Research Project

Artificial Intelligence

Radboud University 

Improving Few-Shot Segmentation Using Negative Class Instances

Author:
Lucas Puddifoot
S1015527

Internal supervisor:
Dr U. Güçlü
Donders Centre for Cognition
u.guclu@donders.ru.nl

External supervisor:
Mathijn Wilkens
Senior Data Scientist at
Capgemini
mathijn.wilkens@capgemini.com

Acknowledgements

I would like to thank Mathijn Wilkens for his continued support throughout the research and o ering many insightful comments and guidance. Also I would like to thank Umut Gu clu for helping me shape the premise of the project.

Abstract

Existing Few-Shot Semantic Segmentation (FSS) networks only rely on training with query images containing the class of interest, which hinders their performance in scenarios where the class is absent, compromising the network's ecological validity. To rectify this, we propose a novel training loop that incorporates negative instances during training, providing the networks with exposure to a broader range of class instances. After testing this approach, our results show that not only did the models trained on our novel training loop improve in cases where the class of interest was absent from the query image, the models also improved when the class of interest was present.

Contents

Chapter 1

Introduction

Human beings possess a remarkable ability to rapidly learn and recognize new objects from just a few examples, providing machines with this ability has been a long held ambition, however, as of yet unrealised.

Traditional computer vision networks have been modeled after the structure and function of biological vision systems, which feature cells that detect light in receptive fields [3]. However, due to the limited availability of training data and the lack of computational power required to train deep networks, the development of computer vision stalled until the advent of AlexNet in 2012, created by Krizhevsky et al. [5]. The AlexNet architecture improved the error rate on the ILSVRC-2012 competition from 26.2% to 15.3%, igniting a surge of deeper networks incorporating various layers such as max pooling or sub-sampling, in addition to leveraging the computational power of graphics processing units to expedite training.

As networks have become deeper, the risk of over fitting has increased [2], and the time required to train them has dramatically increased to the point where it can take weeks on multiple high-end graphics cards to complete training. Furthermore, the depth of these networks necessitates having extremely large datasets to prevent over fitting, which significantly reduces the usefulness of computer vision in business applications. Often in the real world, there is a scarcity of available data, and the available data may not be properly annotated. Consequently, data scientists must dedicate a disproportionately large amount of time to creating a dataset suitable for training the network. Given that more than a decade has elapsed since the significant breakthrough of AlexNet, it begs the question: is there a solution to these challenges?

One potential solution for addressing this challenge lies in the utilization of transfer learning. Transfer learning refers to the process of adapting a pre-trained model, initially trained for a specific task, to perform a related task [13]. An example of transfer learning involves the use of a classifier initially trained to distinguish between images of dogs and cats, which is subsequently modified to discriminate between images of horses and birds. Given the inherent similarities between the two sets of classes, it is likely that numerous image features are shared between them. Consequently, the amount of training required for the adapted network to excel at the new task is considerably lower than training a network

entirely from scratch. Nonetheless, it is crucial that the original network has been trained on a su ciently diverse data set to accurately classify images in its new task, thereby necessitating a substantial amount of data [15].

The existing state-of-the-art Few-Shot Segmentation (FSS) network demonstrates a remarkable mean Intersection over Union (IOU) of 69% on the PASCAL-5i data set, despite having access to only a single support image for reference [10]. This achievement is quite impressive considering the limited amount of information available for inference. Conversely, a conventional segmentation network achieves an IOU of 89%, which appears signi cantly higher in comparison. However, it is essential to note that these two performances cannot be directly compared, as the regular segmentation network attains this score through training on nearly 7.5 thousand images, while the FSS network relies on a single support image during inference.

The notable performance achieved by the FSS network indicates a potential mitigation of the data scarcity problem during inference. The FSS approach allows for the successful segmentation of a class with as few as ve reference images, thereby potentially reducing the dependency on extensive annotated data sets and diminishing the associated costs and e orts involved in their creation. Nonetheless, it is crucial to conduct further research to ascertain the applicability and generalizability of few-shot segmentation in diverse data sets and various application scenarios. This will help establish its e ectiveness and reliability in broader contexts beyond the speci c data set it is trained on.

The training process of Few-Shot Segmentation (FSS) models warrants close attention. Notably, FSS models are exclusively trained on query images that contain the class of interest, and are not exposed to query images that lack this class. Consequently, when tasked with segmenting a query image, the network assumes the presence of the speci ed class, as this is how it was trained. This is in contrast to real-world scenarios, where the presence of the class of interest is not guaranteed. For example, a network that is trained to detect people on a rail track using video footage from a camera will often not encounter people in its input. Furthermore, the use of a limited amount of data for training implies that obtaining su cient data of the speci c class that the network must detect in the real world can be challenging. Therefore, it is likely that not every query image will contain the class of interest that is present in the support images.

To illustrate the aforementioned issue, an example is presented using HSNet, shown in gure 1.1. The network was pre-trained on the PASCAL-5i dataset, using fold three as the validation fold. Firstly, an image of a mug was presented to the network, and a mask in blue was used to delineate the mug. Subsequently, an image of a mug was presented to the network, which successfully segmented the mug with a high degree of accuracy. The segmentation mask of the network is shown using red. This demonstrates the network's ability to segment novel classes, since none of the training images contained a mug.

To further evaluate HSNet's performance when the class of interest is absent in the query image, an image without a mug was presented. The resulting segmentation demonstrates that when the class of interest is not present in the image (negative class instance), HSNet struggles to segment the image, often segmenting arbitrary parts of the image instead of leaving the segmentation mask empty. This highlights the limitations of the training process, as the network has not been exposed to query images that do not contain the class

of interest.

After conducting an extensive literature review, it becomes evident that the matter of ecological validity in FSS networks has not received su cient attention, highlighting a signi cant gap in the current research. Addressing this gap is crucial, particularly considering the practical deployment of FSS networks in non-academic settings where the class of interest may be absent. In such scenarios, the ability of FSS networks to function effectively becomes essential, as it can enhance performance without requiring additional training data.

Moreover, as segmentation algorithms become increasingly pro cient in determining the presence of a speci c class in an image, they can be used interchangeably as classi cation networks [17]. Consequently, for FSS networks to be valid and applicable in real-world situations, they must possess the capability to handle negative class instances. This consideration further emphasizes the need to explore the issue of ecological validity in FSS networks, ensuring their robustness and reliability in practical applications beyond academic contexts.

This research aims to address this gap by proposing a novel training method for FSS networks to enable them to recognise when the class of interest is not present in the query image. It is crucial to achieve this without signi cantly reducing the network's performance in positive class instances, allowing for wider applicability of FSS networks in practical settings. The proposed approach has the potential to enable non-academic institutions, including companies, to utilize FSS networks for detecting low-occurrence classes in images.

Another practical application of a Few-Shot Segmentation (FSS) network is its use as a proof of concept. If an FSS network can produce results that are approximately in line with the requirements of a speci c task, it can serve as evidence that segmentation algorithms are capable of addressing the problem at hand. This proof of concept can then serve as a basis for further research and investment of time and resources into re ning the task.

Moreover, FSS networks can facilitate the creation of datasets for the desired class more e ciently. By running an FSS network on an entire dataset, a rough outline of all instances of the class of interest can be generated. Subsequently, a human expert can review the segmentation outputs and re ne them to create pixel-perfect training data. This curated dataset can then be used to train a regular segmentation algorithm. This process accelerates the dataset creation process and reduces the manual e ort required, allowing for faster training of a conventional segmentation algorithm.

Based on the literature, as well as the investigation that I have carried out during my internship I propose the following research question: \How can the performance of HSNet be maintained in positive class instances while also enabling it to accurately recognize and handle negative class instances?" From this research question I generate the following two hypotheses:

- Hypothesis 1: The augmented training pipeline will result in a model exhibiting a reduced false positive rate.
- Hypothesis 2: The augmented training pipeline will result in a model that exhibits improved overall IOU.

(a) Positive class instance, there is a mug in the support image and in the query image

(b) Negative class instance, there is a mug in the support image but not in the query image

Figure 1.1: Comparison between positive and negative class instances with the current version of HSNet

Chapter 2

Preliminaries

This chapter serves to provide a comprehensive foundation for the subsequent research discussion by describing the key aspects of few-shot learning, the utilized data, and the functioning of HSNet during both training and inference stages. The chapter begins by o ering a detailed explanation of the underlying principles of few-shot learning, shedding light on its fundamental workings. Additionally, the chapter encompasses a thorough examination of the speci c dataset employed in this experiment, PASCAL-5i. Furthermore, an in-depth overview of HSNet is presented, covering its functionality during both the training and inference processes. By delving into these preliminary aspects, this chapter establishes the necessary groundwork for comprehending the subsequent research ndings and analysis presented in this paper.

2.1 Few-Shot Learning

FSS networks commonly operate using a technique known as "episodes," which are essentially a package of support and query images [14]. In this approach, the network receives a support image that features an instance of the new class, delineated by a lter or "mask" [6]. Additionally, a query image is provided, which the network must segment to isolate the new class. Subsequently, both the support and query images are fed into a "backbone" network, which refers to a pre-trained convolutional neural network (CNN) with its classi cation layers removed [12]. Figure 2.1 contains a high level diagram an episode and masking.

The backbone network extracts features from di erent layers of the support and query images. These range from simple shapes such as circles and lines to more complex textures such as fur or faces [16]. Following this, the features obtained from the support image are compared to the features derived from the query image. By identifying resemblances between the class of interest's features and corresponding areas in the query image, the network e ectively performs segmentation, thereby isolating the class from the rest of the query image [9]. Figure 2.2 contains a high level diagram of how a backbone creates feature maps from an input image. Here the stick person is comprised of four features: A circle for its head, a vertical line for its body, a horizontal line for its arms and a caret for its legs.

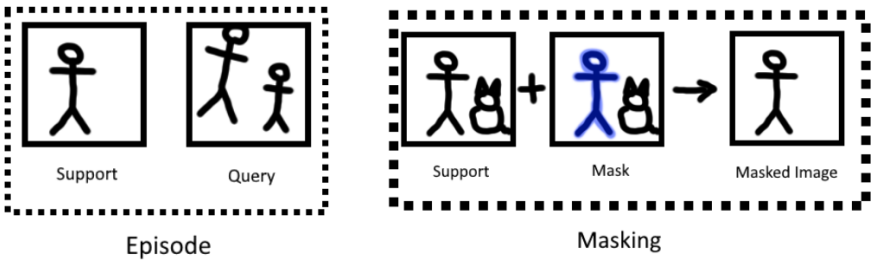

Figure 2.1: High level diagram of an episode and masking.

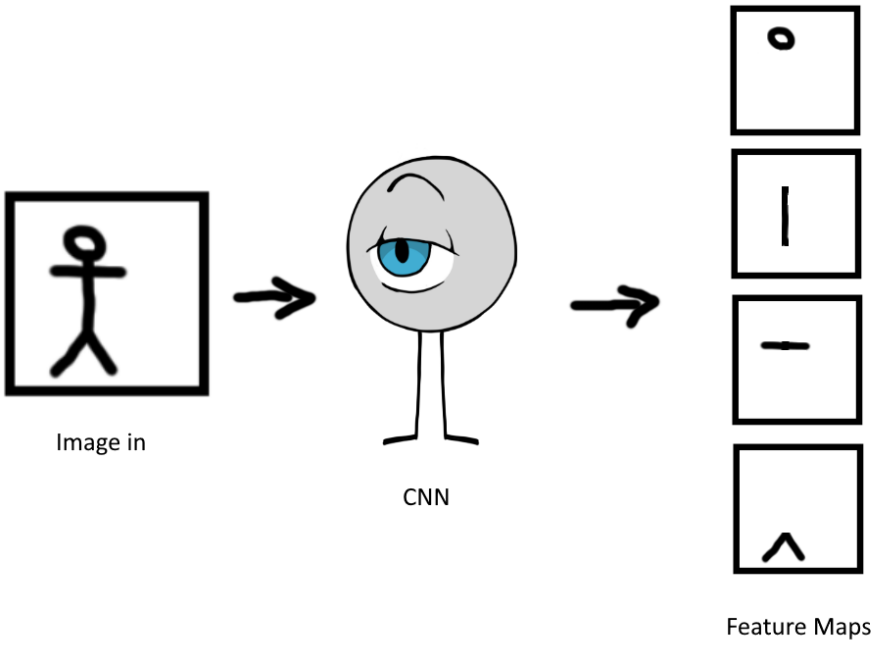

Figure 2.2: High level diagram of a CNN decomposing an input image into its features.

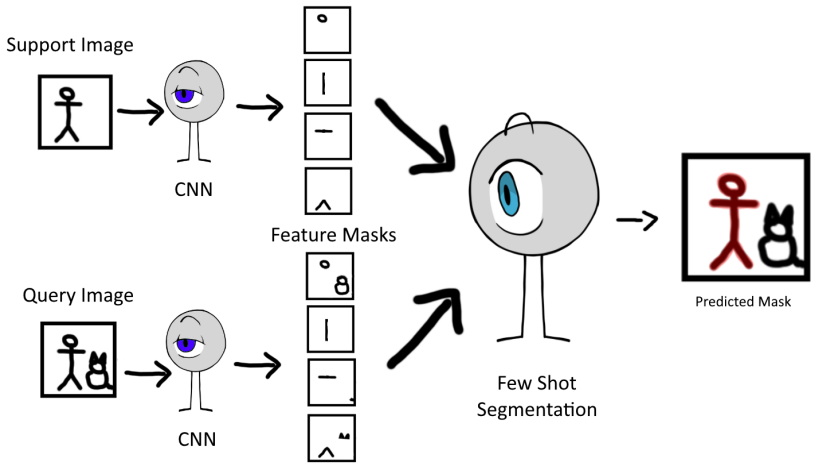

Figure 2.3: High level overview of the full process of few shot segmentation, from the input episode to the nal segmentation output. Here we can see that the features extracted are contained in both images, but the stick person contains one of each feature while the cat has a round body and no legs.

One key characteristic of FSS is its capacity to generalize to new images. This is possible because the network is trained on an extensive collection of episodes, each episode comprising of a distinct pair of support and query images [7]. Through this training process, the network acquires the ability to recognise the essential features that make up a class from the support image, in the query image and segment the class out of the image.

Moreover, FSS networks are not learning to recognize speci c classes based on the features learned over a large number of examples, but rather learning to recognize indicative patterns and features that are from the support images in the query image. Therefore, it can be said that FSS networks are not learning classes, but learning where query images look similar to support class instances [9].

Some FSS models also use various techniques like meta-learning, attention mechanisms, and memory networks to improve their ability to learn from a few examples and adapt to new classes [9, 4].

In summary, as illustrated in Figure 2.3, FSS networks employ a pre-trained CNN to extract features from both support and query images, facilitating feature comparison and subsequent segmentation of new classes in the query image by identifying similarities. Their remarkable ability to generalize from a limited number of examples to novel images and classes is achieved through extensive training on numerous episodes. Furthermore, FSS networks have the potential to enhance their performance through the incorporation of other techniques.

2.2 PASCAL-5i

The dataset utilized for this research is referred to as PASCAL VOC 2012 [1]. This dataset comprises a total of 9,993 images, which are categorized into 20 distinct classes. To facilitate the evaluation process, these classes are further divided into four separate "folds," with each fold consisting of ve classes. The classes within the dataset encompass a diverse array of object settings, ensuring a comprehensive coverage of di erent contextual scenarios. Speci cally:

- Person: person
- Animal: bird, cat, cow, dog, horse, sheep
- Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
- Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

Each image within the dataset is accompanied by a segmentation mask that delineates the assignment of each pixel to a speci c class. Pixels that do not belong to any of the 20 classes are considered background pixels and are not included in any mask. However, it is important to acknowledge a limitation of the dataset regarding the consistency of class de nitions. For instance, the horse class may sometimes include the saddle and bridle, while other instances may not. Similarly, the person class exhibits considerable variation in terms of whether the clothes are encompassed within the mask. Despite these drawbacks, the PASCAL dataset is extensively employed in the eld of FSS and researchers must accept these limitations in order to conduct research within this domain.

During the training phase, the model is presented with three out of the four folds of the PASCAL VOC 2012 dataset for training purposes. The remaining fold is reserved for validation and testing. Given that few-shot segmentation focuses on one class at a time, images that contain multiple classes undergo a modi cation in their masks to accommodate this requirement. Speci cally, the masks are adjusted to create a binary segmentation mask that isolates the class of interest. All pixels corresponding to the class of interest are assigned a value of one, while the rest of the mask is set to zero. This process ensures that the resulting binary mask solely delineates and segments the desired class within the image.

2.3 HSNet

The "Hyper Correlation Squeeze Network" (HSNet), which was presented in October 2021 by Min et al., emerged as a cutting-edge model for few-shot segmentation. This approach garnered signi cant attention in the eld, as it achieved remarkable performance advancements on various datasets [8].

The model operates by comparing each pixel in the query image with every pixel in the support image. This comparison creates a four-dimensional tensor that represents the similarity between the pixels. The tensor is then progressively decoded to generate a mask,

which accurately separates the novel class from the rest of the query image. In essence, HSNet utilizes a pixel-wise comparison between the support and query images to produce a precise mask speci cally isolating the novel class within the query image.

During the training process, a series of image episodes are inputted into a backbone CNN, generating pairs of feature maps | one for the support image and another for the query image. While several networks can be employed to generate these feature maps, RESNET-101 achieved the best results. To ensure the exclusion of irrelevant activations, the feature maps corresponding to the support image are masked using the support mask that indicates where the class instance is in the support image.

Next, the model examines each pair of feature maps to construct a four-dimensional correlation tensor. This tensor quanti es the pixel-wise similarity between the features extracted from the support image and the query image. This analysis is performed for all the feature maps collected from the various layers of the backbone, and the resulting correlation tensors are merged to form an encoded context encompassing the entire correlation information. Consequently, the encoded context contains the degree of similarity between each pixel in the support image and every pixel in the query image at each level of the backbone network.

Finally, the encoded context undergoes a decoding process using 2D convolutions and upsampling operations. This transformation results in the generation of a probability map for each pixel within the query image, indicating the estimated probability of it belonging to the class of interest. During the network training phase, this pixel-level probability can be utilized to assess the network's accuracy, enabling the propagation of error signals back through the network to enhance its segmentation performance on future images. During testing, each pixel is assigned to the class with the highest probability, thereby generating a mask that precisely identi es the location of the novel class within the query image. Figure 2.4 contains a high level diagram of how HSNet works using the feature maps extracted in gure 2.3

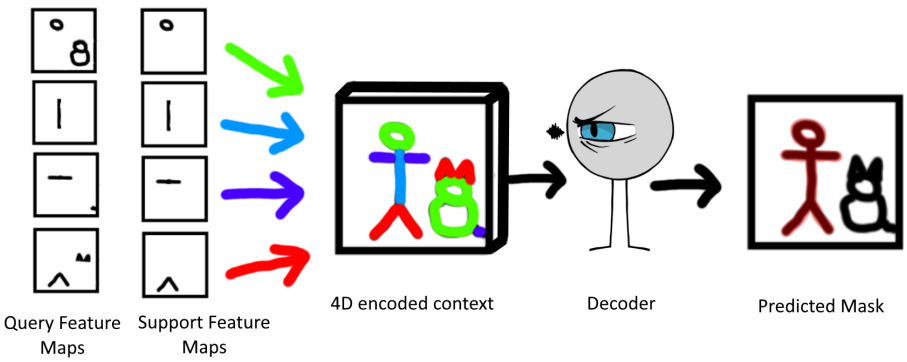

Figure 2.4: High level overview of the decoding stage of HSNet. Here the class of interest is the person, so the 4D encoded context will contain that a person is made of one instance of each of the features extracted. The decoder brings this encoded context down to a two dimensional mask on the query image.

Chapter 3

Method

3.1 Training

Currently, FSS models are typically trained using the following approach: First, the model is presented with an episode consisting of a support image and a query image. The model then generates a logit mask for the query image, indicating its prediction of the location of the class of interest. Subsequently, the model can backpropagate the cross-entropy loss, which measures the discrepancy between the predicted mask and the actual mask, throughout the network.

It is worth noting that FSS make an implicit assumption that the class of interest in the support image will have at least one instance present in the corresponding query image. This assumption arises from the sampling methodology employed for query and support images. FSS datasets are accompanied by a metadata file that specifies the classes contained within each image. During training, a specific class is chosen as the class of interest. Subsequently, the next query image containing that class is selected, and the metadata file is consulted to identify potential support image candidates. A random image featuring the class of interest is then chosen, ensuring it is not the same as the query image, and the network predicts the class segmentation accordingly. As a result, the network is trained to expect the presence of the class of interest in every query image. However, it is important to acknowledge that this assumption often does not hold true in real-world scenarios. In practical applications, there are often instances where the class of interest is absent from the query image.

Consequently, I propose a novel training loop for Few-Shot Segmentation (FSS) networks, designed to address the aforementioned limitation. Prior to sampling each episode, a random value between zero and one is uniformly generated. This random value corresponds to the ratio of positive to negative class instances and to achieve a balance between these two cases, a threshold of 0.5 is established. After this the following steps are carried out:

- Generate a random value between 0 and 1
- If the randomly generated value exceeds or equals 0.5, the episode is sampled con-

ventionally. This means that a query image containing the same class as the support image is selected, following the existing methodology.

- However, if the randomly generated value falls below 0.5, an image that does not contain the class of interest is chosen as the query image.

In order to select support images not containing the class of interest, the sampling function is modified to select a different class randomly from the remaining classes. Consequently, the query image will possess an empty ground truth mask since there is no instance of the class of interest present in the image. Therefore, the correct mask in this case is no mask at all.

Nevertheless, this modification introduced challenges later in the training pipeline. One such issue arises when evaluating the model's performance in segmenting the class of interest. The original masks provided by the dataset lack an outline of the class, which is added to the mask during evaluation. Consequently, it becomes necessary to include metadata indicating whether an episode represents a positive or negative class instance when passing the batch of episodes through the training pipeline. Subsequently, functions can identify and appropriately handle episodes that are negative class instances during evaluation.

Currently HSNet uses the IOU to calculate which weights make the model perform better or worse. This may be a valid approach when only positive class instances are sampled, however, with the introduction of negative class instances this is no longer a valid model selection metric.

This is because the IOU is calculated by summing the total number of true positive pixels, which are pixels correctly classified by the model as part of the segmentation mask. This sum represents the intersection between the predicted mask and the ground truth mask. The IOU value is then obtained by dividing this intersection value by the sum of true positive, false negative, and false positive pixels, which represents the union of the two masks.

In the case of a negative class instance, where there is no ground truth mask, the numerator of this fraction will always be zero since there is no intersection between the predicted mask and a non-existent ground truth mask. Therefore, the IOU value for negative class instances will always be zero.

This circumstance not only leads to the decrement of the average IOU score of all training instances, but also renders it impossible to discern the performance of models when confronted with negative class instances. Given that the primary focus of the research lies in investigating the performance of models specifically on negative class instances, alternative evaluation metrics must be utilised.

As a result, the average validation loss was selected as the metric to determine whether to save a model. The rationale behind this choice is that the loss of a model is computed by comparing the ground truth mask with the predicted mask. When considering the negative class instance, where the ground truth mask consists entirely of zeroes, any predicted mask produced by the model will contribute to an increase in the loss. Consequently, this allows for differentiation among models in terms of their performance on unseen data.

3.2 Models

For my research, I employed three distinct versions of a Few-Shot Segmentation (FSS) model. The first variant, denoted as the "base" model, is based on the model developed by Min et al. [8]. To ensure consistency, I utilized the pre-trained weights provided in the official GitHub repository associated with the paper[1]. Additionally, for this model, I utilized fold 3 of the dataset as the validation set during training.

The second model, referred to as the "scratch" model, was trained entirely from scratch using the novel training loop that I designed. This model is exactly the same architeture as the other two models, however all the connections between neurons were randomised. Then the model was trained exclusively using the novel training loop. Similar to the base model, fold 3 was utilized as the validation and test set.

Lastly, I propose the introduction of a third model referred to as the "updated" model. This model is based on the same architecture and weights as the base model proposed by Min et al. However, it undergoes further training for an additional 40 epochs using the novel training loop described earlier. By doing so, we can evaluate whether models that have already converged are capable of being trained to handle negative class instances, despite not being originally trained for this purpose.

Maintaining consistency in the models used for experimentation is crucial when studying the impact of changes in the training process. By ensuring that the models themselves remain unchanged, the effects of the modified training approach are isolated and any confounding factors that could arise from variations in the models are minimised.

This approach provides several benefits. Firstly, it allows for conclusions to be drawn that can be generalized beyond the specific models used in the research. The findings and insights gained from the study can potentially be applied to other models within the FSS domain, as the fundamental characteristics and behavior of the models are kept constant.

Additionally, by keeping the models consistent, any observed differences or improvements in performance can confidently attributed to solely the variations in the training process. This enhances the validity and reliability of the research findings, allowing for a more accurate evaluation of the impact of the modified training approach on the FSS models.

Overall, maintaining the same models while experimenting with changes in the training process is a robust approach that allows for more rigorous analysis and more meaningful conclusions that can be applied to a broader range of models.

3.3 Evaluation

After all the models were trained, they were evaluated. The evaluation consisted of running the models for 1000 episodes of inference. Each model was given one support and query image pair and tasked with segmenting out the class of interest from the support image. As the support images are randomly sampled from a large collection it is unlikely that any of the episodes were duplicated. In addition, this this is how Min et al. perform

[1] https://github.com/juhongm999/hsnet

the evaluation [8]. Half of all episodes were negative class instances, without the class of interest in the query image. The other half were positive class instances which did contain the class of interest. For each inference, several metrics were collected about the inference that allow for calculation of how well each model performed. The most important of which were the IOU to calculate the performance of the model in positive class instances and the amount of false positive pixels to calculate how the model dealt with negative class instances.

Currently, in the eld of FSS, researchers employ hold-one-out cross-validation to obtain an average performance measure for their networks. However, due to time constraints in my research, I made the decision to utilize only the third fold as the test set. This particular fold was consistently used for testing across all models, and none of the models had been exposed to it during training. Therefore, since none of the models have encountered this fold previously, it serves as an e ective means of evaluating the performance across all models. Additionally, by xing the random seed, each network is exposed to the same negative class instances with the same support images. Consequently, it cannot be argued that the di erent networks had relatively "easier" episodes to infer from, as each network underwent training with an identical set of episodes.

Chapter 4

Results

4.1 False Positive Rate

Firstly, it is important to demonstrate that training the model in the new pipeline does not signi cantly alter the false negative rate during positive class instances. Figure 4.1 visually depicts the distribution, indicating some change in this aspect.

To address the rst hypothesis, *The augmented training pipeline will result in a model exhibiting a reduced false positive rate*, we constructed Figure 4.2, which illustrates the distributions of false positive rate per model per class instance. In the updated model, the mean false positive rate was 0.0101 and 0.0107 for the positive and negative class instances, respectively. Moreover, the standard errors of the mean (SEM) were 0.001 and 0.0015, indicating a slightly greater variation in the negative class instances. Conversely, the scratch model exhibited a mean of 0.0144 for positive class instances and 0.0097 for negative class instances, with SEMs of 0.0013 and 0.0017, respectively, indicating slightly higher variation compared to the updated model. Notably, both the updated and scratch models outperformed the base model, which had mean values of 0.0466 and 0.1798 for positive and negative class instances, along with SEMs of 0.0032 and 0.0083. From this we can conclude that the rst hypothesis has been achieved in both positive and negative class instances. For a comprehensive overview of all the values, refer to Table 4.1, where the lowest means of the positive and negative class instances are highlighted in bold.

	Positive Class Instance		**Negative Class Instance**	
Model	*Mean*	*SEM*	*Mean*	*SEM*
Updated	**0.01006**	0.00102	0.01069	0.00154
Scratch	0.01437	0.00130	**0.00967**	0.00165
Base	0.04659	0.00320	0.17978	0.00831

Table 4.1: The mean and SEMs of the false positive rate per model per class instance.

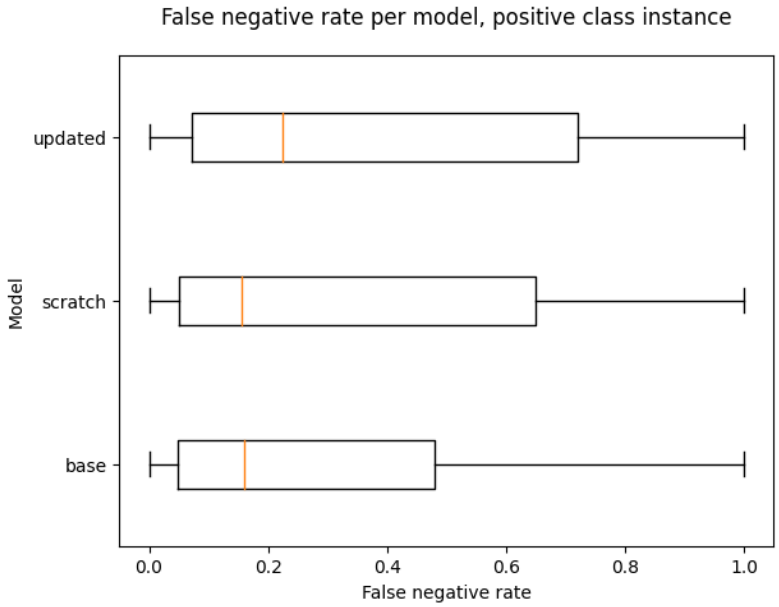

Figure 4.1: Shows a box plot of the false negative rate per model during positive class instances. The box plot was created using matplotlib.plot, with the boxes showing the quartiles and the whiskers extending to the most extreme, non-outlier data points. The box plot shows that some performance is lost regarding the recall of the models but this does not signi cantly change when the model is trained with the novel training pipeline.

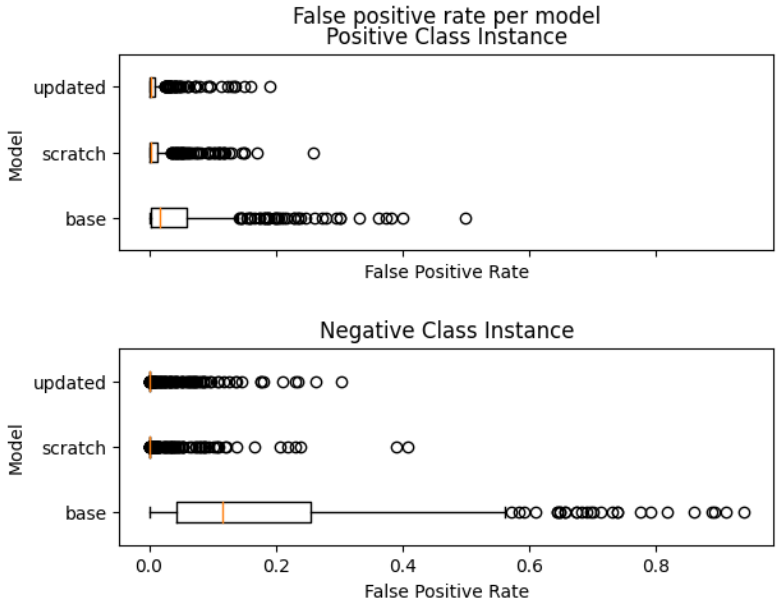

Figure 4.2: These two graphs show the distribution per model of the false positive rate for each class instance. In both class instances the scratch and updated model achieve a lower false positive rate.

4.2 False Positive Rate - Analysis

To illustrate what these results mean, consider the following example. Suppose there are two class instances, one consisting of 50 pixels and the other of 5000 pixels. If the model incorrectly predicts approximately 10% of the pixels for each class instance, it would yield false positive sums of 5 and 500, respectively. Although the model made the same percentage of incorrect predictions for both instances, the second value signi cantly outweighs the rst due to the di erence in class instance sizes. In order to address the impact of image size, metrics incorporating the rates of each type of error were employed, speci cally using the error rates of the models. These metrics account for the size of the ground truth mask by using it as the denominator, resulting in a ratio of errors rather than their absolute values. This approach facilitates meaningful comparisons among the results, as it prevents larger class instances from disproportionately in uencing the analysis of smaller ones.

With the chosen metric rationalised, we now turn our attention to analyzing Figure 4.1. Upon initial inspection, this graph may suggest that the models trained using the novel training loop exhibit inferior performance. However, the situation is not as dire as it may appear. The mean false negative rates for the base, scratch, and updated models are 0.293, 0.343, and 0.384, respectively. Considering the standard deviations of the false negative rates all exceed 0.3, it is no surprise that neither of the newer models demonstrate a signi cantly superior performance compared to the base. Hence, it can be concluded that the introduction of the novel training loop has not led to a substantial change in the

models' performance in terms of the false negative rate.

Nevertheless, a striking disparity emerges in terms of the false positive rate across the models. Firstly, focusing on the top graph of Figure 4.2, it illustrates the false positive rate per model for positive class instances. Notably, this scenario represents the only case on which the base model was trained. Remarkably, the mean false positive rate of the models trained using the novel training loop surpassed the base model's performance by a factor of 5. This compelling evidence suggests a substantial improvement in the false positive rate for positive class instances. Consequently, in line with hypothesis 1, which posits that *The augmented training pipeline will result in a model exhibiting a reduced false positive rate,* we can con dently a rm that this hypothesis holds true for positive class instances.

Lastly, we arrive at the most remarkable ndings of this research. Examining the false positive rate of the models in negative class instances reveals a stark contrast. The two models incorporating the novel training loop exhibited average false positive rates that were on average 18 times lower than those of the baseline model. Furthermore, the SEMs were reduced by a factor of ve. This signi cant improvement provides conclusive evidence in support of the rst hypothesis, demonstrating that the utilization of the novel training loop enables FSS models to e ectively handle negative class instances within query images during inference. Although it would have been ideal for the mean false positive rates to be zero, indicating awless identi cation of negative class instances, the results signify a monumental leap forward in terms of the ecological validity of the models.

4.3 IOU Probability Density

The second hypothesis we seek to investigate is *The augmented training pipeline will result in a model that exhibits improved overall IOU*. To test this hypothesis, the following gures were generated. Firstly, Figure 4.3 displays the results for the background IOUs of the models. The base model exhibited a mean value of 0.858, whereas the scratch and updated models demonstrated mean values of 0.902 and 0.895, respectively. It should be noted that as the IOU can only be calculated in positive class instances each of these graphs contain averages from only positive class instances.

Furthermore, Figure 4.4 presents the histogram and probability density plot illustrating the foreground IOUs of all the models. The base model exhibited a mean value of 0.557, while the updated model achieved a mean IOU of 0.570, and the scratch model obtained 0.592.

It is worth noting that the models trained with the novel training loop displayed a higher frequency of zero scores. Speci cally, the base model had only 10 zero scores, whereas the updated model and scratch model exhibited 67 and 72 zero scores, respectively.

4.4 IOU Performance - Analysis

The metrics discussed to describe the performance of the model are able to capture the di erence between the positive and negative class intances. However, this does not tell us anything about how well the model performed at segmentation. To measure this, we

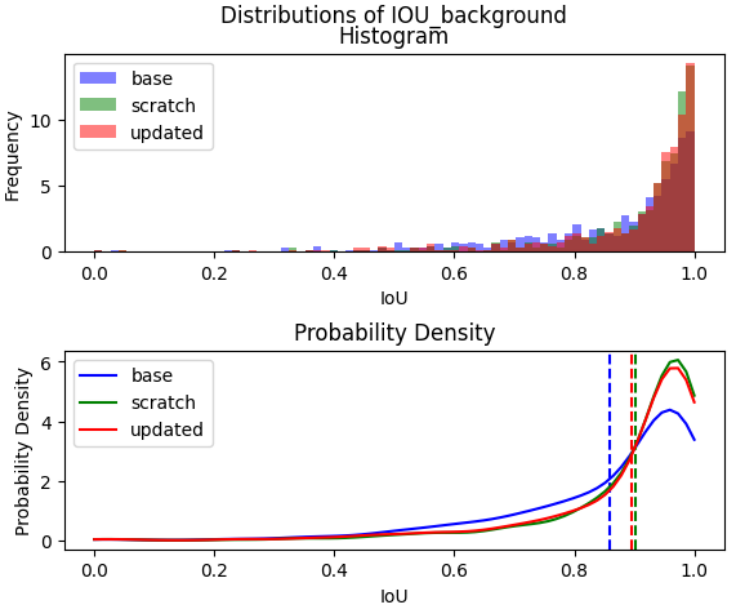

Figure 4.3: Above: a histogram of the background IOU of each inferred episode. Below: a probability density diagram created using scipy's *stats.gaussian_kde* function. The vertical dotted lines show where the mean values lie for each model.

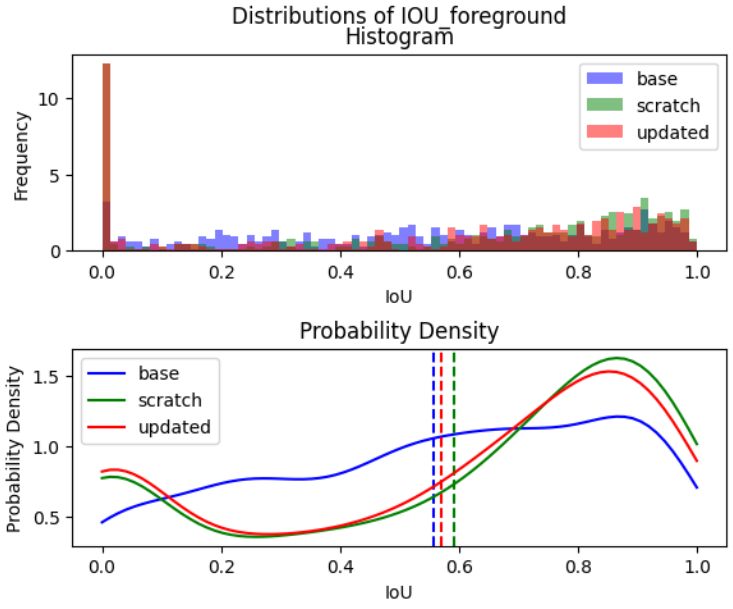

Figure 4.4: Above: a histogram of the foreground IOU of each inferred episode. Below: a probability density diagram created using scipy's *stats.gaussian_kde* function. The vertical dotted lines show where the mean values lie for each model.

can use the IOU of the model. The IOU metric was selected due to its widespread usage in FSS research. It o ers a robust measure of the model's ability to accurately recognize the class of interest while accounting for the size of the ground truth mask. This enables meaningful comparisons across class sizes, as the performance is evaluated relative to the object's segmentation size.

However, as described previously, the IOU is not applicable in negative class instances. This is because the numerator of the IOU, the intersection of the ground truth and the predicted mask, will always be zero because there is no ground truth mask. Therefore, the only two values that the IOU can be in a negative class instance are one, when the predicted mask is empty and the fraction is ($\frac{0}{0}$ = 1) or zero when the predicted mask is any value over zero ($\frac{0}{1}$ = 0). Therefore, all calculations of the IOU are made using positive class instances only.

Figure 4.3 reveals the most substantial increase in IOU was observed in the background category. However, it is important to acknowledge the limitations of this approach in analyzing the models' performance. While achieving a high background IOU is commendable, it should be noted that the majority of the images predominantly consist of background pixels. Consequently, it can be inferred that the model is more likely to attain a higher score due to the prevalence of background pixels. When considering the background and class of interest pixels as separate classes, a signi cant class imbalance in favor of background pixels becomes evident. Thus, it becomes crucial to closely examine the change in foreground IOU to address the second hypothesis.

Nonetheless, the base and scratch models exhibited an increase of nearly ve percent in background IOU, indicating improved performance compared to the base model. To assess the statistical signi cance of these results, a Mann-Whitney U test was conducted. This was done as the three distributions were non Gaussian, yet retained similar characteristics. The resulting p-values for the scratch and updated models were found to be 1.41e-07 and 3.05e-06, respectively, con rming the signi cance of the ndings with respect to the background IOU.

Turning our attention to the foreground IOU, we observed a slightly smaller overall increase in IOU. The scratch model exhibited a 3.5 percent improvement, while the updated model showed a modest increase of only 1.3 percent. It is evident that the scratch model achieved a higher overall performance compared to the updated model. However, it is noteworthy that both models demonstrated improvements over the baseline model. Thus, we can accept hypothesis two regarding the overall improvement in foreground IOU if we are able to reject the null hypothesis.

To this end, a Mann-Whitney U test was employed to examine the signi cance of the improvements in foreground IOU. The resulting p-values for the scratch and updated models were found to be 0.011 and 0.188, respectively. Consequently, the scratch model's p-value falling below the signi cance threshold of 0.05 allows us to conclude that the improvements achieved through the novel training method are statistically signi cant. However, the relatively high p-value for the updated model suggests that its improvements were not statistically signi cant. It is important to note that this does not imply that the model did not improve; rather, it indicates that the extent of improvement was not signi cant enough to reach statistical signi cance.

However, it is crucial not to take these improvements at face value and instead delve deeper into the results. In Figure 4.4, the top histogram demonstrates a noticeable spike at zero value for both the scratch and updated models. This spike occurs when the model fails to detect the class of interest in the query image, resulting in an empty mask being submitted. This phenomenon explains why the error rates appear signi cantly better, while the improvement in IOU may not seem as substantial. The IOU is calculated as an average across all score, including zero scores. This issue is not encountered by the base model since it is exclusively trained on positive class instances and therefore never produces an empty predicted mask. Consequently, it can be argued that while we have tackled the challenge of performing poorly on negative class instances, we have introduced a new problem where the model occasionally fails to detect the class of interest in the query image. Conversely, even with this new problem pulling the average down substantially, the novel models still outperform the baseline model.

Chapter 5

Discussion

The newer models trained with the novel training loop exhibit signi cantly better performance compared to the base model. The improvements achieved in reducing false positive rates demonstrate the e ectiveness of the augmented training pipeline in addressing the challenge of negative class instances. While there is room for further re nement, the advancements made by the newer models represent a substantial leap forward in performance and highlight the potential of the proposed approach.

Translating the false positive error rate of approximately 0.01 to the number of pixels provides a context for evaluating its impact. With an error rate of 0.01, we can estimate that, on average, around 40 by 40 pixels are falsely classi ed as belonging to the class of interest in an image without any actual instances of that class. It is important to note that all of the images that the networks were trained with had a size of 400 by 400 pixels.

When comparing this size to the dimensions of the class instances in the data set, it becomes evident that all class instances are larger than this size threshold. This observation suggests that imposing a minimum segmentation size requirement for a valid solution could potentially enhance the model's performance in negative class instances. By disregarding smaller false positive segments, the model may be able to achieve even higher accuracy in identifying negative class instances. However, should the class of interest be small then having this minimum threshold would mean that the model would be unable to identify small class instances.

In addition, the improvement of the newer models over the base model in terms of false positive rates in positive class instances raises an important question: why do the newer models exhibit better performance, despite the base model being exclusively trained in this domain? One possible explanation is related to the presence of negative class instances in the training data.

By including 50% of the query images without the class of interest during training, the newer models were exposed to instances where the class of interest is absent. This exposure likely resulted in an implicit increase in the con dence threshold of the networks. In other words, the models became more cautious in classifying pixels as belonging to the class of interest, requiring stronger evidence before considering them part of the segmentation mask.

However, this hypothetical change in con dence threshold is not the sole reason for the observed improvement, as manually raising the threshold did not yield the same signi cant improvement. This suggests that the enhanced performance of the newer models is not solely attributed to a change in the decision threshold but also to the complex interactions and adaptations that occur during the training process.

Also,the novel training loop includes a higher proportion of background pixels across all training examples. This is because each training query image either contains a class instance where around a third is the class of interest, or a negative class instance where the whole image is background. Therefore, it makes sense that the model is more prone to false negatives as the model has been trained on a much higher proportion of background pixels.

On the other hand, while there is an overall increase in performance in terms of IOU, it is not as dramatic as the substantial improvement observed in the negative class instance case. This could be attributed to several factors. Firstly, it is possible that the models used in this research have already reached a certain upper limit in terms of their performance. Despite the improvements brought about by the novel training loop, there may be inherent limitations within the model architecture or training methodology that prevent further substantial increases in IOU.

It is worth noting that IOU is widely used as a performance metric in the eld of FSS, but it is not the only measure of model performance. Other evaluation metrics, such as pixel accuracy or mean average precision, could provide further insights into the models' performance and their ability to accurately segment class instances.

In addition, there is a signi cant spike in the number of images with a 0 IOU score for the models trained with the novel training loop. This spike indicates instances where the models were unable to recognize and segment the class of interest in the query images. This issue may have arisen due to the introduction of query images that may not contain instances of the class of interest in the training process.

In contrast, the base model, which was exclusively trained on query images containing the class of interest, does not face this particular challenge. As a result, the base model is more likely to provide a segmentation mask, even if it does not accurately capture the class of interest.

When the features of the class instance in the query image di er signi cantly from the support image, the newer models may struggle to identify and segment the class of interest. Consequently, an empty mask is returned, indicating the inability to nd the class of interest in the query image. While the spike in 0 IOU scores highlights a limitation of the novel training loop, it also re ects the models' ability to identify cases where the class of interest is absent or not well-matched in the query image. This awareness is valuable for achieving greater robustness and accuracy in class instance segmentation tasks, particularly in real-world scenarios where the presence of the class of interest may not be guaranteed.

Further analysis on which classes are most e ected by this problem show that the classes in question are "person", "dining table" and "dog", with "person" having nearly double the errors of the other two. These classes exhibit signi cant diversity in appearance, characteristics, and context, which can pose challenges for accurate class instance segmentation.

For example, "person" instances can di er in clothing, skin color, and pose, while "dining tables" can vary in cloth coverings and shapes. Similarly, "dog" instances encompass diverse breeds, sizes, and colours. These variations pose challenges for accurate segmentation. Moreover, inconsistencies in the dataset regarding the inclusion of certain attributes, like a person's hat, in the segmentation mask further contribute to the challenges of e ectively capturing the class instances.

5.1 Limitations

The research relied on a single dataset, which is widely used in the FSS domain. While this dataset was carefully chosen, incorporating additional datasets can enhance the validity and robustness of the ndings. Assessing the performance of the novel training loop across diverse datasets can provide a more comprehensive understanding of its bene ts and ensure consistent improvements in di erent scenarios. However, currently I see no reason that the ndings found in this research are not applicable to to other datasets and settings.

In addition, a common challenge regarding computer vision networks is computation time for training as the resources required are often quite large making any research into the eld costly. Accordingly, this research was only able to investigate a single network and a single dataset as it took over a month to train the two models.

Another limitation that impacts all FSS models is that the models are only able to work on images that their backbones can extract meaningful features from. In this research HSNet with a backbone of RESNET-101 was used. This backbone was trained on ImageNET and therefore has seen 100,00 images in training. While this may seem like a lot, all these images fall into the same distribution of having been taken by a camera and reduced to 469x387 pixels. Therefore, if we were to use this image encoder on a di erent kind of image, for example one that is much larger or a CT scan/x-ray, it would produce no meaningful features and FSS networks can't segment anything meaningful.

To address the need for a metric that encompasses performance in both positive and negative instances, and allows for meaningful model comparisons, it would be valuable to explore alternative evaluation metrics. As mentioned earlier, metrics like IOU and false positive rate were utilized in the research to assess the performance of positive and negative class instances, respectively. However, it is worth considering additional metrics that capture both aspects simultaneously, potentially through a uni ed evaluation framework.

5.2 Future Research

While the research results demonstrate promising advancements in the eld of FSS, it is important to acknowledge certain limitations. Firstly, the study exclusively focused on one network to evaluate the novel training loop. While consistent results were obtained, it remains uncertain how other models would perform with this approach. Exploring di erent network architectures can provide valuable insights into the generalizability and e ectiveness of the novel training loop across various models. Currently, the research does not

indicate that a different model architecture would change the results of this research, however in order to validate that the results of this research are applicable to the field as a whole, future research should perform the same experiment with different networks.

Another limitation arising from the resource constraints is the inability to explore different variations of the training loop. Specifically, the selected ratio of positive to negative class instances was set at 1:1 to achieve class balance. However, it remains unknown whether altering this ratio could lead to improved performance, allowing the network to effectively handle negative class instances while accurately recognizing positive class instances.

Additionally, the research did not investigate the impact of data augmentation on the performance of the HSNet model. Considering that many other FSS networks incorporate data augmentation techniques during training, exploring the effects of data augmentation on HSNet's performance could have provided valuable insights and comparisons to other FSS models.

Modifying the selection of negative class instances in the novel training loop could offer opportunities for improvement. Incorporating techniques like hard negative mining, could enhance the model's ability to distinguish class features and better generalize to semantically similar but different instances [11]. For instance, if the class of interest is "car" and the query image contains a "bus", exposing the network to such instances could potentially enable it to identify more subtle differences and further refine its segmentation capabilities. This approach would require carefully designing the selection mechanism to ensure the appropriate level of similarity and diversity in negative class instances for effective learning. Exploring these modifications to the negative instance selection process may offer valuable insights into enhancing the performance and discriminative capabilities of FSS models.

Chapter 6

Conclusions

In this research, our primary focus was to address a critical issue prevalent in the field of FSS. We observed a common trend among existing FSS networks where the class of interest was always present in the query image during training. This approach, although yielding satisfactory results during inference when the class of interest was present, posed a significant challenge in scenarios where this was not the case, thus compromising the ecological validity of the network.

Recognizing the need to rectify this limitation, we embarked on an investigation aimed at mitigating the negative impact on network performance during inference. Our key objective was to devise a training strategy that would expose the networks to negative instances, thereby enabling them to accurately handle such cases during inference. To achieve this, we developed a novel training loop that incorporated negative instances during the training process, fostering a more comprehensive understanding of both positive and negative class instances.

Based on our hypotheses, we anticipated that implementing this novel training loop would yield two significant outcomes: firstly, a reduction in the false positive rate of the FSS network HSNet in negative class instance scenarios, and secondly, an improvement in its overall performance during positive class instances. To test these hypotheses, we trained two models: one from scratch, utilizing the novel training loop, and another by applying the novel training loop to a pre-trained HSNet for 40 epochs, further refining its capabilities. These two models were then compared to a baseline pre-trained HSNet

In order to evaluate the effectiveness of the proposed models, we employed various metrics to comprehensively assess their performance. Our findings revealed that both the scratch model and the updated model surpassed the baseline model in terms of handling both negative and positive class instances. These improvements validated the efficacy of our novel training loop in enhancing the network's accuracy and robustness across various scenarios.

Based on our objectives and the outcomes of our research, we can confidently assert that we have successfully achieved both of our intended goals. The models trained using our novel training loop have exhibited superior performance compared to their baseline counterparts, demonstrating enhanced accuracy in both positive and negative class instances.

While these results indicate promising prospects for applying our methodology to other networks, further research should be conducted to validate and corroborate our findings.

If future investigations confirm the effectiveness of this approach on different networks, it is highly recommended that the FSS field as a whole adopts this novel training loop. By embracing this methodology, not only can we expect improved results on benchmark datasets, but we can also address the issue of ecological validity within FSS networks. The inclusion of negative instances during training enables the networks to better handle real-world scenarios where the class of interest may be absent, thus aligning the network's performance with real-world requirements.

In conclusion, our research has demonstrated the successful application of the novel training loop, leading to improved performance in FSS networks for both positive and negative class instances. While further research is necessary to validate these findings across different networks, the potential benefits of adopting this methodology are significant. Transitioning towards the novel training loop has the potential to enhance the overall effectiveness and ecological validity of FSS networks, ultimately advancing the field as a whole.

Bibliography

[1] Mark Everingham, Luc Van Gool, CKI Williams, John Winn, and Andrew Zisserman, *The pascal visual object classes challenge 2012 (voc2012)*, Results, 2012.

[2] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al., *Recent advances in convolutional neural networks*, Pattern recognition **77** (2018), 354{377.

[3] David H Hubel and Torsten N Wiesel, *Receptive fields and functional architecture of monkey striate cortex*, The Journal of physiology **195** (1968), no. 1, 215{243.

[4] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang, *Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation*, arXiv preprint arXiv:2206.09667 (2022).

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM **60** (2017), no. 6, 84{90.

[6] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim, *Adaptive prototype learning and allocation for few-shot segmentation*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8334{8343.

[7] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu, *Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9424{9434.

[8] Juhong Min, Dahyun Kang, and Minsu Cho, *Hypercorrelation squeeze for few-shot segmentation*, Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6941{6952.

[9] Archit Parnami and Minwoo Lee, *Learning from few examples: A summary of approaches to few-shot learning*, arXiv preprint arXiv:2203.04291 (2022).

[10] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia, *Hierarchical dense correlation distillation for few-shot segmentation*,

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23641{23651.

[11] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee, *Stochastic class-based hard example mining for deep metric learning*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7251{7259.

[12] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang, *Singular value ne-tuning: Few-shot segmentation requires few-parameters ne-tuning*, arXiv preprint arXiv:2206.06122 (2022).

[13] Lisa Torrey and Jude Shavlik, *Transfer learning*, Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, 2010, pp. 242{264.

[14] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng, *Panet: Few-shot image semantic segmentation with prototype alignment*, proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9197{9206.

[15] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang, *A survey of transfer learning*, Journal of Big data **3** (2016), no. 1, 1{40.

[16] Qi Zhang, *A novel resnet101 model based on dense dilated convolution for image classi cation*, SN Applied Sciences **4** (2022), 1{13.

[17] Yu Jin Zhang, *A review of recent evaluation methods for image segmentation*, Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat. No. 01EX467), vol. 1, IEEE, 2001, pp. 148{151.