POLITECNICO DI TORINO

Master's Degree in DATA SCIENCE AND ENGINEERING

Master's Degree Thesis

Deep Learning in Mars Exploration/Navigation: Rover Multi-Task Learning and Drone Semantic Segmentation on Synthetic and Real Datasets with Domain Adaptation Techniques

Supervisors

Prof. MARCELLO CHIABERGE

ILARIA BLOISE

LUCA MANCA

Candidate

MIRAKRAM AGHALAROV

December 2022

Abstract

Mars exploration is one of the trending topics in all fields of technology in which Computer science, mechatronics, aerospace engineering have combined workforce to execute these tasks efficiently. Computer vision side has responsibility to keep the mission alive and protect rovers from any occasions that give damage. That is why, high amount of research, accurate datasets and efficient ways should be found and implemented. We divided our work into 2 parts: Multi-task learning on rover study with domain adaptation, Semantic Segmentation on drone study with domain adaptation.

Navigation is the main objective of rover study in which a Deep Learning model should identify the different types of terrain to help in the definition of a safe path for rover exploration. Semantic segmentation is used for this purposes while AI4Mars dataset accuracy was not sufficient to identify the objects. That is why, synthetic datasets are used for this project and domain gap has been decreased by the customized 3 step unsupervised domain adaptation method. Special augmentations have been applied and ablation study has been performed to understand the effect of the parts in determined architecture. To increase the efficiency, we set another main objective for rover study to utilize multi-task learning method to check the feasibility of combination of semantic segmentation and multi-label image classification. As synthetic and AI4Mars datasets do not contain any classification label, additional real dataset (PanCam) have been used. We expected that some limitations of synthetic dataset will be compromised by using image classification dataset which will help to generalize better.

The same approach for semantic segmentation of rover study have been applied to the drone study in which synthetic dataset is different as applications are slightly different. We collected real samples from flight logs of Ingenuity helicopter from NASA. We selected 50 images and labeled them semantically. After training with synthetic images as source and unlabelled real dataset as target images, efficiency of domain adaptation method have been proved by comparisons with baseline performances for both studies.

Table of Contents

List of Tables

List of Figures

Acronyms

AI

Artificial Intelligence

UDA

Unsupervised Domain Adaptation

NN

Neural Networks

CNN

Convolutional Neural Networks

GPU

Graphical Processing Unit

CPU

Central Processing Unit

RF

Receptive Field

MER

Mars Exploration Rover

Chapter 1

Introduction

Mars exploration is one of the greatest challenges of humankind. Searching for new habitats, observation points, new resources forces to spend vast amount of money on research and development of Mars technology. Building durable machines for harsh environment, extreme temperatures and high radiation of Mars affects to the speed of development processes to be much slower. That is why, number of human-made machines in Mars are quite low and they should be preserved from any kind of damage for completing the mission cycle. Deformation of wheels, beaching can be caused by wrong recognition of surface. Sharp corners of rock can severely damage the Rover performance and even completely destroy its operations. That is why, it is high priority to keep rover far from dangers, control wisely and troubleshoot easily.

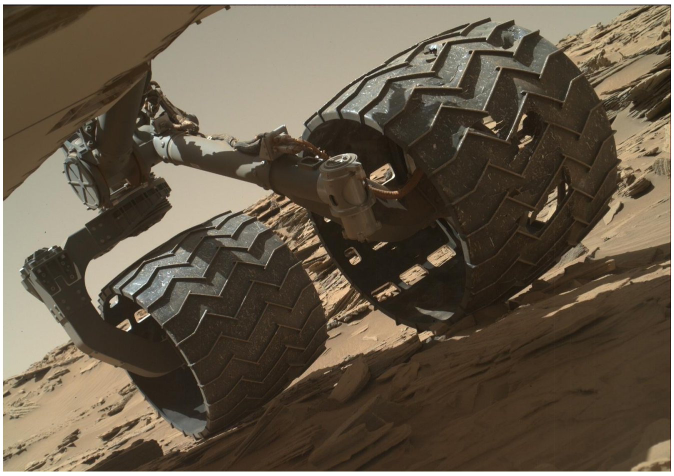


Figure 1.1: Wheel of Curiosity rover: Got damage from surface of Mars

With the recently landed rover Persevarence, communication time between earth and Mars is changing between approximately 4-24 minutes depending on the position of planets. Orbiters enable signals to transmit from rover on Mars to the DSN (Deep Space Network) on earth. This structure gives advantage to keep the power consumption balanced as orbiters are much closer to rover than DSN which does not require high distance antennas in transmitter and receiver. With these issues, high accuracy navigation systems tend to be autonomous as real-time control is nearly impossible to manage.

Persevarence Rover has also brought its powerful drone ingenuity to the surface of mars. After some successful flights, it is expected to increase Rover lifecycle and exploration accuracy by using flights with Ingenuity. That is why similar autonomous navigation system should be developed for better mission planning.

Autonomous control systems have got huge improvement stage as computational power of machines had increased drastically. Earlier control mechanisms like analog control had been replaced by digital control like Fuzzy logic in some parts of industry. Increasing complexity of input data and image processing had created necessity of development in ML and AI side. Development on machine learning then allowed more complex structures like neural networks to improve fast. Deep learning notion started from this stage. By using deep learning, scientist got access to the more fine details in the data, or in other words, deeper features for high performance regardless of input complexity. New structures were defined and still many more are being created according to the specific application, data types by considering how deep the features are hidden in given information.

Computer vision is a part of Deep learning in which input datatypes are mostly images. Collecting and learning the features from each pixel and group of pixels has high computational power and is a time demanding activity. Those tasks are the ones of the most sensitive to the underfitting and overfitting because of vast variety in image sources. Different type and generations of cameras, post-processings, lens types (fisheye, zoom, monochrome and etc.), environment differences require to build robust architecture for consistent performance.

As input structure is getting quite complex, output structures also changed according to the application purposes. Image classification was used to define the class that image is representing. With multi-label classification, existence of object can be analyzed on the images. Object detection algorithms were utilized when more information is needed about objects presented in input like dimensions, location in the image. Semantic segmentation and instance segmentation has been developed to assign classes to each pixel in image in order to have fine-grained details. Each task requires different kind of structures but generally they have similar fundamental processes like feature extraction, convolutions. Moreover, high computational cost should be also taken into consideration because extensive amount of weights are used in these tasks and in all other fields of deep learning. When low number of parameters is used for training, accuracy get affected, otherwise latency increases. Those issues explain main trade-off for any deep learning applications. Latency is very important for autonomous tasks in which decisions should be given fast. These tasks are included in autonomous cars, emergency systems, collision avoidance systems, hazardous industries. Considering the core computational power in electronics of cars, robots or any other IoT system (Internet of Things), latency is much higher compared to the devices equipped with GPU.

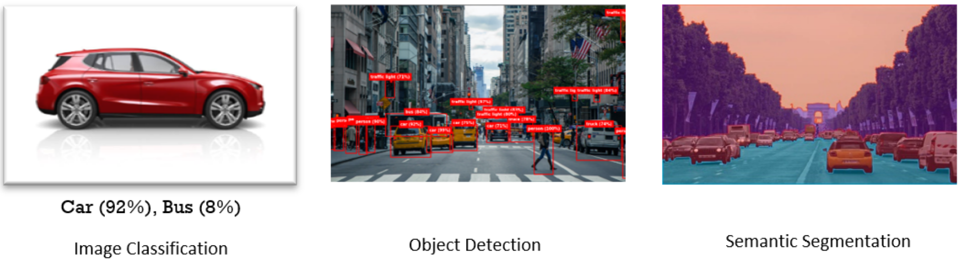


Figure 1.2: Differences between some of the task from computer vision

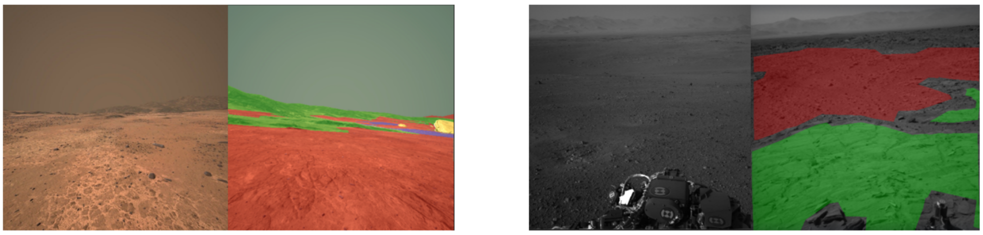
Recently many methods are created to decrease power consumption and lower the latency by balancing performance. Multi-task learning is new phenomenon to be used for simplification purposes. This terminology explains a method that 1 network can combine 2 tasks at the same time. Image Classification and Segmentation or depth estimation can be used very effectively if datasets share the similar features which allows to share some phases of process without having additional model. This enables to save from power consumption and latency in comparison with execution of both tasks while it can be slower for each task individually as model complexity is increasing. In Mars exploration, considering the absence of power resources, in order to increase information from 1 image, it is necessary to test the possibility of adding scientific purpose analysis alongside with navigation in the same network in addition to the autonomous control algorithm.

Until today, 5 rovers have landed to the mars surface. Over last 15 years, plenty of images were collected and stored for further post-processing. Most of them are open to the public for research and development purposes. Considering that each rover had different generation technology, there are some domain differences regarding the cameras, angles, viewpoints. This issue challenges scientists in terms of robustness and expected behaviour with new generation cameras.

In order to carry out navigation task, autonomous driving applications should be referenced. Those papers have prioritised usage of semantic segmentation tasks for better and accurate path plannings. Considering similarity in purpose, semantic segmentation is the most suitable choice for navigation in order to avoid hazardous surfaces. Semantic Segmentation is pixelwise classification task. In other words, model should make prediction for each pixel in given image with predefined classes.

Recently, many research and development were carried out and some datasets are labeled for this purpose. However, some datasets were annotated by crowd-sourcing approach which were not verified by experts. That is why annotation accuracy can be quite low and misleading. The main reason behind this issue is that semantic labels are the most time demanding to record. This is main challenge for Semantic segmentation.

In order to get rid of labeling procedure, synthetic environment can be used. In order to do that, 3 dimensional environments are created, they are optimized according to the real scenes, images are taken in different perspectives and labels are created automatically according to the configuration of environment. These datasets can be used for trainings afterwards. Despite highly accurate annotations, considering noiseless, ideal scenes of 3D environments, strict domain shift appears. In this case new method and architecture should be taken into consideration. Those methods are called domain adaptation which brings additional complexity to the training procedure in order two decrease the gap between to datasets from different domains.



Synthetic Images

Real Images

Figure 1.3: Comparison of Synthetic and real images with their labels

Another kind of solution for this problem is that unsupervised training should be carried out. Moreover, semi-supervised training procedures are proven to be one of the best performing architectures. However, computational costs for unsupervised trainings make development process difficult. Recently developed methods include unsupervised training for domain adaptation to increase generalization capability better.

In order to increase model efficiency, it is planned to check possibility of using multi-task learning in this application. However, there are no multi-labeled dataset for Mars applications. That is why, it is necessary to take extra action to be able to detect both kind of labels. So at the end, model should have multi-label classification output and semantic segmentation result. 2 different and task specific labelled datasets are required in order to train both heads.

In this project, it is planned to have 2 parts which includes Rover and Drone study.

In Rover study, semantic segmentation task for navigation and multi-label image classification task for scientific analysis should be carried out in one network by using multi-task learning. Three datasets are given for this purpose in which two of sources are real while remaining one is synthetic. Two real datasets are corresponding for each task while synthetic dataset does not have labels for multi-label classification which makes project very challenging. For this study, procedures are listed below.

- Extensive data analysis should be carried out and correct augmentation methods should be selected. Datasets contain uncommon characteristics in comparison with datasets used nowadays widely.
- Semantic segmentation model and corresponding classification architecture should be defined. Considering training resources and to maximize latency, lighter but effective models will have more focus.
- After careful research, domain adaptation method should be designed. Different domain adaptation techniques have special compatibility with task-specific datasets like autonomous driving. This issue narrows selection of different methods over DA papers.
- Multi-task learning approach to be applied on domain adaptation procedures.

In Drone study, the same approach used in the rover study will be utilized but with only semantic segmentation part. Here the main problem remains in real data part which is not published as a dataset. Images can be used from flight logs and should be post-processed because of fisheye distortion of camera on the drone.

- Synthetic data should be analyzed and real data should be retrieved from NASA website.
- Every sample from real dataset should be processed to remove fisheye distortion coming from lens.
- Second step is annotation of part of the data from real dataset.
- Apply the same segmentation model and domain adaptation tecnhique over this dataset.

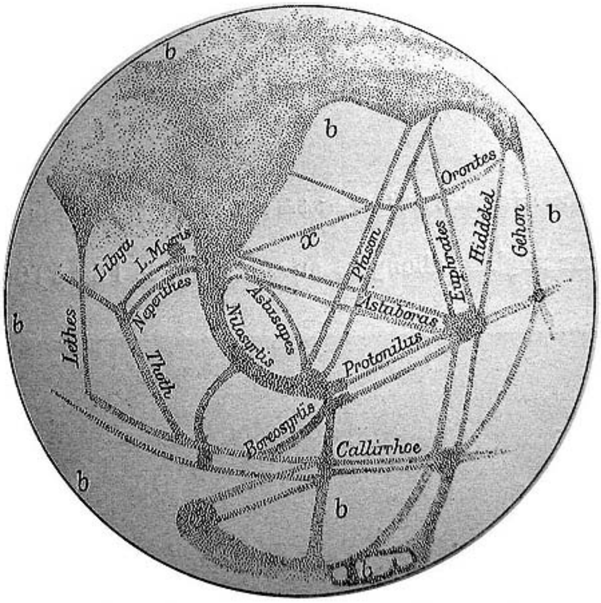
Chapter 2

Background

2.1 Mars Exploration

Enthusiasm which started from 4000 years ago by encountering the existence of Mars, had led the generations to pay attention on this red planet in 17th century after invention of telescope. Galileo Galilei was the first scientist who initiated observation of Mars by using telescope which is followed by many other scientists. In the second half of 19th century, Italian scientist Giovanni Schiaparelli published first map of Mars planet. According to him, Mars had different sectors which are connected by canals. During that time, this information had created huge resonance in which people and scientists thought that these canals were created artificially by generation living in this planet. Few years later, the same observation attempt has done by Percival Lowell in order to analyze those canals better and he failed. However, Eugene M. Antoniadi proved those canals to be optical illusion of telescope and created new detailed map of red planet in the beginning of 20th century. Many years of misunderstanding in this topic created a “Martian aliens“ terminology.

Main reason for attention on other planets from very early times was existence of believe in interconnection between events, materials, substances and formation of stars, resources in other planets. In other words, cosmos was one huge opportunity to understand earth and history of it. Observations of events happened in galaxy, rotations of planets were main clues for this purposes. Discovery of elliptic movement of earth around sun, approximate locations and paths of all other planets were defined by combination of observations, assumptions and mathematical calculations. As an example, in 1600s, great mathematician Johannes Kepler was able to correct his measurements and publish 3 laws for planetary movements according to accurate Mars observations of great astronomer Tycho Brahe. Tycho was not sharing his data about planets and was trying to determine dependency equation for paths of



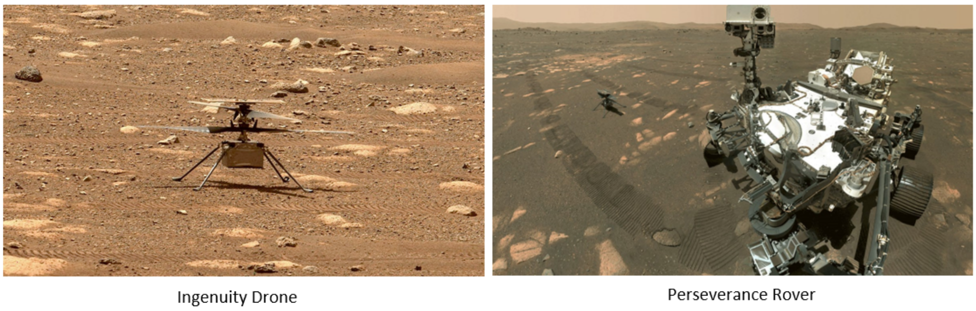
Dessin de Mars fait le 4 juin 1888, par M. Schiaparelli, à Milan.

Figure 2.1: Map of Mars defined by Italian scientist Schiaparelli in 1888: Contains canals among different sectors which were understood as artificially created

Mars by himself despite lack of ability on mathematical side. However, after death of astronomer, imperial scientist Kepler got access to his data and realized that his assumptions on planetary movements did not coincide with observed rotations as he was insisting on perfect circular movement. After some research, he discovered elliptic movement law around sun and later on, Isaac Newton also exploited this knowledge for gravitational law for earth and solar system. That is why, exploration of other planets could be also named as analysis of Earth.

However, there are other special reasons for selection of the Mars as main objective. Apart from closeness and accessibility of red planet from earth, there are also some other scientific reasons to spend vast amount of money and time in this planet. According to the compatibility of atmospheric structure, visibility, durability of machines and possible renewable sources to extend mission life more, Mars has best conditions among all other planets. From 1965, after some failed fly-by attempts from Soviet Union and USA, first machine “Mars 2“ landed, in other words shunted into Mars in 1970. “Mars 3” was next attempt after failure in previous generation but it operated only 18 seconds after landing. First successful operation happened in 1976 by the machine named “Vikings-1” which also sent first picture from the Mars soils. Different kind of satellites, rockets were sent to Jupiter, its satellite Europa. However, best possible conditions for research, observation, still was in Mars. Jupiter had deceiving look from earth that it has crucial elements for sustainable life on its soil but Voyager 1 had sent information about thick cloud level and dangerous object around biggest planet after getting speed by using fly-by technique from Mars.

First moving remote controlled object landed on Mars by using Pathfinder machine which contained Sojourner Rover in 1998. This moving robot did not have a chance to travel, but this mission sent 16500 images to the earth. In 2004, Spirit and Opportunity rovers (which are called twins) started their mission and they operated much more longer than predefined time. Opportunity Rover operated 15 years before strong dust storm and found elements of water which is potential microbial life source while Spirit beached in soft part of soil. Curiosity and Persevarance rovers are the latest in which Persevarance brought Mars drone named Ingenuity.


Ingenuity Drone

Perseverance Rover

Figure 2.2: Perseverance Rover and its Drone Ingenuity. Image taken after 44 Sols (Mars days) of landing

In modern missions, with the availability of accurate technologies and simulations, these rovers allow scientists to analyze atmosphere of Mars better, changing conditions, better observation, exploration of resources. Long durability is the priority in every Mars mission due to the launch and development costs for the new rover. That is why, many scientists try to prevent these machines from extra damage except from harsh conditions of Mars which includes high radiation, lack of energy resources, dust storms. Considering the number of machines sent to Mars in last 50 years, every part of image and data is needed to be used effectively to increase efficiency and safety to prevent the same fate that Spirit rover faced.

Artificial Intelligence used on modern rovers is giving strong power to these robots in extra-ordinary place to know the path, to select best possible landing area, most energy efficient way of completion of missions. Ingenuity drone is huge step towards the future which will have big role in mission plannings for current and further rovers being developed. Considering the thinness of atmosphere in Mars, Ingenuity drone equipped with very strong propeller which allows to lift itself to 10-13 meters of height. For comparison, drone is facing the challenge which is the same as the helicopter is trying to fly at 30km height in Earth in which world record is approximately 12km. This kind of powerful machine has been tested in special facility on Earth. Last flight logs proved that drone is able to operate in harsh conditions of Mars and from 2022, new mission is going to be assigned after recovering the communication due to the positioning of the planets.

2.2 Deep learning

20th century had many first steps in different parts of technology. Increasing demand forced technology to develop and improvement let the people to think wide and needs became more complex. This loop is still continuing and in recent years, each innovation enables new generation. That is why, complexity of technology can be related to the complexity of demand. Building the computers for simple operations, allowed to increase heaviness of computations step by step which increased speed of research.Building a framework for integral and derivative computations, increasing the storage, minimising the hardware size or in other words, enabling the smaller technology (like micro, macro, nano) on transistors had its effect on industry. For example, widely used PID controller has switched from pneumatic transmission to the electrical with the development of different generations of computers.

Artificial intelligence terminology initially appeared in first half of 20th century with the name of “heartless robotics”. In the middle of that century, possibility of this terminology has been proved by mathematician Alan Turing who could not test it practically due to the generation of computers. According to him, fundamental configurations of computers should be changed in a way that they can store the commands given to it. Financial requirements to run a computer was also incredibly high that only certain companies and universities could do research over this topic. After some years, development of storage device, allowed his theory to be tested. After certain years number research increased and this terminology started to develop.

Meanwhile, mathematician Lutfi Zade established basement of fuzzy logic which was based on way of human thinking. Normally, control systems were imagined like 1 or zero. Basically, answer for the question if sky is cloudy could be either yes or no. However Zade claimed that human can answer as it partially, mostly, fully cloudy. Additionally control schemes after deciding the situation can vary according to the level of clouds. So fuzzy logic eliminates the idea of yes or no and creates more detailed situation for control. After establishment of this idea, different applications in robotics field adopted fuzzification approach.

After introduction of machine learning, Deep learning terminology also started to improve due to the increasing capability of computers. Having more complex architecture of input pushed technology to the level that they are able to process these information. At some point, simple machine learning algorithms were insufficent for difficult features to be extracted. That is why, multi-level architectures were developed which can process deeper features from information given as input. That is why, they started to be called as Deep Learning algorithms. Computer Vision, Text and Voice Processing has been transferred into deep learning side for greater performance. By applying different hyper-parameters to the network, information in various depth can be easily analyzed and result can be extracted.

After developing gradient-based back-propagation, neural networks have started to be trained and used in the constraints created by computational capability. However after introduction of GPUs and high speed CPUs, speed of research increased which led to try to solve more complex problems. This advantage have created new challenges as deeper neural network started to be utilized. Gradient vanishing problem is one example for this issue.

2.3 Neural Networks

Briefly explaining, main fundamental notion of deep learning is neural network which consists of different neurons and activation functions inside each layer. First and last layers are called as input and output layers respectively while middle layers are considered as hidden layers. According to the number of layers inside neural network, information from input can be processed in different depth for finer features. Each neuron contains number of weights and bias which is followed by activation functions. This special output function inside layer is breaking non-linearity between different layers. If this function is not utilized, even NN structured by 1k neurons can be easily described as 1 neuron which is not capable to extract deep features from the input. Activation functions in output layers, are softmax, sigmoid which is highly dependent on the application chosen.

Training of NN or most Machine Learning models are similar and works with gradient based backpropagation. Forward pass of input is finalized by loss function to define distance from target. According to the gradient calculation on the basis of loss functions, states are stored. In backpropagation, these states are used to update the weights used in forward propagation. The same loop is starting until predefined threshold or maximum number of epochs.

As mentioned before deeper networks started to create new problems like gradient vanishing which was interrupting hours or days of training. So new architectures, connections, modules has been developed for different type of applications.

2.4 Computer vision

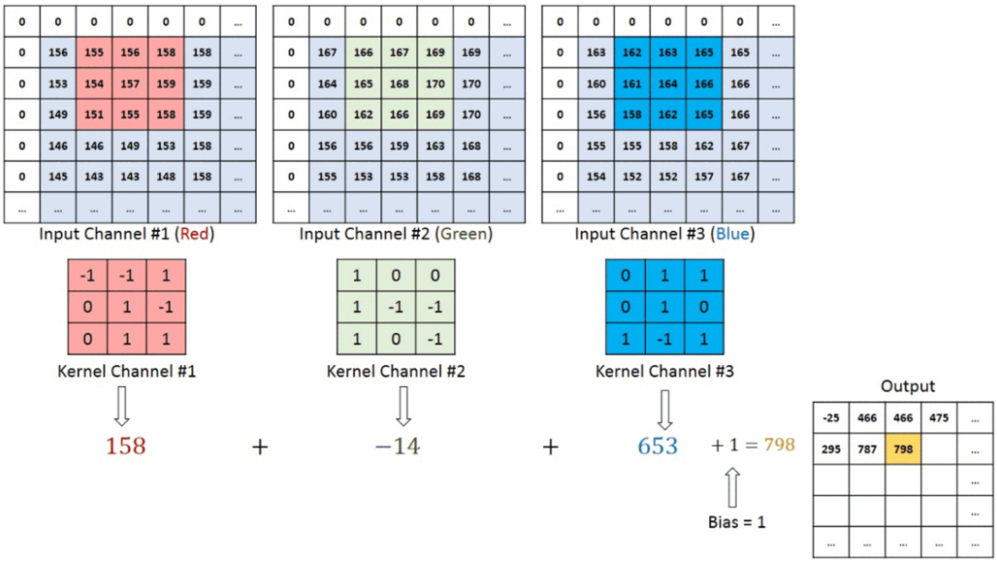
Computer vision is very complex application to be handled by simple algorithms or low computational ability devices. The reason is that inputs in this case mostly are images and they combine vast number of pixels. If we look at normal RGB image as matrix, in normal resolution dimensions will be 3x512x512 which makes more than 786k possible features for only 1 sample. That is why, traditional neural network does not work well for these kind of applications.

Convolutions have main roles for these operations. Convolution was firstly introduced in digital signal processing that it can combine 2 different signals to create new one.

$$(f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$

In image processing, if we consider the image as first signal, the second signal will be the filter for extraction of the necessary feature. Basically these filters are main trained parts which are considered as weights of network. Architectures based on convolutions are called as Convolutional Neural Networks (CNN).

In CNN based feature extractors, every layer which consists of convolutions has number of filters to be trained. Image is scanned by using sliding window that has parameters of kernel size and stride, and each window is convolved with a filter of weights to create feature map from the image. According to the parameters, spatial size and depth are changing which defines one of the main hyperparameters for the architecture.



0	0	0	0	0	0	...
0	156	155	156	158	158	...
0	153	154	157	159	159	...
0	149	151	155	158	159	...
0	146	146	149	153	158	...
0	145	143	143	148	158	...
...	...	...	...	...	...	...

Input Channel #1 (Red)

0	0	0	0	0	0	...
0	167	166	167	169	169	...
0	164	165	168	170	170	...
0	160	162	166	169	170	...
0	156	156	159	163	168	...
0	155	153	153	158	168	...
...	...	...	...	...	...	...

Input Channel #2 (Green)

0	0	0	0	0	0	...
0	163	162	163	165	165	...
0	160	161	164	166	166	...
0	156	158	162	165	166	...
0	155	155	158	162	167	...
0	154	152	152	157	167	...
...	...	...	...	...	...	...

Input Channel #3 (Blue)

-1	-1	1
0	1	-1
0	1	1

Kernel Channel #1

1	0	0
1	-1	-1
1	0	-1

Kernel Channel #2

0	1	1
0	1	0
1	-1	1

Kernel Channel #3

-25	466	466	475	...
295	787	798		...
				...
				...
...	...	...	...	...

Figure 2.3: Processing RGB images with CNN architecture

Mainly CNNs contain additional different layers like pooling, FC (Fully connected layer) and also activation layers as traditional neural networks. Feature map extracted from convolution operations are fed to the activation functions to break linearity before the max pooling layer which serves to decrease the spatial size. Having high spatial size with deep network increases the number of weights in the network. This situation is mainly avoided due to the different reasons: High memory consumption, possibility of overfitting, redundant features.

2.4.1 Image Classification

Image Classification is the simplest variant of computer vision which can be explained easily. Simple image classification network contains convolutional and fully connected layers. Considering in general, as first operation, features are extracted from the images by using convolutions. Most necessary features are fed to the neural networks and according to the specific application, output layer generates certain amount of values. For binary classification, output layer consists of only 1 neuron activated by sigmoid function. Figure (To be corrected) shows general looking of very small and simple CNN image classification model.

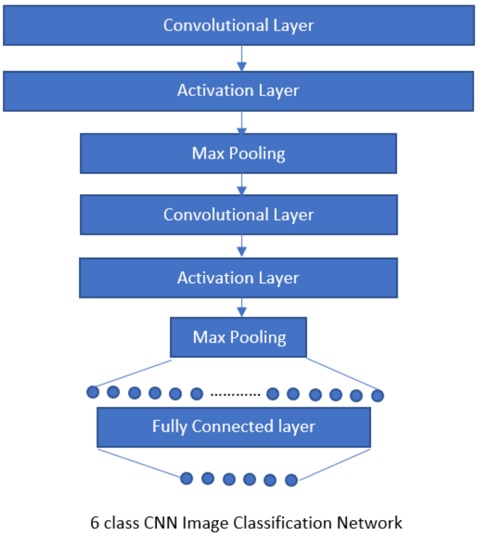


Figure 2.4: Basic Image classification architecture: CNN feature extractor followed by fully connected layers

First fundamental developments have happened over image classification networks due to the simplicity to label and richness of the content. In early times of CNN, VGG [1] type of feature extractors had create fame. However gradient vanishing and hunger to have more parameters for higher accuracy were the main issues to be analyzed. Residual Neural Networks [2] was one of the biggest improvements to remove this problem and decrease the depth of the network without hurting the performance. Main principle of ResNet was to have weightless skip connections to the next layer from the previous layers. This connection helped to remember the input features better than VGG and remove the natural effect of convolution.

2.4.2 Semantic Segmentation

In image classification, main purpose of the application was to get generalized information about picture. In other words, model should give the result that shows what is in the image in general. However for more complex tasks, requirements are more complex. Semantic segmentation is a process that can express the information about exact location and area of the object in pixel level. This task should give an output in the same size as input image which is very difficult after down-sampling for feature extraction. So that is why, Information processing between different types of image processing are strictly different which lead to the variety in structures.

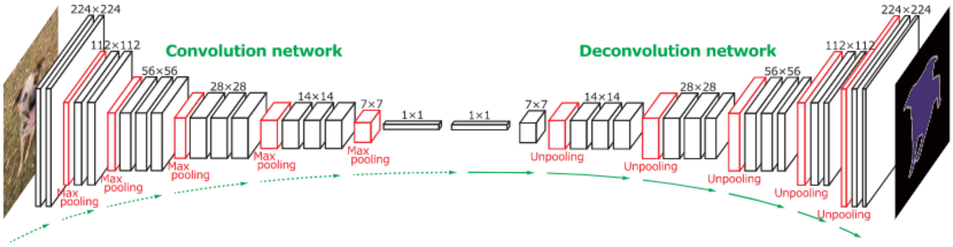


Figure 2.5: Encoder-decoder type Semantic segmentation network example. Input and output have the same spatial dimensions.

Semantic segmentation structures have the capability to give the accurate output with high resolution. This means that max pooling layer or another downsampling methods are obstacle for the Semantic Segmentation. However, otherwise number of parameters will be such high that can not be handled easily. That is why, for this application different methods were developed to handle downsampling and decrease the information loss to be transferred to the output of the network. For example, Atrous convolutions were developed by Google scientists to have flawless upsampling. Other methods suggesting 2 path networks, proper connections between downsampling and upsampling layers were also competitive. However as image classification, main context features is extracted in the same way and the method. Deeplab versions are heavily based on ResNet architectures.

Metrics for semantic segmentation is also different in order to get better performance measurement. Mean Intersection over Union has great advantage to have information about performance. Predicted pixels that coincide with ground truth is written in numerator while union of all these pixels are considered in denominator.

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN}$$

One of the main challenges for semantic segmentation is labeling procedure. As it is mentioned before, each pixel should represent a class of the object. It means that, edges of the objects should be annotated very precisely. Therefore, this procedure takes more than 20 mins per image for accurate labelling. Considering the size of sufficient datasets, it can take months to build a dataset.

2.4.3 Loss Functions

As deep learning chapter explains, loss function is crucial part of learning. Performance of the training and inference is highly dependent on loss function. That is why, there are diversity because of the requirements for each computer vision application. For example Image classification algorithms have simple classification output. Cross Entropy Loss is favorite to be used in this kind of application.

Considering the basic loss function called residual or mean squared error, penalizing scheme for worse prediction is not sufficient for better training. However logarithmic graphs can be utilized to learn faster and better. That is why, we are going to use probability term which will be used to define the loss. Softmax layer is highly beneficial for this purpose which can define the probability function for each class. It is better than using argmax because it can calculate better gradients. Cross entropy is based on probabilities which uses logarithmic function to find an error between ground truth and prediction.

$$CE_{loss} = -\sum_{c=1}^{N} y_c log(p_c)$$

where probabilites Y and P shows ground truth and prediction for class c respectively.

From another point of view, semantic segmentation can be olsa trained with cross-entropy loss as the structure allows to use softmax. However, it can lack in this application because unbalanced dataset is very common for semantic segmentation in which cross entropy loss can not perform well. Focal Loss is developed on the cross-entropy to deal with minority classes better.

Alternatively, dice loss is among the selection because of its decent results. Dice loss is calculated according to the intersection of pixels in numerator as product of them gives output for intersection. Denominator sums probabilities of all pixels.

$$D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}$$

P and G are the probabilities of classes for each pixel i.

2.5 Unsupervised Domain Adaptation

Generating the artificial datasets by using 3D softwares let to generate the labels automatically which eliminate challenge for the real datasets. However it creates domain difference because 3d environment are too ideal for the real scenario. That is why different methods are used to leverage gap between 2 different domains. They are called unsupervised domain adaptation method as there is no indicator between two sets to teach the network.

- Pixel-Wise Domain adaptation methods are perfoming in input space. Before the model or any other types of domain adaptation methods, this approach can decrease the domain gap and remove ideally noiseless environment. They can perform in both online and offline before training to build a augmentor to transfer the features from real environment to the synthetic. However, these method are not so efficient if they used alone, it can be utilized in combination with different domain adaptation method to enhance the perfomance

- Adversarial type of domain adaptation methods are based on having additional small models for discriminating the features whether they are coming from target (real) or source (synthetic) dataset. They can be either used in the end of the model or in the middle after the backbone. Sometimes using multi-layer adversarial learning can help to increase the performance. Main aim of the discriminators is to fool the generator (model) in order to make the training based on more on target dataset. This approach decreases the overfitting on source side of the dataset and performs better in case of high number commonalities between target and source dataset.

- Prototypical Domain adaptation approach can perform very well in case of computational availability. Features extracted from backbone, are analyzed and prototypes are created for each class and dataset. According to the distances between prototypes of the same class, new common space for new prototypes are created and backbone is forced to decrease the gap between common space and dataset specific prototypes.

- Self training has high memory requirements while it has shown great performance in recent years. Apart from unsupervised domain adaptation, self-training is used to increase the performance of the model in supervised learning.

Main principle of this approach is that 2 identical architectures and weights are initialized while they are called student and teacher. Student network aims to imitate teacher architecture by comparing the outputs. Teacher network generates the output named pseudo-labels which are considered as ground truth for student network. In domain adaptation, self-training is used as 1 step and can be initialized with other similar semantic segmentation dataset not to hurt the performance. Random initialization in this method can mislead the student as it is recommended to get high confidence labels. Higher batch size is highly beneficial to increase the confidence by generalization

Taking everything into consideration, unsupervised domain adaptation methods are highly computation demanding methods. There are some methods that achieved to decrease the memory consumption but still decreased version requires a lot of time for 1 iteration and decent amount of memory.

Chapter 3

Literature Review

3.1 Semantic Segmentation

Semantic segmentation is pixel-wise classification task which requires to have a different configuration. Models are mainly divided into 2 different types: Encoder-Decoder and 2-path structure. Having encoder and decoder structure allows to extract feature by decreasing output size in each layer while decoder tries to get full image representation by upsampling the output in different ways.

One of the first models were introduced to be successful in semantic segmentation is UNet [3], which has U-shaped structure with interconnections between encoder and decoder. This allowed to use input features more in decoding stage to recover pixels better. However, facing with long inference time and low efficiency created need for new models to be developed in both U-shaped way and also with different novel structures. DeepLab [4] was one of the sharp improvements in this path and till now in most papers benchmark tests are carried out with this structure. Main novelty in deeplab was bringing ASPP modules into play by using atrous convolutions to recover pixel representations. This allowed us to eliminate skip connections from each layer of encoder and decrease memory usage during inference which was main step to real-time tasks.

Additional models like BiseNet [5], Fast-SCNN [6], CabiNet [7] are the ones which is extensively used for latency focused applications like autonomous driving. BiseNet [5] had different approach on problem and achieved to increase inference speed without decreasing the accuracy. They used 2 paths, in order to keep the trade off between 2 main opposite purposes:

- Keep receptive field high to capture more global information.

- Keep number of parameters low, to increase speed of inference.

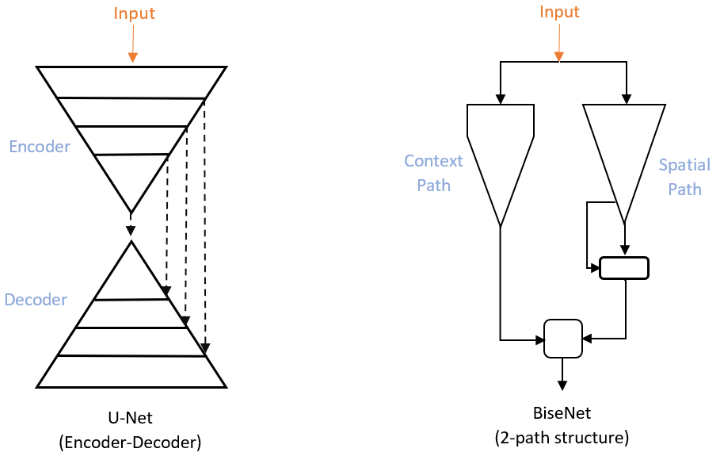


Figure 3.1: Example models from encoder-decoder and 2-path structure

The main problem was that, when receptive field is high, number of parameters increase drastically to keep the same feature extraction capability. However decreasing the RF size was giving us poor results. That is why, BiseNet [5] is structured by 2 parallel tasks in which 1 path is responsible to keep receptive field high without carrying out feature extraction for classification, while 2nd path was working as standard encoder with low RF. 1st path is named as "context path" which consists of lightweight model like Xception[8], while the 2nd path has the name of "spatial path" structured by standard ResNet [2] architecture. By using multi-level attention refinement modules, information from both paths are collected and combined in feature fusion module. That is why the BiseNet model has great advantage of perfomance on predicting both small and large sized objects. Similarly Fast-SCNN [6] and CabiNet [7] also focused on trade off that 2-path networks trying to deal with.

Strong development on CNN based networks for image operations is outperformed by transformer based structures. Recently, after introduction of Vision transformers [9], many discussions have arised about robustness and efficiency of these new models. However, development and tests carried out proved that vision transformers and their follow up works are robust and efficient in domain adaptive environments. ViT [9] backbones are data-hungry structures which demands sufficient amount of data to be trained. DEiT [10] transformers added token distillation to make training feasible with less data. LeViT [11] made progress in real-time applications and improved dense prediction tasks by applying dynamic size position embeddings. In simple semantic segmentation tasks, transformer based networks show minimum of 5% mIoU advantage over CNNs while this difference was more than 10% in domain adaptation tasks.

Segformer [12] is one the most used semantic segmentation network which consists of mixed transformer (MiT) in encoder side and CNN network in decoder. The same encoder is used with ASPP structure in DaFormer [13] architecture which shows 2% mIoU advantage over Segformer.

According to the resources to be used, we should select light models. That is why, models applicable to real-time applications have more attention to be used in our work. However considering the domain adaptation capability of transformer models and huge difference with CNN based backbones, trade-off between performance and memory should be managed.

SegFormer [12] (at the same time Daformer shares the same backbone) has 6 different backbones to be applied in different applications. First one- MiT-b0 has lightest configuration to be used in latency focused application while MiT-b5 shows great accuracy with more complex structure.

Models	Advantages	Disadvantages
U-Net	Skip Connection between layers	Low inference speed High memory consumption Relatively low accuracy
DeepLab	ASPP module applied Better memory consumption Sufficient accuracy	Not domain robust Not applicable to real-time scenario
BiseNet	2-path network to eliminate trade-off for RF selections Applicable to real-time High accuracy	Less stable in domain shifts High model complexity
SegFormer	Higher accuracy Better robustness in case of domain shifts Better management in memory-performance trade-off	Bigger datasets needed that CNN based networks not fully applicable to the real-time application
DaFormer	Decoder is specially designed for domain adapative training Highest accuracy among transformer based models	High number of samples needed for training

Table 3.1: Detailed comparison between different types of semantic segmentation models

3.2 Domain Adaptation

Semantic segmentation models are highly sensitive to the domain shifts which can be classified into 4 groups. Label shift, Concept shift, Conditional shift, Covariate shift. We are mainly dealing with covariate and conditional shifts that occur when we switch to live and dynamic environment. We see the label shift (sometimes called target shift) when the class distributions between datasets of different domains are varying. Standard differences, grouping conventions lead to the concept shifts for different countries and cultures. There are many studies for each of these shifts.

Unsupervised Domain Adaptation techniques are divided into different sections.

Feature level and pixel level adaptation. Domain level adaptation is for transformation of features from target images to the source images or creating the generalized domain space for further training. Feature level domain adaptation are used during training in order to fool feature extraction phase to be more robust on generalizing.

3.2.1 Prototypical Learning

During recent years few - shot learning frameworks were consistently giving high results specially on small datasets. After some few researches, Transferrable Prototypical Learning started to be used for modifying domain spaces for datasets. Cross-domain Prototypical learning paper [14] had great achievements over many methods and included also Instance contrastive learning.

Bi-directional Contrastive learning for Domain Adaptive Semantic Segmentation paper [15] is based on prototypical approach in order to eliminate domain gap in feature level and it is best performing method among the ones that apply the same approach. By using bi-directional contrastive learning more compact intra-class pixels are generated while inter-class features are getting separated very well. Paper adopts a method to leverage the disadvantages of Self-training by using dynamic pseudo labels. Static pseudo labels have problems with high potential overfitting on teacher network results. Having the slower learning tendency of teacher network is one of the reason for this problem.

3.2.2 Pixel-level Domain Adaptation

Another type of unsupervised domain adaptation happens in input stage which is pixel level adaptations. Main purpose of this type of methods is that before training any model, some features which represents reality of images should be transferred from real (target) domain to the synthetic (source) domain. AdvStyle [16] is one of the best performing architectures that exploit augmentation type of domain adaptation during UDA training. Fourier Domain Adaptation [17] is based on transferring high frequency and amplitude target features into source images by using fast fourier transform.

ProCST - Boosting Semantic Segmentation using Progressive Cyclic Style - Transfer paper [18] is image level domain adaptation to be utilized with any UDA. This method is training pyramidal network which decreases the domain gap before training in order to increase performance on UDA. Method uses to generators which are for source to target translation and vice-versa. Discriminator is used to understand differences between source and fake source images and the same for target dataset. At end of pyramidal network composed by generators we have segmentors for supervised learning. This way leads to have generator network composition which gives output of synthetic images with real domain features.

Method has been tested and proved that it decreases pixelwise domain gap better than other same level domain adaptation methods

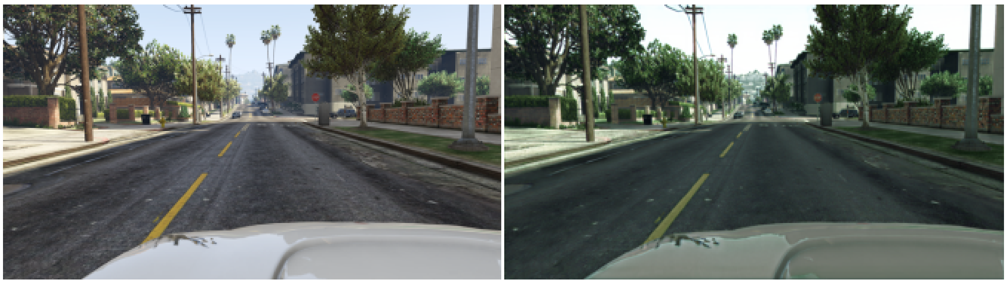


Figure 3.2: ProCST performance on decreasing domain gap between 2 different datasets. Left one original source image. In the right: Modified source image after training with target images

3.2.3 Self-training based Domain Adaptation

In recent years, considering the development in unsupervised training, self-training also affected the domain adaptation field. Having additional steps related to ST improved feature extraction capability on target images in which no labels are present.

Unsupervised Contrastive Domain Adaptation for Semantic Segmentation paper [19] introduces the novel method to initialize and train UDA. For robustness of pseudo-labels generated by teacher network, they defined pre-training stage before self-training.

Fully supervised training is followed by generated target image pseudo-labels to be used in contrastive pre-training. End of this stage is finalized by applying fine-tuning by using full supervision. Self training stage has the same structure but in parallel.

According to the results, the method outperforms state-of-art methods by large margin. Contrastive learning part has been adopted from SimCLR [20] paper and additional 2 MLP layers are used for this operation. As other papers mentioned, self training has the problem of overfitting. High confidence pseudo labels are used for this part of training which small part of dataset. In order to increase the pseudo label size, threshold for confidence should be lowered but this can lead to misclassifications. That is why the paper introduces pseudo-label expansion mechanism which is based on prototypes

DaFormer paper [13] is Self training based method which focuses on developing semantic segmentation network architecture and unsupervised domain adaptation method. Main novelty is bringing new architecture on transformers based network.

From Domain adaptation method side, authors implemented some novelties like Rare-Class sampling to increase performance on minority classes, Feature Distance Computation to preserve feature extraction capability during transfer learning. Method consists of different stages stated below:

- Standard Supervised training on labeled synthetic dataset.
- Feature Distance Computation between backbone and pretrained model with real images.
- Self training with teacher student method on augmented and original real images to increase feature extraction capability.

For the last step, in order to eliminate problem of late update in teacher network which leads to poor reliability and confidence on pseudo-labels, authors applied Exponential Moving Average update on teacher network

3.2.4 Adversarial Domain Adaptation

DecoupleNet: Decoupled Network for Domain Adaptive Semantic Segmentation paper [21] introduces a novel method that tries to eliminate many drawbacks of Self-training and overfitting on source dataset. Main advantage for this work is that they focused on task entanglement on semantic segmentation and adversarial learning. As all tasks are done simultaneously on single network, semantic segmentation task is not prioritized during training. This problem leads to have poor performance.

Solution of this paper is following:

- Adds 2 small generators Gsrc and Gtgt before the model G which consists of 2 parts G shared encoder and C segmentor
- After initial generators shallow features are calibrated according to the adversarial loss by using discriminator
- Results of C segmentor should be used for supervised cross entropy loss and discriminator for adversarial learning in output space.
- Self Discrimination should occur by using new Auxiliary Classifier on target domain

Main advantage for this network structure is that Target pipeline does not have to perform well on source dataset. The reason is that during inference, G * Gtgt should be used which makes more part of training to be focused on target set.

3.2.5 Comparison

Unlike the autonomous driving applications, our problem has some challenges in which it can create huge differences between UDA methods. That is why, extensive analysis should be carried out in order to increase the performance on this specific application.

On Rover study, considering the camera angles that datasets introduced, feature exchange between these 2 domains are getting difficult. In order to have more accuracy, classwise domain adaptation methods can be successful in closing the gap between features. Considering the power of prototypical learning in feature space, those papers have more advantages on this problem. Meanwhile, this learning type needs extensive computations in order to close the gap between feature prototypes.

On self training side, DaFormer [13] has found a way to increase confidence on pseudo labels by updating teacher network with EMA of student network. Meanwhile DecoupleNet [21] is using Auxiliary 1 layer classifier to keep performance high. Unsupervised Bidirectional Contrastive learning [15] approach updates the network only in 200 operation which can lead to inferior performance.

Models	Advantages	Disadvantages
DaFormer	Novel Semantic Segmentation Architecture Feature Distance Comparison Rare class sampler More confident pseudo-labels	Pre-trained model is needed Static pseudo-labels can lead to poor results
DecoupleNet	Better focused on target feature extraction Task entanglement Auxiliary classifier for ST	Increased training complexity Low confidence on pseudo labels
BDCL	Dynamic pseudo-labels Class-wise feature alignment	High computational costs
UCDA	Pre-training for better initialization Pseudo-label expansion Contrastive learning for ST	Decoder is specially designed for domain adapative training Highest accuracy among transformer based models
ProCST	Performs well on decreasing gap in pixel level Can be trained separately	Additional UDA method should be used

Table 3.2: Detailed comparison between different methods for Unsupervised Domain Adaptation

3.3 Mars Exploration

Dataset selection is affecting every aspect of our work. Recently many papers were published about mars exploration fields. [22] has discovered a way to use AI4Mars [23] efficiently for Semantic Segmentation. However, considering the difficulty of labeling, crowd sourcing approach is used to annotate images. This issue leads to get less accurate and inconsistent annotations. That is why, AI4Mars dataset has problems of inaccurate labeling. In order to solve this problem partially, [22] used less labels of AI4Mars in order to get higher accuracy by exploiting Semi-supervised approach. Considering robustness of feature extraction backbone trained by contrastive learning, they achieved quite impressive results. After contrastive learning step, they modified task specific head to be focused on semantic segmentation and fine tuned only the header. This structure is applicable for encoder-decoder based structures. However, main limitation for this paper is that contrastive learning requires 4k batch size for better results. This amount of images demand extensive and wide usage of GPU which is main constraint for us. Luckily, authors of SIMCLR paper [20] which is main idea behind the method used in [22], has published pre-trained weights for different ResNet [2] backbones with ImageNet dataset. Despite high performance with this backbone, considering the differences between ImageNet and AI4Mars data, high computation requiring contrastive learning on AI4Mars was expected to be better performing method.

As we are exploring the usage of multi-task learning on our application, we have to use also labels for image multi-label classification part. There are many datasets that can be utilized in which selection criterias should be taken into account. Detailed comparison for classification datasets are represented in table 2.3.

Models	Advantages	Disadvantages
PanCam	Processed images with high resolution Stable sizings Already Augmented View from 2 Rovers	3k unique images Grayscale colorings Underrepresented classes
MSI Curiosity	Color images 6k unique images Vision from 3 different cameras	Only 1 Rover perspective Nore underrepresented classes Low and unstable resolution
MSL MastCam	Stable and high resolution	Only 1 camera view Anomalious images 3k unique images

Table 3.3: Comparison between available classification datasets

3.4 Multi-task learning

Multi-task learning has quite extensive usage on semantic segmentation tasks which can be also exploited with depth estimation and other tasks. Sometimes it is becoming crucial to detect 2 tasks with the same network due to the energy consumption on real-time applications. Additionally multi-task learning enables model to be more robust because of generalization. Overfitting vanishes in this kind of learning.

Semi Supervised Multi-task Learning for semantic segmentation and depth estimation [24] utilizes 2 different dataset which are labeled for either tasks. In that paper, adversarial method helps to improve generalization for shared parts by using different modes. [25] has also exploited semi-supervised approach because of effectiveness in state-of-art papers. However main novelty here is application of cross-channel attention module which increases generalized and shared feature extraction capability.

According to our application, novel multi-task learning methods can be applied, however considering that synthetic data is utilized, domain adaptation method has more importance. Domain adaptation methods discussed in section 2.2 proved that these implementations are already focused on feature extraction generalization in which [24] and [25] had main attention. Therefore, considering complexity of domain adaptation implementations, novelties in multi-task learning frameworks are redundant for our application.

Chapter 4

Datasets

4.1 Rover Study

Feasibility study of multi-task learning on the images of rover requiring to have the datasets for both classification and segmentation. However, Mars datasets do not contain both labels at the same time. Therefore different datasets should be chosen and model should be generalized on these sets.

4.1.1 Semantic Segmentation

As semantic segmentation labels are the outputs of most difficult annotation processes, present real datasets are not accurate and synthetic ones have completely different domains. In our application, AI4Mars and synthetic dataset from AIKO are the sets going to be used.

AI4Mars

Dataset contains unique 16500 grayscale images in which 322 are annotated for test. Validation set have been generated algorithmically to have 200 samples in the set. Unlike the train set, test set contains 3 versions of labelling which allow to choose the number of agreement between different labelers. It means that objects boundaries are more accurately selected if 3 people agreed on the annotation. Reason for the agreement for test is that, whole dataset is labelled by crowdsourcing approach in which many people envolve in the process. As train set is very huge, there is only 1 labeller for each image while for more realistic and accurate testing, test set has been approved by numerous people. Crowdsourcing approach has advantages of fast labelling. However main drawback is that, annotations are not accurate as most of the data is not labelled by domain experts. Moreover, when more people participate in the process, many standards of labelling is created and training dataset is becoming inconsistent. Unfortunately listed disadvantages are present in AI4Mars. Some of the validation images do not contain any label as none of the labellers had agreement over annotated regions.

Dataset contains 5 segmentation classes:

- Soil: The same as soil in the Earth, solid clean surface or with small clasts which has ideal conditions to run the rover over it.
- Bed Rock: Smooth Rocks those do not require the climbing and energy consumption to pass over it. Does not hurt the rover.
- Sand: Standard sand like the one in the earth which has the risk of beaching if rover drives over it. Spirit rover is beached in the sand during the storm
- Big Rock: Rock that have sharp edges and potential to severely damage the rover parts. Curiosity rover has the damages in its tyres because of big rocks.
- Background (null): any object or area outside of the 30 meters range starting from the rover. Rover parts are also considered as null object

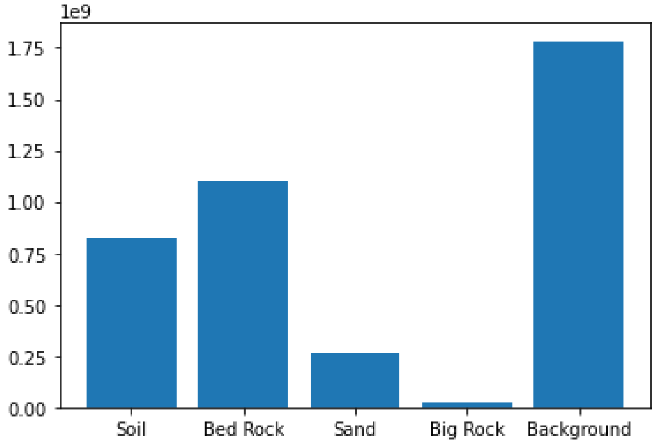


Figure 4.1: Class Imbalance for AIMars based on number of pixels

As figure 4.1 represents, “Big rock” class has significantly less samples than any other classes. The main reason for this imbalance is that big rocks are not occupying large areas inside the image. Additionally, they are not present in most of the images as they are more rare objects. It is common issue to have small objects as minority classes in semantic segmentation. However, “Big Rock“ is one of the crucial classes that needs to be taken into account.

Images are collected from different rovers: Curiosity, Spirit and Opportunity which can provide robust model training. These rovers are equipped with 2 different generation cameras and images are taken in different SOLs (Mars day) and years.

Whole dataset is analyzed by publishers and outliers like full sky images, dark and useless images have been removed. Despite relatively low brightness of the dataset, all images have been post processed, brightness is equalized and resolution has been defined stable at 1024x1024.

Figure 4.2: Example from AI4Mars dataset: Red: Soil, Green: Bed Rock, Blue: Sand, Yellow: Big Rock. Picture in the middle represents original image without label

Synthetic Images

Dataset contains unique 18000 images, 1125 and 375 images are chosen to be validation and test sets accordingly by respecting the class balance. All images are RGB, in resolution of 1024x1024 and contains very accurate labellings as they are generated by 3D environment softwares. Alongwith labels and images, special type files were provided for measuring of depths which indicates the meters far away from camera. All annotations are covering whole area without 30 meter range limit.

As this synthetic dataset has been imitated by AI4mars dataset, main classes are present in this dataset too: Soil, Bed Rock, Sand, Big Rock, Background (null). These classes represent the object as it is defined in real dataset while only 1 class more defined for scientific operations. It is “RSTA“ class which is the metal object for storage of substances from Mars surface for scientific purposes. In mission plannings, these metals are left by older rovers to be collected by new landers for analysis. However, amount of pixels for this class is highly negligible that we included into background class.

“Big Rock“ has similar imbalance problem as it is described in AI4Mars. Unfortunately because of algorithmic labeller error, very few big rocks are annotated in a wrong way as they are present in figure 4.4.

Main drawbacks of this dataset is that, rover parts are not represented in any of the image. Additionally every image has the same position of camera which is

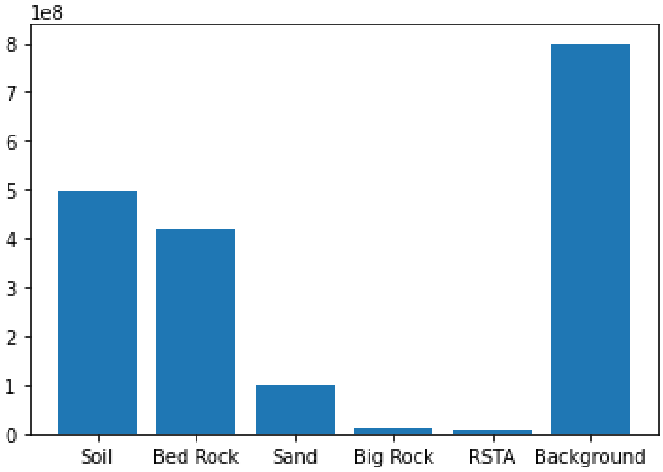


Figure 4.3: Imbalance analysis of synthetic rover dataset

horizontally pointed. However in real scenario images are taken in different angles with the soil. That is why, this dataset creates a challenge.

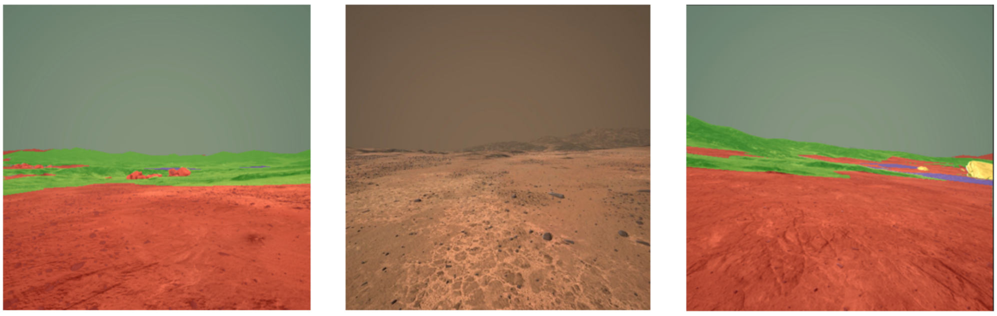


Figure 4.4: Example from synthetic rover dataset: Red: Soil, Green: Bed Rock, Blue: Sand, Yellow: Big Rock. Picture in the middle represents original image without label

Comparison

As these 2 datasets are considered for the same purpose, domain adaptation method aims to decrease the gap between these datasets. Therefore, having the comprehensive comparison between real and synthetic datasets has utmost importance in choosing the proper domain adaptation method.

Most of the domain adaptation methods are developed mainly on the automous car application in which mainly datasets like CamVid, CityScapes for real domain, GTA5, IDDA for synthetic domain are used. Despite the domain difference, these sets have the main commonalities like the same shooting angle, shapes of the object, standards in the signs, traffic lights. However, synthetic rover and AI4Mars datasets do not share these common characteristics. First and most important point is the shooting angle as mentioned before. Almost 70% of the real images contains, rover parts, which can create many problems during the training.

In order to match the real dataset, some pre-processing steps is determined for the synthetic dataset. First of color images are converted to the grayscale style. All images are downscaled to the 512x512 as image classification dataset meets only this requirement. Random crops are defined for the synthetic image to create different angle effects.

Additionally there is also context shift in the datasets which is unsolvable issue except considering it in the metrics. Specially there are contradictions with bed rock and soil classes. In synthetic dataset Bed rock is considered whole high area including soil like surface on it while AI4Mars has the bed rock classes labelled only the seen part of the rock.

4.1.2 Multi-label Classification

As Synthetic rover dataset provided by AIKO does not contain any label for the classification part and it is time consuming operation, classification dataset is needed for training. Therefore PanCam image classification dataset is utilized for the purpose.

3000 unique images from PanCam dataset have been separated originally for train, validation and test with 1800, 462 and 742 respectively. Each training set image has been augmented 20 times by the publishers of the dataset. It is advantageous because it is difficult to augment image classification image in which there is risk of loosing the features as location of the object is not known. The same issue gives disadvantage that, careful and limited augmentations does not contribute to the robustness of the model. Images are in 2 different resolutions: 512x512 and 1024x1024. All images are processed to have stable lightnings and outliers are removed originally.

25 classes have been introduced in full dataset and 1 image can contain more than 1 label. Classes are:

- Rover Deck, Arm Hardware, Other hardware, Rover parts, PanCam Calibration target: these 5 classes are representing different parts of the Rover.
- Rover Tracks: These are created by the movement of the rover, track of the tyre on soil
- Soil trench, bright soil, Soil, Nearby surface: These classes show different types of soils

- RAT Brushed target, RAT Hole, Artifacts: different types of drills on soil for scientific purposes by rover
- Rock Outcrop, Float Rocks, Rocks, Linear and round Rocks: Various types of rocks exist in Mars
- Spherules, Clasts: Names of very small stones on the soil
- Distant Vista: Very far mountains and surfaces
- Sky: Class exists if camera shoots in an angle that sky is seen
- Astronomy: Miscellanious class
- Dunes/Ripples: Sand

3 classes are significantly underrepresented, that is why, we removed those classes from the dataset during training: Astronomy, RAT brushed target, Arm Hardware. Other classes are simplified according to the similar meaning that they are representing. For example, Rock types are simplified and only 1 class created named Rocks. After some analysis, bright soil exists always on rover tracks which means they can be simplified. At the end remaining 12 classes are: Rocks, Rover tracks, Soil, Clasts, Spherules, distant vista, sky, Rover parts, artifacts, Dunes/ripples, Soil trench, RAT Hole. After simplification image classification results have improved as sample size per class have increased.

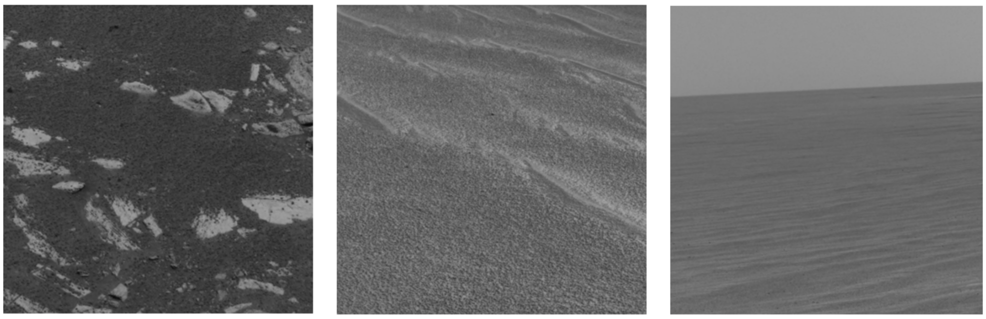


Figure 4.5: Example images from PanCam

4.1.3 Merged Validation Set

As it is mentioned before, there is no dataset that contains the labels for both tasks in Mars application. That is why, to test the model accuracy better, we took equal number of samples from AI4Mars and PanCam datasets and annotated them for both tasks. This can help us to determine the robustness of the multi-task learning model as both datasets have different characteristics. Unlike AI4Mars, PanCam dataset contains close up images for soil, sand and rocks which creates some differences in samples.

After simplification of the PanCam classes, It became faster to label 50 AI4mars images with multi-label classification. We utilized label-studio for 50 PanCam images to label semantically. Overall 100 images from test sets of both dataset have been taken and without further pre-processing,

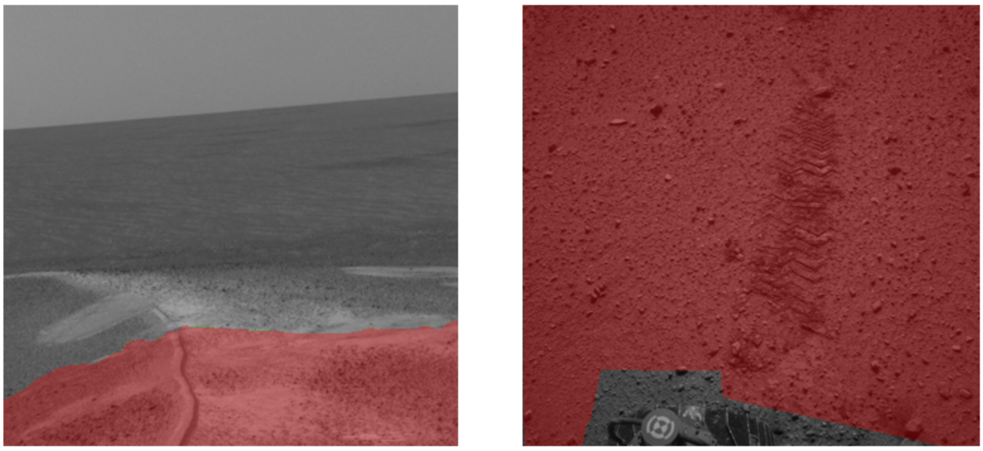


Figure 4.6: Example from merged test set. Left image shows sample from PanCam, right image is from AI4Mars

4.1.4 Data pre-processing

Variety in characteristics of mentioned datasets is not negligible and should be considered during pre-processing step. First of all, real images are all in grayscale format. That is why, synthetic images should be converted to the same format to get better results in real world scenario. Image resolution have been changed to the 512x512 due to 2 reasons:

- Some of the PanCam images are in 512x512 resolution. Removing these images would severely damage the performance of classification as number of samples are low.

- High resolution brings high memory usage during training in GPU. Due to the limited resources, every way to decrease GPU usage has been tested. Lower resolution like 256x256 has been used for optimization and code testing while it was insufficient to increase the performance. That is why optimal value for training has been selected 512x512

For synthetic images to decrease the domain gap, some noise and normalizations should be applied to the images. Gaussian blur is selected to be the most robustness giving augmentation to increase the performance. Horizontal flip, random brightness adjusting are additions for the generalizations.

Additionally, due to the strong class imbalance for “Big Rock“, we had to create custom augmentation method. In this method, we divide the image into 9 overlapping sliding windows. Then in each window, “Big Rock“ class is checked and compared with other windows whether number of pixels for minority class is maximum or not. Maximum occupation is selected to increase the number of pixels for big rocks.

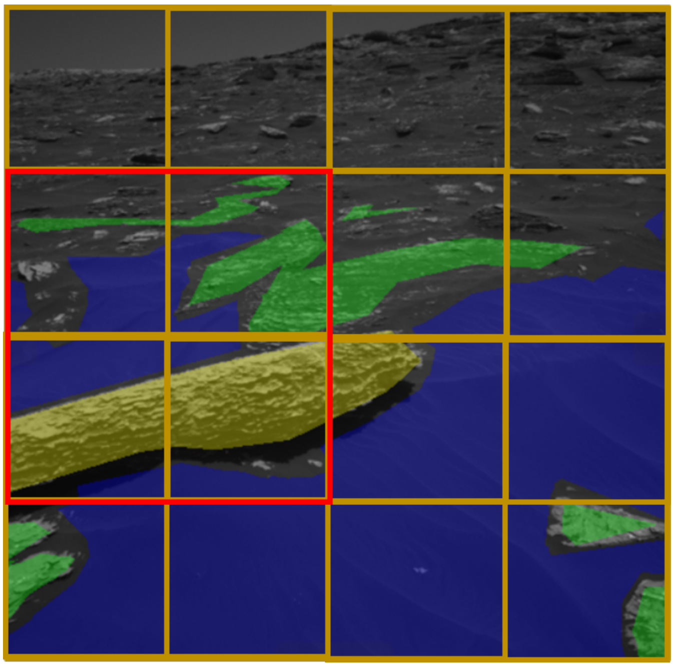


Figure 4.7: Custom Augmentation: Red Window shows the selected region

4.2 Drone Study

Semantic segmentation on Ingenuity drone is one of the newest applications to be tested which creates the challenge in lack of resources. In other words, there is no labeled real images that can be helpful to plot baseline results, improvement, validation.

4.2.1 Drone Synthetic Images

The dataset contains 5950 unique images in which 200 images were allocated for test and 100 for validation set. All images are in 1024x1024 resolution and colored format. 4 different types of samples have been provided: samples taken from 5m, 9m, 11m, 14m. It is expected to give the model more robustness as the real flights do not have all stable heights. All images are labelled semantically by using 3D annotation generator.

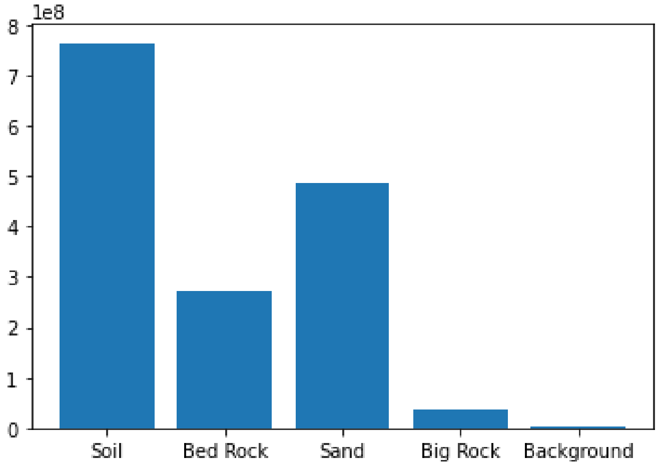


Figure 4.8: Class imbalance ratios in drone synthetic dataset

The same classes as in rover has been used for this dataset while there is no null class in this application. Images are resized to 512x512 match with real dataset.

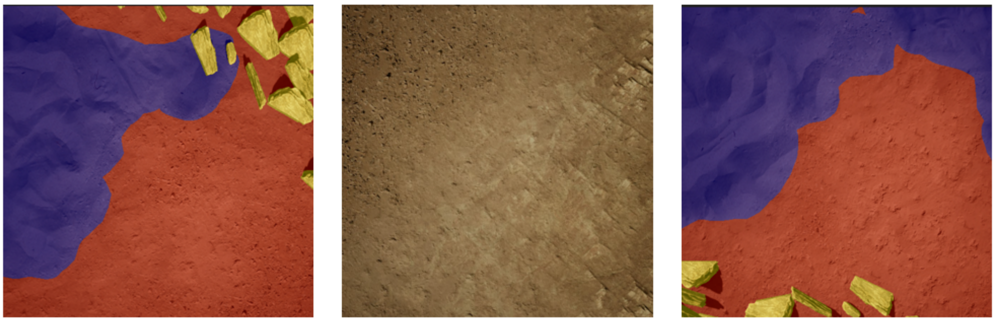


Figure 4.9: Examples from Synthetic drone images: Middle picture represent original image. Labelling standard is the same as the one with rover

4.2.2 Real Images from NASA Ingenuity

Despite the lack of labels, there was only way to get real images which was the flight logs of ingenuity. These images were published in NASA website, according to the ID and day of the flight. Moreover, flights happened in mostly 10 meters height while some of them was in 12 and 8 meters. 4590 grayscacle images were collected from the website in 640x480 resolution. As they are captured by the camera with fisheye lens, all images have certain distortions as they are depicted in the figure 4.10.

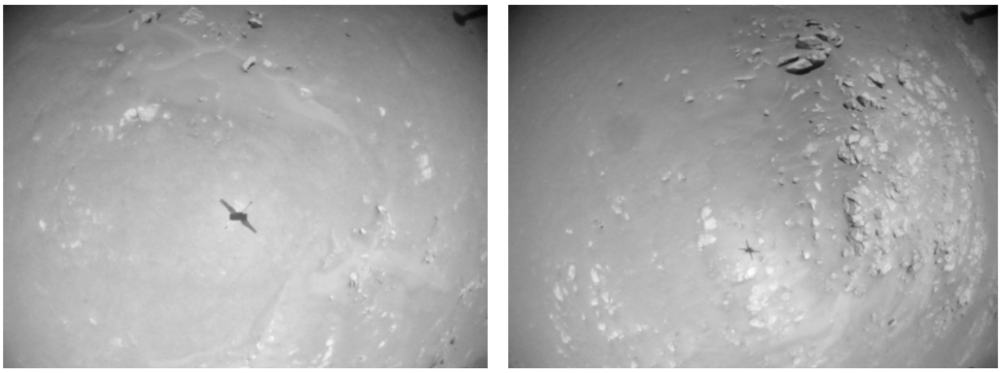


Figure 4.10: Example images from Ingenuity with fisheye distortion

Average flight distance was around 300 meters and average number of pictures per flight was 180. This means that dataset contains images of every 1.5 meters on the surface of Mars. In some flight logs, distance was short and this rate was lower and vice versa. So at the end, some images can look very similar as 1.5 meters distance is not so much distinguishable from 10 meter high. However, at the same time, synthetic images are also designed in the same way. That is why, performance expectation is very high.

However considering synthetic dataset has no distortion, these images is not suitable in default version. That is why, fisheye distortion has been removed by using “DeFisheye“ library and resized into 480x480. Remaining distortion is expected to be adapted during domain adaptation technique.

In order to validate the model over the real dataset, we selected 50 post processed samples to annotate them in the same way as in rover study. Label studio is used to label the sample. Labelled and post processed examples are given in the figure 4.11.

Images are used in 512x512 by upsampling 32 pixel in each dimension. We expected that there will be no information distortion in this operation.

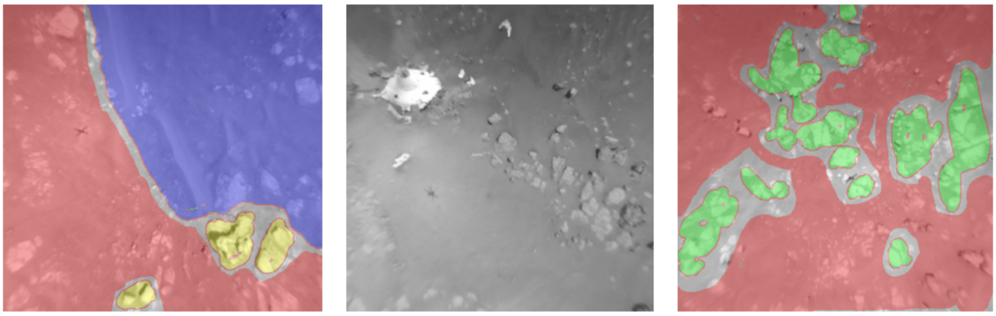


Figure 4.11: Examples from post-processed and annotated dataset

Chapter 5

Methodology

5.1 Model Architecture

According to the papers in recent years, transformer-based Segformer models are presenting pretty high mIoU over all other semantic segmentation models. Especially because of its robustness to domain changes, Segformer is main selection of this project.

Segformer is encoder-decoder structure that consists of transformer based backbone and CNN based head for specific task. Backbone constructed by Mixed 4x4 transformers that have wide selection of types due to its depth in each layer and consequently its efficiency. Lightest and fastest encoder is considered as MiT-b0 and gives compromising results due to its architecture while MiT-b5 has largest number of parameters for highest accuracy. Regardless of number of parameters, each backbone has the same architecture, processing the information in the same manner and variety is because of number of modules.

As all transformer-based backbones, every information fed to the network is divided into different patches. Vision transformers [9] requires 16x16 non-overlapping patches for more dense prediction while MiT working on 4x4 overlapping patches. In other words, according to the model, each image consists of overlapping windows as CNN but with bigger receptive field. After overlap patch embedding stage which separates the patches from image, self-attention layer starts to create attention map for features. Each attention mechanism consists of multiple scale-dot product attention layers. Efficient self-attention is new variant to decrease the computational cost in this stage while the purpose is kept the same.

$$attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

As explained in [26], attention layers projects the output of patch embedding to the 3 matrices called query (Q), key (K) and value (V). Main purpose of this operation is to find similarity between the patches. This approach firstly developed in natural language processing in which the connection between tokens (words) had huge importance. Segformer and also other transformer based architectures finds the connection between different part of the features in image and increases the performance.

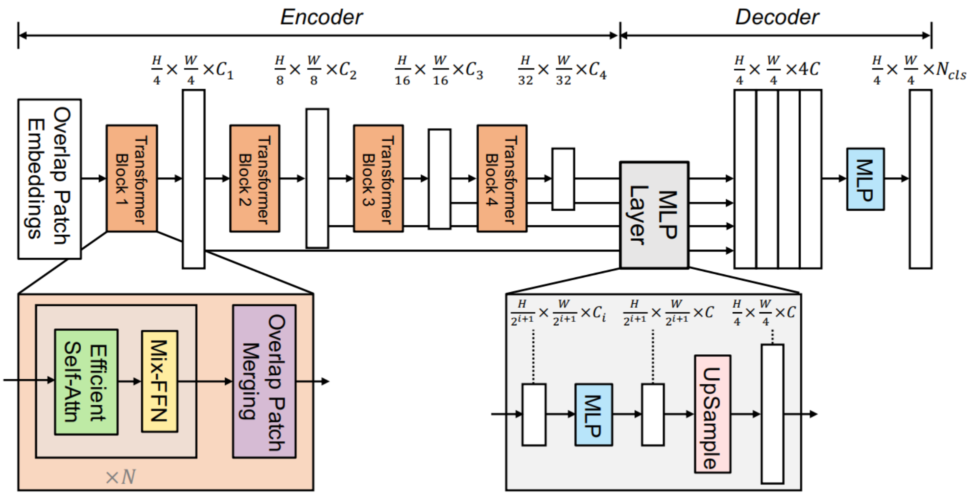

Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
$\times N$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

Figure 5.1: Figure shows the structure of simple Segformer model which is based on encoder-decoder architecture. The image is taken from Segformer paper

Unlike the traditional transformers, Segformer uses mix-FFN (mixed feed forward network) to obtain dynamic patch embeddings. By this way, with different resolutions, performance does not vary too much during testing and numbering of the patches are data oriented. In the final stage of 1st layer, patches are combined and fed to the next layer and to the decoder directly. Having 2 outputs of the layer increases the performance because decoder gets the information at different depths and spatial resolutions.

Segmentation head consists of lightweight CNN network. Unlike DeepLab, there is no atrous convolution mechanism. 4 layer input with high depth is combined and high resolution spatial information is created.

In our cases, we are using the Segformer MiT-B0 for every tests. Drone application needs only segmentation network, therefore no modification is necessary for that part while multi-task learning requires to add the part for the classification as described in the figure 5.2.

Classification head consists of convolution layers and neural networks. After extensive downsampling by using max pooling and depthwise convolution with 1x1 kernels, amount of information has been decreased. According to the test carried out on PanCam dataset, we decided to keep very small amount of information as it

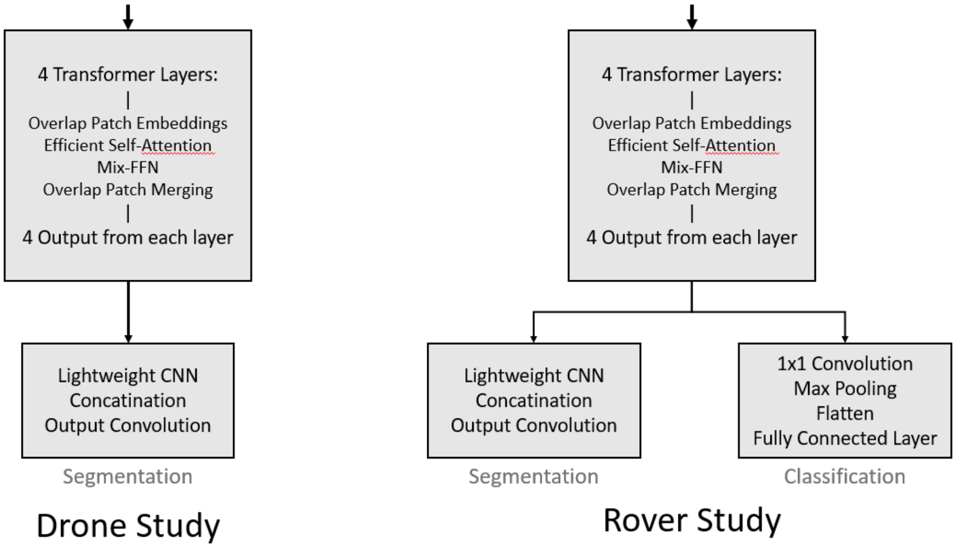


Figure 5.2: Architectures for Drone and Rover study

prevents the overfitting. From another side, we found a balance not to decrease the performance by eliminating many details from input. In overall 6560 parameters have been calculated for classification head.

5.2 Domain Adaptation Technique

According to the literature review, best performing methods and parts were analyzed. After testing and failing in adversarial learning based DecoupleNet [21], Self-training based network seemed to be promising according to the advantages it gives. However, due to the availablity of the details of methods, most clear and beneficial steps have been chosen to apply for our dataset. Another challenge was the characteristics of the dataset which was completely different from the autonomous car application as camera angles are changing in real situation for rovers. That is why, benchmark results for each method can not be considered as a reference to Mars application and comprehensive analysis should be made to enhance the perfomance.

Considering the ideal and noiseless scenes of synthetic dataset, as an augmentation, high range brightness change, horizontal flips and gaussian blur is applied on the image. This gave a chance to increase the robustness of the model training and decrease the gap with noisy real environment. Another augmentation is specially designed for minority class “Big Rock“. In this modification, we divided the image into the windows and determined the area in which big rock is mostly present. To achieve the purpose we used labels of the synthetic images to define occupation of the big rock in the image. So that selected window is cropped in the predefined resolution of 512x512 in order to compensate the class imbalance.

Our customized domain adaptation method have 3 steps. While the first step is based on completely segmentation task, remaining ones try to decrease the domain gap between synthetic and real datasets. All weights are updated simultaneously not to loose an effect of the change in the previous steps.

5.2.1 1st step: Fully Supervised training

Rover Study

1st step of training is fully supervised learning over labeled datasets. As source dataset contains only semantic segmentation labels, for classification side training, we are using PanCam dataset. So in this step, combined batch from 2 datasets is fed to the network and each half of the dataset is responsible on training appropriate part of the networks (e.g in images, 4 from synthetic trains segmentation, 4 from PanCam trains classification side) This approach helps us, to generalize on statistics of batch normalization layers well.

As it is mentioned, PanCam classification dataset is based on multi-label classification. That is why, samples can have more than one label. That is why, to compare the ground truth and the prediction, Binary Cross Entropy Loss has been selected. In the other hand, according to the tests over synthetic datasets, dice loss have been performing much better in measurement of the distance from the ground truth.

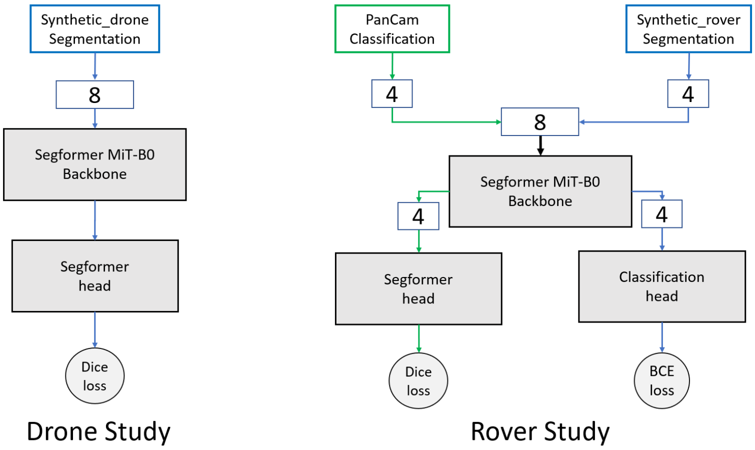


Figure 5.3: Left Image for Drone: Simple segmentation training with 1 synthetic dataset. Right Image for Rover: Fully supervised training jointly with classification and segmentation datasets. 4 samples from each dataset are concatinated before the backbone. After backbone they get seperated for appropriate parts.

Drone Study

Only the 1st step is different for the drone study as only segmentation task is needed for this task. Structure is pure segformer MiT-B0 architecture and the segmentation dataset is synthetic provided by AIKO. The same dice loss is utilized for the training and 8 batch size is used as it is tested to be optimal hyperparameter. They are described in figure 5.3.

5.2.2 2nd step: Contrastive learning on backbone

As a 2nd step of training, unlabeled target dataset for rovers from AI4Mars, PanCam and for drone from NASA Ingenuity images is fed to the network for contrastive learning as published in SimCLR. In this learning, the same image is augmented several times, and different versions of this 1 image are considered as label for each other. So the model is trained in a way that it should be able to distinguish if these versions are generated from the same images or not. Very high batch size is beneficial for this training to increase the performance.

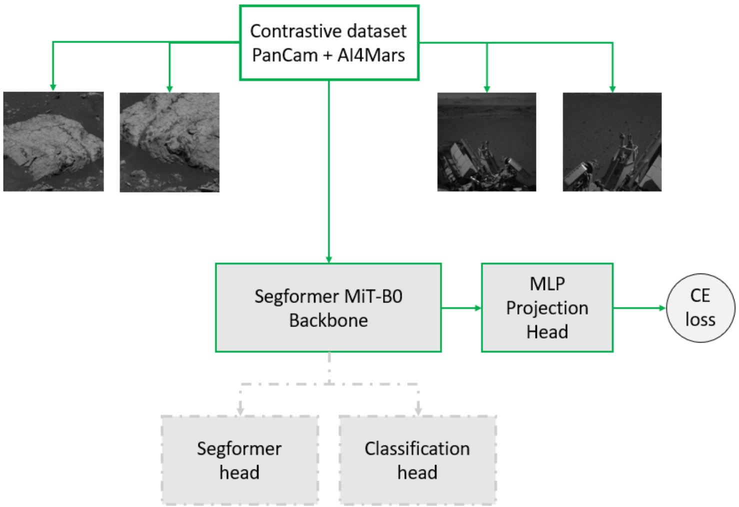


Figure 5.4: Architecture for step 2 (contrastive learning). In case of number of views is 2, 2 variants per sample are generated for self supervised learning.

In more details, MLP header is attached to the backbone which projects features into 128x128 dimensions. Each image is changed according to the predefined number of views which identifies the number of versions for 1 image. Contrastive learning should keep the features for different augmentations of the same image together while that of other images should be repelled from these features. The same applies for all images in order to create detailed feature extraction scheme.

At the end of the network, cross-entropy loss is utilized to measure the difference and save the states.

5.2.3 3rd step: Self-training on output space

3rd step contains the self training part as described in [13]. In initizialization 2 identical networks with similar weights were defined. It is also called mean teacher method as it is relied on Exponential Moving Average method. 1 of the architectures is called teacher which generates the pseudo labels without calculation of gradients. Another model is called student which is the one that is updated with other steps too and calculates the gradients according to the inputs. Another difference is that teacher network takes only original images as input while the student model have strong augmentations of the same image in the input space. In the end of the step, pseudo-label is considered as ground truth for the prediction coming from the student.

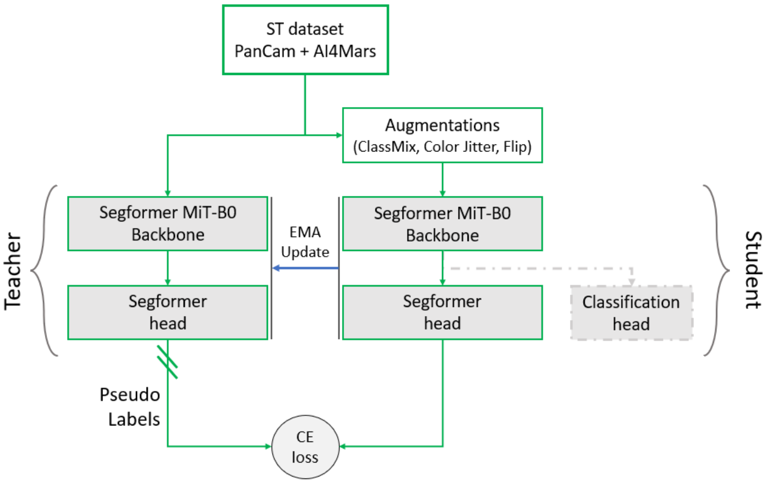


Figure 5.5: Description of the 3rd step: ST dataset contains only unlabeled data which is utilized for generating pseudo labels and augmenting the image to apply mean-teacher method.

Considering that teacher network generates the pseudo labels which are considered as ground truth and data is unlabelled, it is not appropriate to randomly initialize the model weights. Daformer [13] applied ImageNet weights as it is done in [27] for different CNN based model. ImageNet is dataset that has completely different features than mars datasets, because soil, sand and rocks are not labelled as it is needed for our application. Sharp corners, same structures make the situation easy for ImageNet users to adopt it for further trainings. That is why, despite low labelling accuracy of AI4Mars, only way to initialize weights was to train the model offline. By this way, we can be sure that, teacher model will generate much better results.

2 main critical point should be mentioned:

- Teacher network is updated according to the EMA of student at each epoch. This allows to keep the confidence of the pseudo labels high which prevents to have false direction from teacher network. Exponential moving average method is defined as:

  $$O(\theta) = \lambda C(\theta) + (1 - \lambda) J(\theta)$$

  Where C is the parameters of teacher, J is new parameters of student and alpha is hyperparameter to give weight for the models.

- Special augmentation is applied to the input of the student network. Apart from horizontal flip and color jitter, ClassMix is also used which is segmentation based data augmentation technique.

Main principle of ClassMix is to mix the features of unlabelled dataset by preserving the object boundaries. In order to do that, pseudo labels are also used to define the classes inside the image. 2 samples are taken from the dataset and they are fed to the teacher network. Pseudo labels from the teacher is masked by binary filters composed of half 1s and half 0s. Then, according to the masked labels, 2 images are combined to create new mixed image by respecting the objects inside samples. In other words, 1 object from 1 sample is taken and put to another image to create completely new image. For example, car from road is taken and have been put on the sand which creates new image contains car on the sand (randomly).

5.2.4 Motivation for selection of Domain adaptation steps

As it is mentioned before, DecoupleNet is adversarial based method and it was unsufficient as they don't work well in the feature space. Considering the huge differences between synthetic and real datasets, in output space it was very difficult to close the gap to fool the generator. That is why, after tests we did not utilize the adversarial architectures. Self training based methods pay more attention in training of the real datasets which generates better trainings and can eliminate the some part of disadvantages of different characteristics of the dataset.

Mean teacher self-training method is also used to increase the performance of the supervised training and make the model more robust by artificially increasing the dataset size. Additionally, drawbacks of self-training like low confidence on pseudo labels, late updates have been solved in this mean teacher approach and proved to be one of the best performing techniques. So by this way, backbone and segmentation head is not overfitting on synthetic dataset.

Contrastive learning from SimCLR [20] have showed perfect improvement over supervised trainings. Usage of this method in one of the Mars Image segmentation [22] papers was another prove that this method is highly beneficial despite the computational costs. Moreover, forcing the backbone to increase the ability of feature extraction over real dataset would be advantageous for multi-task learning purposes. Having the classification labels of PanCam dataset will increase the chance to understand the features of rover parts existing in real image but not in synthetic images.

5.3 Metrics

For semantic segmentation of both studies, mean intersection over union is considered best matching metrics as mentioned in Background section. However considering the anomalies present in the AI4Mars dataset, we added tolerance to limit them. This unwanted situation comes from the fact that there are negligible number of pixels which remains after agreement of 3 annotators. Existence of this pixel verifies that there is certain class in the picture while in reality that class should not be annotated. That is why, mIoU estimation have been used principally but by using slight modification over calculation. Only difference is that IoU for each class is calculated if and only if certain amount of pixels exceeds some threshold in label and prediction.

For classification purposes, F1 score is one of the robust metrics that can be good performance indicator in case of strong class imbalance. All optimization of the architecture and overfitting has been done on measurements of F1 score.

Chapter 6

Results

6.1 Configuration

Due to the resource limitation, we avoided extensive optimizations and hyperparameter search to save time. Low resolution test at 256x256 and optimization have been carried out in RTX 2060 6GB GPU while full resolution training was on Google COLAB and 2 distributed RTX 2080 8GB GPUs. Therefore batch sizes were kept low at 8 although both self-training and contrastive learning require 128 for optimal training and performance at full resolution. Additionally, for supervised training 4 samples from classification and 4 from segmentation have been fed to the network.

Adam optimizer was used with learning rate of 0.0001 which decreases by polynomial rate at each epoch. Dice loss have been used for supervised semantic segmentation training while cross entropy is used for other steps. Binary cross entropy function have been used for supervised multi-label classification training due to the reason that activation functions for output layer have been determined as sigmoid.

6.2 Measurements

6.2.1 Rover Study

In order to understand the effect of the method in each of the dataset and analyze better, we trained in each configuration. First tests have been carried out on AI4Mars test dataset for segmentation as first and main optimizations was focused on navigation side. Model trained on Synthetic dataset and tested on AI4Mars results have been shown in table 6.1

Main aim of unsupervised domain adaptation method is to prevent overfitting on

Training type	*Test Dataset*	*Soil IoU*	*Bed Rock IoU*	*Sand IoU*	*Big Rock Iou*	*mIoU*
No UDA	AI4Mars	19.89	0.25	23.38	0.98	20.94
With UDA	AI4Mars	**49.94**	**45.71**	**33.17**	**2.43**	**42.14**

Table 6.1: Comparison of domain adaptation method usage for AI4Mars dataset in semantic segmentation

synthetic dataset. Therefore, we tested performance of the model on the synthetic dataset which is represented in table 6.2

Training Type	*Test Dataset*	*Soil IoU*	*Bed Rock IoU*	*Sand IoU*	*Big Rock IoU*	*mIoU*
No UDA	Synthetic	89.43	62.96	70.42	45.81	86.91
With UDA	Synthetic	89.76	49.97	24.97	3.78	42.57

Table 6.2: Comparison of Domain adaptation method on test set of sythetic dataset in semantic segmentation

Training Type	*Test Dataset*	*Soil IoU*	*Bed Rock IoU*	*Sand IoU*	*Big Rock IoU*	*mIoU*	*F1 Score*
No UDA	AI4Mars+MER	26.39	41.99	6.76	1.1	25.38	76.98
With UDA	AI4Mars+MER	60.22	45.05	7.85	0.36	50.23	81.19
No UDA	Merged Set	17.21	32.16	14.06	11.16	23.94	57.48
With UDA	Merged Set	50.51	42.92	1.28	5.45	41.92	65.13

Table 6.3: Comparison of Domain adaptation method on test set of different datasets in multi-task learning

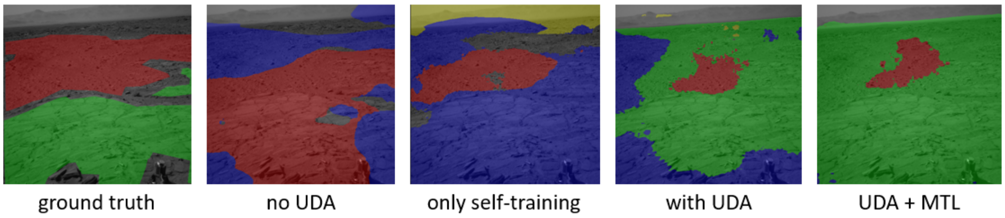
Supervised Learning	*Augmentation*	*Self-training*	*Contrastive Learning*	*Multi-task learning*	*mIoU*
✓					11.26
✓	✓				20.94
✓	✓	✓			32.74
✓	✓	✓	✓		42.14
✓	✓	✓	✓	✓	50.23

Table 6.4: Ablation study for Rover

The same experiments have been performed for multi-task learning. However, as there is no classification labels for synthetic dataset, PanCam dataset has been used alongside the Synthetic dataset. Performance have been tested in 2 test datasets which are, AI4Mars + PanCam test datasets and merged validation set for multi-task learning as shown in table 6.3.

We performed ablation study to check the effectiveness of the defined method on AI4Mars test dataset with models trained on synthetic training set in table 6.4.

Visualized ablation study has been represented in figure 6.1



ground truth

no UDA

only self-training

with UDA

UDA + MTL

Figure 6.1: Test carried on different levels of training for rover

6.2.2 Drone Study

We applied the same implementation for the semantic segmentation with rover study but with different datasets. Model trained with synthetic drone images with and without unsupervised domain adaptation have been tested with real and synthetic test sets as shown in table 6.5

Training Type	*Test Dataset*	*Soil IoU*	*Bed Rock IoU*	*Sand IoU*	*Big Rock IoU*	*mIoU*
No UDA	Real Ingenuity	76.73	89.41	0	27.04	58.37
With UDA	Real Ingenuity	51.26	44.08	60.23	47.36	58.35
No UDA	Synthetic	89.26	64.37	85.94	80.75	85.36
With UDA	Synthetic	77.44	69.53	93.23	83.66	84.62

Table 6.5: Comparison of Domain adaptation method on test set of different datasets in segmentation of drone

The same kind of ablation study has been tested on drone datasets to prove the effect of the unsupervised domain adaptation method in table 6.6.

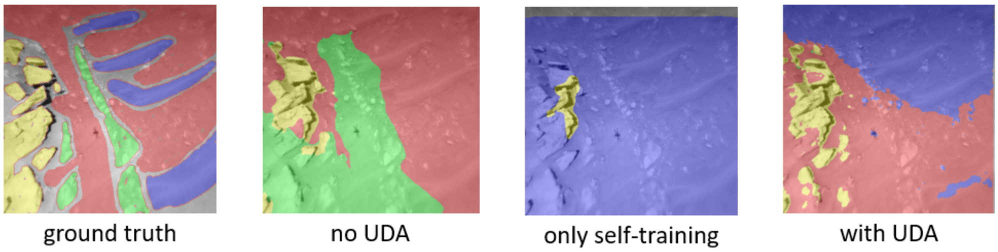
Supervised Learning	*Augmentation*	*Self-training*	*Contrastive Learning*	*mIoU*
✓				26.16
✓	✓			58.37
✓	✓	✓		43.49
✓	✓	✓	✓	58.35

Table 6.6: Ablation study for drone

We can see the effect of the contrastive learning addition which is named as full UDA in the figure 6.2

6.3 Discussion

Starting from the overall results of AI4Mars in table 6.1, progress of the customized UDA can be seen easily as there is significant increment in mIoU. The same progress



ground truth

no UDA

only self-training

with UDA

Figure 6.2: Visualization of the results at different levels of training for drone

was expected from the drone part in table 6.5 but mIoU stayed the same. However, the main improvement was in performance of minority classes which were "Sand" and "Big rock". Training with complete synthetic dataset is not sufficient to detect sand as it was seem very artificial in synthetic set. That is why, efficiency of customized UDA can be proven by balanced result after UDA utilization in drone study.

From the table 6.3 it can be seen that utilization of MER PanCam Dataset is beneficial for generalization of backbone as semantic segmentation result without domain adaptation had increased in comparison with table 6.1 . Domain adaptation method has been affected by MTL positively, and helped to increase the overall mIoU in every dataset. The reason is that, as classification side is trained by real dataset, backbone is forced more to extract features of real dataset.

Ablation study for each task, shows the roles of each step in the pipeline. According to the tables 6.6 and 6.4, contrastive learning has great support on generalization of the backbone side to the self training which prevents overfitting in the header of the segmentation.

However, there is overall lack of performance which is generated by different problems:

- "Bed Rock" and "Soil" classes have different annotations in synthetic and real datasets. In synthetic dataset, bed rock is considered whole high area which contains bed rocks and soil inside. On the other hand, AI4Mars is annotated in a way that only visible part of bed rock is assigned like that. So, it creates strong context shift. In drone testings, these difference was more clear as represented in figure 6.2 and 6.1. Bed rock is taking more space and includes soil on it.

- Rover parts are not represented in synthetic dataset. However AI4Mars contains a lot of rover parts in the dataset.

- Due to the memory constraints, we could not utilize full capacity of the contrastive learning and self-training. These methods require greater batch sizes. With low resolution it has been tested and proved that higher batch size gives better results while in full resolution it is difficult to implement.

- In rover study, synthetic dataset had class imbalance for big rocks. That did not help to eliminate the disadvantage that we were facing in AI4Mars.
- Images are generated in the same shooting angle in which it is not suitable for the real dataset. That is why it decreases the robustness of the model
- Annotation of merged validation and drone validation sets have been done by us. A lot of mistakes can exist as it is very difficult to distinguish soil and sand or different combinations

Chapter 7

Conclusion

In all domain adaptation methods, main objective of generating synthetic datasets is to eliminate the problems of real dataset in the cost of domain gap. For example, if "big rock" class has the problem of imbalance, synthetic dataset should focus on increasing sample size. However, our synthetic dataset had created more challenges that performance of the model is severely and negatively affected. Context shift, imbalance issue, image angles are the main problems. By analyzing main problems in unsupervised domain adaptation method, we selected best performing and most efficient parts that can be beneficial for our project. We customized the unsupervised domain adaptation steps and by selecting one of the best performing semantic segmentation model, we were able to decrease domain gap in both applications. However, as it is mentioned before, we suffered a lot in rover study due to the dataset problems.

Taking everything into consideration, from segmentation side, having problems of context shift, decreases performance but considering in general it does make too much difference for the purpose. Because, soil and bed rocks do not contain any risk for the rover. Context shift between risky and safe object is dangerous and unacceptable. That is why, despite low performance, results can be called as decent. The issue is applicable for both of the studies. From classification side, there is slight overfitting on the classification head while segmentation side got advantage of generalization of backbone.

In general, considering all problems, risks, available resources and limitations, model performance is decent and can be further improved. Moreover, both tasks have sufficient latency due to light backbone selection for real-time application while low latency application is not objective of this project. High efficiency can be obtained with further minor optimizations on method and wide computational resources.

Chapter 8

Future Improvements

There are some points which should be considered for future improvements. First of all, as it is mentioned in discussions of lower performance, batch sizes should be increased to 32 or 64 for self-training and contrastive learning which will increase the performance definitely.

Supervised Contrastive Learning approach is improvement over self-supervised contrastrive learning and outperforms it. By using the labels of synthetic dataset and possibly AI4Mars, backbone efficiency can be increased.

In order eliminate problem related to absence of rover parts in synthetic dataset, similar approach with Classmix can be applied to transfer rover parts from AI4Mars to the synthetic set, by using labels of both dataset and depth indication labels. This augmentation type change can increase the robustness and performance of big rock class which is failing in predicting rover parts.

Rare Class Sampler from DaFormer [13] should be analyzed and can be applied to the our dataset as big rock class is minority. We thought that these changes to the architecture and hyperparameters should theoretically increase the performance significantly.

Bibliography

[1] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. DOI: 10.48550/ARXIV.1409.1556. URL: https://arxiv.org/abs/1409.1556 (cit. on p. 12).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: https://arxiv.org/abs/1512.03385 (cit. on pp. 13, 18, 24).

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597 (cit. on p. 17).

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. DOI: 10.48550/ARXIV.1706.05587. URL: https://arxiv.org/abs/1706.05587 (cit. on p. 17).

[5] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. *BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation*. 2018. DOI: 10.48550/ARXIV.1808.00897. URL: https://arxiv.org/abs/1808.00897 (cit. on pp. 17, 18).

[6] Rudra P K Poudel, Stephan Liwicki, and Roberto Cipolla. *Fast-SCNN: Fast Semantic Segmentation Network*. 2019. DOI: 10.48550/ARXIV.1902.04502. URL: https://arxiv.org/abs/1902.04502 (cit. on pp. 17, 18).

[7] Saumya Kumaar, Ye Lyu, Francesco Nex, and Michael Ying Yang. «CABiNet: Efficient Context Aggregation Network for Low-Latency Semantic Segmentation». In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2021, pp. 13517–13524. DOI: 10.1109/ICRA48506.2021.9560977 (cit. on pp. 17, 18).

[8] François Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. 2016. DOI: 10.48550/ARXIV.1610.02357. URL: https://arxiv.org/abs/1610.02357 (cit. on p. 18).

[9] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* 2020. DOI: `10.48550/ARXIV.2010.11929`. URL: `https://arxiv.org/abs/2010.11929` (cit. on pp. 18, 37).

[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. «Training data-efficient image transformers & distillation through attention». In: *CoRR* abs/2012.12877 (2020). arXiv: `2012.12877`. URL: `https://arxiv.org/abs/2012.12877` (cit. on p. 18).

[11] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. *LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference.* 2021. DOI: `10.48550/ARXIV.2104.01136`. URL: `https://arxiv.org/abs/2104.01136` (cit. on p. 18).

[12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers.* 2021. DOI: `10.48550/ARXIV.2105.15203`. URL: `https://arxiv.org/abs/2105.15203` (cit. on p. 19).

[13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. *DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation.* 2021. DOI: `10.48550/ARXIV.2111.14887`. URL: `https://arxiv.org/abs/2111.14887` (cit. on pp. 19, 21, 23, 42, 51).

[14] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto L. Sangiovanni-Vincentelli. «Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation». In: *CoRR* abs/2103.16765 (2021). arXiv: `2103.16765`. URL: `https://arxiv.org/abs/2103.16765` (cit. on p. 20).

[15] Geon Lee, Chanho Eom, Wonkyung Lee, Hyekang Park, and Bumsub Ham. *Bidirectional Contrastive Learning for Domain Adaptive Semantic Segmentation.* 2022. DOI: `10.48550/ARXIV.2207.10892`. URL: `https://arxiv.org/abs/2207.10892` (cit. on pp. 20, 23).

[16] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. *Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation.* 2022. DOI: `10.48550/ARXIV.2207.04892`. URL: `https://arxiv.org/abs/2207.04892` (cit. on p. 20).

[17] Yanchao Yang and Stefano Soatto. *FDA: Fourier Domain Adaptation for Semantic Segmentation.* 2020. DOI: `10.48550/ARXIV.2004.05498`. URL: `https://arxiv.org/abs/2004.05498` (cit. on p. 20).

[18] Shahaf Ettedgui, Shady Abu-Hussein, and Raja Giryes. *ProCST: Boosting Semantic Segmentation Using Progressive Cyclic Style-Transfer.* 2022. DOI: `10.48550/ARXIV.2204.11891`. URL: `https://arxiv.org/abs/2204.11891` (cit. on p. 20).

[19] Feihu Zhang, Vladlen Koltun, Philip Torr, René Ranftl, and Stephan R. Richter. *Unsupervised Contrastive Domain Adaptation for Semantic Segmentation*. 2022. DOI: 10.48550/ARXIV.2204.08399. URL: https://arxiv.org/abs/2204.08399 (cit. on p. 21).

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. DOI: 10.48550/ARXIV.2002.05709. URL: https://arxiv.org/abs/2002.05709 (cit. on pp. 21, 24, 44).

[21] Xin Lai, Zhuotao Tian, Xiaogang Xu, Yingcong Chen, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. *DecoupleNet: Decoupled Network for Domain Adaptive Semantic Segmentation*. 2022. DOI: 10.48550/ARXIV.2207.09988. URL: https://arxiv.org/abs/2207.09988 (cit. on pp. 22, 23, 39).

[22] Edwin Goh, Jingdao Chen, and Brian Wilson. *Mars Terrain Segmentation with Less Labels*. 2022. DOI: 10.48550/ARXIV.2202.00791. URL: https://arxiv.org/abs/2202.00791 (cit. on pp. 23, 24, 44).

[23] R. Michael Swan, Deegan Atha, Henry A. Leopold, Matthew Gildner, Stephanie Oij, Cindy Chiu, and Masahiro Ono. «AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars». In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021, pp. 1982–1991. DOI: 10.1109/CVPRW53098.2021.00226 (cit. on p. 23).

[24] Yufeng Wang, Yi-Hsuan Tsai, Wei-Chih Hung, Wenrui Ding, Shuo Liu, and Ming-Hsuan Yang. *Semi-supervised Multi-task Learning for Semantics and Depth*. 2021. DOI: 10.48550/ARXIV.2110.07197. URL: https://arxiv.org/abs/2110.07197 (cit. on p. 25).

[25] Nitin Bansal, Pan Ji, Junsong Yuan, and Yi Xu. *Semantics-Depth-Symbiosis: Deeply Coupled Semi-Supervised Learning of Semantics and Depth*. 2022. DOI: 10.48550/ARXIV.2206.10562. URL: https://arxiv.org/abs/2206.10562 (cit. on p. 25).

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762 (cit. on p. 37).

[27] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. *ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning*. 2020. DOI: 10.48550/ARXIV.2007.07936. URL: https://arxiv.org/abs/2007.07936 (cit. on p. 42).