly achieved using supervised methods that can accept either single images or video frames as input. Methods applied in the marine domain can benefit from additional modules for horizon line estimation and water-land boundary detection. Joint architectures are also investigated for the sake of memory and time efficiency. Whether an FCN-based, a UNet-based, or a transformer-based architecture is deployed for joint learning, a feature encoder is usually shared among the vision tasks and the dedicated decoders are separated, with the possibility of adding cross-task attention between the decoders. In addition, we discuss the synthetic and real datasets found in both the aerial and marine domains. There exist a number of annotated datasets in the aerial field. However, datasets that contain both depth and semantic segmentation annotations (that can be used for training joint networks) are limited. Few datasets were found in the marine field, and most of them do not contain annotations of depth and segmentation. In the end, research techniques found for tackling the gap between the synthetic and real domains are discussed. The methods include image-style transfer, image-mixing, and few-shot prompting.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

2.1 Scene Analysis using UAVs

Applications of UAVs are countless and they are expanding across diverse fields. In military and public safety, they are used for surveillance, attack, border patrol, firefighting, and search and rescue. In commerce, they are used for package delivery, infrastructure inspection, traffic monitoring, and surveying for various industries [44]. In agriculture, they can be used for crops irrigation, applying fertilizers, and mapping and monitoring the fields [45]. Most of the mentioned applications require heavy vision tasks like object detection and tracking, classification, depth estimation, optical flow estimation, motion detection, semantic segmentation, etc. Such tasks work on a pixel level and require a large amount of computational resources to be achieved. Therefore, there is a crucial need in the aerial field to perform such tasks using memory and time-efficient approaches. One of these approaches is the use of multi-tasking deep joint architectures. Such architectures can share parts among the different vision tasks, leading to effectiveness in the utilization of computational resources and lower inference time. Unfortunately, much of the work in outdoor scene analysis has been driven by advancements in the automotive field, leaving aerial scene understanding comparatively under-investigated. Besides, data-hungry deep networks require an abundance of annotated datasets for training. Such annotated data are hard to be found in the real domain due to the burdensome annotation process. At the same time, recent advancements in computer graphics have remarkably improved the realistic appearance of simulation environments. Hence, there is a recent shift in the scientific community towards training neural networks on synthetic data that can be obtained in big amounts with minimal effort and automatically annotated by the simulation engine [46; 47; 48; 49].

2.2 Monocular Depth Estimation

After investigating multiple works in the literature addressing the task of monocular depth estimation, it was found that this task can be performed using either classical methods or deep learning methods. While classical methods require minimal number of data and deep learning methods require an abundance of data for training, the accuracy and performance gain achieved using deep learning methods make them a better candidate in many cases for performing monocular depth estimation [50]. One of the well-known classical methods for depth estimation is Structure from Motion (SfM). SfM focuses on the detection and tracking of feature points along given video frames and, then, using factorization to predict their 3D positions in space. This, however, generates sparse depth maps that contain 3D information

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

for only the feature points selected. While interpolation techniques can be used to obtain a continuous depth map, SfM is often limited by the availability of feature correspondences that can be difficult to be captured in low-textured image regions, repetitive patterns, occlusions, and lighting changes. This leads to noise and missing parts in the output depth map [51], see Figure 2.1 for an example.

Instead of matching features across frames and geometric triangulation, **deep learning methods** can handle the above challenges by employing priors learned from a variety of training datasets. The architectures used for monocular depth estimation can rely on single images or image sequences as input.

- In **single-image** based architectures, the input to the neural network is a single image and the output is its predicted depth map. Some of the works found in this category are: the GAN-based architecture presented in [52], the transformer-based AdaBins [53] that approached the depth estimation as a classification problem (bins of depth ranges) rather than regression, the U-Net based architectures proposed in [54; 55; 56], and the standard Fully Convolutional based networks (FCN) presented in [57; 58; 59]. However, image-based methods suffer from the scale-ambiguity problem, where the scale of objects cannot be inferred accurately using a single image, and this leads to errors in the metric depth estimation. An attention-based network proposed in [60] was used to solve such a problem by embedding the camera parameters as additional input to the network to reason over the physical size of objects and learn scale priors. Nevertheless, the network is huge and requires a big amount of data for training.

- **image-sequence** based methods overcome the scale-ambiguity problem by making use of the temporal relation between input video frames. This helps to better understand the scale of the objects appearing in the images and leads to more accurate depth predictions. Some of the works used ConvLSTM or ConvGRU to extract temporal relation between input video frames [61; 62; 63; 64], some used an additional pose estimation network to infer camera poses from input video frames and help the original depth estimation network [50; 51; 65], some used a modified U-Net architecture for capturing spatio-temporal relation [66], and others used GAN or GCN [67; 68; 69]. In their paper, ManyDepth [9] shows the superiority of using multiple frames by comparing the predicted depth map from single frame input versus multiple frame input using the same model.

Whether depth estimation is done using single images or multiple frames, the training of the network can be achieved in a supervised or unsupervised fashion:

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

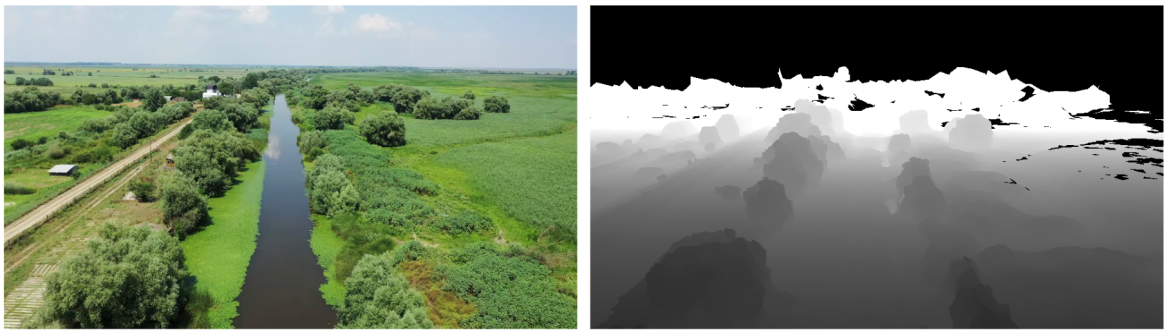


Figure 2.1 Output depth map using SfM method [1] resulting in holes and unknown regions in black

- In **supervised methods** [2; 70; 71; 72; 73], depth maps ground truth should be provided to the network, where the network tries to learn the mapping function from images to depth maps by minimizing the cost between predictions and the ground truth. The main challenge for these methods is the scarcity of available annotated datasets that cover different scenarios and environments. Some of the most used datasets in the literature are KITTI [74] for autonomous driving applications, NYU-Depth-v2 [75] for indoor scenes, and the synthetic MidAir [76] and TartanAir [77] datasets for aerial outdoor scenes. Recently, there exist foundation models, like DepthAnyhting [17; 78], that are concerned with zero-shot depth estimation on any input image thanks to the huge amounts of data used for training such models.

- **Self-supervised methods** remove the limitation of requiring ground truth depth maps. They can be trained to predict the depth using as a supervision signal nearby video frames [9; 79; 80] or synchronized stereo image pairs [11; 26]. Methods using synchronized stereo pairs require datasets that contain left and right stereo images and are trained to minimize image reconstruction losses. The methods using video frames require the scenes to be static, and there should be overlapping regions among the nearby frames to allow a pose estimation network to predict the pose transformation correctly. These requirements limit the application of such methods on dynamic scenes or videos of low frame rate, where there is little or no overlapping between the consecutive frames.

In [81], they particularly handle images that contain reflective surfaces (water) and reformulate the depth estimation self-supervised problem to rely only on intra-frame features instead of inter-frame features as in stereo and adjacent-frames based methods. The reflection image in the water is regarded as another view and, thus, allows the application of multi-frame

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

based methods on a single image. A segmentation network is used to detect the water mask region, a depth U-Net-based network is used to predict the depth map, and a photometric re-projection loss [82] is employed. The challenge of weak texture in images received from USV is addressed in [83] by the use of multi-scale feature fusion

Challenges: Few MDE methods mentioned their inference time, and no benchmarking was found in the literature comparing the inference time of different MDE methods. Also, very few methods were benchmarked on low-altitude aerial datasets [2; 84]. Foundation models like DepthAnything [17] are generally large in size ($> 20M$ params) and may not be suitable for hardware deployment.

2.3 Semantic Segmentation

Semantic segmentation is widely addressed in the literature, in particular in the autonomous driving field [12; 13; 14; 15; 16] using the CityScapes [30] dataset for benchmarking.

Semantic segmentation networks are normally trained in a supervised fashion, except for few cases [85; 86]. A variety of architectures were found in the literature that can accept either single images or video frames as input. Unlike depth estimation, image-based semantic segmentation is not affected by the scale-ambiguity problem. This is because the prediction of the semantic class does not depend on the scale of the objects. Some state-of-the-art examples of **image-based methods** are ICNet [87], ERFNet [88], PSPNet [89], SegFormer [90], BiSeNet [91], MANet [92], Image Cascade Network [93], and RefineNet [94].

In contrast, **video-based methods** in semantic segmentation [95; 96; 97; 98; 99; 100; 101; 102] focus on speeding up the semantic segmentation computation on individual frames by propagating the semantic segmentation prediction done on previous frames to successive video frames. Such propagation can be realized using light optical flow networks or interpolation techniques leveraging the overlapping regions between successive frames. This is done to speed-up inference time and to ensure temporal consistency among predicted semantic maps. However, the mIoU reported using video-based methods was usually lower than that of image-based methods [97].

A segmentation network, WaSR, is proposed in [103] for Unmanned Surface Vehicles in marine environments. The architecture is composed of a contracting path (encoder) and an expansive path (decoder). The encoder is an adaptation of the low-to-mid level parts of DeepLab2 [104]. The decoder, instead, is a fusion of visual and inertial information. In particular, the horizon line is estimated by the camera-IMU projection [105], and a binary mask distinguishing pixels below and above the horizon is constructed. Such an IMU mask

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

is fused with the encoder features and it serves as a prior estimation of the water edge location and improves the water segmentation in the output. In [106], they further enhance the segmentation of marine scenes by not only considering the boundary line between water and the shore, but also the boundaries between water and water obstacles. This is achieved by implementing a boundary feature extraction stream that takes as input the GT boundary image, and the extracted features are fused with the semantic feature stream during training time to strengthen its performance.

Recently, the foundation model SegmentAnything [18] has gained significant attention due to its notable results in zero-shot semantic segmentation. It is trained on a huge amount of data covering several domains and objects. However, SegmentAnything is dedicated for mesh segmentation, not class-based segmentation.

Challenges: Few semantic segmentation methods [107; 108; 109] were benchmarked on low-altitude aerial datasets.

2.4 Joint Vision Architectures

The idea of a joint or multitasking architecture has been tackled in multiple works in the literature for various vision tasks. In [110], a joint architecture is implemented for image segmentation and classification, while in [111], they developed a joint network for motion estimation and segmentation.

The main advantages of using joint architectures compared to single dedicated architectures are the efficiency in computational time and GPU memory (for hardware deployment, memory and speed are critical factors) as well as the possibility of enhancement in the accuracy of the predictions of each task, owing to the shared learning process. We hereby mention some of the works explored in the domain of joint depth estimation and semantic segmentation.

Multiple works [112; 113; 114; 115] used a shared encoder part in the joint architecture and two separate decoders dedicated for semantic segmentation and depth estimation. In [116], a multi-tasking transformer (Swin transformer) was developed where the encoder is shared, and each task has its own separate decoder. The encoder and the decoders have pyramidal structures with skip connections between the encoder and decoder levels. The architecture was tested on 6 tasks (depth, semantic segmentation, surface normals, edge detection, keypoints, and reshading). The results obtained show significant improvements compared to the single-task transformer. However, the model has a lot of parameters (231 million), and thus requires a big GPU memory.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

A Spatial-Channel Multi-task Prompting framework based on vision transformers was proposed in [3]. With the help of task prompts and attention modules, the network learns task-generic, task-specific representations and cross-task interaction in the same network without using multiple networks. It produced state-of-the-art results on NYUD-V2 [75] and PASCAL [117] datasets. However, the architecture appears somewhat complicated, and no information is provided about its inference time.

A modular design was proposed in [115] where depth estimation and semantic segmentation were estimated separately using any chosen open-source methods, and then both predictions enter a joint refinement network (CNN) to produce better estimations. A similar idea was proposed in [118] where multiple off-the-shelf networks were used to predict various vision tasks. The initial predictions of different tasks were then combined together through a distillation unit to make refined predictions. The predictions were distilled on different scales and then were aggregated to make the final predictions. While such methods help in enhancing the output accuracy compared to the initial predictions, the main disadvantage is the required additional computational time for the refinement phase.

2.5 Aerial Datasets

There are various available datasets in the aerial domain. Such datasets can be captured in indoor, outdoor, urban, or non-urban environments. They can be synthetic or real, and they can have different view angles: forward-view or top-view (also called nadir-view). We hereby list the available datasets that are annotated with ground truth depth maps and semantic segmentation maps. In Table 2.1, we provide a summary of the main characteristics of the datasets.

Aerial Datasets for depth and semantic segmentation:

- *MidAir* [76]: a synthetic dataset collected using the AirSim simulator. It contains around 420K frames captured in different trajectories in 2 outdoor non-urban natural environments in different weather and light conditions. Its semantic segmentation annotation has 14 classes: {animals, trees, dirt ground, ground vegetation, rocky ground, boulders, empty, water plane, man-made construction, road, train track, road sign, and others}.

- *SynDrone* [119]: a synthetic dataset collected using Carla simulator at various heights. It contains around 72K frames in 8 different environments. It has annotation of

Finanziato dall'Unione europea NextGenerationEU · Ministero dell'Università e della Ricerca · Italiadomani · Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

28 semantic classes: {building, fence, pedestrian, pole, roadline , road, sidewalk, vegetation, cars, wall, traffic sign, sky, bridge, railtrack, guardrail, traffic light, water, terrain, rider, bicycle, motorcycle, bus, truck, others}

- *SkyScenes* [38]: a synthetic dataset collected using Carla simulator. It contains around 33.6K frames, and it covers a wide variety of pitch angles, and heights in 8 different environments (same towns as in SynDrone). It has an annotation of 28 semantic classes: {building, fence, pedestrian, pole, roadline , road, sidewalk, vegetation, cars, wall, traffic sign, sky, bridge, railtrack, guardrail, traffic light, water, terrain, rider, bicycle, motorcycle, bus, truck, others}.

- *VALID* [120]: a synthetic dataset collected using AirSim simulator. It contains around 6.7K frames collected in 6 different environments, and it has 30 semantic segmentation classes.

- *WildUAV* [121]: a real dataset containing around 1.5K frames captured in an outdoor non-urban environment. Its semantic segmentaton annotation contains 16 labels: {sky, deciduous tree, coniferous tree, fallen trees, dirt ground, ground vegetation, rocks, water plane, building, fence, road, sidewalk, static car, moving car, people, and empty}.

- *DroneScape* [122]: a real dataset collected in 8 different outdoor urban and non-urban environments. It contains around $> 10K$ frames and it has 8 semantic segmentation classes: {Land, Forest, Residential, Road, Little-objects, Water, Sky, and Hill}.

- *UAVid* [123]: a real dataset containing around 410 frames captured in an outdoor urban environment in China. The ground truth depth maps are provided for only some frames of the test set. Its semantic segmentation annotation has 8 classes: {building, road, static car, tree, low vegetation, human, moving car, background clutter}.

Image samples from the datasets are shown in Figure 2.2.

Aerial Datasets for depth estimation:

- *TartanAir* [77]: a synthetic dataset collected using the AirSim simulator. It contains more than 1 million frames collected in indoor and outdoor environments with various view angles. While this dataset contains semantic segmentation annotation, it is mesh segmentation, not class segmentation.

- *ESPADA* [124]: a synthetic dataset collected using the AirSim simulator in 49 different scenes. It contains around 80K frames.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

- *UrbanScene3D* [125]: a dataset containing both synthetic and real images. The synthetic data were collected using the AirSim simulator. It contains around 128K frames collected in urban environments. As in Tartan Air, the semantic segmentation in this dataset is mesh segmentation, not class segmentation.
- *UseGeo3D* [126]: a real dataset containing 829 frames collected in semi-urban environments.

Image samples from the datasets can be found in Figure 2.3.

Aerial Datasets for semantic segmentation:

- *SynthAer* [127]: a synthetic dataset collected using the modeling tool Blender [128] in an urban environment at 30-50 meters height. It contains 765 frames and 8 semantic labels: {Building, Vehicle, Road, Footpath, Sky, Tree, Vegetation, Wall}
- *Aeroscapes* [108]: a real dataset collected in urban and rural environments with various altitudes and view angles. It is composed of 3269 frames and its semantic segmentation annotation has 12 classes: { background, person, bike, car, drone, boat, animal, obstacle, construction, vegetation, road, sky }
- *Ruralscapes* [129]: a real dataset collected in semi-urban environments. It contains 1127 frames and its semantic segmentation annotation has 12 classes: {forest, residential, land, sky, hill, road, church, fence, water, person, car, haystack}
- *VDD* [130]: a real dataset collected in urban and semi-urban environments. It contains 400 images and its semantic segmentation annotation has 7 classes: {wall, roof, road, water, vehicle, vegetation, others}
- *UDD* [131]: a real dataset collected in urban and semi-urban environments (same environments in VDD). It contains 200 images and its semantic segmentation annotation has 6 classes: {facade, road, vegetation, vehicle, roof, others}
- *FSI, Flood Satellite Imagery* [132]: a real dataset collected in semi-urban environments where some of them were flooded. It contains 261 images, and we are only interested in the images without flood. Its semantic segmentation annotation contains 25 labels: {background, grass, trees, roof, vehicle, chimney, secondary structure, swimming pool, power lines, window, road, flooded, forest, street light, water, garbage bins, trampoline, satellite antenna, parking area, solar panel, under construction, boat, sports complex, industrial site, water tank.}

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

- *ICG* [133]: a real dataset collected at low-medium heights in semi-urban environments. It contains 600 images with semantic segmentation annotation of 20 classes: {tree, rocks, dog, fence, grass, water, car, fence-pole, other vegetation, paved area, bicycle, window, dirt, pool, roof, door ,gravel, person, wall, obstacle}

Image samples from the datasets can be found in Figure 2.4.

2.6 Maritime Datasets

Different from aerial datasets, the datasets available in the maritime domain are limited.
Marine datasets annotated for depth estimation or semantic segmentation:

- MassMIND [134]: contains 2916 diverse Long Wave InfraRed (LWIR) images captured around the Boston Harbor in Massachusetts, USA, in various weathers, seasons, and time of the day. Each image is annotated with pixel-level semantic segmentation across 7 classes: Sky, Water, Bridge, Obstacles, Living Obstacles, Background, Self. Resolution is 640×512.

- MaSTr1325 [135]: consists of 1325 diverse images captured in the Gulf of Koper, Slovenia, with a real USV. It includes various weather conditions and times of day. Each image is per-pixel semantically segmented with one of four labels: Obstacles and Environment, Water, Sky, or Unknown. In addition, GPS and IMU data are provided and time-synchronized with the images. Resolution is 512×384.

- MODD2 [136]: an extended version of MaSTr1325 dataset containing 11675 stereo frames with a resolution of 1278×958 pixels. The dataset was recorded over a period of several months to ensure diversity.

- MODS [137]: A dataset of 81k images acquired in the Slovanian coast. Their dataset annotations were performed in a very strict procedure to ensure quality and accuracy. The segmentation classes are 3: Sky, Water, and Others. Most of the obstacles present in the dataset are in the range of 15-20 meters away from the USV.

- MIT[1]: a multi-modal sensor dataset for mobile robotics research in the marine domain. It is composed of videos acquired from optical and IR cameras, point clouds acquired from 3D Lidar, RADAR images, and IMU data. The data was captured during various

[1] Available on this link

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Table 2.1 Available aerial outdoor datasets and their main characteristics. D means depth maps, S means semantic segmentation maps, and Loc means camera location information. top-view means pitch angle close to 90°. The top part of the table lists datasets with both depth and semantic annotations. The middle part contains datasets with only depth annotation, and the bottom part contains datasets with only semantic annotation.

Dataset	Type	img/vid	Annotation	View	Size	Environment	Height	#Classes
MidAir	synth	vid	D+S+Loc	fwd	420*K*	rural	low-med	14
SynDrone	synth	vid	D+S+Loc	fwd + top	72*K*	urb+semi-urb	low-med-high $(20, 50, 80m)$	28
SkyScenes	synth	vid	D+S+Loc	fwd + top	6.7*K*	urb+semi-urb	low-med $(15, 35, 60m)$	28
VALID	synth	vid	D+S	top	6.7*K*	urb+semi-urb	med-high $(20, 50, 100m)$	30
WildUAV	real	vid	D+S+Loc	top	1.5*K*	rural	low-med $(50m)$	16
Dronescape	real	vid	D+S+Loc	fwd	$> 10K$	urb+rural	high	8
UAVid	real	vid	D+S	top	410	urban	high	8
TartanAir	synth	vid	D+Loc	fwd+top	1*M*	rural+urb	low-med	-
ESPADA	synth	vid	D+Loc	top	80*K*	urb+rural	med-high$(30 - 100m)$	-
UrbanScene3D	synth+real	vid	D+Loc	fwd+top	128*K*	urb	low-med	-
UseGeo3D	real	vid	D+Loc	top	829	semi-urb	high$(80m)$	-
SynthAer	synth	vid	S	fwd+top	765	semi-urb	med	8
Aeroscapes	real	vid	S	fwd+top	3.3*K*	urb	low-med $(5 - 50m)$	12
RuralScapes	real	vid	S	fwd	1.1*K*	semi-urb	med-high	12
VDD	real	img	S	top	400	urb	med-high $(50 - 120m)$	7
UDD	real	img	S	top	200	urb	med-high $(60 - 100m)$	6
FSI	real	img	S	top	261	semi-urb	med-high	25
ICG	real	img	S	top	600	semi-urb	low-med $(5 - 30m)$	20
TopAir	synth	vid	D+S+Loc	top	10*K*	rural+urb	low-med-high$(5 - 100m)$	9

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

missions at sea with several obstacles and weather conditions. Point clouds can be used as ground truth for depth estimation tasks.

See Figure 2.5 for samples from the datasets.

There are other marine datasets without depth or semantic segmentation annotations that can be used only for qualitative evaluation.

Non annotated Marine datasets:

- ABOships [138]: a dataset only annotated with bounding boxes belonging to one of 11 categories.The data is composed of 10k images acquired in the Aura River and the port of Turku in Finland.
- Singapore Maritime Dataset (SMD)[2]: a dataset of optical and IR video frames captured at various times of the day and weather conditions. Obstacles are annotated with bounding boxes.
- SeaShips [139]: It contains images of sea ships annotated with bounding boxes. The images are taken from the shore around Hengqin Island, Zhuhai city, China.
- MarDCT [140]: A dataset captured in Venice, Italy, from a top view. Images contain different boats, with each image containing only one boat in the center.
- Marvel [141]: A large-sized dataset of marine vessels images. The images are collected from the internet, and each image contains only one vessel.

See Figure 2.6 for samples from the datasets.

2.7 Synthetic-to-Real Domain Shift

Many works addressed the synthetic-to-real domain shift problem in either monocular depth estimation or semantic segmentation. However, most of these works, driven by the current trend in the autonomous driving field, focused on the ground vehicle domain. Only few papers [38; 39; 119; 121; 143] investigated the synthetic-to-real domain shift in the aerial field, and they mostly focused on semantic segmentation. In the following, a summary of the reviewed works is presented.

In the automotive field: The techniques used for improving the synthetic-to-real generalization in depth estimation or semantic segmentation were found to be either image mixing

[2]Available on this link

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

techniques [144; 145], or image style transfer techniques [35; 36; 37; 146]. In image mixing, the synthetic images used for training a neural network are mixed with parts or elements from the real images (for example, cars, persons, or buildings). This enhances a bit the realistic appearance of the synthetic images, and hence, improves the generalization to the real domain during testing.

In image style transfer, generative models like GANs or Diffusion models are used to transform the appearance of the synthetic images to look more realistic, and consequently, when used to train a deep model, it enhances its generalization to the real domain.

The widely-used benchmarks for the evaluation of synthetic-to-real domain shift in the automotive field are SYNTHIA-to-Cityscapes and SYNTHIA-to-Mapillary [30; 147; 148].

In the aerial field: The main works found addressing the synthetic-to-real problem [38; 39; 119; 143] focused on semantic segmentation.

In [119], the DeepLabV3 network [149] was trained on the SynDrone dataset and evaluated on the real datasets: UAVid, Aeroscapes, ICG, and UDD. In addition, the effect of changing training and testing towns or altitudes was demonstrated, and it was shown that even when using the same synthetic dataset, changing the environment between training and testing or changing the altitude leads to a drop in performance.

In [38], 3 different semantic segmentation architectures were trained on synthetic datasets (SkyScenes and SynDrone), and the synthetic-to-real performance was evaluated on 3 real datasets: UAVid, Aeroscapes, and ICG. It was found that networks trained on SkyScenes generalize better to real datasets than those trained on SynDrone, due to the variability of heights and pitch angles in SkyScenes as well as the better representation of human and vehicle classes. It was also found that changing the height and pitch angles between the synthetic and the real datasets greatly affects the generalization accuracy. It was found as well that augmenting part of the testing real data with the synthetic data during training improves the model's generalization performance.

In [143], a study on the perceptual and structural complexity of datasets (SkyScenes and DroneScapes) was presented, and it was suggested that the gap between a synthetic and a real dataset can be quantified by measuring the difference between their corresponding perceptual complexity scores.

In [39], a synthetic dataset similar to RuralScapes [129] was generated and used to train a semantic segmentation network using zero-shot and low-shot regimes. In low-shot, it was found that by only adding a small percentage $(1-5\%)$ of the real Ruralscapes to the training synthetic data, the evaluation accuracy on Ruralscapes was boosted.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

In [150], they divide the UDA methods into three levels: on the input, on the feature representation, and on the output. Methods working at the input level include image-style transfer techniques that convert the style of the input images to the style of the target real domain before training on the input images. The core idea of adaptation at the feature-level is to force the feature extractor of the network to predict domain-invariant latent representations from the source and target domains. After that, the network classifier should be able to classify both the source and target representations correctly by relying solely on the supervision from the source. Adaptation at the output level can benefit from adversarial strategies applied to the output low-dimensional space spanned by the segmentation maps. A domain discriminator is trained to distinguish the source and target domains from the given predicted maps, and the segmentation network has to fool the discriminator by aligning the distribution of the predicted labels across the two domains. In [151], an unsupervised domain adaptation semantic segmentation network is proposed and tested on remote sensory images. The approach focuses on storing invariant features of the source and target domains using a memory module.

For depth estimation, in [121], some experiments were performed where depth estimation networks were trained on selected images from the MidAir dataset [76] and tested on the real WildUAV dataset [121].

In the marine domain: The work in [152] addressed the synthetic-to-real domain shift in object detection YOLO [153] by adding a limited number of annotated real data to the synthetic training data, while a synthetic marine environment was proposed in [154] to support the research in this area.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

(a) MidAir [76]

(b) SynDrone [119]

(c) SkyScenes [38]

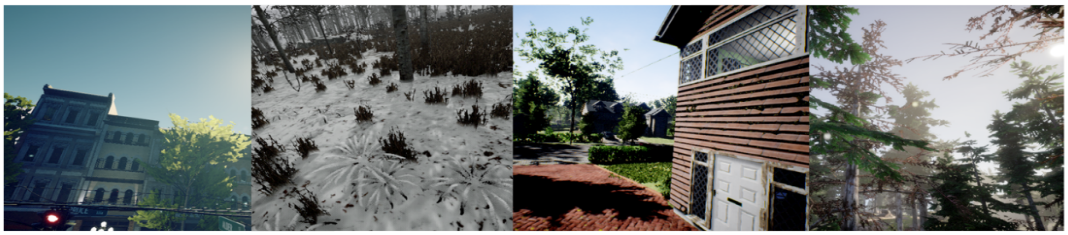
(d) WildUAV [121]

(e) DroneScapes [122]

Figure 2.2 Sample images from different aerial datasets that contain depth and semantic segmentation annotation

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*



(a) TartanAir [77]

(b) ESPADA [124]

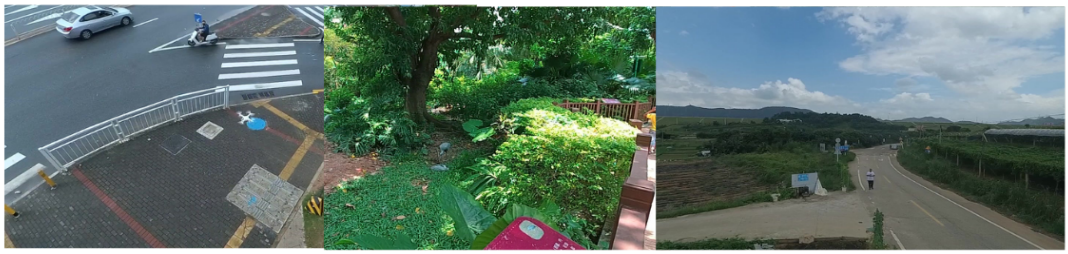
(c) UrbanScene3D [125]

(d) UseGeo [126]

Figure 2.3 Sample images from different aerial datasets that contain only depth annotation

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*



(a) AeroScapes [108]

(b) RuralScapes [129]

(c) UDD [131]

(d) FSI [132]

(e) ICG [133]

Figure 2.4 Sample images from different aerial datasets that contain only semantic segmentation annotation

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

(a) MassMIND [134]

(b) MaSTr1325 [135]

(c) MODD2 [136]

(d) MIT

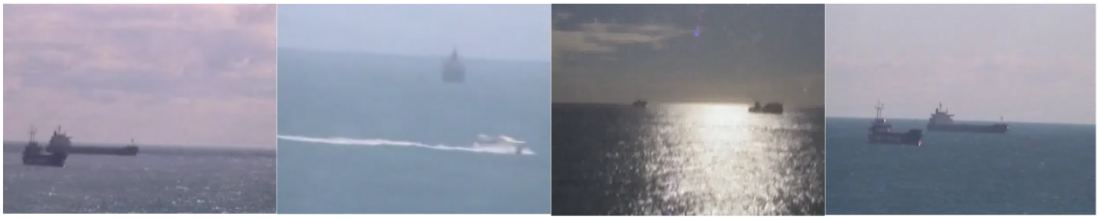
Figure 2.5 Sample images from different marine datasets that are annotated with either depth or semantic segmentation

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

(a) ABOships [138]

(b) SMD [142]



(c) MarDCT [140]

(d) Marvel [141]

Figure 2.6 Sample images from different non-annotated marine datasets

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 3

Proposed Co-SemDepth

Summary

In this chapter, we first give a brief overview of M4Depth, our backbone network for depth estimation, then describe our developed M4Semantic segmentation network. After that, we merge the two networks and shed light on our proposed joint Co-SemDepth architecture. To summarize, M4Depth [2] is a network with a pyramidal structure that incorporates motion and video frames in the process of supervised training to enhance the depth estimation. M4Semantic is an adapted version of M4Depth modified to be dedicated to semantic segmentation. Co-SemDepth is designed by merging the encoder part in M4Depth and M4Semantic, and separating the dedicated decoders for the two tasks.

In addition, the setup used to conduct the experiments on Co-SemDepth is explained. Then, we discuss the experiments conducted using the developed joint architecture Co-SemDepth to validate its effectiveness and competency against other state-of-the-art methods. Results reveal that using Co-SemDepth is more time and memory-efficient than using single dedicated architectures. In addition, Co-SemDepth is notably faster than other joint architectures (and single ones), while being competent in the accuracy of both depth estimation and semantic segmentation.

3.1 Architecture Design

3.1.1 Backbone Depth Network

The starting point of our work is the M4Depth network proposed in [2]. Our motivations to choose such a network are that, to the best of our knowledge, it produces the current top

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

results in monocular depth estimation on the synthetic MidAir [76] aerial dataset, and its model weights and code are publicly available. In addition, the encoder-decoder modularity of their architecture makes it a convenient candidate to be transformed into a joint architecture. Moreover, M4Depth has a unique architecture that integrates camera motion data to enhance the depth estimation. We adopt the M4Depth architecture as it is, without changes, except for the number of layers we use 5 instead of 6 layers, and this will be justified in the Experiments section.

The architecture of M4Depth [2], see Figure 3.1, is an adaptation of the standard U-Net encoder-decoder network [155] trained to predict parallax maps to be then transformed into depth maps. The authors define parallax as a function of perceived motion, thus it can be seen as a general form of stereo disparity for an unconstrained camera baseline. Similar to disparity, parallax can be related to the depth of points in space appearing in the image. They train the network to estimate parallax instead of depth directly to make it robust to unseen environments. This is because estimating depth from input images make the network tied to the training data distribution while estimating parallax is more robust and is less dependent on the training data distribution. One of the unique things about M4Depth is that it incorporates motion information in the process of supervised training and exploits such information for enhancing the depth prediction.

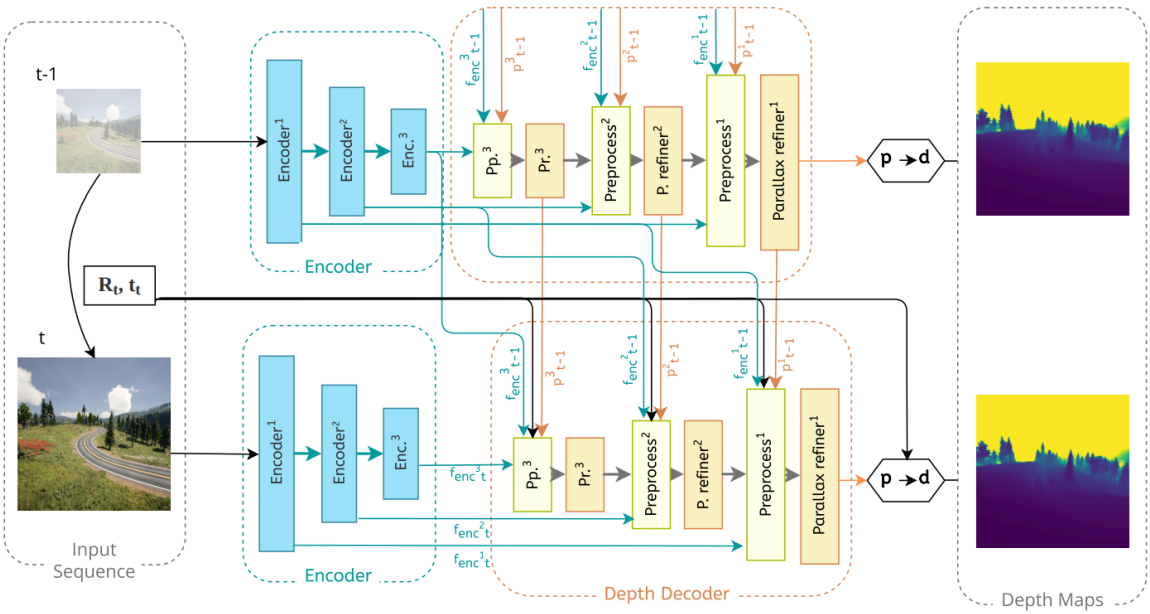


Figure 3.1 The architecture of M4Depth. It is fed by two consecutive frames and the camera motion. Each convolution layer is followed by a ReLU activation unit. The depth map is shown in the viridis scale. Details of the modules are given in Figure 3.2

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

The network takes as input a sequence of n video frames (we choose $n = 3$) and the camera transformation T between every two consecutive frames. At each time step t, the encoder takes a new video frame I_t and extracts image features at different scales using its pyramidal structure. Each encoder level is composed of two convolutional layers and a domain-invariant normalization layer (DINL) [156] to increase the network robustness to varied colors and luminosity conditions.

Then, the decoder takes the feature maps at different resolutions obtained by the encoder at time t, the features extracted from the previous frame $(t-1)$, the parallax map predicted at time $t-1$, and the camera motion transformation T_t to predict the parallax map of the current frame I_t. This parallax ρ_t is then transformed into a depth map d_t following the pixel-wise transformation proposed in Equation 3.1.

$$\rho_{i_t j_t} = \frac{\sqrt{(f_x t_x - t_z i_V)^2 + (f_y t_y - t_z j_V)^2}}{z_{i_t j_t} z_V + t_z} \tag{3.1}$$

Where $z_{i_t j_t}$ is the depth of the point P located in pixels (i, j) at time t, f_x and f_y are the focal lengths of the camera, t_x, t_y, and t_z are the camera translation between two consecutive frames, i_V and j_V are the image pixel coordinates of the point P in the plane of a virtual camera V whose origin is the same as the camera at time t but with the orientation of the camera at time $t-1$.

Each level of the decoder is composed of a preprocessing unit and a parallax refiner. The preprocessing unit is responsible for preparing the input to the parallax refiner at this level, refer to Figure 3.2 for an overview of the modules included in the preprocessing unit. Specifically, this unit performs the following operations:

- **Upscaling:** It upscales the parallax map ρ^{L-1} and the parallax features f_ρ^{L-1} estimated from the parallax refiner of the previous level by a multiple of 2 to match the resolution of the current level.

- **Split and Normalize:** the split layer subdivides the feature 3D matrices into K submatrices to decouple the relative importance between them. The normalize layer normalizes the features of each submatrix, allowing to leverage the information embedded in them that have low magnitudes due to the Leaky ReLU activation.

- **Spatial Neighborhood Cost Volume (SNCV):** It describes the two-dimensional spatial autocorrelation of the feature map. Each pixel of the cost volume is assigned the cost of matching the feature vector located in the same location in the feature map with its neighboring feature vectors within a given range, where the cost of matching two

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

vectors x_1 and x_2 of size N is defined as the correlation:

$$cost(x_1, x_2) = \frac{1}{N} x_1^T x_2 \tag{3.2}$$

In this way, the network should be invariant to changes in the feature vectors if they lead to the same cost volume. Thus, this makes the network focus more on the spatial structure of the image rather than the values of the features themselves, and, consequently, the network becomes robust and generalizable.

- **Parallax Sweeping Cost Volume (PSCV):** It is computed from two consecutive feature maps f_t and f_{t-1} and a parallax estimate ρ_t coming from the upscaled version of the estimated parallax map of the previous level. Each pixel in the cost volume is assigned the value of matching the feature vector in the same location in f_t with the corresponding feature vector in f_{t-1} after being reprojected to the current time t by the use of ρ_t and a range of values close to it. By searching through candidates in the range around ρ_t, it is possible to assess which parallax value at each pixel leads to the best feature matching, and thus it is more likely to be associated with this pixel.

- **Recompute Layer:** It recomputes the parallax values estimated in the previous time step using the camera transformation matrix and the warping operation [157] to provide a first estimate of the parallax values at the current step. Such a first estimate serves as a hint to the parallax refiner. Refer to the Appendix for a detailed explanation of the warping operation.

The parallax refiner, 3.2, is a stack of 7 convolutional layers responsible for giving an estimate of the parallax map at each level, given as input the preprocessed data generated by the preprocessing module.

Depth loss: The network is trained in an end-to-end fashion, and a scale-invariant L_1 loss is used to compute the loss between the predicted depth map $\hat{d}_i$ and the ground truth d_i. The loss is computed at each decoder level and then accumulated through a weighted sum across all levels, refer to Equation 3.3:

$$L_{depth} = \sum_{l=1}^{M} \frac{1}{N_p^l} \sum_{d_i^l} 2^{l+1} |log(d_i) - log(\hat{d}_i)| \tag{3.3}$$

where M is the number of decoder levels and N_p^l is the total number of pixels in the image at level l.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

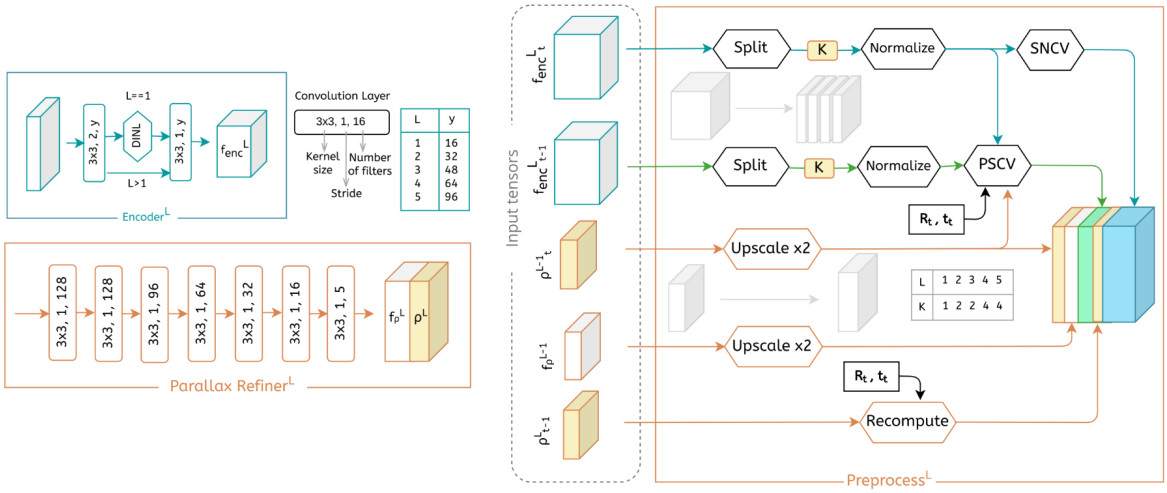


Figure 3.2 An illustration of the modules included in M4Depth. The preprocessing unit does not contain any learnable parameters. Feature vectors are subdivided into K sub-vectors using the split layer for subsequent parallel processing. The value of K at each level L is provided. For the encoder, the number of filters y at each level L is provided.

3.1.2 M4Semantic Network

Inspired by M4Depth, we propose a similar but simpler architecture for semantic segmentation depicted in Figure 3.3. Similar to M4Depth, our architecture is composed of an encoder and a decoder. The encoder is a stack of multiple levels and has a pyramidal structure where the resolution of the feature map is decreased while proceeding forward through the levels. The feature map predicted at each level is passed to its corresponding decoder level. Each decoder level is composed of a preprocessing unit and a semantic refiner in the place of the parallax refiner in M4Depth. The preprocessing unit prepares the input to the semantic refiner, and the semantic refiner at each level gives an estimate of the semantic segmentation map at a specific resolution. The resolution of the semantic map is scaled-up proceeding forward through the decoder levels.

In Figure 3.4 we show the modules used in our architecture. It can be noted that the modules in M4Semantic are less than the ones used in M4Depth. The encoder at each level is composed of 2 convolutional layers. In the first level, DINL [156] is added after the first convolution to increase the network's robustness to varied colors and luminosity conditions. ReLU activation is applied after each convolutional layer, and the resolution is decreased by a factor of 2 after each level. The pyramidal structure of the encoder helps in extracting both coarse and fine (global and local) features from the input image.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

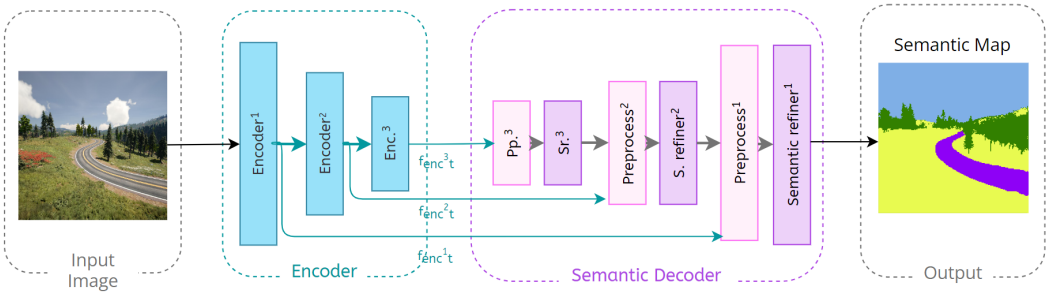


Figure 3.3 Our M4Semantic Architecture. It is composed of an encoder and a decoder module with a pyramidal structure. Each level of the decoder is composed of a preprocessing unit and a semantic refiner.

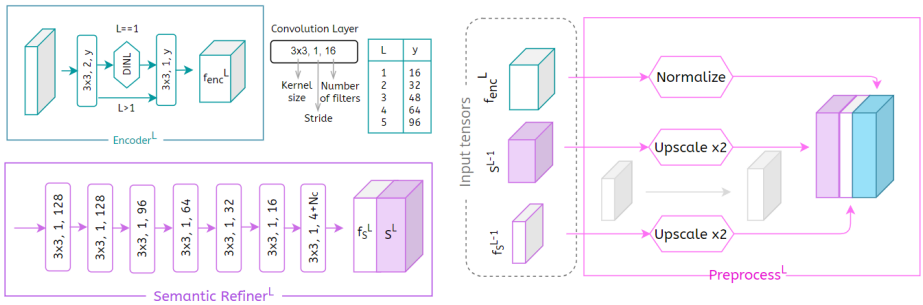


Figure 3.4 An illustration of the modules in our M4Semantic architecture. N_c is the number of semantic classes

The preprocessing unit at each decoder level is a pure computational unit with no parameters to be trained. It performs two operations:

- It upscales the semantic map S^{L-1} and the semantic features f_S^{L-1} estimated from the semantic refiner of the previous level by a multiple of 2 to match the resolution of the current level.
- It normalizes the feature map f_{enc}^{L} received from the encoder. This allows to leverage the information embedded in the feature map that has low magnitudes due to the Leaky ReLU activation applied after each layer of the encoder.

Similar to the parallax refiner, the semantic refiner at each level is composed of a stack of convolutional layers. The last convolutional layer has a depth of 4 (depth of the semantic features map) + N (the number of semantic classes). The output of the semantic refiner is a predicted semantic features map and an estimated semantic segmentation map. We apply Softmax activation on the predicted segmentation map to obtain a probability score for each class on every pixel.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Different from M4Depth, our M4Semantic architecture works on single images. We removed the time dependency in the semantic decoder because this produced 2 times faster results than the one with time dependency, with a slight drop in accuracy. Refer to the Architecture Study in section 3.3.3 for further clarification.

Semantic loss: The standard categorical cross-entropy loss is used on the predicted semantic maps at each level. The ground truths are resized using Nearest Neighbor interpolation to match the resolution of the predicted semantic maps at intermediate levels. Then, the losses are aggregated through a weighted sum, as shown in Equation 3.4.

$$L_{semantic} = \sum_{l=1}^{M} \frac{1}{N_p^l} \sum_{p_t^l} -log(p_t) \tag{3.4}$$

where M is the number of decoder levels, N_p^l is the total number of pixels in the image at level l, and p_t is the softmax score for the target class.

3.1.3 Joint Co-SemDepth Network

To merge the two previously described networks, M4Depth and M4Semantic, we adopt a multi-tasking shared encoder architecture [158; 159; 160; 161; 162]. The depth estimation and semantic segmentation networks share the encoder part for feature extraction, but each of them has its own decoder for their corresponding map prediction. An overview of our joint architecture is in Figure 3.5.

Loss Function: Our joint network is trained in an end-to-end fashion. The loss function for our architecture is defined as the weighted summation of the depth and semantic segmentation losses:

$$L_{total} = L_{depth} + w * L_{semantic} \tag{3.5}$$

where w is a weighting factor whose value was set after experimentation to 0.15. A discussion on the choice of w is reported in Section 3.2.2.

3.2 Experiments Setup

3.2.1 Datasets

It was stated before that there is a limited availability of annotated real datasets in the aerial domain compared to other application domains. In particular, they often lack appropriate ground truth for joint depth estimation and semantic segmentation tasks. For this reason, we

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

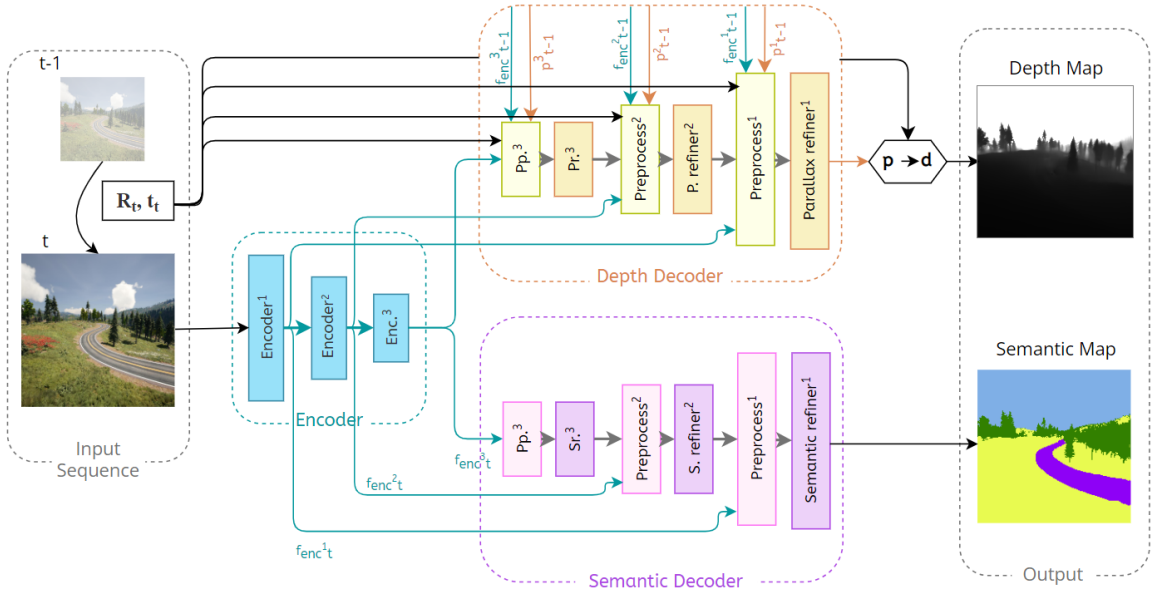


Figure 3.5 Our proposed joint architecture Co-SemDepth. It predicts both depth and semantic segmentation maps. It is composed of a shared encoder and two decoders. The encoder and the depth decoder are the same presented in [2]. The semantic decoder makes use of the encoded feature maps to give an estimate of the semantic segmentation map. The depth and semantic maps get scaled up as they go forward through the successive levels of the decoders. The number of shown levels here is 3, while in our experiments, we use 5 levels. Depth map is shown in gray scale

conduct our experiments on the joint architecture using the synthetic MidAir [76] dataset, and we use the real dataset AeroScapes [108], which contains only semantic segmentation annotation, for the validation of our semantic segmentation network. See Figure 3.6 and Figure 3.7 for sample images and annotations from both datasets.

MidAir [76] is a synthetic dataset collected using the AirSim simulator [32], consisting of 420K forward-view RGB video frames captured at low altitude in outdoor unstructured environments with various weather conditions. It contains annotations of depth maps, semantic segmentation, surface normals, stereo disparity, and camera locations. Hence, this dataset is suitable for training and testing our joint Co-SemDepth architecture.

We adopt the train-test split used in [2], but we select 8 trajectories that cover a variety of conditions to create validation data, and we provide the data split on our Github page. In the evaluation, the depth values are capped at 80.0 meters. We resize images to a resolution of 384x384. In the original semantic annotation of MidAir, there are 14 semantic classes: Sky, Animals, Trees, Dirt Ground, Ground Vegetation, Rocky Ground, Boulders, Empty, Water, Man-Made Construction, Road, Train Track, Road Sign, and Others. Since several classes

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

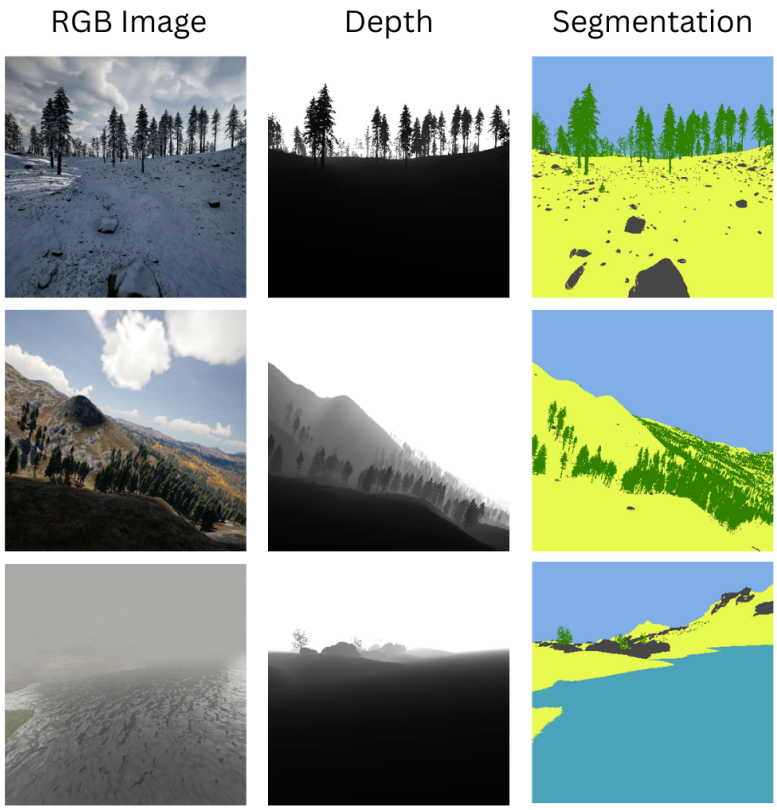


Figure 3.6 Sample images and annotations from the MidAir dataset

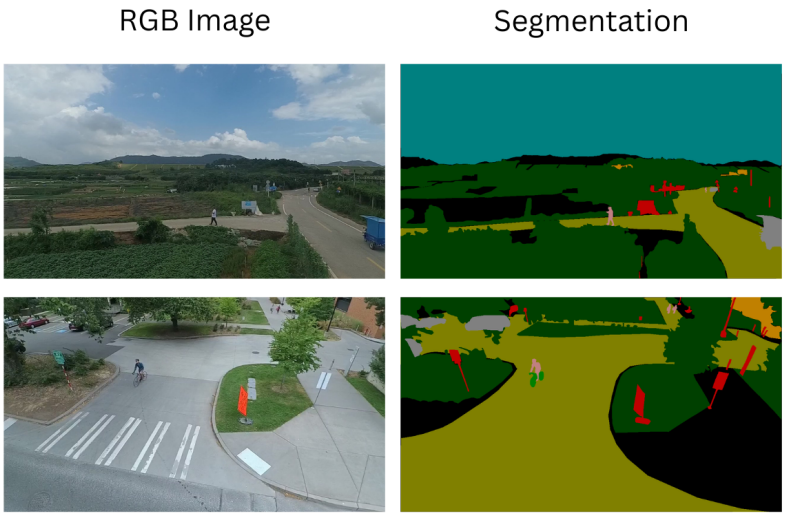


Figure 3.7 Sample images and annotations from the AeroScapes dataset

are visually indistinguishable and some of them are very small, we map them to a smaller set of 7 semantic classes: Sky, Water, Land, Trees, Boulders, Road, and Others. Specifically,

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

we considered Ground Vegetation, Rocky Ground, and Dirt Ground as Land, and Animals, Empty, Train Track, and Road Sign in Others.

Aeroscapes [108] is a real dataset collected using drones at low-mid altitude in various outdoor environments. It consists of 3,269 images with an 80%-20% train-test split and a resolution of 1280x720. This dataset contains only semantic segmentation annotation. For this reason, we can not use it for the training of our joint architecture. However, we use Aeroscapes for the training and testing of the individual M4Semantic network.

3.2.2 Implementation Details

We adopt Adam optimizer with the default momentum parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) and a fixed learning rate of 10^{-4}. We apply image augmentation of random rotation, flipping, and changing color (contrast, brightness, hue, and saturation) during training, and we train with a batch size of 3 and a number of epochs $= 60$. After training, we choose the checkpoint that produced the best validation results for evaluation on the test set.

Our workstation has 16GB RAM, an Intel Core i7 processor, and a single NVIDIA Quadro P5000 GPU card running CUDA11.4 with CuDNN 7.6.5 and Ubuntu OS. Due to its memory-limited resources, depth and semantic maps are predicted at a resolution equal to half the input resolution, and then Nearest Neighbour interpolation is applied on the output maps to scale up their resolution to the original size. As reported in [163], decreasing the image resolution can slightly decrease the accuracy; however, it gives the advantage of reducing the computational runtime and memory footprint.

To quantitatively evaluate the depth prediction results, we consider the commonly used evaluation metrics in prior works [2; 9; 24]. These include the linear root mean square error (RMSE) (Equation 3.6), the absolute relative error (Equation 3.7), and accuracy under a threshold (Equation 3.8).

$$RMSE = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p} |\hat{d}_i - d_i|} \tag{3.6}$$

$$AbsRelErr = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p} |\hat{d}_i - d_i|} \tag{3.7}$$

$$max(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}) = \delta < T \tag{3.8}$$

where $\hat{d}_i$ is the estimated depth, d_i is the ground truth, N_p is the number of pixels in the image, and T is the set thresholds: $(1.25, 1.25^2, 1.25^3)$.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

For semantic segmentation, we adopt the commonly used mean Intersection over Union $mIoU$ metric (Equation 3.9).

$$mIoU = \frac{TP}{TP+FP+FN} \tag{3.9}$$

where TP, FP, and FN are, respectively, the number of true positives, false positives, and false negatives at the pixel level. The Inference Time (Inf. Time) is computed in milliseconds per frame (ms/f).

In Figure 3.8, the loss curves for M4Depth and M4Semantic networks are shown. The curves show the difference in the range of loss values between the two tasks due to the difference in the loss function used: L1 loss and categorical cross-entropy loss. We incorporate a weighting factor $w = 0.15$ to force the loss values for semantic to lie within the same range of losses for depth, and thus ensure a comparable contribution for the two losses during training of the joint model.

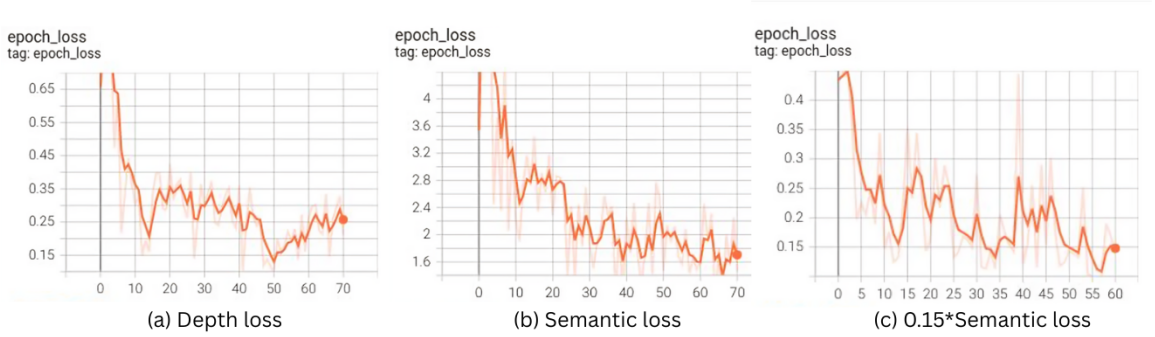


Figure 3.8 Loss curves showing the difference in the loss range of values for the two tasks during training. After multiplying the semantic loss (b) by a weighting factor of 0.15 (c), its range of values became comparable to the depth loss (a).

3.3 Results

In this section, we discuss the experiments we conducted to validate the effectiveness of our joint architecture on various datasets. First, we evaluate the effectiveness of using our joint architecture compared to using the two single architectures. Then, we benchmark our model against other state-of-the-art single and joint methods. Finally, we make a comparative study of architectural design alternatives. Code is available on our Github repository: https://github.com/Malga-Vision/Co-SemDepth

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

3.3.1 Joint vs Single Architectures

We conduct experiments to compare the performance of our joint architecture Co-SemDepth with the two single ones: M4Depth and M4Semantic. Each architecture is trained equally for 60 epochs. The results are reported in Table 3.1. We can notice that the accuracy values (in terms of depth and semantic metrics) of the joint architecture are close to the single ones. Depth is a bit better using Co-SemDepth, and this can signify that adding the shared encoder that extracts features related to both depth estimation and semantic segmentation has helped to enhance the results of depth estimation, probably because it has enriched the features entering the parallax decoder part with features related to semantic segmentation. While for semantic segmentation, the results using M4Semantic ($mIoU = 76.8\%$) are a bit better than using the joint architecture $mIoU = 75.44\%$. From this, we can say that the shared encoder in the joint architecture did not enhance the results of semantic segmentation; instead, using a dedicated encoder for extracting solely the semantic segmentation features led to a better accuracy.

While it was found in other works [158] that depth features can enhance the semantic segmentation accuracy, it was mentioned in [118] that multi-tasking architectures can perform poorly compared to dedicated single ones. We thereby think that this enhancement or degradation can differ depending on the dataset used and its semantic classes.

The inference time of the joint architecture (49.6 ms/f) is *lower than* the sum of m4Semantic and m4Depth ($44.9 + 9.8$ ms/f). Moreover, the number of parameters of the joint architecture (5.2 Million) is less than the sum of the two single ones (3.06 + 2.61 Million) by around 500K parameters.

The above signifies that using our joint architecture Co-SemDepth is more effective in terms of computational time and memory footprint than using the two single architectures while achieving very close accuracies. The trade-off between accuracy and computational cost in the multi-tasking architectures was previously discussed in [118], and it can differ depending on the environment. The high inference time of the depth branch compared to the segmentation branch can be due to the added computations of parallax in the depth branch and the additional modules for computing the cost volumes SNCV and PSCV, which are not present in the segmentation branch.

During inference, Co-SemDepth required only *6.2GB of GPU memory* while running M4Depth and M4Semantic together required 14.6GB of GPU memory. This makes Co-SemDepth compatible to run on microcontrollers that have only 8GB RAM and that are widely used in robotics hardware due to their affordable cost.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Table 3.1 Evaluation of our joint vs single architectures for depth estimation and semantic segmentation on the MidAir dataset.

Architecture	Output	Params(M)	Inf. Time (ms/f)	Semantic	Depth				
				mIoU ↑	RMSE ↓	RelErr ↓	$\delta 1$ ↑	$\delta 2$ ↑	$\delta 3$ ↑
M4Depth	D	3.06	44.9	-	6.99	0.109	92.0%	95.4%	97.0%
M4Semantic	S	2.61	9.8	**76.80%**	-	-	-	-	-
Co-SemDepth	D+S	5.2	49.6	75.44%	**6.70**	**0.096**	**92.3%**	**95.7%**	**97.2%**

3.3.2 Benchmarking

Benchmarking on MidAir:

We compare the performance of Co-SemDepth with other open-source state-of-the-art methods. We compare it with both single and joint architecture methods. Table 3.2 summarizes the training parameters used for each method. For each method, we fix the input image size to 384x384 and the number of training epochs to 60 or a maximum of 80k iterations (for Segformer and TaskPrompter). Maximum depth is set to 80 meters. Other parameters are kept as the default.

- For FCN, we implement FCN-32S [164] and we use two backbone networks; VGG16 [165] and MobileNetV2 [166].
- For SegFormer [14], SegFormer-B0 is used. It should be noted that we had difficulties training and testing higher versions of SegFormer on our server due to their large size and complexity.
- For TaskPrompter [167], the vision transformer Base model was set as the backbone with an embedding dimension of 384 and a number of channel heads equal to 8. All the other values were kept as the default.
- For DepthAnything [17], we use the small model of DepthAnything-V2 because it has the lowest number of parameters and to be compatible to run on our machine. The trained model dedicated to outdoors, "Outdoor Virtual KITTI2", was selected, and zero-shot prediction was performed.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Table 3.2 Training parameters used for benchmarking the baseline methods on MidAir.

Method	optimizer	lr	sched wt decay	max iter	epochs	batch
FCN	Adam	1×10^{-4}	-	-	60	4
ERFNet	Adam	5×10^{-4}	1×10^{-4}	-	60	6
SegFormer-B0	Adam	6×10^{-5}	1×10^{-2}	80K	-	4
RefineNet	Adam	1×10^{-4}	1×10^{-4}	-	60	8
TaskPrompter	Adam	1×10^{-5}	1×10^{-6}	80K	-	3

Table 3.3 Benchmarking Co-SemDepth on MidAir against other state-of-the-art methods in both depth estimation (D) and semantic segmentation (S). The top part reports single depth estimation methods, the middle part for single segmentation methods, and the bottom part for joint methods. * means the depth metrics values were reported in [2]. @ means zero-shot prediction.

Method	Output	Params(M)	Inf. Time (ms/f)↓	Semantic	Depth				
				mIoU ↑	RMSE ↓	RelErr ↓	$\delta 1$ ↑	$\delta 2$ ↑	$\delta 3$ ↑
MonoDepth2*	D	14.8	23.9	-	12.35	0.394	61.0%	75.1%	83.3%
ST-CLSTM*	D	15.04	35.3	-	13.69	0.404	75.1%	86.5%	91.1%
ManyDepth*	D	46.3	82.9	-	10.92	0.203	72.3%	87.6%	93.3%
PWCDC-Net*	D	9.4	25.8	-	8.35	0.095	88.7%	93.8%	96.2%
DepthAnythingV2@	D	24.8	75.2	-	33.37	0.640	12.3%	25.6%	39.7%
FCN(VGG16)	S	14.7	58.3	72.93%	-	-	-	-	-
FCN(MobileNetv2)	S	2.2	60.5	69.82%	-	-	-	-	-
ERFNet	S	2.07	19.1	77.40%	-	-	-	-	-
SegFormer-B0	S	3.8	49.1	75.10%	-	-	-	-	-
RefineNet	D+S	3.0	74.2	72.70%	9.74	0.200	74.9%	89.0%	94.5%
TaskPrompter	D+S	126	120.6	**80.20%**	9.80	0.250	50.0%	78.3%	90.0%
Co-SemDepth	D+S	**5.2**	**49.6**	75.44%	**6.70**	**0.096**	**92.3%**	**95.7%**	**97.2%**

The benchmarking results are reported in Table 3.3. From the table, we can clearly notice that our method outperforms the other joint networks in depth accuracy and inference time. While its semantic segmentation accuracy is lower than TaskPrompter, Co-SemDepth has a notably lower inference time and model size, making it more convenient for hardware deployment. The inference time of the network can be further enhanced by converting the model to ONNX[1] or using tensor RT[2].

Compared to the single depth estimation networks, Co-SemDepth could maintain its superior accuracy in depth estimation that was reported in [2]. This indicates that transforming M4Depth to the joint Co-SemDepth did not have a negative effect on its depth estimation

[1] https://onnx.ai/onnx/intro/converters.html

[2] https://docs.nvidia.com/deeplearning/tensorrt/latest/getting-started/quick-start-guide.html

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

performance compared to the state-of-the-art. In addition, Co-SemDepth has a notably smaller number of parameters compared to the other single depth estimation methods.

For the single semantic segmentation networks, Co-SemDepth has a competitive mIoU with the others, only slightly inferior to ERFNet, which is, in any case, a dedicated architecture, not a joint one.

For further analysis of semantic segmentation performance, we compute the per-class IoU for each method. The per-class IoU evaluation can be found in Table 3.4. It can be noted that the high mIoU of TaskPrompter is mostly caused by its good detection of the class "Others" (20% better than other methods). This can be justified by the fact that the "Others" class contains a variety of objects (train track, road sign, animals, and others), and such variety requires a model of high capacity to learn it well (TaskPrompter has 126M parameters).

Table 3.4 Per-Class IoU Evaluation of Co-SemDepth architecture and other baseline methods on MidAir.

Method	Sky	Water	Trees	Land	Boulders	Road	Others	**mIoU**
FCN(VGG16) [168]	88.56%	83.12%	75.50%	82.30%	29.97%	88.04%	54.60%	72.93%
FCN(MobileNetv2) [168]	87.82%	82.42%	73.42%	81.28%	26.57%	84.64%	46.09%	69.82%
ERFNet [13]	91.50%	87.64%	82.48%	85.10%	40.63%	90.90%	63.70%	77.42%
SegFormer-B0 [14]	90.54%	88.58%	79.70%	83.33%	30.13%	92.19%	61.20%	75.10%
RefineNet [160]	89.70%	81.60%	79.70%	82.20%	30.15%	91.36%	54.10%	72.69%
TaskPrompter [167]	90.60%	87.10%	79.10%	85.30%	41.00%	95.30%	83.20%	**80.23**%
Co-SemDepth	90.10%	86.40%	79.95%	82.60%	33.10%	94.60%	59.25%	75.44%

Visualization: The qualitative visualization of the depth map predictions of M4Depth and other methods on MidAir was already done in [2]. Here, we do a qualitative visualization of the semantic map predictions of the different methods. The outputs can be found in Figure 3.9. We can notice that all networks could capture the overall semantic layout of the input images; the location of trees, roads, water, land, and sky. However, Co-SemDepth and ERFNet are remarkably better in capturing the details (notice the trees in the first row and the train track in the second row). Compared to Co-SemDepth, ERFNet and TaskPrompter are better in capturing some of the faraway objects (for example, the distant boulders and bushes in the third row). While TaskPrompter has a higher *mIoU* than Co-SemDepth, Co-SemDepth is notably better at recognizing the details of the trees, road, and rocks than TaskPrompter.

Generalization: We conduct some experiments for assessing the generalization capability of the network. The datasets TartanAir [77] and WildUAV [121] were used for testing. *TartanAir* is a synthetic dataset captured using AirSim in a multitude of outdoor and indoor environments with depth and mesh segmentation annotations. It also includes camera pose information. For this reason, it could be used for testing Co-SemDepth generalization. How-

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

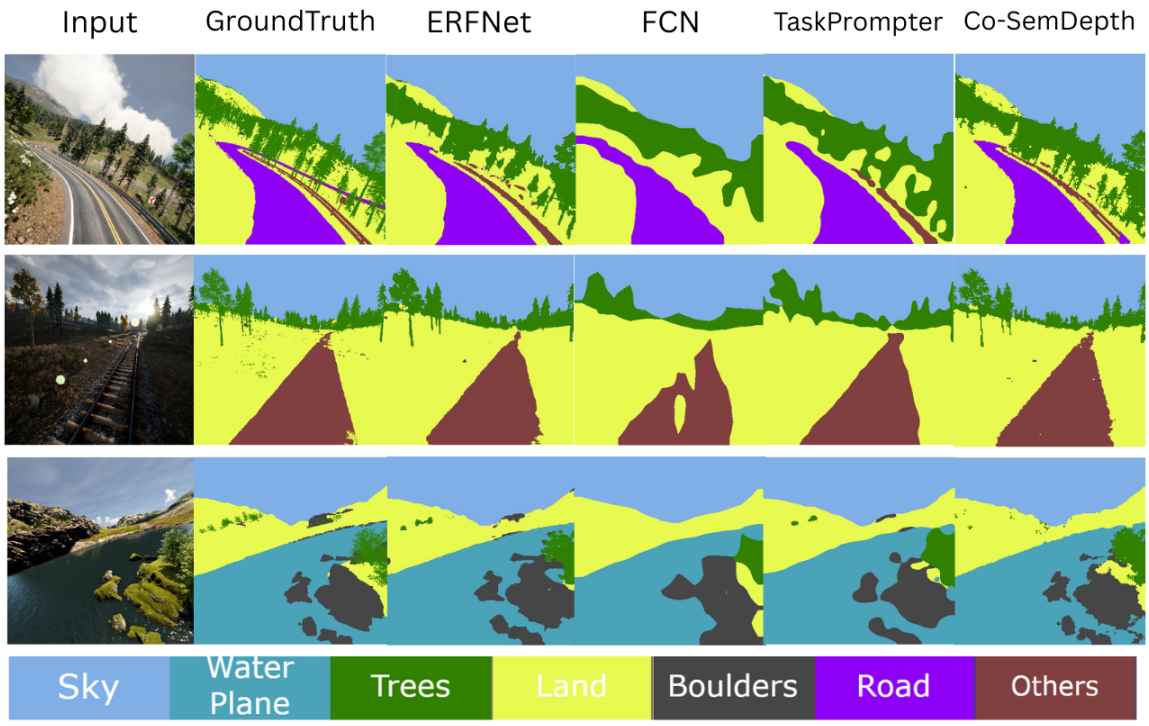


Figure 3.9 Qualitative evaluation of the semantic map predictions of ERFNet, FCN MobileNet, TaskPrompter, and Co-SemDepth respectively, on sample images from MidAir dataset.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

ever, the evaluation was done only qualitatively due to the absence of per-class segmentation in the dataset. The trajectories of the two environments: Gascola and SeasonForestWinter were used in the evaluation due to their similarity with the scenes of MidAir. On the other hand, *WildUAV* is a real dataset captured in an outdoor rural environment. It contains both depth and per-class semantic segmentation annotations, and it includes camera pose information. Therefore, it could be used to test the generalization capability of Co-SemDepth both quantitatively and qualitatively. All the trajectories of WildUAV were used due to their similarity with MidAir environments.

In Figure 3.10, a visualization of the prediction on sample images from the synthetic TartanAir [77] and the real WildUAV [121] datasets using Co-SemDepth (trained on MidAir) is shown.

The semantic segmentation prediction of TartanAir images went well, and trees, land, rocks, and sky were correctly segmented. However, for depth estimation of the first image, there was no big difference in the depth values predicted for trees and land. It is apparent that the network was able to predict the depth of the sky correctly, but it was unable to distinguish the relative depth of the trees and land, and they were all predicted with very close values. For depth estimation of the second image, the network was able to distinguish between the depths of the close and far trees and the land. However, due to the strange color of the sky in the winter, the network predicted part of the sky wrongly with near depth values.

In WildUAV, in the third image, the semantic segmentation was overall correct except that the vehicle was detected as a rock. The depth estimation of the third image could distinguish between the relatively far distance of the road compared to the tree area. In the fourth image, there is an apparent confusion between Water and Sky in the semantic segmentation. This can be due to the visual similarity of these two classes. In addition, the training data contained only forward camera-view images where the sky is always present in the top part of the images, while in WildUAV, the images are captured from a nadir view where the sky is not present in the images. For this reason, it can be noticed that the network predicted the top part of the lake as Sky while the bottom part was predicted correctly as Water. The depth of the water area was predicted with a far distance in the depth output, as the water was treated wrongly as sky.

Quantitatively, the generalization of the model is assessed on WildUAV. Co-SemDepth was trained on the synthetic MidAir+TopAir datasets and then tested on WildUAV. The 16 semantic classes of wildUAV were mapped to our set of 7 classes as the following: sky $\rightarrow$ sky, water $\rightarrow$ water, {deciduous trees, coniferous trees, fallen trees}$\rightarrow$ trees, {dirt ground,

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

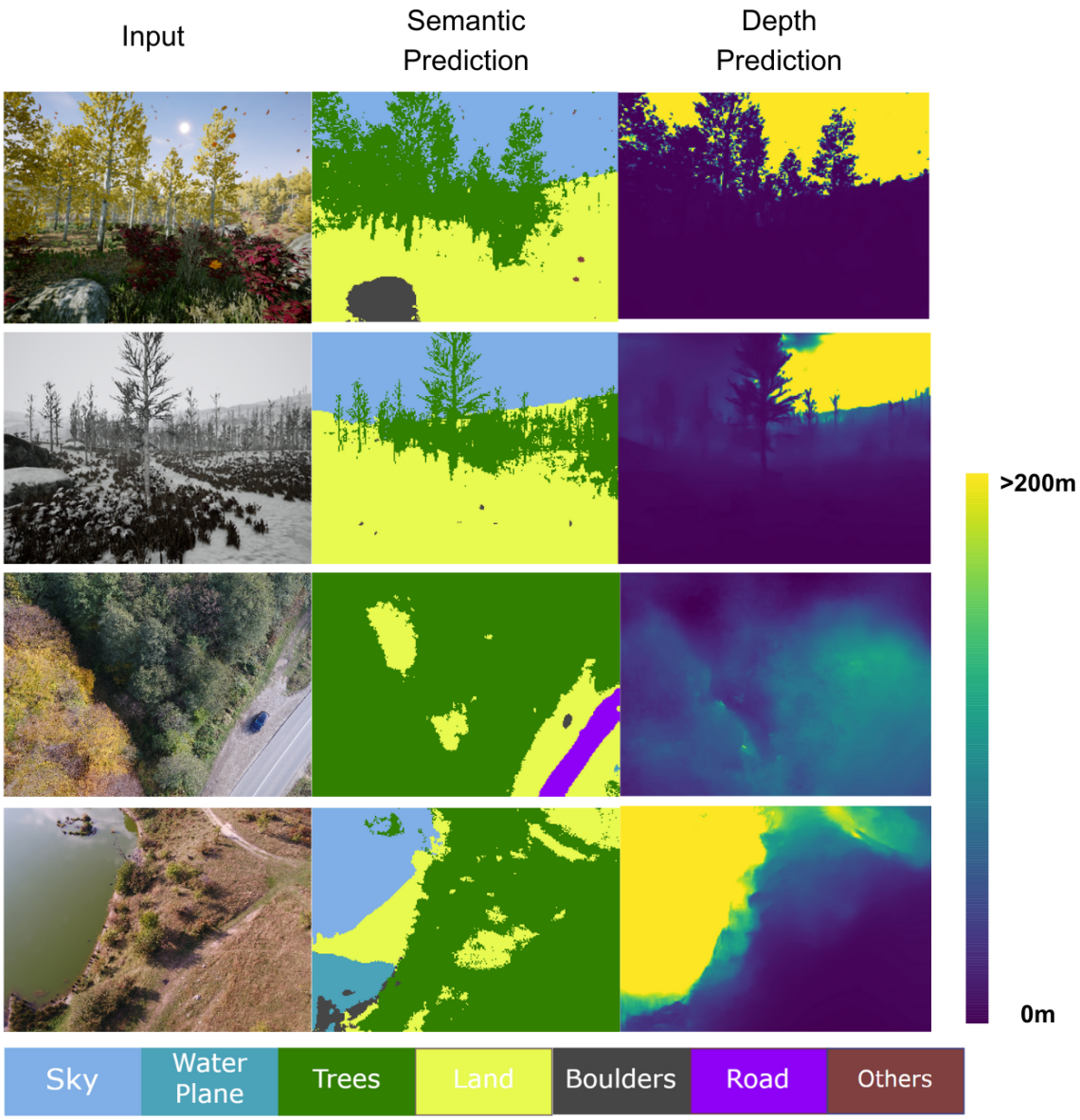


Figure 3.10 Qualitative evaluation of the zero-shot generalization performance of MidAir-trained Co-SemDepth on sample images from TartanAir (top and 2nd row) and WildUAV (3rd and bottom rows).

ground vegetation}→land, rocks → rocks, {road, sidewalk}→ road, {static car, moving car, building, fence, people, empty}→ others.

The depth estimation results can be found in Table 3.5, and the semantic segmentation results in Table 3.6. While the depth estimation results are acceptable, the semantic segmentation results are only acceptable for trees and land classes.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Table 3.5 Depth estimation results on WildUAV obtained using Co-SemDepth tained on MidAir+TopAir

RMSE ↓	AbsRelErr ↓	$\delta 1 \uparrow$	$\delta 2 \uparrow$	$\delta 3 \uparrow$
14.9	0.303	58.7%	80.7%	91.5%

Table 3.6 Semantic segmentation results on WildUAV obtained using Co-SemDepth tained on MidAir+TopAir

Sky	Water	Trees	Land	Rocks	Road	Others	mIoU
-	0.9%	32.8%	34.8%	-	14.9%	0%	16.7%

Aeroscapes:

As mentioned in the Experiments Setup, the AeroScapes dataset is used to evaluate the performance of the individual M4Semantic network because the dataset only contains semantic segmentation annotations. The M4Semantic results are reported in Table 3.7 and compared with other methods. Our network was trained for 200 epochs with a batch size of 3 and a learning rate of 10^{-4} for the first 70 epochs and then decreased to 10^{-5}.

M4Semantic was implemented on TensorFlow as one whole model that can be trained in an end-to-end fashion without separation between the weight files of the encoder and the decoder. This led to a more compact code but limited us from pretraining the encoder separately on Imagenet, as was done in the other methods. Nevertheless, we could produce a competitive mIoU compared to the others (50.4% compared to the best 52%), as can be seen in Table 3.7. In Figure 3.11, a visualization of the output semantic segmentation maps predicted by M4Semantic on samples from the AeroScapes dataset is presented. It can be seen that the network can capture the layout of the input images and detect well the roads, trees, sky, and background areas.

3.3.3 Ablation experiments

We conduct architecture study experiments on M4Semantic to highlight the importance of the addition or ablation of different modules. The results can be found in Table 3.8. From the top part of the table, we can notice that using 5 levels produced the highest *mIoU* (76.8%) while keeping the inference time relatively low (9.8 ms per frame).

In the bottom part, Original means M4Semantic (5 level). In Original+{SNCV}, we used the Spatial Neighbourhood Cost Volume module used in the decoder of M4Depth,

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

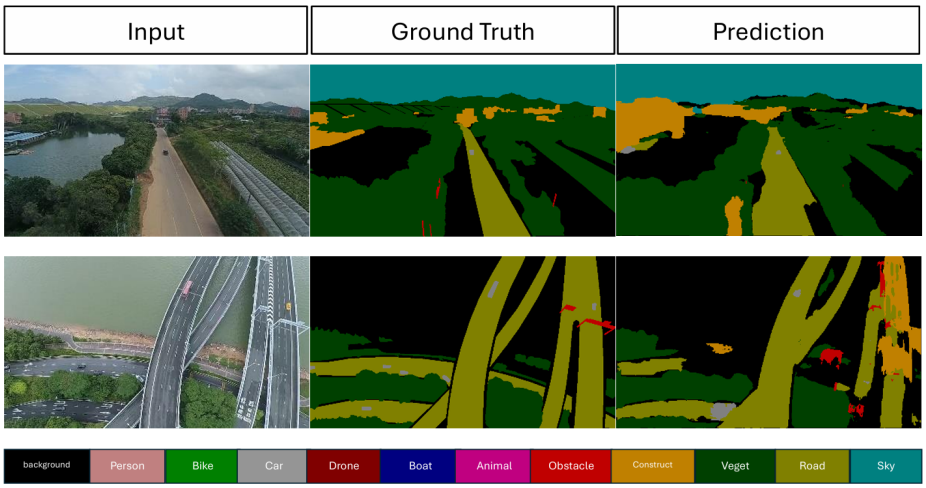


Figure 3.11 Visualization of the predicted semantic maps using M4Semantic on sample images from the Aeroscapes test data.

Table 3.7 Comparison of our M4Semantic architecture with other semantic segmentation methods benchmarked on Aeroscapes. P means pretrained on other datasets and S means trained from scratch

Method	Params(M)	P/S	Open-Source	mIoU ↑
FCN-8S [109]	14.7	P	No	43.12%
FCN-16S [109]	14.7	P	No	44.52%
FCN-32S [109]	14.7	P	No	45.51%
FCN-ImageNet-4S [108]	14.7	P	No	48.96%
FCN-Cityscapes [108]	14.7	P	No	49.55%
FCN-ADE20K [108]	14.7	P	No	51.62%
FCN-PASCAL [108]	14.7	P	No	52.02%
M4Semantic (Ours)	2.61	S	Yes	50.40%

see Figure 3.2, on the encoded feature maps instead of adding the normalized feature maps directly in the preprocessing unit. Such a module measures the two-dimensional spatial autocorrelation of the scene and improves the performance in depth estimation. However, using such a module in M4Semantic did not improve the performance in semantic segmentation and, moreover, it increased the inference time. So, we decided to discard it in semantic segmentation.

In Original+$\{S^{t-1}\}$, we test the addition of time dependency on previous frames in M4Semantic. At each decoder level, the semantic segmentation map predicted from the previous frame S^{t-1} is used along with the camera motion information and the ground truth depth map to warp it and give an initial prediction of the semantic map of the current frame. For more details about the warping operation, see the Appendix. We concatenate such a

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

warped map with the outputs of the preprocessing module, Figure 3.4, at each decoder level in order to act as an initial hint for the semantic segmentation map at the current time step. While such a technique achieved a higher mIoU 78.4%, the inference time increased due to the added warping computation. Also, the warping of the segmentation map requires the ground truth depth map, and this is not guaranteed to be available in reality. For these reasons, we decided to discard such a module in the semantic segmentation network.

Given this architecture study and the one done on the single depth estimation architecture [2], we choose the number of levels of Co-SemDepth to 5 levels, and we adopt the architecture depicted in Figure 3.5.

Table 3.8 Evaluation of our M4Semantic architecture on MidAir with the addition (+) or ablation (-) of different modules. The top part evaluates choosing a different number of levels. The bottom part was performed on M4Semantic (5level).

Architecture	Inf. Time (ms/f) ↓	mIoU ↑
M4Semantic (4level)	9	75.1%
M4Semantic (5level)	9.8	**76.8**%
M4Semantic (6level)	10.9	74.9%
Original-{DINL}	9.5	75.6%
Original-{Normalize}	9.7	74.1%
Original+{SNCV}	17.7	71.9%
Original+{S^{t-1}}	16.7	78.4%

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 4

Synthetic-to-Real Analysis

Summary

In this chapter, the goal is to assess the synthetic-to-real performance of joint networks on aerial datasets. First, a detailed description of our collected *TopAir* dataset is provided in section 4.1. Then, a brief recounting of the methods used in the synthetic-to-real experiments is presented. The setup used to conduct the experiments is introduced. Then, multiple experiments are discussed to test the effect of different training modalities on closing the gap between the synthetic and real domains.

To summarize, TopAir is a synthetic aerial dataset collected using AirSim in a variety of outdoor environments, comprising $\sim 10K$ video frames with annotations of depth, segmentation, and camera translation-rotation data. To assess the synthetic-to-real domain shift, two networks with different backbone architectures are employed in the experiments for comparison: Co-SemDepth (U-Net backbone) and TaskPrompter (ViT backbone). The obtained results show that, generally, Co-SemDepth is more robust in the depth estimation generalization, while TaskPrompter (which has high capacity) is better in semantic segmentation. In the zero-shot learning scheme, changing the synthetic data used for training affects the generalization performance of the network depending on the similarity of the scenes between the synthetic and real data. In the few-shot learning scheme, both depth estimation and semantic segmentation benefit from the addition of a small percentage of real data to the training of the network.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

4.1 TopAir Data Collection

As was discussed in Chapter 2, only a few datasets are available in the aerial field that contain both depth and semantic segmentation annotations. Such datasets do not cover all possible variations of outdoor environments, camera viewpoint, altitude, tilting angle, weather conditions, and lighting. We seek to contribute to enriching the available aerial data, especially those annotated, by collecting a new synthetic aerial dataset covering as much variations as possible. To this purpose, we use *AirSim* simulator integrated with *UnrealEngine4* to collect the data and call it *TopAir*. AirSim [32] is an open source simulator for drones, cars, and more that is built on the game development engine UnrealEngine [41]. The simulator is well-documented, and it has multiple modes for controlling the drone or the vehicle in the simulation. In our data acquisition, we use the Computer Vision mode that simplifies the physics of the drone and treats it as a camera moving in space. By using this mode, we could navigate using only the keyboard without the need for a particular remote controller of actual drones. We adjusted the settings to collect RGB images, depth maps, and semantic segmentation maps. The setting values we tuned can be found in Table 4.1. All other settings were kept as the AirSim default.

Table 4.1 Settings adjusted of AirSim during data acquisition

Setting	Value
SimMode	ComputerVision
ImageType	0, 3, 5
RecordOnMove	True
RecordInterval	0.05 (Sec)
Width	384
Height	384
FOV_Degrees	90
AutoExposureSpeed	100
AutoExposureMaxBrightness	0.64
AutoExposureMinBrightness	0.03

As the name *TopAir* suggests, the data have been collected with a nadir (top) camera viewpoint where the tilting (pitch) angle is close to or equal to 90 degrees. The motivation to use this viewpoint, particularly, is the limited available synthetic data with the top view; only SkyScenes contains top view images, while MidAir, TartanAir, and SynDrone contain mostly images with forward or oblique view. The dataset is available at: https://huggingface.co/datasets/yaraalaa0/TopAir

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

During the collection, the camera is set at low, mid, and high altitudes varying from 10 to 100 meters above the ground. A multiplicity of UnrealEngine environments, rural and urban, were used for the data collection. They cover a variety of weather and daytime conditions, as well as object and building styles and layouts. These environments are:

- *Africa* (3 trajectories): an African forest containing mostly trees and sandy ground. The lighting is around sunset.
- *City Park* (7 trajectories): a park in the city that has trees, grass, roads for cars and pedestrians, lakes of greenish water, kids' areas, bridges, and playgrounds. The lighting in the environment is around midday.
- *Oak Forest* (3 trajectories): forest of mid-height trees and muddy ground. There exist several small rocks in the ground. Lighting is around the afternoon.
- *Rural Australia* (4 trajectories): an environment representing a rural area in the deserts of Australia. The environment has roads, where cars and vehicles were placed manually, and trees of different heights on the side of the roads.
- *Assetsville Town* (15 trajectories): an environment that represents an urban scenario where there are multiple houses, vehicles, shops, government buildings, police stations, farms, and people. Some houses, buildings, cars, and persons were inserted in the environment to enrich its diversity. The lighting here is early to midday sunlight.
- *Accustic Cities* (3 trajectories): Modern high buildings with roads in the middle and a lake of brownish water with a bridge on it. Lighting is strong around midday.
- *Downtown City* (7 trajectories): a small downtown city with low-height buildings and roads in the middle. Tables and decorations are placed on the sides and center of the pedestrian roads. Some cars were inserted manually in this environment. The lighting is around the afternoon.
- *Edith Finch* (2 trajectories): A small house with a lake and a truck in front of it and tall trees around. Lighting is around evening (moonlight
- *Landscape Mountains* (1 trajectory): Rocky mountains and rocky ground with a lake of bluish water in the middle. Lighting is around midday.
- *Nature Pack* (1 trajectory): a small natural environment of a waterfall and trees with grassy land. Rocks rest on the sides of the waterfall. The lighting in this environment is midday.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

- *Neighbourhood* (8 trajectories): a rural neighbourhood area with many houses of blue roofs. Some houses have swimming pools. Cars of different colors are on the roads. The lighting is afternoon sunlight.

An additional advantage of TopAir is that it is a light-weight dataset that has an image resolution of 384×384 and a total size of only 4GB, making it convenient for downloading and training on resource-restricted machines. Samples of RGB images from the collected TopAir data can be found in Figure 4.1.

The dataset comprises 54 trajectories containing, in total, 10,385 frames. Each RGB frame is annotated with a semantic segmentation map, a depth map, and camera transformation (location+orientation) data. To generate the semantic segmentation annotations, the object classes in each environment were limited to a set of 9 classes by manually renaming the elements in UnrealEngine to represent these classes. While such a process required a significant effort, it had to be done only once for each environment. After that, we could collect as much data as needed from the environment. The semantic segmentation classes we used are:

1. *Sky*: the sky
2. *Water*: includes all water surfaces, including lakes, seas, or swimming pools
3. *Trees*: all trees and vegetation above the ground, whether short or tall, excluding grass
4. *Land*: all ground types (dirt ground, rocky ground, grass, sand, etc)
5. *Vehicle*: all types of ground vehicles (car, trucks, bus, etc) excluding bikes
6. *Rocks*: boulders and rocks (including rocky mountains)
7. *Road*: Lanes, streets, paved areas on which cars drive, or sidewalks for pedestrians
8. *Building*: buildings, residential houses, and constructions, including bridges
9. *Others*: any other object that is not included in the above classes

The settings of AirSim were adjusted to collect the needed annotations (depth + segmentation + location) automatically while navigating through the environments. A median filter of size 10 was applied to all the generated depth and segmentation maps to smooth their appearance and remove the noise. Such a filter computes and assigns the median value to all the pixels included in a square window of size 10×10, and this helps in removing the noisy

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

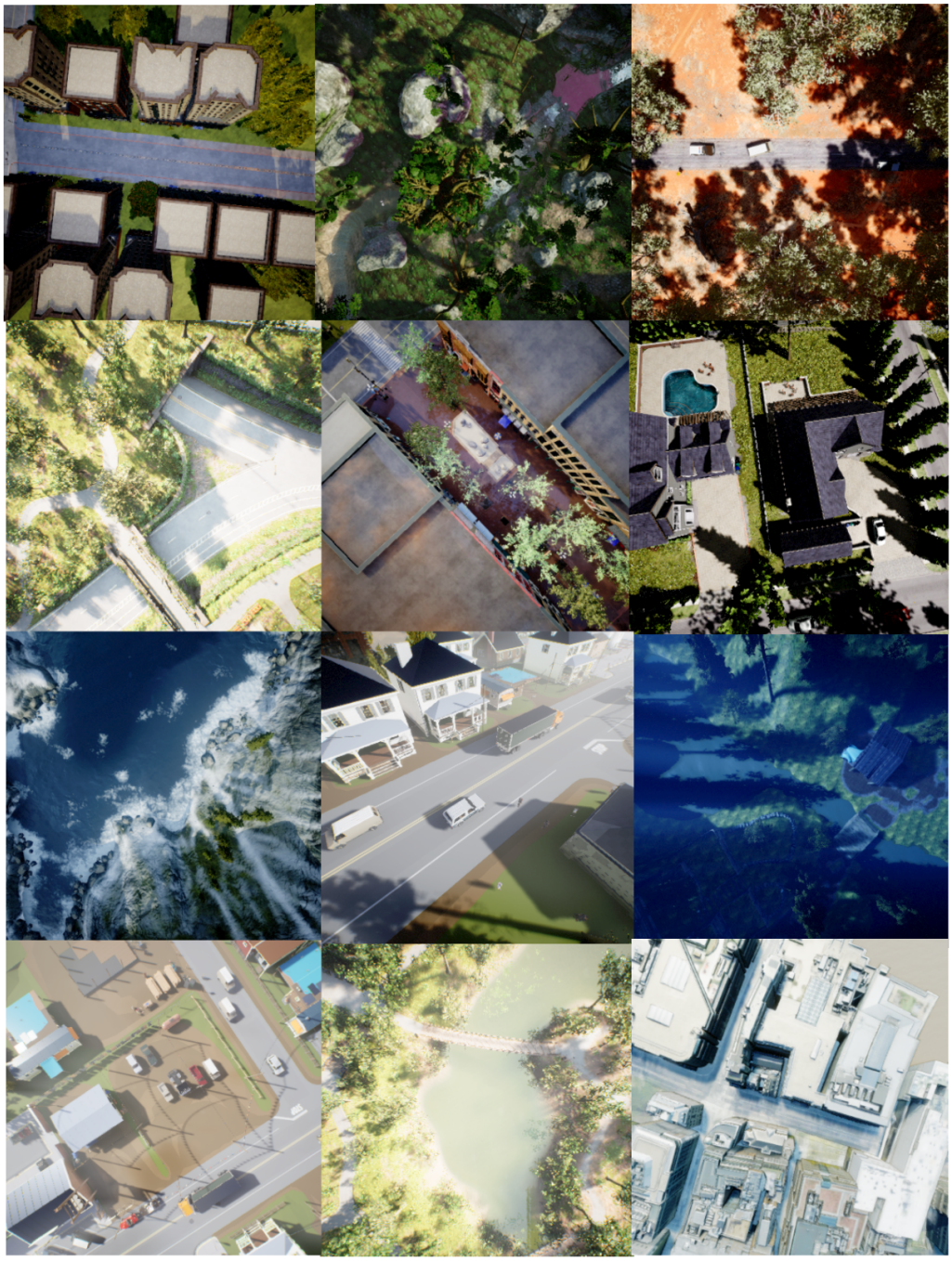


Figure 4.1 Sample images from the collected TopAir dataset showing the variety of environments used and the variety of altitudes

non-frequent pixels from the segmentation map (they get assigned the most frequent value within the 10×10 window). See Figure 4.2 for a demonstration of the maps before and after applying the filter.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

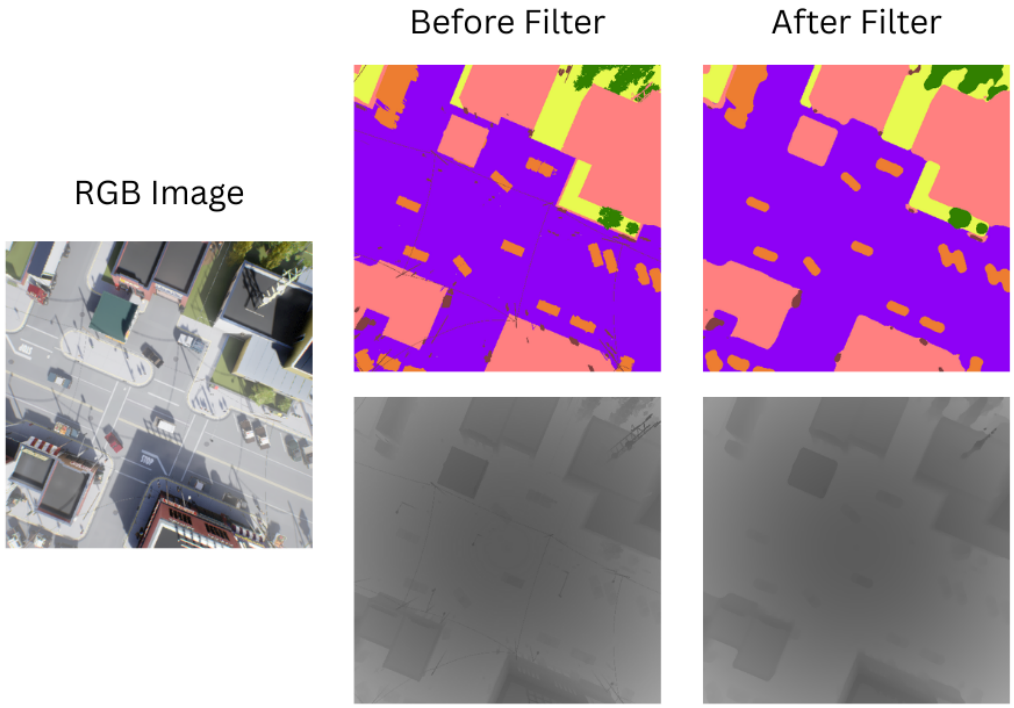


Figure 4.2 Applying a median filter on the collected depth and semantic segmentation maps of TopAir to smooth the appearance and remove unnecessary details

There was a problem faced in the AirSim segmentation of water regions, where they were treated as transparent objects, and they were identified as "Land" instead of "Water" in the output semantic segmentation maps. For this reason, we had to manually post-process such generated maps, using a photo editor *PhotoPea* [169], to paint "Water" areas that were not appearing in the segmentation maps, see Figure 4.3.

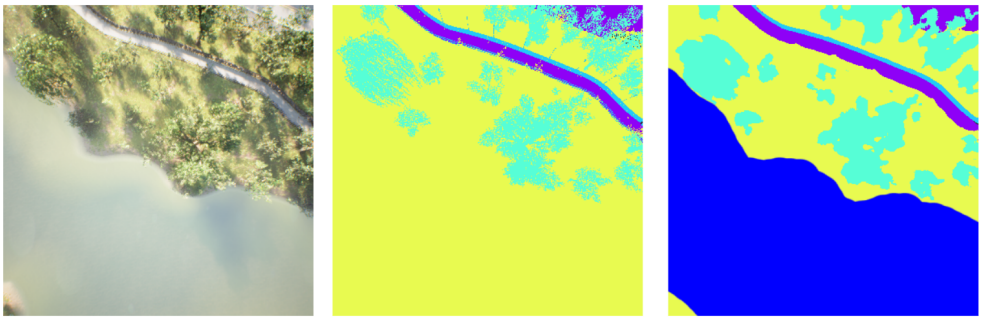


Figure 4.3 Curating the semantic segmentation maps of the collected TopAir data by manually painting the Water region after applying a median filter

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

The per-pixel distribution of the semantic segmentation classes in TopAir is depicted in Figure 4.4. Overall, it shows the diversity of the classes present in the dataset. Trees, Land, Roads, and Buildings have relatively high pixel count because they occupy big spaces in the images. On the other hand, vehicles, Rocks, and Others occupy small spaces, and for this reason, their pixel count is relatively low. Sky and Water appear only in a small number of images, and they have around 50 million pixels each.

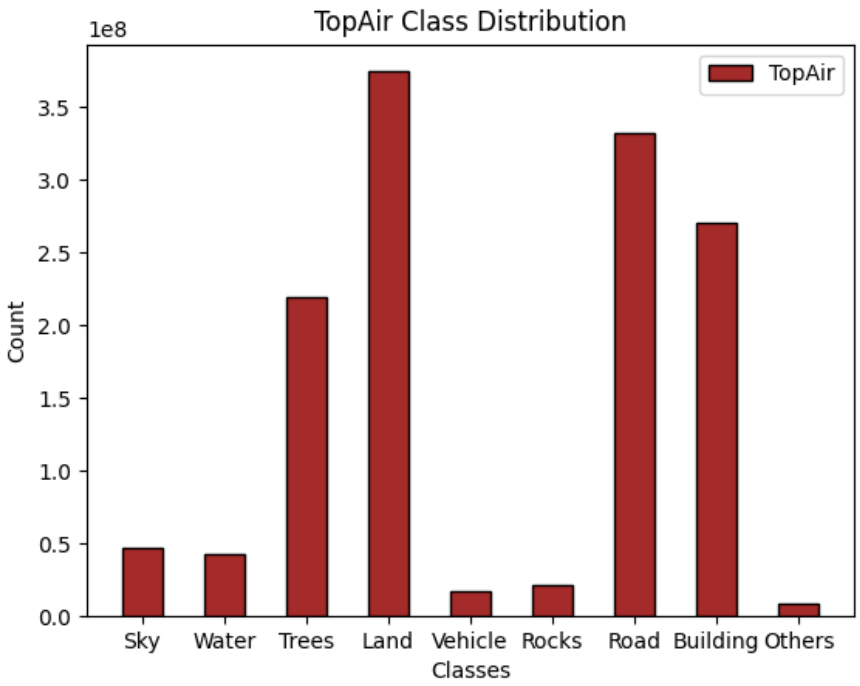


Figure 4.4 TopAir per-pixel semantic segmentation classes distribution

The per-pixel distribution of the depth values is depicted in Figure 4.5. It can be noticed from the figure that there is a diversity and a good representation of all depth values ranging from 10 to 100 meters in the collected data. The most frequent depth values range from 30 to 70 meters, while the least frequent are the edge values 10, 90, and 100 meters.

The data was collected at a frame rate of 20 FPS, and in Figure 4.6 we demonstrate the convention of the reference frames of the camera and the world used in the simulation.

4.2 Adopted Methods

In the synthetic-to-real analysis, we focus on evaluating the synthetic-to-real generalization performance of joint architectures because of their memory and time efficiency compared to

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

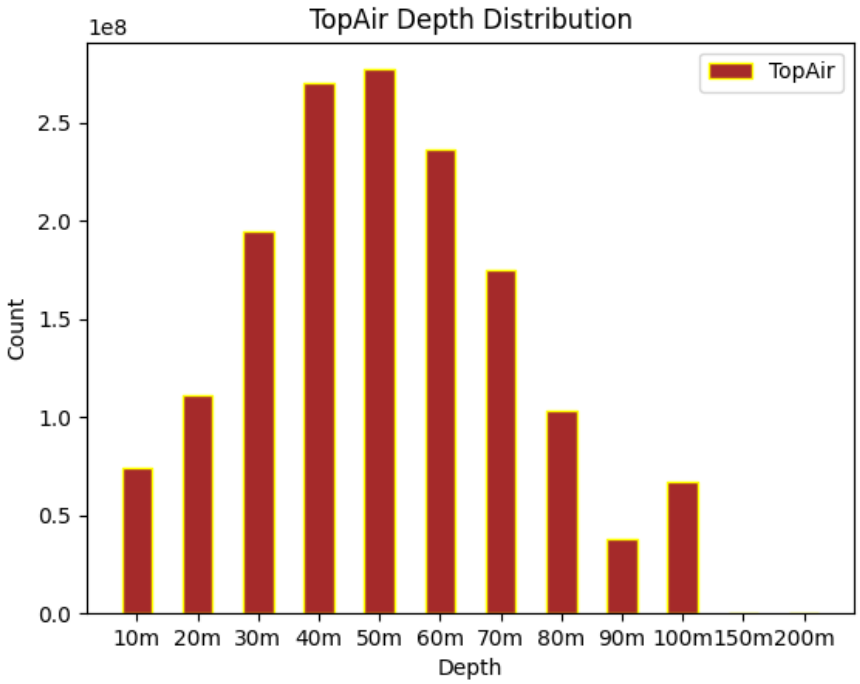


Figure 4.5 TopAir per-pixel depth values distribution

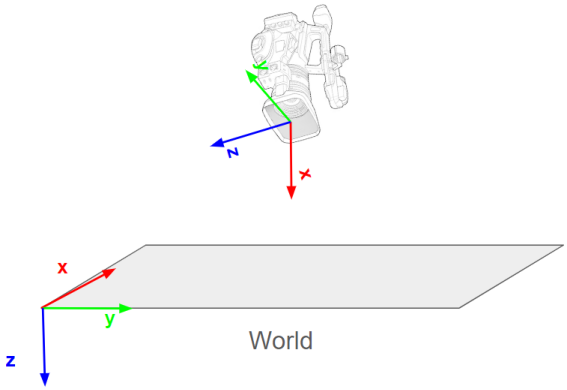


Figure 4.6 An illustration of the convention of the camera and world reference frames used in AirSim while collecting TopAir data

single architectures. Specifically, we consider two joint methods: Co-SemDepth [42] and TaskPrompter [3]. The motivation to use these two networks, particularly, is that the first network, Co-SemDepth, is a light, small network that is more suitable for deployment on resource-limited hardware platforms,while the second network, TaskPrompter, is relatively big and suitable for offline testing or deployment on only high-performance hardware

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

platforms. Since the architecture of **Co-SemDepth** was previously described in the previous chapter, chapter 3, we hereby give a brief description of the TaskPrompter's architecture.

The main idea of **TaskPrompter** [3] is to design a multi-task architecture that can jointly model (i) task generic, and (ii) task specific representations, and (iii) cross-task interactions. While in Co-SemDepth and several other multi-tasking works [112; 113; 114; 115], the task-generic representations are extracted using the shared encoder module, and the task-specific learning is done using a separate decoder module for each task, in TaskPrompter, the three objectives are learned in each network layer in an end-to-end manner. It is a Spatial-Channel Multi-task Prompting framework based on vision transformers. They design a set of spatial-channel task prompts and learn their spatial and channel interactions with the input image tokens in each transformer layer. With the help of task prompts and attention mechanisms, the network learns task-generic, task-specific representations and cross-task interactions in the same network layer without the need to add dedicated network modules for learning them.

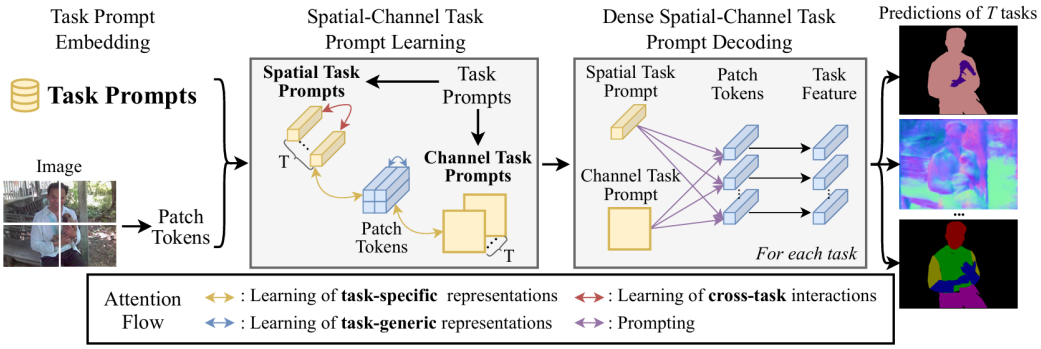


Figure 4.7 Illustration of the spatial-channel multi-task prompt learning framework used in each layer of the transformer in TaskPrompter. Task generic representations are extracted from the image patch tokens, while task-specific representations are extracted from the relations between spatial task prompts, channel task prompts, and image patch tokens. The cross-task interactions are extracted from the spatial task prompts. Taken from [3], courtesy of the authors, in accordance with the policy of ICLR

Each task prompt learns the task-specific representations of one task, while the shared image tokens are learned by the contribution of all task prompts. The interactions between each pair of task prompts contribute to the cross-task relationships. Task prompts are learnable and tuned during training. They are task-specific representations that learn both spatial and channel information of a specific task. After learning the task prompts, the task-generic image tokens, and the cross-task interactions comes the decoding phase to output the final predictions for each task (in our case, we are interested only in two tasks: depth estimation

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

and semantic segmentation). This method produced state-of-the-art results on NYUD-V2 and PASCAL datasets. However, the architecture has a large number of parameters (>100 million parameters for the different versions) and high inference time, as will be revealed in the experiments.

The whole process can be divided to three stages (as illustrated in Figure 4.7): Task Prompt Embedding, Spatial-Channel Task Prompt Learning, and the Decoding phase.

4.3 Experimental Setup

In this section, we discuss the setup of experiments conducted to test the effect of different factors on the synthetic-to-real generalization performance of deep joint architectures in the aerial domain. In our analysis, the following factors are considered:

1. Given a real dataset of interest for testing, how does changing the synthetic datasets used for training change the model performance on real data?
2. How can changing the model architecture affect the synthetic-to-real generalization performance?
3. Whether adding a small number of real data to the training would affect positively the synthetic-to-real generalization of the network (few-shot prediction)?

4.3.1 General Settings

For Co-SemDepth, 5 layers are used, and all the other values of the model are kept as the default. This results in a light model that has only 5.2 million trainable parameters. For TaskPrompter, instead, the vision transformer "Base" model is set as the backbone with an embedding dimension of 384 and a number of channel heads equal to 8. All the other values are kept as the default. This results in a big model that has around 126 million trainable parameters.

The synthetic datasets considered for training are:

- MidAir: contains natural scenes with a forward camera viewpoint.
- TopAir: contains both natural and urban scenes with a top camera viewpoint. The environments used for training and validation are: Neighbourhood, Assets Ville Town, DownTown city, and City Park.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

- SkyScenes: contains mostly urban scenes with forward and top camera viewpoints.
- SynDrone: contains mostly urban scenes with oblique camera viewpoint.

To make the training data balanced in each experiment, regarding camera viewpoint and dataset size, MidAir and TopAir datasets are concatenated and used jointly for training the models, while SkyScenes and SynDrone datasets are used together.

For training on MidAir+TopAir, due to the difference in the number of images in the two datasets, the number of training images per epoch is set to $2\times$ the number of images in the smaller dataset (TopAir). This results in a total number of training frames per epoch = 16868. The sampling ratio from MidAir and TopAir is set to $1:1$, meaning an equal contribution from both datasets during training. For validation, the validation set of MidAir used in [42] is added to the validation set of TopAir. This results in a number of validation frames equal to 5381.

Instead, for SkyScenes+SynDrone, the number of training images per epoch is also set to $2\times$ the number of training images of the smaller dataset (SkyScenes), resulting in a total number of training frames per epoch = 12972. The sampling ratio is set to $1:1$. For validation, we select random trajectories covering a variety of conditions from the two datasets, creating a number of validation frames = 4138.

The evaluation of depth estimation is carried out on the following real data:

- WildUAV: depth values ranging from 0 to 70 meters.
- Dronescapes: depth values ranging from 40 to 500 meters.

Due to the limited number of available real datasets with depth annotation, only the WildUAV and DroneScapes datasets could be used in our experiments for conducting a quantitative analysis of the depth estimation.

In Figure 4.8, the per-pixel distribution of the depth values in the datasets used is demonstrated. From the figure, it can be noted that the range of depth values of WilUAV is close to the range of values in the synthetic datasets used for training. However, the range of depth values of DroneScapes tends to include only large values.

The evaluation of semantic segmentation is done on the real datasets WildUAV, DroneScapes, RuralScapes, AeroScapes, FSI, UDD, and ICG. All of these datasets contain semantic

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

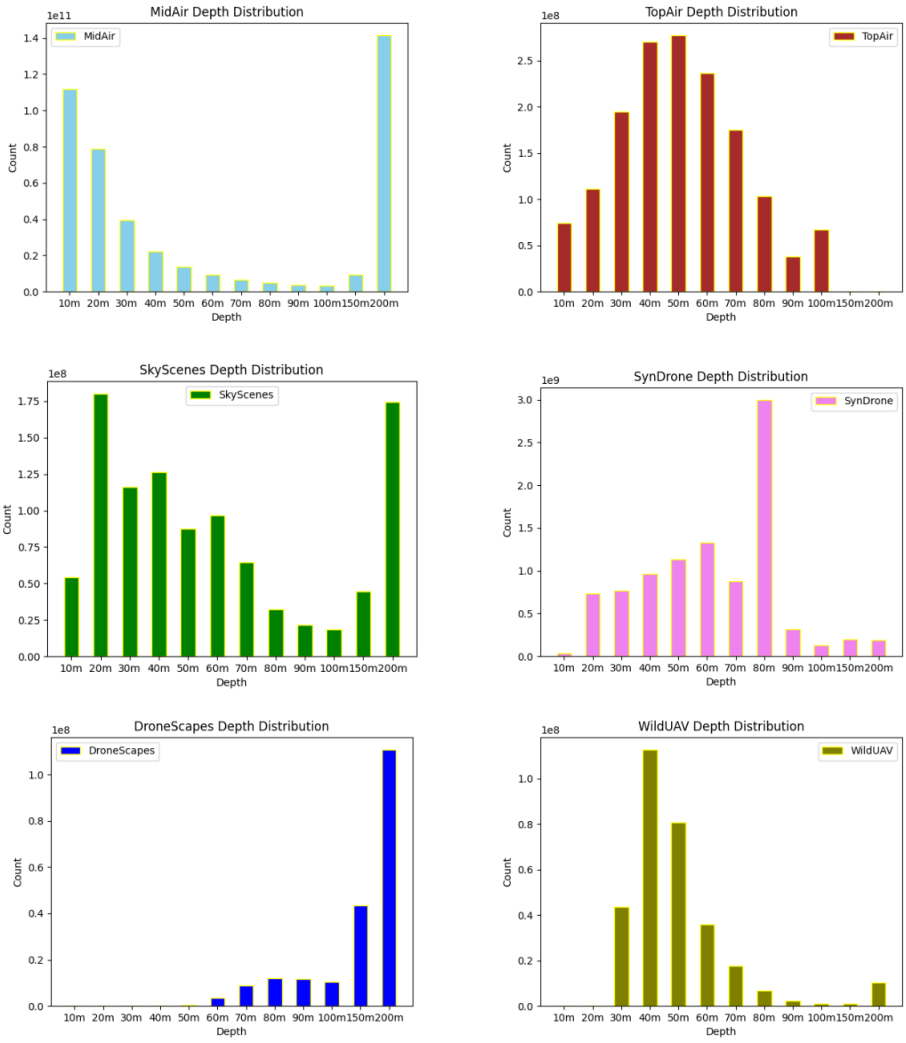


Figure 4.8 Histograms of pixel-wise depth distribution in each dataset. Each pin on the x-axis indicates all depth values less than or equal to the pin value. It can be marked how the range of depth values in WildUAV is more relevant to the depth values of the synthetic datasets used for training, while this does not hold true for DroneScapes.

segmentation annotations, and the evaluation can be done quantitatively and qualitatively. Here is a brief description of the semantic content of each of them:

- WildUAV: 1535 images resized to 528×384 instead of the original 5280×3956 resolution.
- Dronescapes: 1419 images resized to 480×320 instead of the original 960×540 resolution. The environments used are "Gradistei", "Herculane", "Olanesti", and "Petrova".

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

- RuralScapes: 1108 images resized to 420×220 instead of the original 4096×2160 resolution.
- Aeroscapes: 3269 images resized to 640×360 instead of the original 1280×720 resolution.
- FSI: 186 images resized to 300×400 instead of the original 3000×4000 resolution.
- UDD: 112 images resized to 400×300 instead of the original 4000×3000 resolution.
- ICG: 400 images resized to 500×300 instead of the original 6000×4000 resolution.

To tackle the third item in our testing factors, a small part of DroneScapes (*600 images*) is added to the synthetic data during training. The percentage of real to synthetic images in the training is very small (5% *for MidAir+TopAir and* 4.6% *for SkyScenes+SynDrone*), and this reflects the typical situations for neural networks where synthetic data are abundant and a limited number of real, annotated data. However, Co-SemDepth requires as input the camera transformation between frames, and we can not use the transformation data provided in DroneScapes due to the ambiguity in the convention of its reference frames. For this reason, only TaskPrompter is used in our experiments for testing the third factor.

4.3.2 Semantic Classes Mapping

Each dataset is annotated with a unique set of segmentation classes. To unify the semantic segmentation classes used in our experiments, we define a set of 9 classes to include the most common and unique classes of interest in the aerial domain. Specifically the 9 classes are: Sky, Water, Trees, Land, Vehicle, Rocks, Road, Building, and Others. The segmentation classes of each dataset are mapped to our set as listed in Table 4.2. In general, the classes mapping between datasets may not be highly accurate due to the differences in the appearance and class definitions in each dataset.

The pixel-wise class distribution for each dataset after applying the mapping can be found in Figure 4.9. As can be noted from the figure, datasets differ in the semantic classes distribution; some of them have a high representation of buildings and road classes (like SkyScenes, UDD, and ICG) showing that they are mostly urban, while others contain a high representation of land and trees and do not contain buildings (like MidAir and WildUAV) suggesting that they are captured in the wild nature. It can also be noted how the class distribution of our introduced TopAir dataset is nearly representative of all the classes compared to the other datasets. The "Vehicle" class count is low in all the datasets because

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Table 4.2 The mapping of the semantic classes across different datasets. The mapping used for SkyScenes is the same as that done on the SynDrone dataset and is referred to as Sky/Syn.

Common Set	*MidAir*	*TopAir*	*SkySyn*	*Dronescapes*	*WildUAV*	*Aeroscapes*	*Ruralscapes*	*UDD*	*FSI*	*ICG*
0 Sky	sky	sky	sky	sky	sky	sky	sky	-	-	-
1 Water	water	water	water	water	water	-	water	-	water body swimming pool flooded	water pool
2 Trees	trees	trees	vegetation	forest	deciduous tree coniferous tree fallen trees	vegetation	forest	vegetation	trees forest	tree other vegetation
3 Land	dirt ground ground vegetation rocky ground	land	terrain other	land hill	dirt ground ground vegetation	background	land hill	others	grass under-construction sports arena	grass dirt
4 Vehicle	-	vehicle	cars bus truck motorcycle rider train	little-objects	static car moving car	car	car	vehicle	vehicle	car
5 Rocks	boulders	rocks	-	-	rocks	-	-	-	-	rocks
6 Road	road	road	road sideWalk roadLine ground	road	road sidewalk	road	road	road	road parking background	paved area gravel
7 Building	construction	building	building bridge railTrack wall	residential	building	construction	residential church	facade roof	property roof chimney industrial solar panels antenna window secondary structure	roof door window wall
8 Others	others road sign train track animals empty	others	static dynamic fence pedestrian pole trafficSign trafficLight unlabeled other (in town10) bicycle guardRail	little-objects	fence people empty	person bike drone boat animal obstacle	fence haystack person	-	trampoline garbage bins boat street light water tank cables	bicycle dog fence person obstacle fence-pole

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

the vehicles (cars, bikes, etc) normally occupy a small number of pixels compared to the other classes.

4.3.3 Training and Evaluation Settings

For Co-SemDepth, the model is trained using the Adam optimizer for 60 epochs and with a fixed learning rate of 10^{-4}. The batch size is set to 3 and the number of frames per sequence to 4. Table 4.3 summarizes the training parameters of Co-SemDepth.

For TaskPrompter, the model is trained using the Adam optimizer for a maximum of 80K iterations. The batch size is set to 3, the learning rate to 10^{-5}, and a polynomial scheduler is used with a weight decay of 10^{-6}. Table 4.4 summarizes the training parameters of TaskPrompter.

During the training of both models, image augmentations of random rotation, flipping, and changing color (contrast, brightness, hue, and saturation) are applied. The depth values are cropped to a maximum of 200 meters.

It should be noted that when testing TaskPrompter on different datasets, the testing images have to be resized to have a resolution equal to the resolution of the training images, otherwise it gives an error. However, such a problem is not faced with Co-SemDepth, which can accept testing images of various sizes without the need to resize them.

Table 4.3 Training parameters used for training **Co-SemDepth** on MidAir+TopAir and SkyScenes+SynDrone

Train Dataset	optimizer	lr	weight decay	epochs	batch	Train size	Valid size	resolution
MidAir+TopAir	Adam	10^{-4}	-	60	3	16868	5381	384×384
SkyScenes+SynDrone	Adam	10^{-4}	-	60	3	12972	4138	480×320

Table 4.4 Training parameters used for training **TaskPrompter** on MidAir+TopAir and SkyScenes+SynDrone

Train Dataset	optimizer	lr	weight decay	iterations	batch	train size	valid size	resolution
MidAir+TopAir	Adam	10^{-5}	10^{-6}	80k	3	16868	5381	384×384
SkyScenes+SynDrone	Adam	10^{-5}	10^{-6}	80k	3	12972	4138	480×320

The camera motion model used in our experiments with Co-SemDepth is shown in Figure 1.3. For all the datasets that used a camera model with a different orientation of the

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

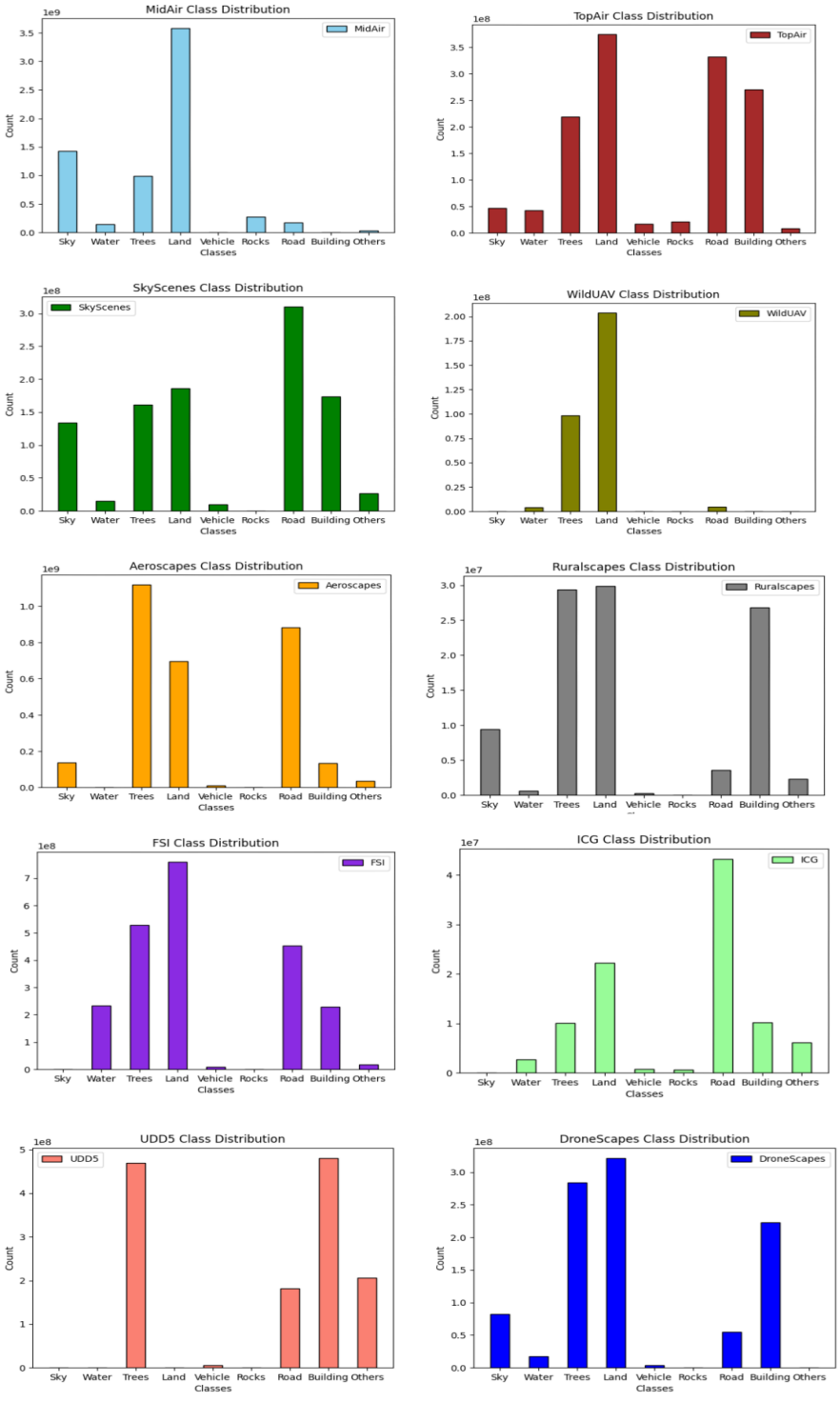
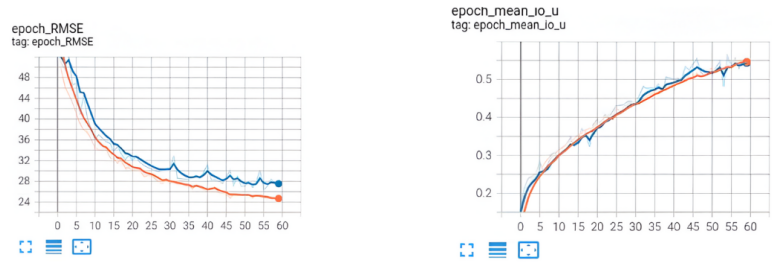


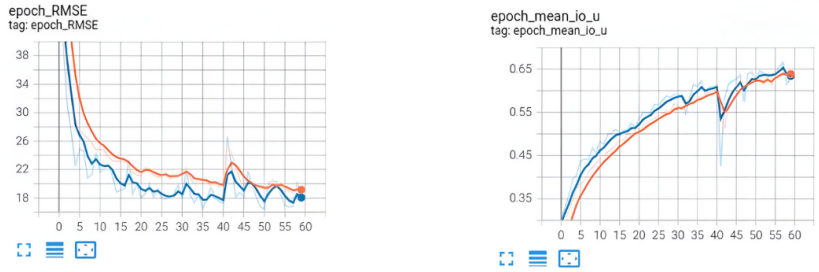
Figure 4.9 Histograms of pixel-wise class distribution in each dataset after mapping to our common set of segmentation classes

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

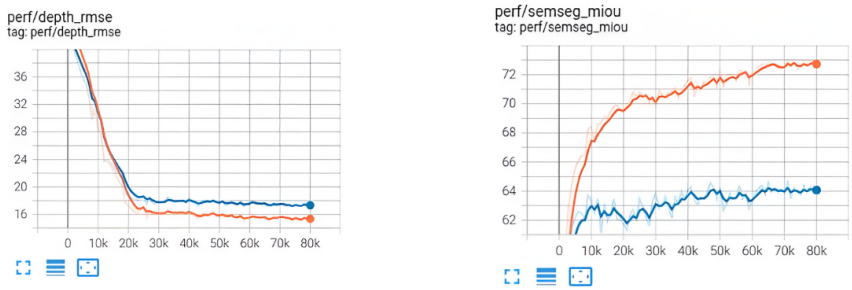
Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*



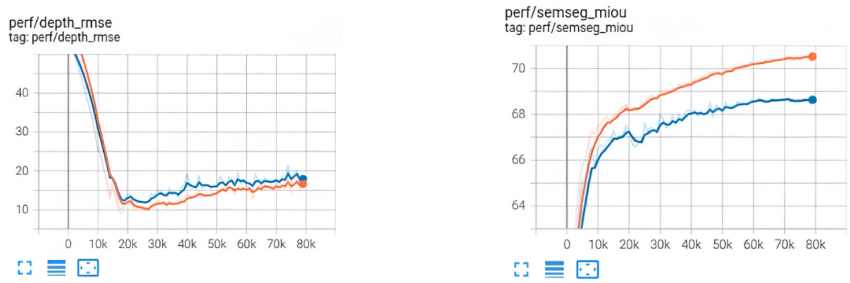
(a) Co-SemDepth training plots on MidAir+TopAir



(b) Co-SemDepth training plots on SkyScenes+SynDrone



(c) TaskPrompter training plots on MidAir+TopAir



(d) TaskPrompter training plots on SkyScenes+SynDrone

Figure 4.10 Plots of the train and validation RMSE and mIoU values of the models during training on MidAir+TopAir and SkyScenes+SynDrone. While TaskPrompter tends to overfit on the training data due to its high capacity, Co-SemDepth's performance is almost equal on the train and validation data (no overfitting). For all the models, the checkpoint that produced the best validation results was selected.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

principal axes, the necessary transformation is applied on the camera orientation to make it similar to the model in the figure. This is specifically necessary for the training and testing of depth estimation using the Co-SemDepth architecture, which requires as input the camera motion data between every two consecutive frames.

To quantitatively evaluate the depth prediction results, the commonly used evaluation metrics in prior works [58; 61; 170] are used. These include the linear root mean square error (*RMSE*), the absolute relative error, and accuracy under threshold.

For semantic segmentation, we adopt the commonly used mean Intersection over Union *mIoU* metric, with a focus on the per-class *IoU* to assess the segmentation performance on each of the 9 classes in our set.

The training and validation plots of the *RMSE* and *mIoU* metrics for the two models can be found in Figure 4.10. From the plots, it can be noted that while TaskPrompter tends to overfit on the training data (its performance on the training set is better than on the validation set), Co-SemDepth's performance is almost equal on the training and validation data (no overfitting). This can be due to the high capacity of TaskPrompter, which makes it tend to memorize the training data, while Co-SemDepth has a limited number of parameters making it capture only important generic features.

On average, the inference time of Co-SemDepth on a single NVIDIA Quadro P4000 GPU is 49.2 ms per frame, and for TaskPrompter, it is 120.6 ms per frame. This shows that Co-SemDepth is more convenient with regard to speed for real-time applications.

4.4 Results

The experiments of synthetic-to-Real performance are divided into two parts: *Depth Estimation Evaluation*, and *Semantic Segmentation Evaluation*.

4.4.1 Depth Estimation Evaluation

In Table 4.5, we assess the depth estimation performance of Co-SemDepth and TaskPrompter on the WildUAV dataset. In general, we can notice that the results obtained using Co-SemDepth are notably better than those of TaskPrompter despite the huge difference in the model sizes. This can be due to the parallax estimation modalities implemented in Co-SemDepth that are designed to allow for better generalization of the model in depth estimation.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

It can also be noted that, for both models, using MidAir+TopAir for training produces better results than using SkyScenes+SynDrone, and this can be justified by the semantic similarity between the scenes of WildUAV (which are taken in nature) and the natural scenes of MidAir and TopAir.

In Figure 4.12, a visualization of the predicted output on sample input WildUAV images is depicted. By a qualitative analysis of the figure, it can be validated that Co-SemDepth trained on MidAir+TopAir produces the closest results to the ground truth compared to the other methods.

Table 4.5 Evaluation of depth estimation performance of joint architectures on WildUAV real dataset. Best results are highlighted.

Test Data	Train Data	Method	Params	RMSE ↓	AbsRelErr ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
WildUAV	MidAir+TopAir	Co-SemDepth	5.2M	**14.9**	**0.303**	**58.7%**	**80.7%**	**91.5%**
		TaskPrompter	126M	21.26	0.459	18.5%	37.5%	54.5%
	SkyScenes+SynDrone	Co-SemDepth	5.2M	18.9	0.388	33.78%	58.6%	73.4%
		TaskPrompter	126M	23	0.48	10.5%	25.2%	46.2%

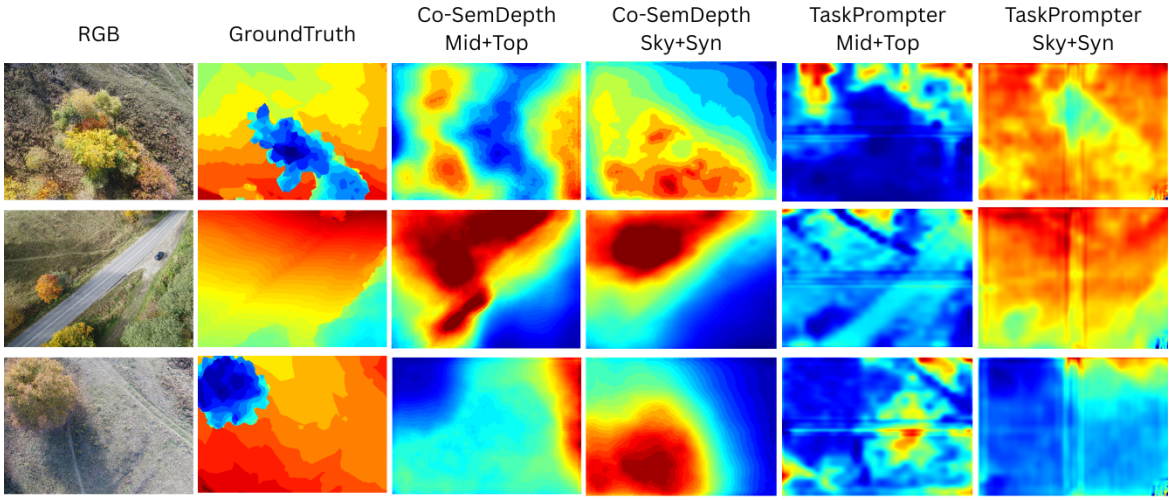


Figure 4.11 Visualization of the estimated depth maps of sample input images from WildUAV using Co-SemDepth and TaskPrompter (displayed in relative scale where red is the highest distance and blue is the lowest). The output of Co-SemDepth is closer to the ground truth than TaskPrompter, and training on MidAir+TopAir (Mid+Top) is better than SkyScenes+SynDrone (Sky+Syn).

Assessment of Adding real data to the training: As explained earlier in this chapter, this assessment is carried out using only TaskPrompter. In the top part of Table 4.6, the depth estimation performance is assessed on WildUAV before and after adding real data (DroneScapes) to the training.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Despite the low percentage of real to synthetic training data ($4.5-5\%$), a big enhancement is achieved in the results for both MidAir+TopAir and SkyScenes+SynDrone training modalities. This strengthens the findings reported in [38; 39], where it was demonstrated the positive effect on the model's generalization of adding different percentages of the same real test data to the training. It further adds to the previous findings that such enhancement can be achieved even if the added data is not taken from the same test data distribution (if we use a different real dataset for finetuning). Output visualization of the results can be found in Figure 4.12. It can be seen from the figure that adding DroneScapes to the training makes the network predict more realistic depth maps than those predicted without adding DroneScapes.

In the bottom part of Table 4.6, the depth is evaluated on DroneScapes dataset before and after adding real data to the training. In this case, the performance of the models trained only on synthetic datasets is very low due to the big difference in the depth values between DroneScapes and the datasets used for training, see the depth distributions of the datasets in Figure 4.8. Despite that, adding a small part of Dronescapes to the training $(4.5-5\%)$ of the model makes a big boost in its performance on the test data.

Table 4.6 Evaluation of depth estimation performance of TaskPrompter on WildUAV and DroneScape real datasets without and with adding real data to the training.

Test Data	Method	Train Data	Finetuning Data	RMSE ↓	AbsRelErr ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
WildUAV	TaskPrompter	MidAir+TopAir	-	21.26	0.459	18.49%	37.46%	54.54%
			Dronescapes	**15.87**	**0.32**	**31.42%**	**62.39%**	**84.27%**
	TaskPrompter	SkyScenes+SynDrone	-	23	0.48	10.5%	25.2%	46.2%
			Dronescapes	18.05	0.354	29.4%	53.86%	73.14%
DroneScapes	TaskPrompter	MidAir+TopAir	-	96.77	0.78	0.1%	0.6%	2.5%
			Dronescapes	76.12	0.485	8.12%	34.53%	55.15%
	TaskPrompter	SkyScenes+SynDrone	-	88.7	0.694	0%	0.02%	2.15%
			Dronescapes	**40.65**	**0.237**	**50.35%**	**80.7%**	**93.49%**

The depth estimation is additionally assessed on sample images from the FSI and ICG datasets. However, due to the absence of ground truth depth maps, the analysis is carried out only qualitatively. Only TaskPrompter can be used for making predictions because CoSemDepth requires camera location data and they are not provided in the mentioned datasets. In Figures 4.13 and 4.14 it can be found a visualization of the output on sample images from FSI and ICG, respectively. For FSI, it is observed that the model trained on MidAir+TopAir produces the best output. This can be justified by the apparent similarity between FSI scenes and the images captured in the *Neighbourhood* environment in TopAir. It can also be noted that finetuning on real images from DroneScapes enhanced the results of the model trained on SkyScenes+SynDrone. For ICG, it can be observed by eye that the output of the model trained on MidAir+TopAir is generally better than the one trained on SkyScenes+SynDrone.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

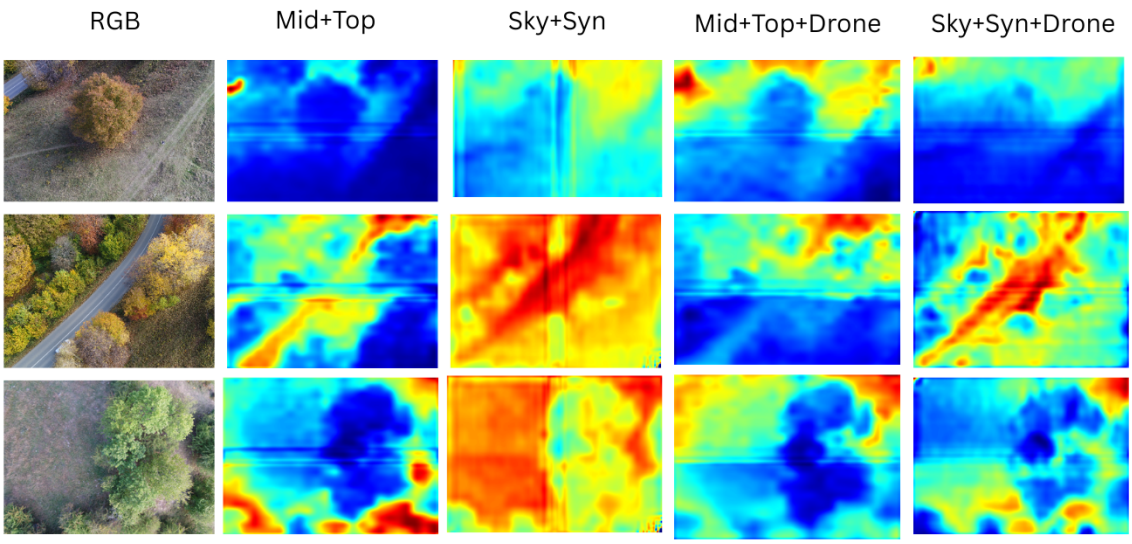


Figure 4.12 Qualitative visualization of the estimated depth maps using TaskPrompter of sample input images from WildUAV before and after adding DroneScapes to the training (displayed in relative scale where red is the highest distance and blue is the lowest). Adding real data to the training enhanced the results.

Unfortunately, finetuning on DroneScapes does not provide considerable enhancements on the outputs in this case.

4.4.2 Semantic Segmentation Evaluation

The semantic segmentation performance of the two models is assessed in Table 4.7 on the real datasets: WildUAV, FSI, UDD, DroneScapes, RuralScapes, AeroScapes, and ICG. From the table, it can be noted that, regarding the model architecture, TaskPrompter in general is better in semantic segmentation than Co-SemDepth. This can be due to TaskPrompter's large capacity that makes it more capable of capturing the small semantic details. Specifically, it is remarkably better at segmenting the small objects ("Vehicle" and "Others").

Regarding the training data, depending on the appearance of the objects and the layout of the scenes in the datasets, the class segmentation in some datasets can be better using MidAir+TopAir for training or SkyScenes+SynDrone in other cases. In particular, the UDD dataset contains only urban scenes, and this makes it more similar to SkyScenes+SynDrone; hence the models trained on SkyScenes+SynDrone give higher accuracy on UDD than the ones trained on MidAir+TopAir. On the other hand, the WildUAV dataset is more similar to MidAir+TopAir due to its natural scenes that are similar to the ones present in MidAir and TopAir; hence, the models trained on MidAir+TopAir perform better in

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

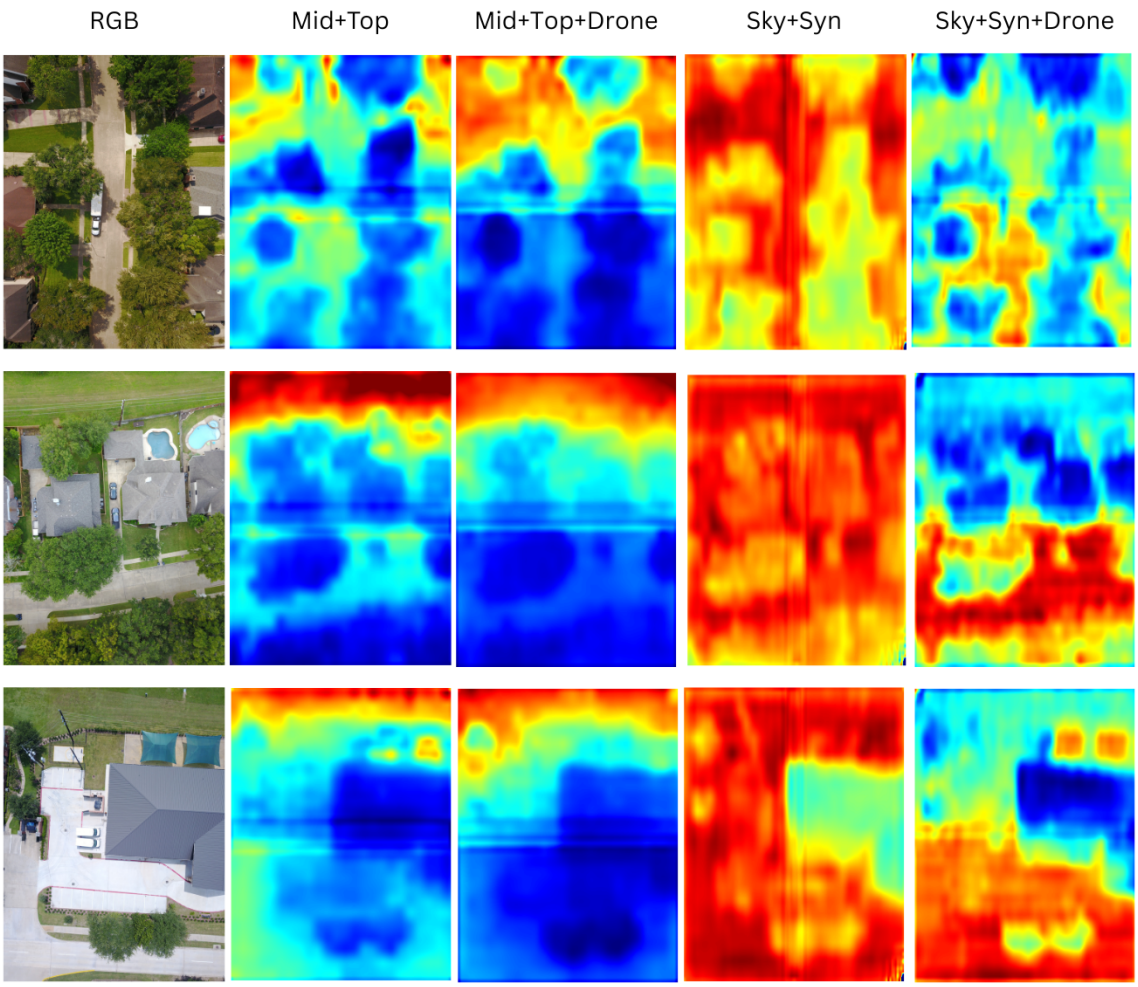


Figure 4.13 Visualization of the estimated depth maps (displayed in relative scale where red is the highest distance and blue is the lowest) of sample input images from FSI using TaskPrompter. The model trained on MidAir+TopAir produced the best output. Finetuning on DroneScapes enhanced the results of the model trained on SkyScenes+SynDrone

this case than the ones trained on SkyScenes+SynDrone. On FSI, the model trained on MidAir+TopAir is more capable in segmenting "Water", "Trees", and "Building", while the model trained on SkyScenes+SynDrone is better in segmenting "Land", "Vehicle", and "Road". On RuralScapes and DroneScapes, the model trained on MidAir+TopAir achieves higher accuracy on the segmentation of "Sky", "Water", and "Land", and the model trained on SkyScenes+SynDrone performs better in the segmentation of "Vehicle", "Road", and "Building". On AeroScapes, training on MidAir+TopAir gives the highest accuracy in class "Land" and "Building", while training on SkyScenes+SynDrone is better in the other classes. The results are also class-dependent between the two training modalities for ICG. Therefore,

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

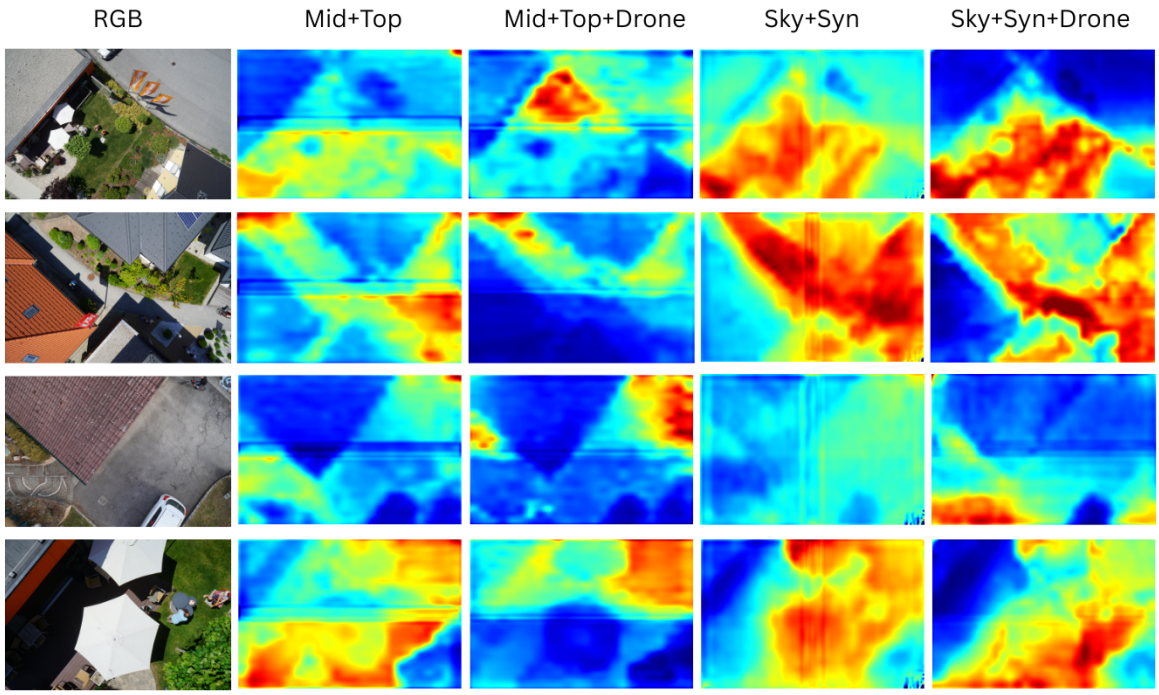


Figure 4.14 Visualization of the estimated depth maps (displayed in relative scale where red is the highest distance and blue is the lowest) of sample input images from ICG using TaskPrompter. The output of the model trained on MidAir+TopAir is generally better than the others. Finetuning on DroneScapes did not enhance the results except in the last row (notice the umbrellas).

what we can conclude from the above is that **there is no specific training dataset that gives the best results in all the cases**. Instead, the synthetic-to-real performance is dependent on the appearance of the objects belonging to different classes in the target test data and how similar they are with the objects in the synthetic training data.

A visualization of the output semantic segmentation maps on sample images from FSI can be found in Figure 4.15. It can be seen that, in general, the models trained on SkyScenes+SynDrone tend to segment most of the pixels as either "Building" or "Road" (see the second and third rows) due to the urban nature of the training data. On the other hand, the models trained on MidAir+TopAir tend to the "Land" and "Trees" classes in the segmentation (see the first and third rows). It can also be noted the big performance gap in the semantic segmentation by varying the model architecture, using TaskPrompter compared to Co-SemDepth, trained on the same datasets.

A visualization of the output semantic segmentation maps on sample images from ICG can be found in Figure 4.16. It can be noted how the semantic segmentation using TaskPrompter is much better than using Co-SemDepth on this dataset. Additionally, by

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

comparing the output of TaskPrompter trained on MidAir+TopAir and TaskPrompter trained on SkyScenes+SynDrone, we can say that using SkyScenes+SynDrone is better in the third row where the vehicle is present in the image; however, in the fourth row, using MidAir+TopAir makes the network better detect and segment the water area. In the first row, MidAir+TopAir is better at segmenting the building.

Table 4.7 Per-Class IoU segmentation evaluation of the joint architectures on a variety of real data. Best results for each dataset are highlighted.

Test Data	Train Data	Method	Params	Sky	Water	Trees	Land	Vehicle	Rocks	Road	Building	Others
WildUAV	MidAir+TopAir	Co-SemDepth	5.2M	-	0.93%	32.8%	34.76%	0%	-	14.9%	-	0%
		TaskPrompter	126M	-	61.66%	**55.36%**	**67.63%**	1.65%	-	**53.5%**	-	3.1%
	SkyScenes+SynDrone	Co-SemDepth	5.2M	-	17.63%	4.2%	14.1%	1.2%	-	1.8%	-	0.04%
		TaskPrompter	126M	-	**62.1%**	51%	62.05%	**46.2%**	-	45.9%	-	**5.9%**
FSI	MidAir+TopAir	Co-SemDepth	5.2M	-	10.1%	40.6%	34.67%	0%	-	9.14%	1.43%	0.01%
		TaskPrompter	126M	-	**49.57%**	**67.36%**	63.7%	3.32%	-	34.1%	**59.83%**	1.73%
	SkyScenes+SynDrone	Co-SemDepth	5.2M	-	6.31%	38.25%	36.54%	4.07%	-	31.65%	17.12%	0.35%
		TaskPrompter	126M	-	32.5%	63.6%	**64.1%**	**19.76%**	-	**37.8%**	56.83%	**3.9%**
UDD	MidAir+TopAir	Co-SemDepth	5.2M	-	-	37.2%	17.1%	0%	-	9.9%	1.7%	-
		TaskPrompter	126M	-	-	56.52%	**13.02%**	3.16%	-	48.53%	70.75%	-
	SkyScenes+SynDrone	Co-SemDepth	5.2M	-	-	46.99%	11.24%	1.62%	-	24.96%	34.01%	-
		TaskPrompter	126M	-	-	**76.32%**	8.9%	**18.4%**	-	**49.77%**	**73.1%**	-
DroneScapes	MidAir+TopAir	Co-SemDepth	5.2M	43.88%	0%	15.66%	37.13%	0.5%	-	10.4%	10.9%	-
		TaskPrompter	126M	**95.1%**	**10.15%**	37.25%	**53.7%**	0.8%	-	22.5%	48.5%	-
	SkyScenes+SynDrone	Co-SemDepth	5.2M	50.74%	0%	39.56%	23.87%	4.52%	-	19.3%	24.8%	-
		TaskPrompter	126M	70.27%	5.77%	**55.79%**	53.5%	**15.04%**	-	**43.1%**	**68.15%**	-
RuralScapes	MidAir+TopAir	Co-SemDepth	5.2M	61.3%	0%	39.55%	26.7%	0%	-	2.42%	0.76%	0%
		TaskPrompter	126M	**92.7%**	**9.13%**	40.14%	**42.12%**	0.2%	-	23.1%	31.45%	6.8%
	SkyScenes+SynDrone	Co-SemDepth	5.2M	69.5%	0.45%	30.1%	23.84%	1.1%	-	10.65%	21.76%	0.97%
		TaskPrompter	126M	86.02%	3.99%	**49.7%**	30.24%	**10.2%**	-	**33.98%**	**56.96%**	**8.7%**
AeroScapes	MidAir+TopAir	Co-SemDepth	5.2M	77.04%	-	33.3%	24.89%	1.3%	-	11.2%	3.63%	0.3%
		TaskPrompter	126M	92.7%	-	37.84%	**17.75%**	8.86%	-	63.73%	**35.7%**	14.1%
	SkyScenes+SynDrone	Co-SemDepth	5.2M	60.56%	-	46.4%	10.63%	6.7%	-	43.6%	12.85%	1.7%
		TaskPrompter	126M	**93.83%**	-	**57.11%**	13.04%	**34.99%**	-	**71.8%**	31.3%	**23.9%**
ICG	MidAir+TopAir	Co-SemDepth	5.2M	-	10.62%	12.7%	20.99%	0.13%	0.44%	18.6%	12.52%	0.08%
		TaskPrompter	126M	-	**30.15%**	27.8%	44.67%	3.04%	**3.4%**	53.83%	**46.53%**	16.2%
	SkyScenes+SynDrone	Co-SemDepth	5.2M	-	5.2%	27.3%	20.5%	3.82%	0%	52.3%	17.65%	1.4%
		TaskPrompter	126M	-	14.54%	**45.52%**	**61.27%**	**45.8%**	0%	**61.7%**	35.04%	**18.88%**

Assessment of Adding real data to the training: As explained earlier in this chapter, this assessment is carried out using only the TaskPrompter model. The effect of adding real data to the training is assessed in Table 4.8 on WildUAV, RuralScapes, Dronescapes, FSI, UDD, AeroScapes, and ICG datasets. Here, due to the gap in the appearance of some classes between the real data used in the training (DroneScapes) and the testing data, the effect of adding real data is not always positive, as it can be observed from the table. As a positive example, the environments of RuralScapes look similar to DroneScapes, and for this reason, adding DroneScapes to the training enhanced the semantic segmentation accuracy on most of the classes in RuralScapes, see Figure 4.17 for a visualization of the output. In WildUAV, instead, there is an enhancement in the detection of "Water", "Trees", and "Land", while the accuracy of "Vehicle" and "Road" is decreased. Datasets like AeroScapes and ICG look completely different than DroneScapes, and this caused a drop in the accuracy after

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

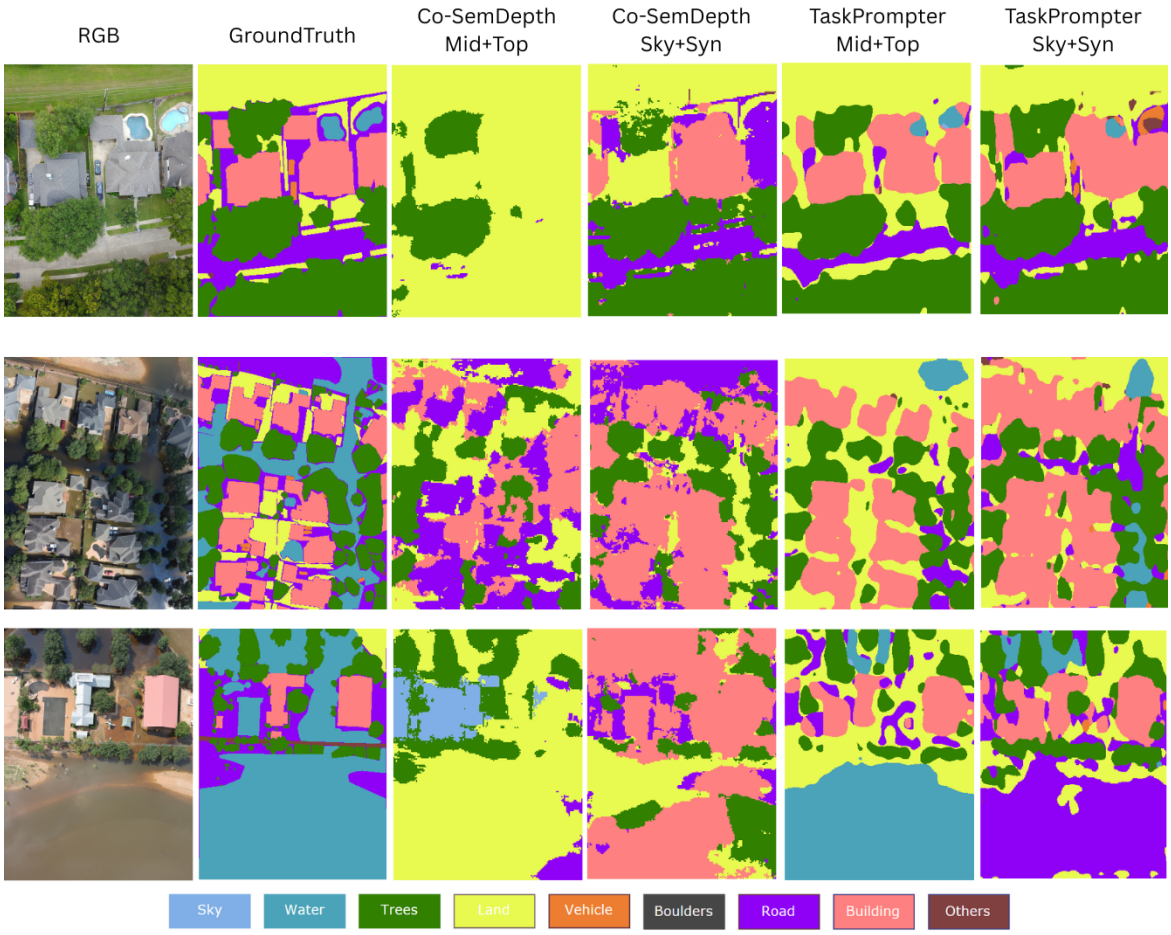


Figure 4.15 Qualitative visualization of the estimated depth maps of sample input images from FSI using different training modalities and architectures. It can be seen that using TaskPrompter trained on MidAir+TopAir gives the closest results to the ground truth. TaskPrompter trained on SkyScenes+SynDrone is the best in the detection of cars (top row).

finetuning the model on DroneScapes. Instead, on DroneScapes, the accuracy on all classes is boosted after the finetuning. On FSI, the accuracy is enhanced on the segmentation of "Water", "Vehicle", and "Road", and the accuracy on the other classes is slightly lower after finetuning. On UDD, only the segmentation of "Vehicle" and "Building" was enhanced after the fine-tuning.

4.4.3 3D Reconstruction

For a further qualitative assessment of the output, we resort to the Open3D library [171] to construct a 3D RGB and semantic map of the scene using the predicted 2D depth and semantic maps. Figures 4.18 and 4.19 show sample outputs constructed using the predicted

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

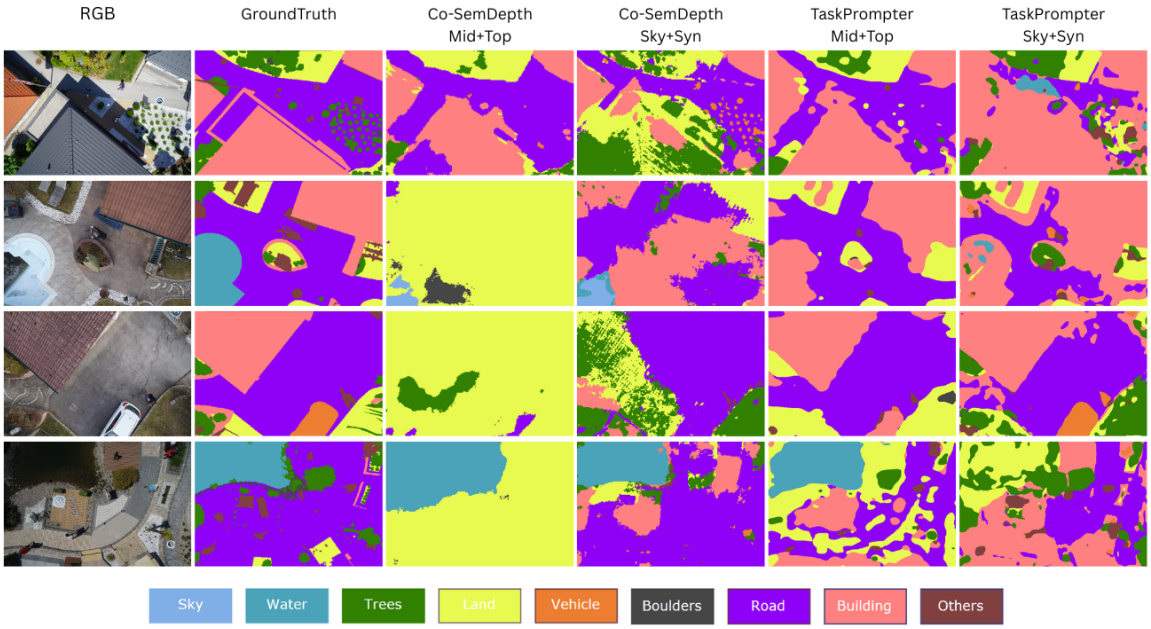


Figure 4.16 Qualitative visualization of the estimated segmentation maps of sample input images from ICG using different training modalities. Best performing method is TaskPrompter trained on MidAir+TopAir except in the detection of cars (third row) where TaskPrompter trained on SkyScenes+SynDrone is better.

Table 4.8 Per-Class IoU segmentation evaluation of joint architectures on WildUAV and RuralScapes without and with adding real data to the training.

Test Data	Train Data	Finetuning Data	Method	Sky	Water	Trees	Land	Vehicle	Rocks	Road	Building	Others
WildUAV	MidAir+TopAir	-	TaskPrompter	-	61.66%	55.36%	67.63%	1.65%	-	53.5%	-	3.1%
		Dronescapes		-	59.36%	56.88%	77.54%	7.62%	-	39.12%	-	2.6%
	SkyScenes+SynDrone	-	TaskPrompter	-	62.1%	51%	62.05%	46.2%	-	45.9%	-	5.9%
		Dronescapes		-	63.9%	47.77%	55.66%	22.89%	-	45.26%	-	5.5%
RuralScapes	MidAir+TopAir	-	TaskPrompter	92.7%	9.13%	40.14%	42.12%	0.2%	-	23.1%	31.45%	6.8%
		Dronescapes		89.61%	14.07%	46.8%	47.65%	2.1%	-	33.6%	46.93%	7.8%
	SkyScenes+SynDrone	-	TaskPrompter	86.02%	3.99%	49.7%	30.24%	10.2%	-	33.98%	56.96%	8.7%
		Dronescapes		91.1%	7.75%	50.3%	30.02%	9.17%	-	39.53%	66.52%	4.5%
DroneScapes	MidAir+TopAir	-	TaskPrompter	95.1%	10.15%	37.25%	53.7%	0.8%	-	22.5%	48.5%	-
		Dronescapes		**95.16%**	**49.98%**	51.06%	**55.94%**	10.2%	-	42.83%	69.1%	-
	SkyScenes+SynDrone	-	TaskPrompter	70.27%	5.77%	55.79%	53.5%	15.04%	-	43.1%	68.15%	-
		Dronescapes		78.98%	35.12%	**56.37%**	52.36%	**26.46%**	-	**45.55%**	**75.3%**	-
FSI	MidAir+TopAir	-	TaskPrompter	-	49.57%	**67.36%**	63.7%	3.32%	-	34.1%	**59.83%**	1.73%
		Dronescapes		-	**52.1%**	58.2%	61.8%	5.3%	-	**37.9%**	57.3%	1.95%
	SkyScenes+SynDrone	-	TaskPrompter	-	32.5%	63.6%	**64.1%**	19.76%	-	37.8%	56.83%	**3.9%**
		Dronescapes		-	39%	61.04%	60.8%	**21.3%**	-	35.6%	54.3%	3.01%
UDD	MidAir+TopAir	-	TaskPrompter	-	-	56.52%	**13.02%**	3.16%	-	48.53%	70.75%	-
		Dronescapes		-	-	50.97%	11.62%	2.95%	-	41.1%	62.6%	-
	SkyScenes+SynDrone	-	TaskPrompter	-	-	**76.32%**	8.9%	18.4%	-	**49.77%**	73.1%	-
		Dronescapes		-	-	75.55%	6.1%	**19.97%**	-	45.1%	**74.2%**	-
AeroScapes	MidAir+TopAir	-	TaskPrompter	92.7%	-	37.84%	17.75%	8.86%	-	63.73%	**35.7%**	14.1%
		Dronescapes		93.5%	-	47.42%	**18.96%**	11.3%	-	63.1%	33.74%	12.6%
	SkyScenes+SynDrone	-	TaskPrompter	**93.83%**	-	**57.11%**	13.04%	**34.99%**	-	**71.8%**	31.3%	**23.9%**
		Dronescapes		93.34%	-	55.4%	16.42%	30.3%	-	63.2%	26.34%	20.34%
ICG	MidAir+TopAir	-	TaskPrompter	-	30.15%	27.8%	44.67%	3.04%	**3.4%**	53.83%	**46.53%**	16.2%
		Dronescapes		-	**31.58%**	30.3%	50.8%	4.6%	1.66%	56.54%	43%	8.6%
	SkyScenes+SynDrone	-	TaskPrompter	-	14.54%	**45.52%**	**61.27%**	**45.8%**	0%	**61.7%**	35.04%	**18.88%**
		Dronescapes		-	12.72%	39.3%	45.86%	37.12%	0%	58.03%	34.95%	14.6%

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

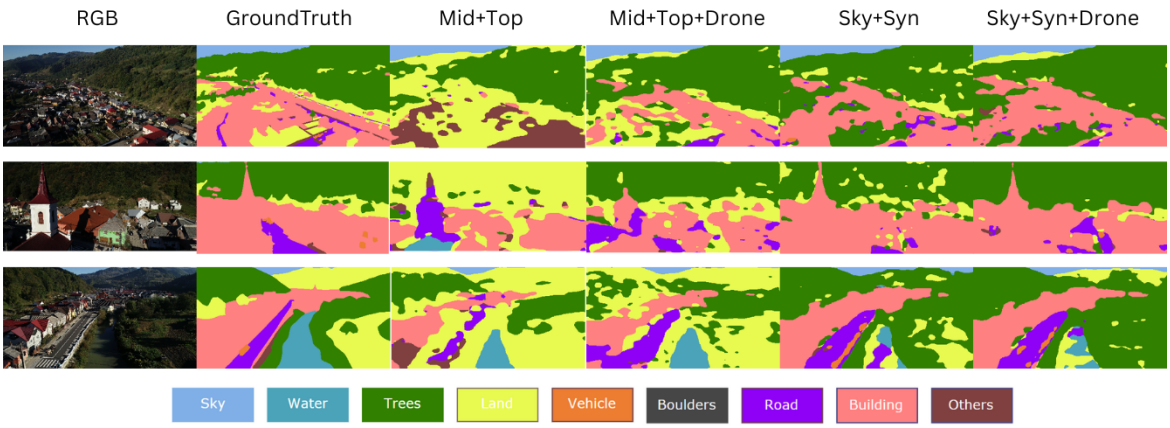


Figure 4.17 Qualitative visualization of the estimated semantic segmentation maps of sample input images from RuralScapes using TaskPrompter before and after adding real data (DroneScapes) to the training

depth and semantic maps of TaskPrompter on sample input images from DroneScapes and FSI datasets. First, an RGB-D image is created using Open3D from the input segmentation map (or RGB image) and depth map. Then, a point cloud is constructed from the RGB-D image using the Pinhole camera default intrinsic parameters of Open3D. Finally, the point cloud is plotted using the visualization tool of the library. The constructed 3D maps can be particularly useful in 3D SLAM navigation of UAVs.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

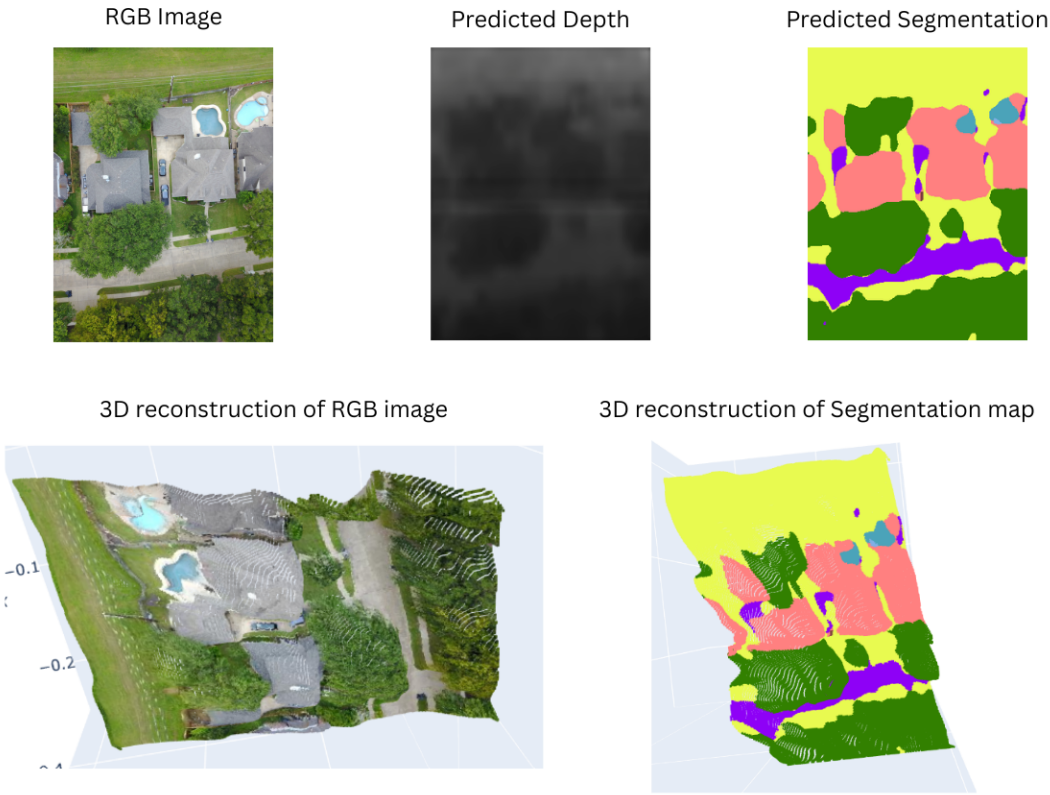


Figure 4.18 3D semantic map reconstruction using the predicted depth and segmentation maps using TaskPrompter (trained on MidAir + TopAir) on a sample image from the FSI dataset. Depth is truncated at 200m.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

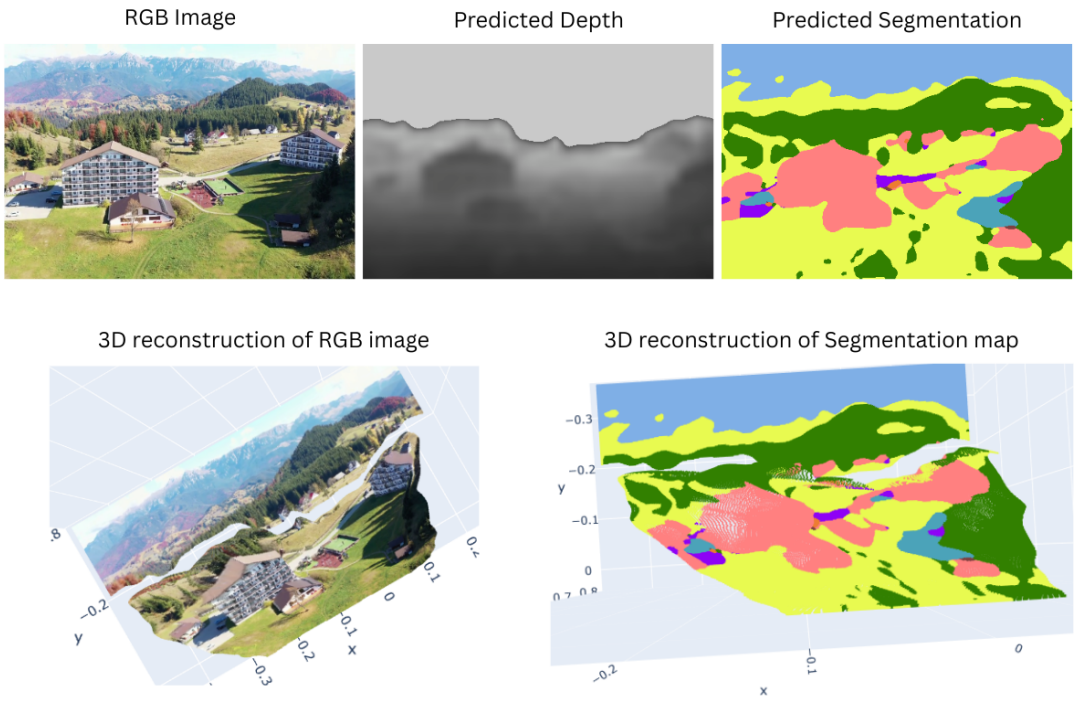


Figure 4.19 3D semantic map reconstruction using the predicted depth and segmentation maps using TaskPrompter (trained on SkyScenes + SynDrone + DroneScapes) on a sample image from DroneScapes test data. Depth is truncated at 200m

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 5

Exploring Image-Style Transfer

Summary

In this chapter, image-style transfer techniques are explored with the aim of transferring from the synthetic style to the real style for aerial images. The adopted methods are Cycle-GAN and Diffusion models. We first give a brief recount of these methods and explain the main principle behind each of them. After that, we present our experimental setup and the parameters used during the training of both methods. In the end, the preliminary results we obtained for the style transfer of images from *TopAir* and *SynDrone* are discussed. While Cycle-GAN focused on altering the colors of the input images to make them look realistic, the Diffusion model changed the content in the input image with realistic objects. Unfortunately, diffusion models change the semantic layout of the input images. However, the obtained results open the door for further research in that area.

5.1 Adopted Methods

For further exploration, we try adopting image style transfer techniques for converting the appearance of the synthetic images to a realistic style. For this purpose, we use image generative models: Cycle-GANs (based on Generative Adversarial Networks GAN) [4] and InST (based on Diffusion models) [5]. In Figures 5.1 and 5.3, the overview of the Cycle-GAN and InST methods is depicted, respectively.

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

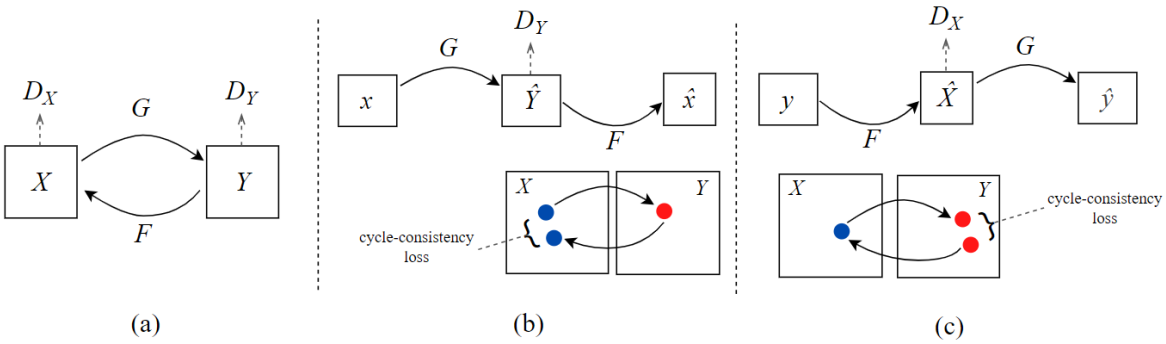


Figure 5.1 Overview of the Cycle-GAN method for image style transfer [4]. The process is composed of two mapping functions G and F (a). The loss used for training of the networks is the addition of two losses: forward cycle-consistency loss (b), and backward cycle-consistency loss (c).

5.1.1 Cycle-GAN

In Cycle-GAN, the goal is to learn the mapping between one style of images and the other without the need to align image pairs for training, because it is difficult to obtain such aligned image pairs in reality. The goal is to train the network using a group of images belonging to the source domain $\{x_i\}_{i=1}^{N}$ and another group of images belonging to the target domain $\{y_j\}_{j=1}^{M}$. While the well-known pix2pix GAN model proposed in [172] produced remarkable results in image style transfer, it requires aligned image pairs for model training. Instead, Cycle-GAN overcomes such a requirement by the incorporation of a cycle-consistency loss. As depicted in Figure 5.1, the method is composed of the synchronous training of two generative networks G and F. G learns the mapping from domain X to domain Y and F learns the inverse mapping from domain Y to X. Each of them has its own discriminator. The discriminator D_X aims to discriminate between images $\{x\}$ and translated images $\{F(y)\}$ while D_Y aims to discriminate between images $\{y\}$ and translated images $\{G(x)\}$ The intuition is that if an image is translated from one domain to the other and back, it should arrive at the starting image. Therefore, the mapping objectives are:

1. forward mapping: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$
2. backward mapping: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

For this purpose, two types of losses are used for the training of the generative networks: adversarial losses for matching the translated image to the data distribution in the target domain and cycle-consistency losses to fulfill the forward and backward mapping objectives.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Adversarial Loss: The adversarial loss for the mapping function $G : X \rightarrow Y$ and the discriminator D_Y can be formulated as the standard formula used for GANs:

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[logD_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))] \tag{5.1}$$

Where the generator G tries to minimize this objective while the discriminator D tries to maximize it. A similar adversarial loss $L_{GAN}(F, D_X, Y, X)$ is introduced for the mapping function $F : Y \rightarrow X$ and the discriminator D_X.

Cycle-Consistency Loss: Using only adversarial loss can guarantee that the generative networks G and F produce outputs belonging to the target domain distribution Y and X, respectively. However, it is not guaranteed that such generated images are the desired outputs. For example, the output can be a random permutation of images in the target domain, and it is not guaranteed that the output holds the same semantic structure as the input image, which is a necessity for synthetic-to-real style transfer. For this reason, it is essential to incorporate another type of loss, Cycle-consistency loss, to force the mapping functions to be cycle-consistent; that is the mapping of an image from source to target domain using G and back to the source domain using F should produce the same input image, for visualization refer to the images in the original paper. By using such loss, the generated output would hold the style of the target domain while being as close as possible to the original image to allow for reconstruction. This is called *forward cycle-consistency*. Similarly, as in Figure 5.1(c), *backward cycle consistency* should hold true. That is, the mapping of an image from the target domain to the source domain using F and back to the target domain using G should lead to the original image.

The cycle-consistency loss can be formulated as:

$$L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\left\|F(G(x)) - x\right\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\left\|G(F(y)) - y\right\|_1] \tag{5.2}$$

Overall Objective: The overall loss function used for training is formulated as the following:

$$L(G, F, D_x, D_y) = L_{GAN}(G, D_y, X, Y) + L_{GAN}(F, D_x, Y, X) + \lambda L_{cyc}(G, F) \tag{5.3}$$

where λ is a weighting factor used to control the relative importance of the two types of losses. By the use of the above loss, the two generative mapping networks G and F are trained to solve: $G^*, F^* = arg_{w_G, w_F} min_{G,F} max_{D_X, D_Y} L(G, F, D_x, D_y)$ where w_G and w_F are the weights of the networks.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

5.1.2 Diffusion models

Diffusion models have recently gained significant attention due to their astonishing and high-quality results in image generation [173; 174]. We hereby explain the principal concept behind such models and how we adopt InST [5] for synthetic to real style transfer.

Diffusion models try to overcome the problem of low diversity in the generated images using GANs. The main idea of a diffusion model is to learn the manifold of interest by the gradual addition of noise to images in the manifold until it becomes pure Gaussian noise and, then, learn to retrieve the original image (return back to the manifold of interest) by progressively removing the noise using a deep neural network [175].

The first step of adding Gaussian noise, also called the forward diffusion process, is a deterministic process and eventually transforms the image into pure Gaussian noise. In the second step, called the reverse sampling process, the deep model learns to remove noise gradually from a pure random noise sample to generate a new and coherent image belonging to the original data distribution. The process is illustrated in Figure 5.2.

For formulation, the denoising model can be parameterized using a function $\varepsilon_\theta(x,t)$ that tries to predict the noise component of a noisy sample x_t. By subtracting the predicted noise component from x_t we reach x_{t-1}. Proceeding with the same denoising process, we can obtain gradually $x_{t-2}, x_{t-3}, \ldots$ until we reach the original sample x_0. To train the function $\varepsilon_\theta(x,t)$, each sample in a minibatch is produced by selecting a random sample x_0, a time step t, and a noise ε, and this produces a noisy sample x_t such that the network tries to predict the noise component ε using this sample. Therefore, the training objective is $\left\|\varepsilon_\theta(x_t,t)-\varepsilon\right\|^2$. It was shown that the denoising distribution can be modelled as a Gaussian distribution $\mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\sum_\theta(x_t,t)$. It was found that fixing the variance $\sum_\theta(x_t,t)$ to a constant is better for sampling with fewer diffusion steps. It is suggested to parameterize the variance $\sum_\theta(x_t,t)$ with a neural network that is trained together with the network $\varepsilon_\theta(x_t,t)$. In the end, the image generation performance (sample quality) is evaluated using the FID score [176] metric that measures the Frechet Inception Distance between the generated sample and the real samples.

InST is a type of conditional image synthesis using Stable Diffusion Models (SDM). It is originally defined as a method to create artistic image(s) from a given example. This is done by combining the content of the input image and the style of an example image. The example styling image is converted to a short text prompt using CLIP. Then, the diffusion-based method generates high-quality and variant artistic images conditioned on the text prompt describing the style. In Figure 5.3, an overview of the method is depicted. During training, the content image x is the same as the style image y. The image embedding of

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

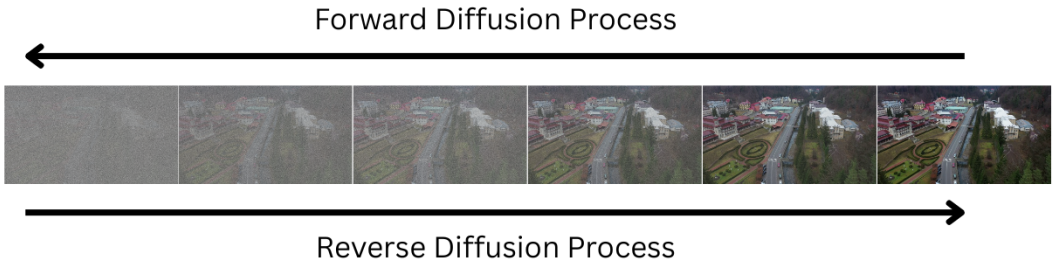


Figure 5.2 a simple illustration of the forward and reverse diffusion processes

image y is obtained using CLIP image encoding. Then, it enters into an inversion attention module to convert the image embedding $\tau_\theta(y)$ into a text embedding v, and it is converted to the standard format $c_\theta(y)$ for caption conditioning SDM. Through a sequential denoising process, the diffusion model conditioned on the image caption $c_\theta(y)$, produces the artistically converted image z. During inference, the textual embedding v of the style image y guides the conversion process of the content image x.

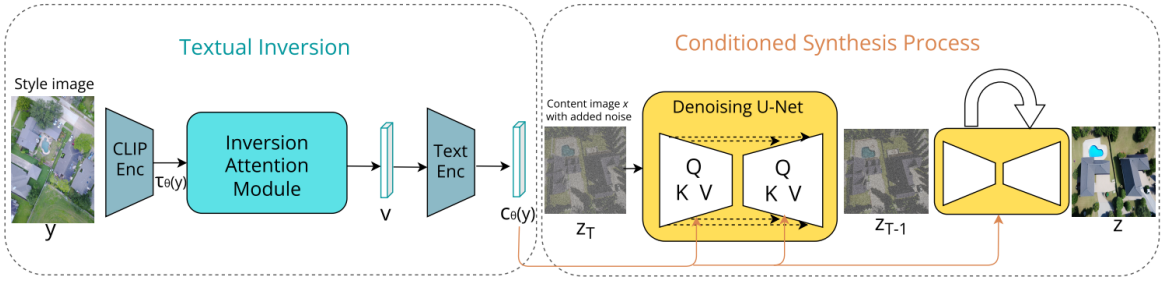


Figure 5.3 Overview of InST method for image style transfer using Stable Diffusion Models (SDM) [5]. CLIP [6] is used to give image embedding of the style image, then it is inserted as a caption conditioning in the standard form of SDM.

5.2 Experimental Setup

For Cycle-GAN, the default architecture was adopted for the generative and discriminator networks[1]. The generative network contains 3 convolutions, 9 residual blocks, two fractionally-strided convolutions with stride $\frac{1}{2}$, and one final convolution to map features to the output RGB image. For the discriminators, the patch-level architecture was adopted [177].

[1]Github repository: https://github.com/tensorflow/docs/blob/master/site/en/tutorials/generative/cyclegan.ipynb

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

For training, the weighting factor λ is set to 10. Adam optimizer is used with a batch size of 1. The learning rate is set to 2×10^{-4} for generative and discriminator networks, and they are trained from scratch. The model is trained for a maximum of 50 epochs, and the best checkpoint is used for evaluation. The resolution of input and predicted images is 256×256. The implementation is done on TensorFlow. Two experiments were run: the first to train the model on the conversion from MidAir+TopAir to real images, and the second to train it on the conversion from SkyScenes+SynDrone to real images. The group of real images used during training contains 209 images randomly selected from the datasets UDD, FSI, AeroScapes, WildUAV, ICG, and RuralScapes. The first group of synthetic images contains 173 images randomly selected from TopAir, while the second group contains 145 images randomly selected from SkyScenes.

For InST, we use the default implementation done by the authors[2]. We have finetuned the model on the same group of 209 real images used above. The configuration file for finetuning is kept as the default. For testing, the checkpoint of the latent diffusion model is used along with the embeddings learned by the finetuned model. Here, during inference, a styling image has to be provided to the network to convert the input image to the target style. For converting synthetic images of the *Neighbourhood* environment of *TopAir*, the styling image is set as a random image from the real dataset *FSI*. For converting the synthetic images of *SkyScenes* and *SynDrone*, the styling image is set as a random image from the real dataset *VisDrone* [178].

5.3 Preliminary Results

We conduct some preliminary experiments in image style transfer to evaluate the quality and realism of the style transfer of our synthetic data used for training into the real style. This is currently an active area of research in the automotive field [35; 36; 37; 146], but we seek to apply it to our synthetic aerial data to see how realistic the output is. In the case of high realism of the output, the converted images used for training can help in closing the gap of synthetic-to-real generalization of the network. This can be a line of research for future improvements, God willing, on the synthetic-to-real performance of the network.

Cycle-GAN:

In Figure 5.4, sample style-converted images from SkyScenes and SynDrone are shown. It can be noticed that Cycle-GAN focuses on changing the colors in the input images to convert them to the realistic style rather than changing the appearance and content of the images.

[2]Github repository: https://github.com/zyxElsa/InST/tree/main

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

The converted output contains noise, and the quality is not satisfactory. In Figure 5.6, sample style-converted images from TopAir are shown. The same analysis done on SkyScenes and SynDrone holds here as well; the network focused only on changing colors, hue, jittering, brightness, and illumination conditions without content modification. The main drawbacks are the noise appearing in the output and the colors of the objects are diffused outside their boundaries.

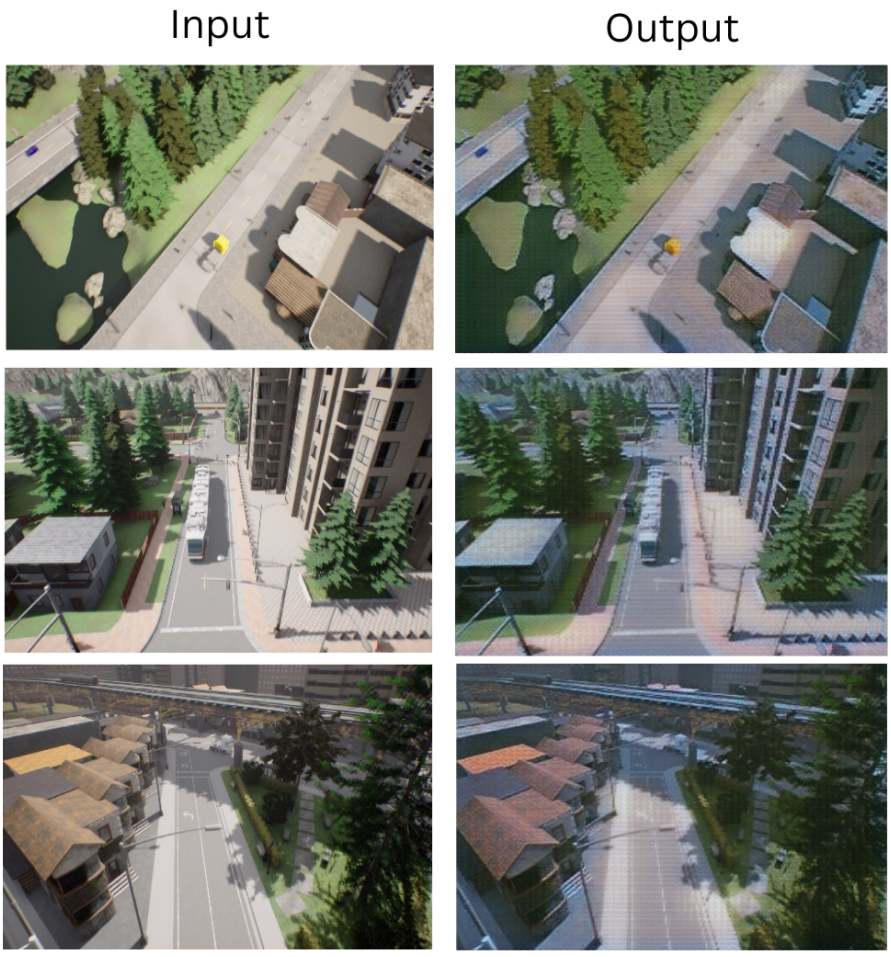


Figure 5.4 Sample output of style transfer of images from SynDrone using Cycle-GAN

Diffusion Model:

In Figure 5.6, sample style-converted images from SynDrone are shown, and in Figure 5.7, sample style-converted images from TopAir are shown. The same input images used in Cycle-GAN are used here for testing the diffusion model. It can be noted that the network here focused on changing the content in the input images to reflect realistic elements instead of the synthetic ones. See, for example, the bus, buildings, roads, and the trees in Figure 5.6.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Figure 5.5 Sample output of style transfer of images from TopAir using Cycle-GAN

However, while the output looks very realistic, it does not hold exactly the same semantic segmentation layout as the input, see the third column in Figures 5.6 and 5.7. In the second row of Figure 5.7, the cars and the swimming pool are removed in the output, and the position

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

and size of the houses have been changed. In the third row, the position of the trees is changed. For this reason, the converted images cannot be used for supervised training of neural networks with the same ground truth depth and semantic segmentation maps belonging to their synthetic pairs because they are no longer valid. However, the obtained results open the door for further research in that area.

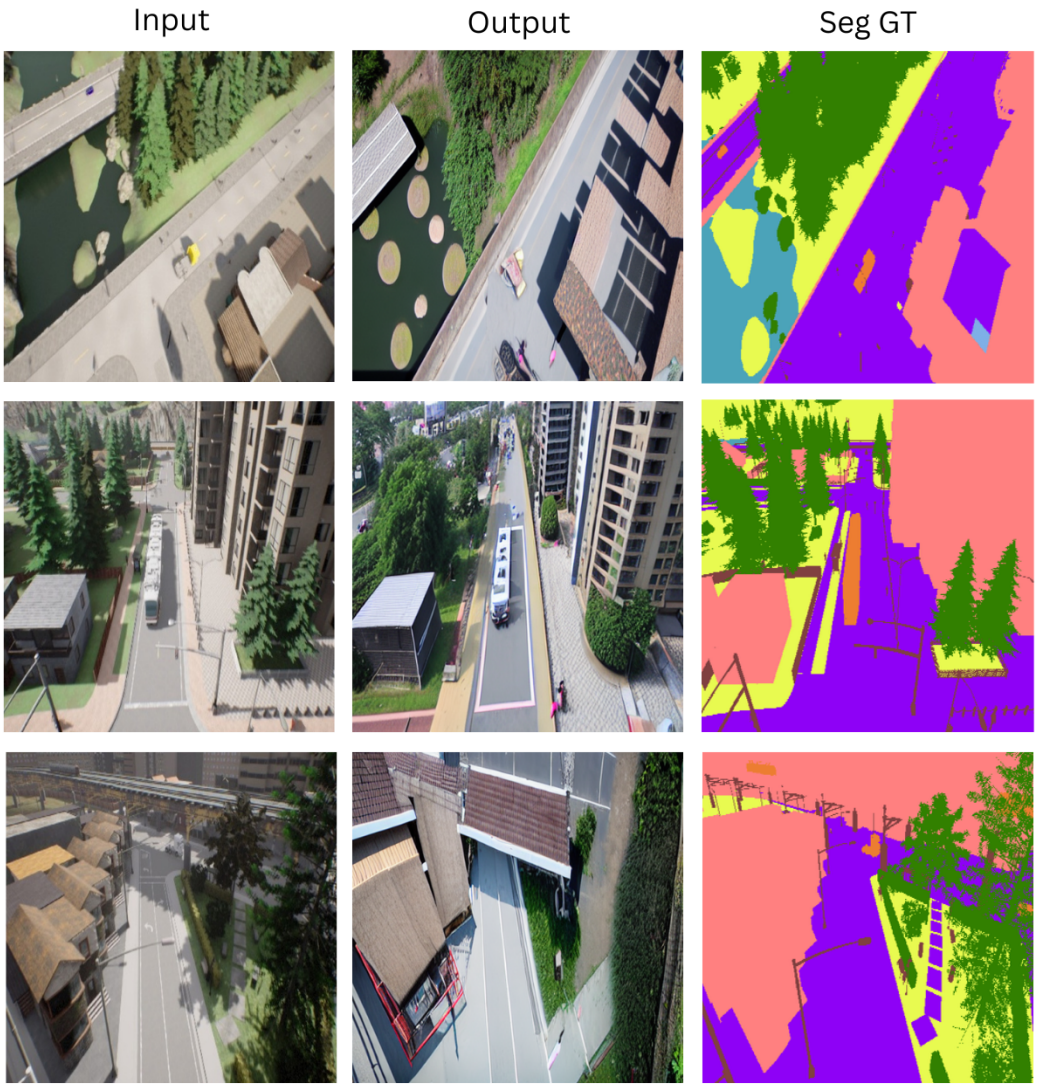


Figure 5.6 Sample output of style transfer of images from SynDrone using Diffusion Model InST. It can be noted that the output images do not hold exactly the same semantic segmentation layout.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

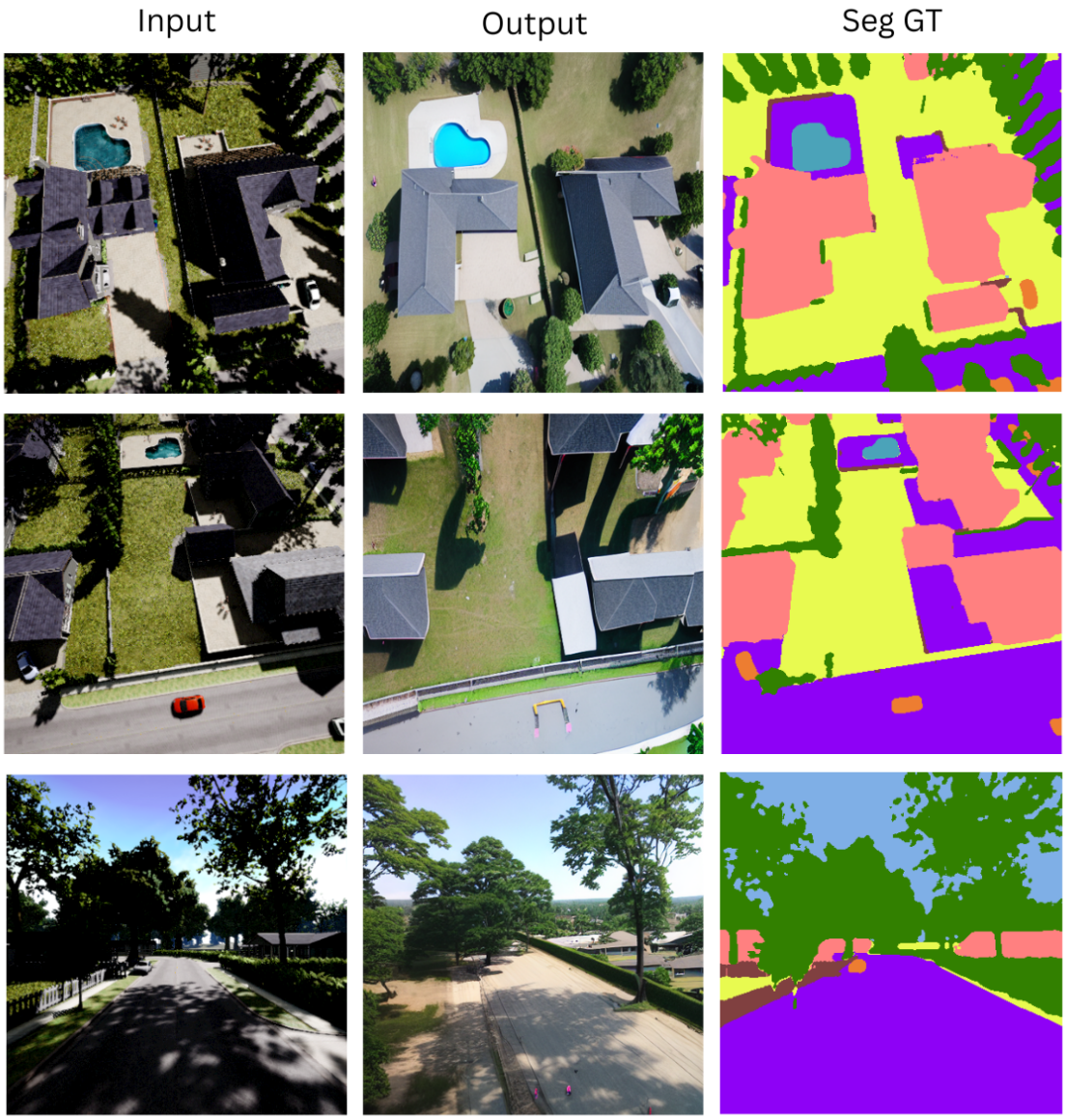


Figure 5.7 Sample output of style transfer of images from TopAir using Diffusion Model InST. It can be noted that the output images do not hold exactly the same semantic segmentation layout.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 6

Application in the Marine Domain

Summary

In this chapter, the methodology adopted to apply in the maritime domain is described, and the collected synthetic marine data *MidSea* is presented. Then, the setup used in conducting the experiments in the maritime domain is revealed, and the experiments done on the marine datasets are analyzed.

MidSea is a dataset (with $\sim 100k$ frames) collected using UnrealEngine in a synthetic marine environment, and it is annotated with depth, segmentation maps, and camera pose information. Experiments in the maritime domain were twofold: one for validating Co-SemDepth on synthetic data (MidSea), and the other for assessing the synthetic-to-real generalization of Co-SemDepth using different techniques, including self-supervised approaches, to enhance the generalization. The results obtained indicate that Co-SemDepth, trained from scratch on MidSea, can perform well on both the MidSea test data and in semantic segmentation on real data from the SMD dataset. However, it did not perform well on the real MIT dataset, even after applying various techniques.

The work in this chapter was conducted in the master's thesis of Gabriele Dellepere under the supervision of the authors of this thesis.

6.1 MidSea Data Acquisition

UnrealEngine5 was exploited to collect annotated marine data in the synthetic Tropical Island environment[1], see Figure 6.1 for a sample of the collected data. Such a dataset was

[1] Available on this link

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

named *MidSea* to reflect its content that was captured in the middle of the sea. It is annotated with depth and semantic segmentation maps as well as camera transformation data. Ships, boats, and buoys were inserted manually in the sea of the environment, and the semantic segmentation classes of objects were limited to 4 classes: Sky, Water, Ship, and Land. In this case, the class *Ship* contained ships, boats, and buoys. For depth, the sea surface in all images appeared as a linear gradient in the 2D depth maps. The data comprises a total of 8 trajectories, containing in total around 98K frames. A train-validation split of $80\% - 20\%$ was used.

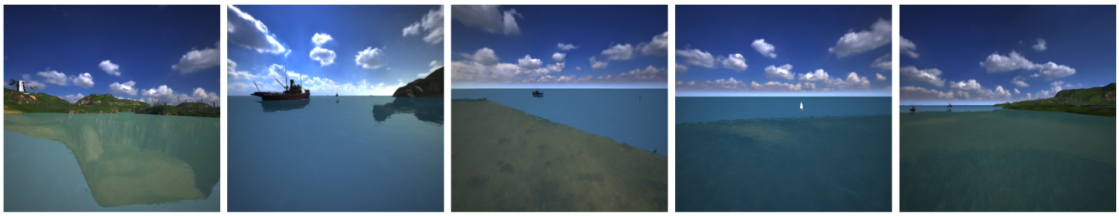


Figure 6.1 Sample images from the collected marine data in the Tropical Island environment

6.2 Coping with few annotated data

For assessing the monocular depth estimation and semantic segmentation in marine scenes, Co-SemDepth was employed for training and evaluation. In addition to the usual supervised training of the network, self-supervised methods and training modalities were adopted on the architecture of Co-SemDepth for assessment and to help the sim-to-real transition of the network. Namely, MoCo [179] and DINO [180] self-supervised approaches were used. MoCo is a method for training the encoder in a self-supervised way. It consists of using two instances of the same encoder (+ a projection head for each of them) for extracting feature maps from sufficiently large unannotated datasets of images. The objective of training the first encoder, through gradient descent and backpropagation, is that the predicted features of an image should be as close as possible to the predicted features of an augmented version of the same image using the second encoder. At the same time, the predicted features of the first encoder should be as far as possible from all other images in the dataset (named "negative examples"). The second encoder is not trained using gradient descent; however, its weights are updated with an exponential moving average of the weights of the first encoder. By using MoCo, the encoder of any neural network can be trained in a self-supervised way by making it invariant to image augmentations and without the need for data labeling. In this

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

work, MoCo is employed for the training of the encoder of Co-SemDepth, and the network performance is compared with the other modalities of training.

DINO is another self-supervised approach for pre-training the encoder of a deep network to extract features from the image regardless of color, illumination, and rotation variations. Similar to MoCo, two versions of the same encoder are used: a student and a teacher. The predicted feature maps of the same image with and without augmentations are compared to each other. The teacher is updated using an exponential moving average of the student weights. The main differences from MoCo lie in the type of augmentations done on the images and in the mechanisms adopted to avoid collapse to constant or meaningless outputs. The augmentations done on the image are both global and local. The teacher network only sees the global augmentations, while the student network sees all types of augmentations, thus forcing the student encoder to match local features and global features and become robust against different scales.

Avoiding collapse means avoiding the encoder student from predicting always a constant map (or maps where always one dimension dominates over the other) to match the teacher output. Such a prediction is called a trivial solution. DINO uses two operations to solve this issue: centering and sharpening of the output. In centering, the moving average of the teacher's logits is subtracted from the output before applying softmax. This step forces balance in the teacher's output. Sharpening means using a small temperature for the Softmax's output of the teacher so that its class prediction becomes more confident in one class. This pushes toward non-uniform outputs. Combining the two operations pushes the network to use all output dimensions and ensures high variability in the output feature predictions.

6.3 Experimental Setup

All the default settings previously described for the validation of Co-SemDepth are kept in the marine experiments[2].The number of epochs in each experiment depended on the convergence speed of the plotted metrics on the validation data to avoid overfitting.

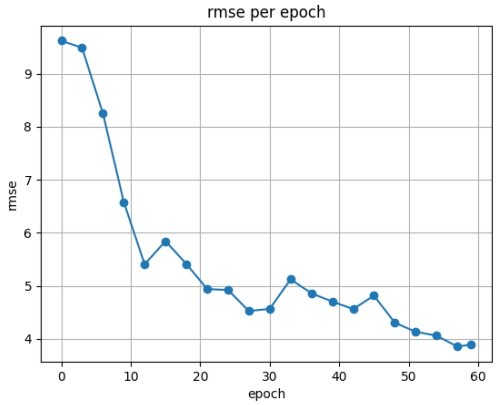
The dataset used for training is the collected synthetic **MidSea** dataset. Image resolution of 384×384 is used. The validation plots of *mIoU* and *RMSE* during training of Co-SemDepth can be viewed in Figures 6.2b and 6.2a.

For testing, two real datasets are used for only a qualitative evaluation due to the lack of depth and segmentation annotations:

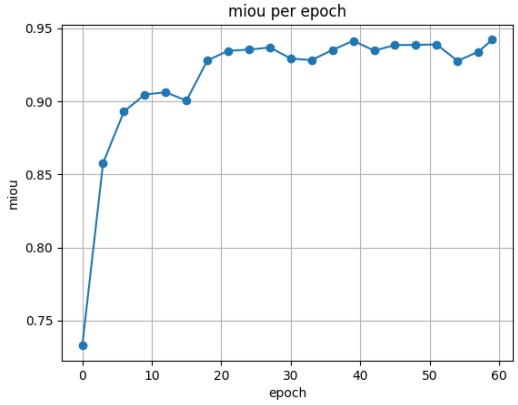
[2]The experiments are conducted using a PyTorch re-implementation of Co-SemDepth. The code can be found here: https://github.com/GabrieleDellepere/thesis.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*



(a) Validation RMSE per epoch



(b) Validation mIoU per epoch

Figure 6.2 Plots of the metrics evaluated on the validation data of MidSea while training Co-SemDepth

- **SMD**: contains clear scenes without water reflections or bad weather. However, the IMU measurements are not available. Hence, only semantic segmentation evaluation can be done on this dataset because the parallax decoder branch requires camera transformation data.
- **MIT**: it contains a variety of real-life data annotated with multiple sensor measurements, including the IMU. Thus, this dataset can be used for predicting both depth and semantic segmentation maps using the Co-SemDepth architecture.

The workstation used has 16GB RAM, an Intel Core i7 processor, and a single NVIDIA Quadro P5000 GPU card running CUDA11.4 with CuDNN 7.6.5 and Ubuntu OS.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

6.4 Results

6.4.1 Co-SemDepth Validation on Synthetic Data

For testing the application of Co-SemDepth in the marine domain, we train and test it on the collected MidSea synthetic marine data. It can be seen from the plots in Figure 6.2 that the network shows promising results when trained on MidSea and tested on the validation split of the same dataset. Both depth and segmentation metrics continuously improve over the training epochs. In Figure 6.3, sample predictions on the validation data can be visualized. It can be concluded that the network did not overfit the training data; however, it can be the case that there is little diversity in the dataset itself. For testing the network on other datasets, we choose the weights of the last epoch, as it gives the best validation performance. In Table 6.1, the quantitative results on the validation set of MidSea can be found.

Method	Semantic Metrics	Depth Metrics				
	mIoU ↑	RMSE ↓	AbsRelErr ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Base CoSemDepth	83.38%	4.24	0.092	81.7%	95.5%	98.5%
CoSemDepth + custom augm.	91.21%	3.59	0.081	88.0%	97.3%	98.9%
Finetuned MoCo	75.0%	9.03	0.265	67.3%	79.1%	87.4%
ViT+DINO encoder	79.11%	9.42	0.221	65.0%	75.2%	83.4%

Table 6.1 Comparison of different methods on the collected Synthetic dataset MidSea. Base CosemDepth includes the default augmentations, while the "custom augm." version includes more aggressive color jittering and z-axis rotation.

Figure 6.3 Sample prediction of depth and segmentation maps on MidSea. Depth is shown in a linear scale, and distance values greater than 255 meters are clamped to 255; that's why the far island disappears in the depth map. Semantic segmentation color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

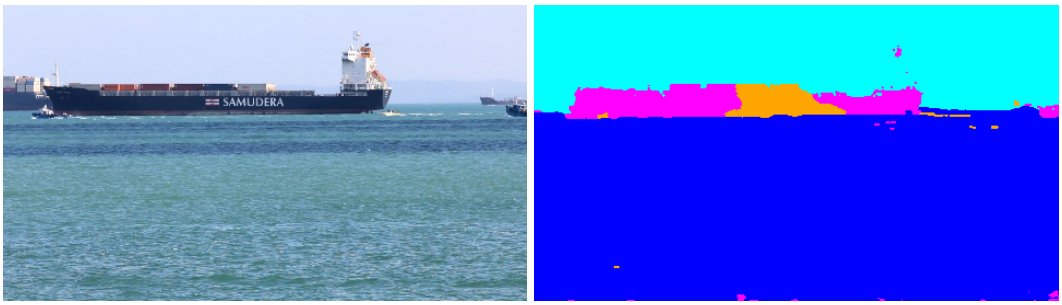
Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

6.4.2 Synthetic-to-Real Performance

Results on SMD:

The synthetic-to-real zero-shot performance of Co-SemDepth is evaluated by training Co-SemDepth on MidSea and testing it on images from SMD real data. Unfortunately, there are no semantic or depth annotations provided in SMD, and for this reason, the evaluation is done only qualitatively. In addition, the lack of camera pose data in SMD forces us to predict only semantic segmentation. In general, the results turn out positive. Sample predictions can be found in Figures 6.4 and 6.5. It can be seen that the network is able to identify the main elements of the scene (i.e. water, sky, and ships) even in hard weather conditions like fog. However, the network can sometimes be confused between ships and land, especially when the ships are far away.

(a) sample prediction 1



(b) sample prediction 2

Figure 6.4 Sample semantic predictions on SMD using Co-SemDepth trained on the synthetic MidSea. Semantic segmentation color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

Results on MaSTr:

In addition, we run tests on the real dataset MaSTr [135]. In this case, we can conduct a quantitative evaluation due to the availability of semantic segmentation annotation in the

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

(a) sample prediction 1

(b) sample prediction 2

Figure 6.5 Sample semantic predictions on foggy images from SMD using Co-SemDepth trained on the synthetic MidSea. Semantic segmentation color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

dataset. The network Co-SemDepth trained on MidSea could achieve 45.84%*mIoU* in the zero-shot testing and 49.41%*mIoU* after fine-tuning on part of MaSTr for 20 epochs. Hence, there is an enhancement; however, it is not huge.

Results on MIT

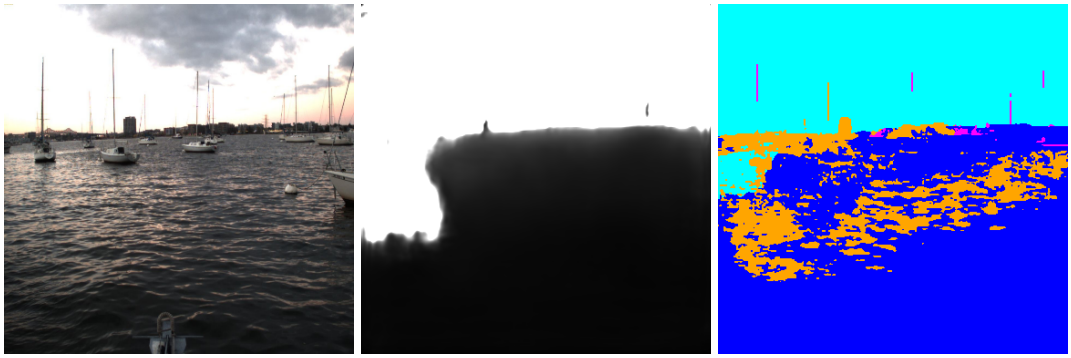
The MIT Sea Grant Perception dataset is considered the most diverse dataset containing marine scenarios and weather conditions. For this reason, it is selected as our high target. **Zero-shot prediction:** It can be seen in Figure 6.6 sample predictions on images from MIT data using Co-SemDepth trained on MidSea. Unfortunately, the network performance here is not excellent. For semantic segmentation, it can be noted that the network can well distinguish between sky and land, especially in good weather conditions. However, when it comes to distinguishing the line between land and the sea, it struggles. Besides, it can be observed that Co-SemDepth is not able to properly identify the ships in the input image, and this is a problem since most of the applications in the marine domain, like exploration or obstacle avoidance, require the detection of the surrounding vehicles. For depth estimation, the situation is not good enough. The network predicts all the sea and land areas as a

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

linear gradient, and it is not able to well detect the horizon line. It can also be noted from Figures 6.6a and 6.6b that the sunlight has an effect on the predicted depth maps. The areas of the sea illuminated with high sunlight are predicted with high depth values (greater than 255 meters).

(a) sample prediction 1



(b) sample prediction 2

(c) sample prediction 3

Figure 6.6 Sample zero-shot predictions on images from MIT dataset using Co-SemDepth trained on the synthetic MidSea. Depth is displayed in a linear scale and clamped at 255 meters. Semantic segmentation color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Finetuning on MaSTr: To overcome the above challenges, Co-SemDepth, which is finetuned on the MaSTr dataset, is tried to predict semantic segmentation on the MIT dataset. In this case, no depth predictions can be performed because the MaSTr dataset does not contain camera pose data. The output results can be found in Figure 6.7. From the figure, the network is more able to detect the sea area. However, it is prone to detecting land areas as sea. For this reason, the results of this solution are also not satisfying, and the problem of undetected vehicles persists.

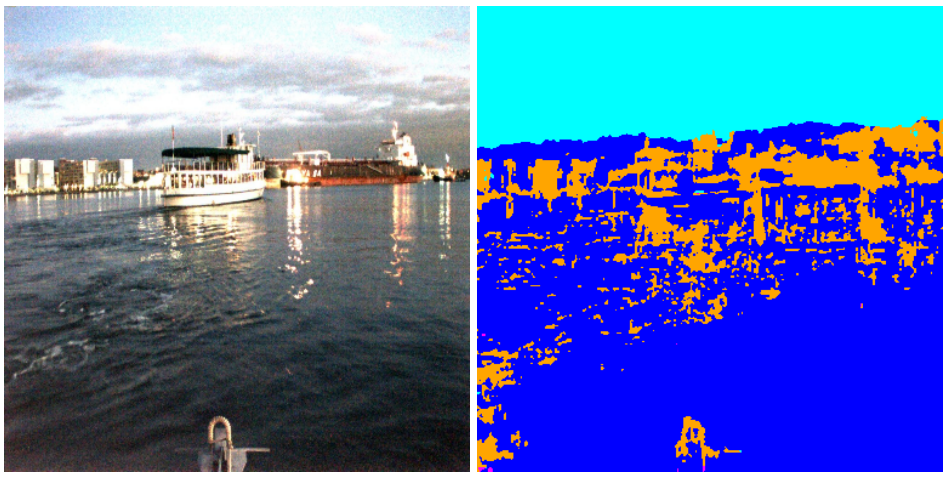


Figure 6.7 Sample prediction on MIT dataset after finetuning Co-SemDepth on MaSTr dataset. Color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

Augmentations: More augmentations are added to the training of Co-SemDepth in the hope of making the images of MidSea more diverse and, thus, making the network generalize better to the unseen data of MIT. In particular, the range of color jittering is increased, and more cropping and z-axis rotations are added. In fact, the metrics obtained on the validation set of MidSea are promising compared to the values obtained with the default augmentations. The results can be found in Table 6.1. Unexpectedly, the results on MIT got worse, see Figure 6.8. The network is completely confused between land and sea; this can be because increasing the color jittering so much makes the color of the sea and land in the training data change significantly, and that leads the network to not be able to distinguish them by color anymore. The network is still unable to detect the ships in the scene well. For depth estimation, the network is able to well detect the horizon line and this is an improvement compared to the model with default augmentations. However, it still cannot distinguish between the distances of the sea surface and vehicles in it. Such augmentations are not used in the following experiments due to the bad segmentation performance.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Figure 6.8 Sample predictions on MIT dataset using Co-SemDepth after adding custom augmentations. Depth is viewed in metric scale. Segmentation color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

MoCo pretraining: The MoCo self-supervised pretraining method is adopted for the training of the encoder of Co-SemDepth on both MidSea and MIT datasets. The MLP used after the encoder to aggregate the features during pretraining is composed of two layers with 2048 and 128 perceptrons with Leaky ReLU activation. The encoder is pretrained for 80 epochs. After that, the whole architecture (pre-trained encoder + decoder) is trained on MidSea with a learning rate one order of magnitude less than the one used for pre-training. The network is trained for 50 epochs. The evaluation results on MidSea validation data can be found in Table 6.1. It can be seen that using the MoCo pretraining method makes the model perform well on MidSea (although the results are inferior to the base network). However, by looking at the output maps on images from the MIT dataset, Figure 6.9, it can be concluded that the results using the MoCo pretraining method are not satisfactory. A probable reason for that can be that MoCo needs a huge number of good negative examples to work well, and in our case, there is little diversity in the data we have. So, the difference between positive examples and negative examples is minimal, not sufficient for a proper training of the encoder.

DINO pretraining: In this experiment, the DINO self-supervised method is adopted for evaluation. However, in this case, we load the encoder weights available on the DINOv3 HuggingFace repository[3]. We discard the encoder of Co-SemDepth and use instead the pretrained ViT encoder of DINO with the decoder of Co-SemDepth. The network is fine-tuned after that on MidSea. From Table 6.1, it can be noted that the results achieved with this approach are better than MoCo. However, they are still inferior to the ones achieved by Base Co-SemDepth.

[3]https://huggingface.co/docs/transformers/main/model_doc/dinov3

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

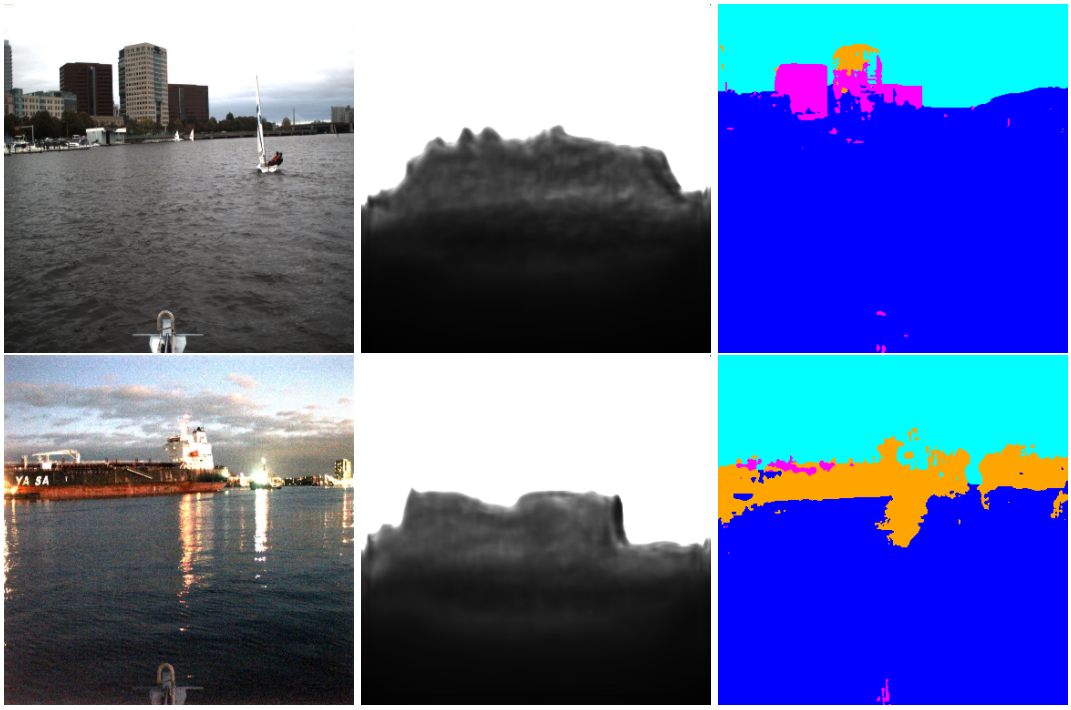


Figure 6.9 Sample predictions on the MIT dataset using Co-SemDepth after finetuning the weights obtained from MoCo pre-training. Depth is viewed in metric scale. Segmentation color index is: Blue=Sea, Cyan=Sky, Yellow=Ground/Buildings/Vegetation, Magenta=Other

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Chapter 7

Conclusions

In this thesis, we presented Co-SemDepth, a fast joint architecture for depth estimation and semantic segmentation on UAV images. Our architecture proved to be lightweight (containing only 5.2 M parameters) and fast (with a frame rate of 20 FPS) compared to the other state-of-the-art methods while achieving better or on-par accuracies (*RMSE* of 6.7 and *mIoU* of 75.4% on MidAir). This makes it suitable to be deployed on UAV hardware for conducting real-time scene analysis. By using our light architecture, it is possible for a UAV to analyze its surrounding scenes autonomously and independently using its onboard microcontroller or single board computer that are limited in memory; our architecture requires only $6.2 GB of GPU memory$. In this case, there will be no need for sending data to a remote server to carry out the analysis. The output depth and semantic maps can provide geometric and semantic understanding of the scene that is necessary to carry out a variety of UAV missions (including 3D SLAM).

Moreover, to help in filling the gap of the scarcity of available annotated datasets in the aerial field, the *TopAir* dataset was introduced. It is a synthetic dataset collected using the AirSim simulator in a variety of UnrealEngine environments at different altitudes and lighting conditions. The dataset contains annotations of depth and segmentation maps as well as camera translation and orientation data for every frame.

In addition, the synthetic-to-real domain shift problem was analyzed. Our architecture (light U-Net-based) and TaskPrompter (large Transformer-based architecture) were used to conduct an extensive analytical study in synthetic-to-real domain generalization of monocular depth estimation and semantic segmentation for aerial robots. The networks were trained on synthetic datasets and evaluated on real data. The study reveals that Co-SemDepth is better in the generalization of depth estimation due to the parallax estimation modalities implemented in the architecture. On the other hand, the results reveal that TaskPrompter is

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

better in semantic segmentation, as it seems that for this task, it is more beneficial to have a complex, large network like TaskPrompter that can capture the semantic potential of the data, especially for segmenting small objects. In general, it was observed that the similarity between the environments (rural or urban) appearing in the synthetic data used for training and the real data used for testing greatly affects the domain shift network performance. Specifically, using the synthetic MidAir and TopAir datasets for training the models led to better generalization to the real data that were captured in rural areas, while using the synthetic SkyScenes and SynDrone datasets was preferable to make the model generalize to real datasets of urban nature.

Additional experiments were conducted to test the effect of adding a small percentage of real data to the synthetic data during training (few-shot learning). The obtained results were positive on some datasets and disappointing on others. As a whole, there is a promising line for further research in that area. 3D semantic maps of the scenes were constructed using the predicted depth and semantic segmentation maps.

Furthermore, we have explored image style transfer techniques using Cycle-GANs and Diffusion models (InST method) for converting the style of the synthetic images to a realistic appearance. From the results, it was noted that Cycle-GAN focused on changing the colors of the pixels in the input images while maintaining the same semantic layout. On the other hand, the Diffusion model changed the elements appearing in the input image to their realistic counterparts. However, the semantic layout of the converted images changed with respect to the ground truth. The obtained preliminary results open the door for further research on closing the gap between the synthetic and real domains.

In the end, experiments were conducted in the maritime domain, trying to address the specific challenges related to this domain in depth estimation and semantic segmentation. A synthetic dataset was collected using UnrealEngine called *MidSea*, and a synthetic-to-real assessment of Co-SemDepth and self-supervised approaches was carried out. The results obtained were good on the MidSea dataset as well as when transferring to the real SMD dataset. However, for the MIT marine dataset that contains various weather conditions, the results were not acceptable using Co-SemDepth trained from scratch. To improve the performance, MoCo and DINO self-supervised approaches were adopted to pretrain the encoder of the network, and image augmentations were increased during training. However, the results were still unsatisfactory after applying these techniques. Future work can be directed towards collecting more diverse synthetic data that covers many challenging marine conditions and use such data for training the network to enhance its synthetic-to-real generalization performance.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

References

[1] C. Griwodz, S. Gasparini, L. Calvet, P. Gurdjos, F. Castan, B. Maujean, G. D. Lillo, and Y. Lanthony, "Alicevision Meshroom: An open-source 3D reconstruction pipeline," in *Proceedings of the 12th ACM Multimedia Systems Conference - MMSys '21*, ACM Press, 2021.

[2] M. Fonder, D. Ernst, and M. Van Droogenbroeck, "Parallax inference for robust temporal monocular depth estimation in unstructured environments," *Sensors*, vol. 22, pp. 1–22, November 2022.

[3] H. Ye and D. Xu, "Taskprompter: Spatial-channel multi-task prompting for dense scene understanding," in *The Eleventh International Conference on Learning Representations*, 2022.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[5] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10156, 2023.

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[7] M. Vel'as, M. Španěl, Z. Materna, and A. Herout, "Calibration of rgb camera with velodyne lidar," 2014.

[8] C. F. Olson and H. Abi-Rached, "Wide-baseline stereo vision for terrain mapping," *Machine Vision and Applications*, vol. 21, pp. 713–725, 2010.

[9] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1164–1174, 2021.

[10] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4009–4018, 2021.

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[11] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 270–279, 2017.

[12] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 405–420, 2018.

[13] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[15] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018.

[16] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 1451–1460, Ieee, 2018.

[17] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

[19] R. Bachmann, O. F. Kar, D. Mizrahi, A. Garjani, M. Gao, D. Griffiths, J. Hu, A. Dehghan, and A. Zamir, "4m-21: An any-to-any vision model for tens of tasks and modalities," *Advances in Neural Information Processing Systems*, vol. 37, pp. 61872–61911, 2024.

[20] T. Specs, "Arduino uno." https://docs.arduino.cc/hardware/uno-rev3/#tech-specs. Accessed: 2025-04-17.

[21] T. Specs, "Arduino portenta h7." https://docs.arduino.cc/hardware/portenta-h7/#tech-specs. Accessed: 2025-04-17.

[22] T. Specs, "Raspberry pi." https://www.raspberrypi.com/. Accessed: 2025-04-17.

[23] T. Specs, "Jetson modules." https://developer.nvidia.com/embedded/jetson-modules. Accessed: 2025-04-17.

[24] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[25] M. A. Wani, F. A. Bhat, S. Afzal, and A. I. Khan, "Basics of supervised deep learning," in *Advances in Deep Learning*, pp. 13–29, Springer, 2019.

[26] T. Huang, S. Zhao, L. Geng, and Q. Xu, "Unsupervised monocular depth estimation based on residual neural network of coarse–refined feature extractions for drone," *Electronics*, vol. 8, no. 10, p. 1179, 2019.

[27] Z. Ghahramani, "Unsupervised learning," in *Summer school on machine learning*, pp. 72–112, Springer, 2003.

[28] X. J. Zhu, "Semi-supervised learning literature survey," 2005.

[29] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *2010 IEEE Computer society conference on computer vision and pattern recognition*, pp. 902–909, IEEE, 2010.

[30] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[31] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1206–1217, 2023.

[32] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*, pp. 621–635, Springer, 2018.

[33] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*, pp. 1–16, PMLR, 2017.

[34] T. Loiseau, T.-H. Vu, M. Chen, P. Pérez, and M. Cord, "Reliability in semantic segmentation: Can we use synthetic data?," in *European Conference on Computer Vision*, pp. 442–459, Springer, 2024.

[35] A. Xiao, J. Huang, D. Guan, F. Zhan, and S. Lu, "Transfer learning from synthetic to real lidar point cloud for semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, pp. 2795–2803, 2022.

[36] C. Zheng, T.-J. Cham, and J. Cai, "T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

[37] Y. Chen, W. Li, X. Chen, and L. V. Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1841–1850, 2019.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[38] S. Khose, A. Pal, A. Agarwal, Deepanshi, J. Hoffman, and P. Chattopadhyay, "Skyscenes: A synthetic dataset for aerial scene understanding," in *European Conference on Computer Vision*, pp. 19–35, Springer, 2024.

[39] C. Hinniger and J. Rüter, "Synthetic training data for semantic segmentation of the environment from uav perspective," *Aerospace*, vol. 10, no. 7, p. 604, 2023.

[40] S. Kale, "Developments in unmanned surface vehicles (usvs): A review," in *Int. Conf. Appl. Eng. Nat. Sci*, vol. 1, pp. 596–600, 2023.

[41] U. Engine, "Unreal engine," *Retrieved from Unreal Engine: https://www. unrealengine. com/en-US/what-is-unreal-engine-4*, 2018.

[42] Y. AlaaEldin and F. Odone, "Co-semdepth: Fast joint semantic segmentation and depth estimation on aerial images," *SAC ACM*, 2026.

[43] F. O. Yara AlaaEldin and M. M. Ata, "A study on synthetic-to-real generalization of depth estimation and semantic segmentation networks on aerial images," *Complex and Intelligent Systems*, 2025.

[44] Y. Li, M. Liu, and D. Jiang, "Application of unmanned aerial vehicles in logistics: a literature review," *Sustainability*, vol. 14, no. 21, p. 14473, 2022.

[45] S. A. H. Mohsan, M. A. Khan, F. Noor, I. Ullah, and M. H. Alsharif, "Towards the unmanned aerial vehicles (uavs): A comprehensive review," *Drones*, vol. 6, no. 6, p. 147, 2022.

[46] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 969–977, 2018.

[47] T. A. Le, A. G. Baydin, R. Zinkov, and F. Wood, "Using synthetic data to train neural networks is model-based reasoning," in *2017 international joint conference on neural networks (IJCNN)*, pp. 3514–3521, IEEE, 2017.

[48] V. Seib, B. Lange, and S. Wirtz, "Mixing real and synthetic data to enhance neural network training–a review of current approaches," *arXiv preprint arXiv:2007.08781*, 2020.

[49] T. Alkhalifah, H. Wang, and O. Ovcharenko, "Mlreal: Bridging the gap between training on synthetic data and real data applications in machine learning," *Artificial Intelligence in Geosciences*, vol. 3, pp. 101–114, 2022.

[50] J. Kopf, X. Rong, and J.-B. Huang, "Robust consistent video depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1611–1621, 2021.

[51] Z. Teed and J. Deng, "Deepv2d: Video to depth with differentiable structure from motion," *arXiv preprint arXiv:1812.04605*, 2018.

[52] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2800–2810, 2018.

[53] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.

[54] H. Xu, F. Li, and Z. Feng, "Mlffnet: multilevel feature fusion network for monocular depth estimation from aerial images," *Journal of Applied Remote Sensing*, vol. 16, no. 2, pp. 026506–026506, 2022.

[55] V.-C. Miclea and S. Nedevschi, "Dynamic semantically guided monocular depth estimation for uav environment perception," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2023.

[56] V.-C. Miclea and S. Nedevschi, "Monocular depth estimation with improved long-range accuracy for uav environment perception," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[57] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, pp. 3372–3380, 2017.

[58] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248, IEEE, 2016.

[59] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[60] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambruș, and A. Gaidon, "Towards zero-shot scale-aware monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9233–9243, 2023.

[61] V. Patil, W. Van Gansbeke, D. Dai, and L. Van Gool, "Don't forget the past: Recurrent depth estimation from monocular video," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6813–6820, 2020.

[62] D. Maslov and I. Makarov, "Online supervised attention-based recurrent depth estimation from monocular video," *PeerJ Computer Science*, vol. 6, p. e317, 2020.

[63] H. Zhang, C. Shen, Y. Li, Y. Cao, Y. Liu, and Y. Yan, "Exploiting temporal consistency for real-time video depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1725–1734, 2019.

[64] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1164–1174, 2021.

[65] V. Licăret, V. Robu, A. Marcu, D. Costea, E. Sluşanschi, R. Sukthankar, and M. Leordeanu, "Ufo depth: 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.

[66] M. Fonder, D. Ernst, and M. Van Droogenbroeck, "Parallax inference for robust temporal monocular depth estimation in unstructured environments," *Sensors*, vol. 22, no. 23, p. 9374, 2022.

[67] A. CS Kumar, S. M. Bhandarkar, and M. Prasad, "Monocular depth prediction using generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 300–308, 2018.

[68] T. Shimada, H. Nishikawa, X. Kong, and H. Tomiyama, "Pix2pix-based monocular depth estimation for drones with optical flow on airsim," *Sensors*, vol. 22, no. 6, p. 2097, 2022.

[69] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, "Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network," *Neurocomputing*, vol. 517, pp. 81–92, 2023.

[70] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single rgb images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3372–3380, 2017.

[71] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248, IEEE, 2016.

[72] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[73] A. CS Kumar, S. M. Bhandarkar, and M. Prasad, "Depthnet: A recurrent neural network architecture for monocular depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 283–291, 2018.

[74] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[75] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images.," *ECCV (5)*, vol. 7576, pp. 746–760, 2012.

[76] M. Fonder and M. Van Droogenbroeck, "Mid-air: A multi-modal dataset for extremely low altitude drone flights," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

[77] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916, IEEE, 2020.

[78] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024.

[79] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 8001–8008, 2019.

[80] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, "Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network," *Neurocomputing*, vol. 517, pp. 81–92, 2023.

[81] Z. Lu and Y. Chen, "Self-supervised monocular depth estimation on water scenes via specular reflection prior," *Digital Signal Processing*, vol. 149, p. 104496, 2024.

[82] H. Jiang, L. Ding, Z. Sun, and R. Huang, "Dipe: Deeper into photometric errors for unsupervised learning of depth and ego-motion from monocular videos," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10061–10067, IEEE, 2020.

[83] F.-Z. Zhang, J.-S. Wang, and J.-Q. Hu, "Unsupervised monocular depth estimation based on multi-scale feature fusion in berthing scenarios," in *2025 IEEE 2nd International Conference on Deep Learning and Computer Vision (DLCV)*, pp. 1–6, IEEE, 2025.

[84] V.-C. Miclea and S. Nedevschi, "Monocular depth estimation with improved long-range accuracy for uav environment perception," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[85] S. Karimijafarbigloo, R. Azad, A. Kazerouni, Y. Velichko, U. Bagci, and D. Merhof, "Self-supervised semantic segmentation: Consistency over transformation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2654–2663, 2023.

[86] W. He, S. Jamonnak, L. Gou, and L. Ren, "Clip-s4: Language-guided self-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11207–11216, 2023.

[87] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 405–420, 2018.

[88] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[89] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[90] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[91] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 325–341, 2018.

[92] P. He, L. Jiao, R. Shang, S. Wang, X. Liu, D. Quan, K. Yang, and D. Zhao, "Manet: Multi-scale aware-relation network for semantic segmentation in aerial scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[93] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 405–420, 2018.

[94] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1925–1934, 2017.

[95] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8866–8875, 2019.

[96] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, "Dynamic video segmentation network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6556–6565, 2018.

[97] M. Paul, C. Mayer, L. V. Gool, and R. Timofte, "Efficient video semantic segmentation with labels propagation and refinement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2873–2882, 2020.

[98] B. Mahasseni, S. Todorovic, and A. Fern, "Budget-aware deep semantic video segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1029–1038, 2017.

[99] J. Li, W. Wang, J. Chen, L. Niu, J. Si, C. Qian, and L. Zhang, "Video semantic segmentation via sparse temporal transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 59–68, 2021.

[100] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8818–8827, 2020.

[101] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 352–368, Springer, 2020.

[102] G. Sun, Y. Liu, H. Ding, T. Probst, and L. Van Gool, "Coarse-to-fine feature mining for video semantic segmentation," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3126–3137, 2022.

Finanziato dall'Unione europea NextGenerationEU · Ministero dell'Università e della Ricerca · Italiadomani Piano Nazionale di Ripresa e Resilienza · Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[103] B. Bovcon and M. Kristan, "A water-obstacle separation and refinement network for unmanned surface vehicles," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9470–9476, IEEE, 2020.

[104] M. Weber, H. Wang, S. Qiao, J. Xie, M. D. Collins, Y. Zhu, L. Yuan, D. Kim, Q. Yu, D. Cremers, *et al.*, "Deeplab2: A tensorflow library for deep labeling," *arXiv preprint arXiv:2106.09748*, 2021.

[105] B. Bovcon, J. Perš, M. Kristan, *et al.*, "Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation," *Robotics and Autonomous Systems*, vol. 104, pp. 1–13, 2018.

[106] L. Zhang, X. Sun, Z. Li, S. Zhang, J. Liu, H. Dong, and D. Kong, "Scene-adaptive semantic segmentation guided by multi-level boundary-semantic-reinforcement for unmanned surface vessels," *Measurement*, p. 118655, 2025.

[107] S. Nedevschi *et al.*, "Weakly supervised semantic segmentation learning on uav video sequences," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 731–735, IEEE, 2021.

[108] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1499–1508, IEEE, 2018.

[109] G. Zheng, Z. Jiang, H. Zhang, and X. Yao, "Deep semantic segmentation of unmanned aerial vehicle remote sensing images based on fully convolutional neural network," *Frontiers in Earth Science*, vol. 11, p. 1115805, 2023.

[110] Q. Xu, Y. Zeng, W. Tang, W. Peng, T. Xia, Z. Li, F. Teng, W. Li, and J. Guo, "Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network," *IEEE journal of biomedical and health informatics*, vol. 24, no. 9, pp. 2481–2489, 2020.

[111] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Joint learning of motion estimation and segmentation for cardiac mr image sequences," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pp. 472–480, Springer, 2018.

[112] A. Mousavian, H. Pirsiavash, and J. Košecká, "Joint semantic segmentation and depth estimation with deep convolutional networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 611–619, IEEE, 2016.

[113] Y. Cao, C. Shen, and H. T. Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 836–846, 2016.

[114] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, "Real-time joint semantic segmentation and depth estimation using asymmetric annotations," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7101–7107, IEEE, 2019.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[115] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother, "Analyzing modular cnn architectures for joint depth prediction and semantic segmentation," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 4620–4627, IEEE, 2017.

[116] D. Bhattacharjee, T. Zhang, S. Süsstrunk, and M. Salzmann, "Mult: An end-to-end multitask learning transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12031–12041, 2022.

[117] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[118] S. Vandenhende, S. Georgoulis, and L. Van Gool, "Mti-net: Multi-scale task interaction networks for multi-task learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 527–543, Springer, 2020.

[119] G. Rizzoli, F. Barbato, M. Caligiuri, and P. Zanuttigh, "Syndrone-multi-modal uav dataset for urban scenarios," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2210–2220, 2023.

[120] L. Chen, F. Liu, Y. Zhao, W. Wang, X. Yuan, and J. Zhu, "Valid: A comprehensive virtual aerial image dataset," in *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 2009–2016, IEEE, 2020.

[121] H. Florea, V.-C. Miclea, and S. Nedevschi, "Wilduav: Monocular uav dataset for depth estimation tasks," in *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 291–298, IEEE, 2021.

[122] A. Nair and N. Mehendale, "Dronescape: A high-resolution drone footage dataset for tree region segmentation," *Available at SSRN 4512595*, 2023.

[123] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.

[124] R. Lopez-Campos and J. Martinez-Carranza, "Espada: Extended synthetic and photogrammetric aerial-image dataset," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7981–7988, 2021.

[125] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, "Capturing, reconstructing, and simulating: the urbanscene3d dataset," in *European Conference on Computer Vision*, pp. 93–109, Springer, 2022.

[126] F. Nex, E. Stathopoulou, F. Remondino, M. Yang, L. Madhuanand, Y. Yogender, B. Alsadik, M. Weinmann, B. Jutzi, and R. Qin, "Usegeo-a uav-based multi-sensor dataset for geospatial research," *ISPRS Open Journal of Photogrammetry and Remote Sensing*, p. 100070, 2024.

[127] M. Scanlon, "SynthAer - a synthetic dataset of semantically annotated aerial images," 9 2018.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[128] R. Caudron and P.-A. Nicq, *Blender 3D By Example*. Packt Publishing Ltd, 2015.

[129] C. Hinniger and J. Rüter, "Synthetic training data for semantic segmentation of the environment from uav perspective," *Aerospace*, vol. 10, no. 7, p. 604, 2023.

[130] W. Cai, K. Jin, J. Hou, C. Guo, L. Wu, and W. Yang, "Vdd: Varied drone dataset for semantic segmentation," *arXiv preprint arXiv:2305.13608*, 2023.

[131] Y. Chen, Y. Wang, P. Lu, Y. Chen, and G. Wang, "Large-scale structure from motion with semantic constraints of aerial images," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 347–359, Springer, 2018.

[132] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[133] D. Ninja, "Visualization tools for semantic drone dataset." https://datasetninja.com/semantic-drone, dec 2025. visited on 2025-12-09.

[134] S. Nirgudkar, M. DeFilippo, M. Sacarny, M. Benjamin, and P. Robinette, "Massmind: Massachusetts marine infrared dataset," *The International Journal of Robotics Research*, 2022.

[135] B. Bovcon, J. Muhovič, J. Perš, and M. Kristan, "The mastr1325 dataset for training deep usv obstacle detection models," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2019.

[136] B. Bovcon, J. Perš, M. Kristan, *et al.*, "Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation," *Robotics and Autonomous Systems*, vol. 104, pp. 1–13, 2018.

[137] B. Bovcon, J. Muhovič, D. Vranac, D. Mozetič, J. Perš, and M. Kristan, "Mods–a usv-oriented object detection and obstacle segmentation benchmark," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[138] B. Iancu, V. Soloviev, L. Zelioli, and J. Lilius, "Aboships-an inshore and offshore maritime vessel detection dataset with precise annotations," *Remote Sensing*, vol. 13, no. 5, p. 988, 2021.

[139] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "Seaships: A large-scale precisely annotated dataset for ship detection," *IEEE transactions on multimedia*, vol. 20, no. 10, pp. 2593–2604, 2018.

[140] D. D. Bloisi, L. Iocchi, A. Pennisi, and L. Tombolini, "Argos-venice boat classification," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2015.

[141] E. Gundogdu, B. Solmaz, V. Yücesoy, and A. Koc, "Marvel: A large-scale image dataset for maritime vessels," in *Asian conference on computer vision*, pp. 165–180, Springer, 2016.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[142] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 1993–2016, 2017.

[143] A. Marcu, "Quantifying the synthetic and real domain gap in aerial scene understanding," *arXiv preprint arXiv:2411.19913*, 2024.

[144] Y. Song, J. Shi, D. Zou, C. Liu, S. Bai, X. Shu, Q. Qian, D. Xu, Y. Yuan, and Y. Sun, "Transformer framework for depth-assisted uda semantic segmentation," *Engineering Applications of Artificial Intelligence*, vol. 137, p. 109206, 2024.

[145] M. Chen, Z. Zheng, and Y. Yang, "Transferring to real-world layouts: A depth-aware framework for scene adaptation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 399–408, 2024.

[146] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2800–2810, 2018.

[147] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

[148] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

[149] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[150] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: a review," *Technologies*, vol. 8, no. 2, p. 35, 2020.

[151] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[152] J. P. Martinez-Esteso, F. J. Castellanos, A. Rosello, J. Calvo-Zaragoza, and A. J. Gallego, "On the use of synthetic data for body detection in maritime search and rescue operations," *Engineering Applications of Artificial Intelligence*, vol. 139, p. 109586, 2025.

[153] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia computer science*, vol. 199, pp. 1066–1073, 2022.

[154] N. S. CMRE, "Synthetic environment for interoperable advanced marine simulation,"

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[155] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[156] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 420–439, Springer, 2020.

[157] C. A. Glasbey and K. V. Mardia, "A review of image-warping methods," *Journal of applied statistics*, vol. 25, no. 2, pp. 155–171, 1998.

[158] Y. Cao, C. Shen, and H. T. Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 836–846, 2016.

[159] A. Mousavian, H. Pirsiavash, and J. Košecká, "Joint semantic segmentation and depth estimation with deep convolutional networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 611–619, IEEE, 2016.

[160] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, "Real-time joint semantic segmentation and depth estimation using asymmetric annotations," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7101–7107, IEEE, 2019.

[161] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 235–251, 2018.

[162] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "Sosd-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, 2021.

[163] L. Chen, Z. Yang, J. Ma, and Z. Luo, "Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1283–1291, IEEE, 2018.

[164] M. Lin, S. Teng, G. Chen, J. Lv, and Z. Hao, "Optimal cnn-based semantic segmentation model of cutting slope images," *Frontiers of Structural and Civil Engineering*, vol. 16, pp. 1–20, 06 2022.

[165] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[166] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[167] H. Ye and D. Xu, "Taskprompter: Spatial-channel multi-task prompting for dense scene understanding," in *ICLR*, 2023.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

[168] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[169] M. Burton, *Unlock Your Creativity with Photopea: Edit and retouch images, and create striking text and designs with the free online software*. Packt Publishing Ltd, 2024.

[170] T. Huang, S. Zhao, L. Geng, and Q. Xu, "Unsupervised monocular depth estimation based on residual neural network of coarse–refined feature extractions for drone," *Electronics*, vol. 8, no. 10, p. 1179, 2019.

[171] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[172] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[173] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[174] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM computing surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[175] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[176] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[177] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[178] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[179] Y. Chen, H. Zhou, and Z. C. Lipton, "Moco-transfer: Investigating out-of-distribution contrastive learning for limited-data domains," *arXiv preprint arXiv:2311.09401*, 2023.

[180] T. Ren, Y. Chen, Q. Jiang, Z. Zeng, Y. Xiong, W. Liu, Z. Ma, J. Shen, Y. Gao, X. Jiang, *et al.*, "Dino-x: A unified vision model for open-world object detection and understanding," *arXiv preprint arXiv:2411.14347*, 2024.

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

Appendix A

Image Warping

Given the pinhole camera model illustrated in figure A.1, a camera is represented by a sensor plane and an optical center at a distance f (focal length) from the sensor plane. Such camera model can be characterized simply by 4 intrinsic parameters grouped in a matrix K called the projection matrix, as follows:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

A point P with coordinates $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ in 3D space can be projected to its pixel coordinates (i, j) on the sensor plane using the camera intrinsic matrix K as follows:

$$\begin{bmatrix} i \\ j \\ 1 \end{bmatrix} = \frac{1}{z} K \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{A.1}$$

The previous relation can be written in homogeneous coordinates as follows:

$$\begin{bmatrix} zi \\ zj \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{A.2}$$

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

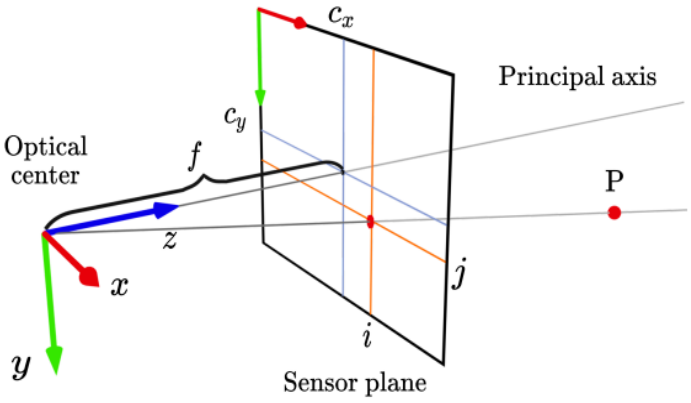


Figure A.1 A diagram of the pinhole camera model with axes and notations [2]

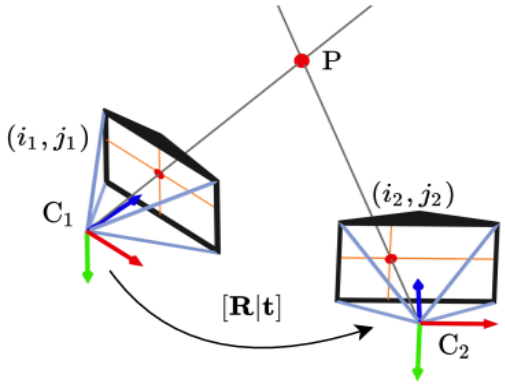


Figure A.2 Moving the camera from position C1 to C2 [2]

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*

When the camera moves from position C1 to C2 (figure A.2), this movement can be expressed using a translation vector t and a rotation matrix R that are grouped in the *transformation matrix T* as follows:

$$ {}^{1}_{2}T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} $$

The 3D coordinates of point P expressed with respect to C1 can, then, be expressed with respect to C2 using the relation:

$$ \begin{bmatrix} x_2 \\ y_2 \\ z_2 \\ 1 \end{bmatrix} = {}^{2}_{1}T \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} \tag{A.3} $$

where:

$$ {}^{2}_{1}T = ({}^{1}_{2}T)^{-1} $$

Given the predefined relations (equations 2 and 3), we can relate pixel coordinates $(i2, j2)$ to pixel coordinates $(i1, j1)$ using the following equation:

$$ \begin{bmatrix} z_2 i_2 \\ z_2 j_2 \\ z_2 \\ 1 \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0 & 1 \end{bmatrix} {}^{2}_{1}T \begin{bmatrix} K^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 i_1 \\ z_1 j_1 \\ z_1 \\ 1 \end{bmatrix} \tag{A.4} $$

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
Università di Genova

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4, componente 1 "*Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università*