Using StyleGAN for Few-Shot Semantic Segmentation

Amirhossein Baghghalaghdampour

Temple University

CIS 5603 - Artificial Intelligence

Abstract

The current trajectory in various computer vision domains involves the development of increasingly deep models to attain superior performance. Although advancements in hardware capabilities have facilitated the creation of deeper architectures, these models necessitate copious amounts of data for effective training and generalization. Furthermore, the procurement of annotated training data can be extremely expensive and demand specialized knowledge, particularly in applications like medicine.

Addressing this challenge, building a data-efficient model that exhibits robust generalization with a limited set of training samples becomes highly advantageous. The focal point of this project is the semantic segmentation of human facial images, aiming to achieve reasonably well generalization using only three annotated instances for training. As part of this approach, we incorporate intermediate features from the Style GAN generator to augment our training data, hypothesizing that these features encapsulate crucial information about the semantic aspects of the generated images.

Our investigation assesses the quality of segmentation masks on five manually annotated images. The outcomes provide insights into the feasibility and effectiveness of utilizing GANs for few-shot image segmentation.

1. Introduction

Semantic segmentation, a cornerstone in computer vision tasks, involves the classification of each pixel in an image into distinct categories, enabling a nuanced understanding of visual scenes. State-of-the-art segmentation models heavily rely on extensive training datasets for robust performance. In contrast, human learning operates differently, as we can decipher the intricacies of an object by observing a few samples and discerning variations through discriminative features.

In this project, we draw inspiration from human learning to address few-shot segmentation, a scenario where only a handful of annotated samples are available. The crux of this approach lies in answering the question of whether leveraging the features embedded in the intermediate layers of generative adversarial networks (GANs) would provide enough prior knowledge for downstream tasks. It can be argued that these intermediate layers progressively assemble different parts of an image, making them indispensable as building blocks of an image.

Several works have hypothesized that these intermediate features can provide more expressive data samples as they contain semantically important clues. We will use this augmentation to guide a separate segmentation network to form a few annotated data samples to evaluate the effectiveness of these features. However, a challenge arises during testing, as obtaining these features for arbitrary images becomes impractical.

To surmount this issue, we employ GAN inversion—a process where backpropagation optimizes the input random vector to generate a desired image. This necessitates a sufficiently large and diverse training dataset for the generative model to learn varied object variations. Nevertheless, this quality is achievable in an unsupervised manner so that there's no need for annotated data.

2. Related Works

2.1 Few-Shot Semantic Segmentation

Few-shot semantic segmentation has gained significant attention due to its ability to perform dense pixel labeling on new classes with only a few support samples. This scenario is particularly challenging as it requires models to generalize to unseen classes with minimal training data. To address this, various methodologies have been proposed in the literature. Prior work typically draws from different approaches to few-shot image classification and semantic segmentation.

Furthermore, few-shot segmentation has been introduced to reduce the dependency on expensive annotations and to segment images given only a few training samples. Some methodologies propose novel frameworks based on meta-learning to tackle the few-shot semantic segmentation task (Tian et al., 2020). Others have introduced effective multi-similarity hyperrelation networks to address the few-shot semantic segmentation problem (Shi et al., 2022). Moreover, there are approaches that reformulate few-shot segmentation as a semantic reconstruction problem, converting base class features into a

series of basis vectors that span a class-level semantic space for novel class reconstruction (Liu et al., 2021).
In addition, there are methodologies that aim to generalize few-shot semantic segmentation to simultaneously recognize novel categories with very few examples as well as base categories with sufficient examples (Tian et al., 2020). Some approaches also incorporate attention-based multi-context guiding for few-shot semantic segmentation (Hu et al., 2019). Furthermore, there are techniques that utilize cross-reference networks to better find co-occurrent objects in images, thus aiding the few-shot segmentation task (Liu et al., 2020).

2.2 Generative Models in Few-shot Learning

Generative models have been increasingly utilized in few-shot learning tasks, particularly in the context of semantic segmentation and object recognition. The use of generative models for few-shot learning has been explored in various domains, including medical imaging, computer vision, and signal recognition. For instance, recent work by Nontawat Tritrong, et al. (2023) has focused on repurposing Generative Adversarial Networks (GANs) for one-shot semantic part segmentation, demonstrating the potential of generative models in addressing the challenges of few-shot learning in the context of semantic segmentation. (My project evaluates the approaches proposed in this work.) Similarly, the work by Kaixin Wang et al. (2019) on PANet highlights the use of generative models for few-shot image semantic segmentation with prototype alignment, showcasing the effectiveness of generative models in addressing the limited data availability in few-shot learning scenarios.

Furthermore, the study by Mehdi Rezagholiradeh et al. (2019) on Reg-gan demonstrates the application of generative adversarial networks for semi-supervised learning, indicating the potential of generative models in leveraging unlabeled data to improve few-shot learning performance. Additionally, the work by Nanqing Dong et al. (2019) on few-shot semantic segmentation with prototype learning, and Yongfei Liu et al. (2021) on part-aware prototype network for few-shot semantic segmentation, further emphasize the significance of generative models in addressing the challenges of limited data availability in few-shot learning tasks, particularly in the domain of semantic segmentation.

Moreover, the research by Mennatullah Siam et al. on one-shot weakly supervised video object segmentation, and the work by Tero Karras et al. (2017) on analyzing and improving the image quality of stylegan, further highlight the diverse applications of generative models in few-shot learning, ranging from video object segmentation to image quality analysis.

3. Methods

3.1 Dataset

In this experiment, we are interested in generating part-segmented masks for human faces, outlining different elements of people's faces. More specifically, we have seven classes constituting various parts of the human face: Skin, hair, eyebrows, nose, mouth, and ear.

Our aim is to test the hypothesis of whether the intermediate features of Generative Adversarial Networks provide enough prior knowledge and guidance for a segmentation network to generalize well on unforeseen images.
In this approach, we use the pre-trained generator of Style GAN v2 on the Flickr Faces HQ Dataset (FFHQ) with a size of 256x256 for more efficiency during the experiment.

We take this pre-trained generator and generate 10 random vectors to be fed into the model to, subsequently, get 10 different images of human faces. During this forward pass, we also keep track of all the generated intermediate features of the model. We concatenate all these features and save them along with the generated images. Figure 1 represents the generated images subsequently used for training the segmentation network.

Figure 1

As the next step in preparing a training dataset, we need to provide annotations for at least the three generated images to supervise a separate segmentation network. Therefore, we manually annotate different parts of these three images, as demonstrated in Figure 2.

Figure 2

3.2 Segmentation Model

After generating our training images, extracting features of the intermediate layers of the Generator, and annotating them, we use this dataset of three images to train a small network for semantic segmentation. The output of the network is seven masks denoting pixel-level probabilities of each class.

The network has a basic architecture designed solely to test the hypothesis. It consists of four intermediate convolutional layers, each with channel sizes of 128, 64, 64, and 32, respectively. Moreover, the second-to-last layer uses atrous convolution with a dilation rate of 2 to provide a wider field of view for the network. Figure 3 demonstrates the basic few-shot segmentation model.

```
class FewShotCNN(nn.Module):
    def __init__(self, in_ch, n_class):
        super().__init__()

        dilations = [1, 2, 1, 2, 1]
        channels = [128, 64, 64, 32]

        channels = [in_ch] + channels + [n_class]

        layers = []
        for d, c_in, c_out in zip(dilations, channels[:-1], channels[1:]):
            layers.append(nn.LeakyReLU(0.2, inplace=True))
        self.layers = nn.Sequential(*layers[:-1])

    def forward(self, x):
        return self.layers(x)
```

Figure 3

4. Results and Evaluations

In order to train the network, we use the Adam optimizer with an initial learning rate of 1e-3, decreasing it by a factor of 10 after 50 epochs. To regularize the weights of the network and avoid overfitting, we apply a regularization factor of 1e-3 to the loss function during training. Finally, we train the network for 100 epochs and keep track of the loss and performance during training.

As for the performance metric, we use Intersection Over Union (IoU), which indicates the overlap between the predicted mask and the ground truth, reporting both class-specific and mean IoU. All the manually annotated images, except for the three images used for training, are considered as the test set and used for computing the IoU. Table 1 shows the results of the 3-shot model's performance.

Table 1

Model	Mean IoU	Background	Skin	Hair	Eye	Eyebrow	Nose	Mouse	Ear
3-Shot	%72.88	%86.3	%80.39	%75.1	%61.01	%62.04	%82.34	%79.04	%56.78

Figures 4 and 5 demonstrate the training loss and validation IoU during the training.

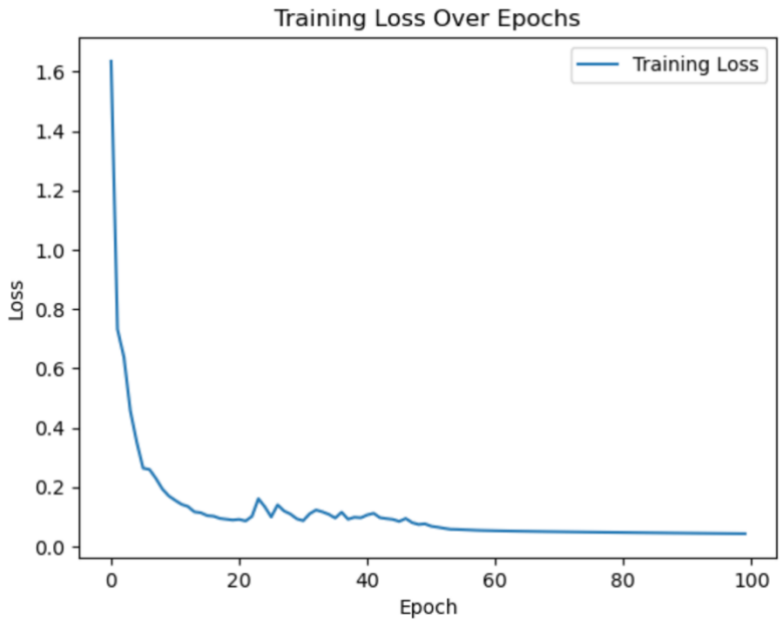

Figure 4

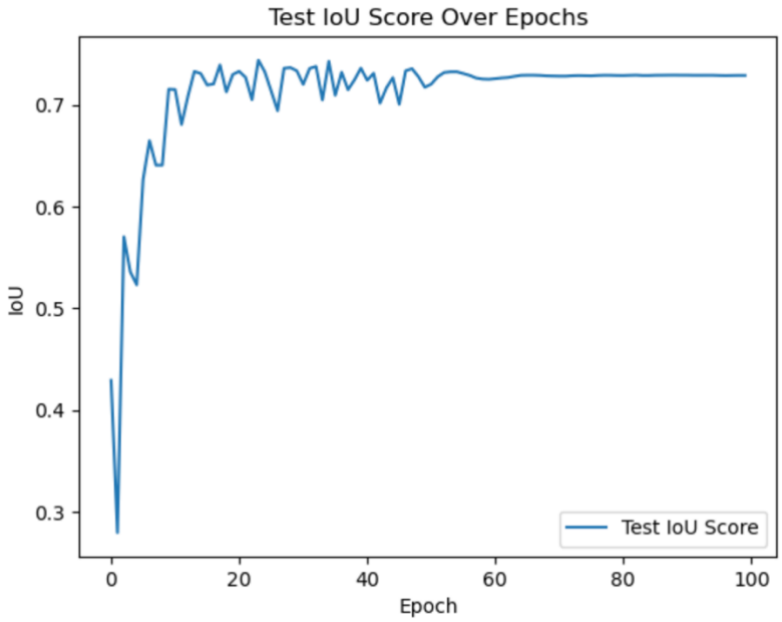

Figure 5

Figure 6 demonstrates the segmentation output for 3 random test images with each color representing a different element in the face.

Figure 6

5. Conclusion

In this exploration, my aim was to delve into various techniques employed for few-shot semantic segmentation as discussed in the relevant literature. Many of these methodologies construct a support set comprising annotated images, serving as prior knowledge for unforeseen classes. However, a novel and emerging concept in this domain is to harness the prior knowledge embedded in the GANs used for image generation to guide downstream tasks in computer vision.

In this project, we successfully trained a basic segmentation network with only three annotated images, utilizing features extracted from the intermediate layers of GANs. While the results are not flawless and exhibit sketchy masks, especially in less common classes like ears, they affirm that these features are discerning enough to act as a prior network for our segmentation model. This approach is applicable to datasets that are sufficiently large for GAN training but lack annotations.

Nevertheless, a limitation of this approach is the necessity to extract GAN features for each input to the segmentation network. An intriguing idea worth exploring is the use of mutual distillation, training another student network with the image as input (rather than GAN features) to learn how to generate a similar output to the original network.

6. References

- Liu, B., Ding, Y., Jiao, J., Ji, X., & Ye, Q. (2021). "Anti-aliasing semantic reconstruction for few-shot semantic segmentation." https://doi.org/10.48550/arxiv.2106.00184)

- Liu, W., Zhang, C., Lin, G., & Liu, F. (2020). "Crnet: cross-reference networks for few-shot segmentation." (https://doi.org/10.1109/cvpr42600.2020.00422)

- Liu, Y., Zhang, X., Zhang, S., & He, X. (2020). "Part-aware prototype network for few-shot semantic segmentation."(https://doi.org/10.1007/978-3-030-58545-7_9)

- Nguyen, K., and Todorovic, S. (2019). "Feature weighting and boosting for few-shot segmentation."(https://doi.org/10.1109/iccv.2019.00071)

- Shi, X., Cui, Z., Zhang, S., Cheng, M., He, L., & Tang, X. (2022). "Multi-similarity based hyperrelation network for few-shot segmentation." Iet Image Processing, 17(1), 204-214. (https://doi.org/10.1049/ipr2.12628)

- Tian, P., Wu, Z., Qi, L., Wang, L., & Shi, Y. (2020). "Differentiable meta-learning model for few-shot semantic segmentation." Proceedings of the Aaai Conference on Artificial Intelligence, 34(07), 12087-12094. (https://doi.org/10.1609/aaai.v34i07.6887)

- Tian, Z., Lai, X., Li, J., Liu, S., Shu, M., Zhao, H., … & Jia, J. (2020). "Generalized few-shot semantic segmentation." (https://doi.org/10.48550/arxiv.2010.05210)

- Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., & Jia, J. (2022). "Prior guided feature enrichment network for few-shot segmentation." Ieee Transactions on Pattern Analysis and Machine Intelligence, 44(2), 1050-1065. (https://doi.org/10.1109/tpami.2020.3013717)

- Adiwardana, D. D. F., Matsukawa, A., & Whang, J. (2016). "Using generative models for semi-supervised learning." In Medical image computing and computer-assisted intervention–MICCAI (Vol. 2016, pp. 106–114).

- Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., & Babenko, A. (2022). "Label-Efficient Semantic Segmentation with Diffusion Models."

- Dong, K., Yang, W., Xu, Z., Huang, L., & Yu, Z. (2021). "ABPNet: Adaptive Background Modeling for Generalized Few Shot Segmentation." In MM (pp. 2271–2280).

- Liu, L., Cao, J., Liu, M., Guo, Y., Chen, Q., & Tan, M. (2020). "Dynamic extension nets for few-shot semantic segmentation." In MM (pp. 1441–1449).

- Rezagholiradeh, M., & Haidar, M. A. (2018). "Reg-gan: Semi-supervised learning based on generative adversarial networks for regression." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2806–2810). IEEE.

- Tritrong, N., Rewatbowornwong, P., & Suwajanakorn, S. (2021). "Repurposing GANs for One-shot Semantic Part Segmentation." arXiv preprint arxiv.2103.04379.

- Shaban, A., Bansal, S., Liu, Z., Essa, I., & Boots, B. (2017). "One-shot learning for semantic segmentation." arXiv preprint arXiv:1709.03410.

- Dong, N., & Xing, E. P. (2018). "Few-shot semantic segmentation with prototype learning." In BMVC (Vol. 3).

- Liu, Y., Zhang, X., Zhang, S., & He, X. (2020). "Part-aware prototype network for few-shot semantic segmentation." arXiv preprint arXiv:2007.06309.

- Wang, K., Liew, J. H., Zou, Y., Zhou, D., & Feng, J. (2019). "Panet: Few-shot image semantic segmentation with prototype alignment." In Proceedings of the IEEE International Conference on Computer Vision (pp. 9197–9206).

- Siam, M., Doraiswamy, N., Oreshkin, B. N., Yao, H., & Jagersand, M. (2019). "One-shot weakly supervised video object segmentation." arXiv preprint arXiv:1912.08936.

- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). "Analyzing and improving the image quality of StyleGAN." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8110–8119).