TEACHING GAN TO GENERATE PER-PIXEL ANNOTATION CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a method for joint image and per-pixel annotation synthesis with GAN. We demonstrate that GAN has good high-level representation of target data that can be easily projected to semantic segmentation masks. This method can be used to create a training dataset for teaching separate semantic segmentation network. Our experiments show that such segmentation network successfully generalizes on real data. Additionally, the method outperforms supervised training when the number of training samples is small, and works on variety of different scenes and classes. The source code of the proposed method will be publicly available.

1 INTRODUCTION

Generative Adversarial Networks (GANs) introduced in Goodfellow et al. (2014) have attracted much attention recently. GANs achieve state-of-the-art results among generative models and have many applications, such as image-to-image translation (Isola et al., 2016; Zhu et al., 2017a; Huang et al., 2018b; Liu et al., 2019), super-resolution (Wang et al., 2018; Ledig et al., 2017), colorization (Isola et al., 2016; Nazeri et al., 2018), texture synthesis (Li & Wand, 2016; Xian et al., 2018), etc.

GAN can generate high resolution images from latent vector with high-level information. This means that intermediate features from GAN also contain high-level information such as position of the objects in the scene and their boundaries. Thus, natural question arises whether it is possible to project features to semantic segmentation mask and generate images along with per-pixel annotation. To tackle this question, we have conducted several experiments which show that the answer to this question is positive.

Our contributions are as follows:

- We propose method for joint image and per-pixel annotation synthesis.
- We show that separate semantic segmentation network trained on synthetic dataset generalizes on real images.
- We show that our method outperforms regular supervised training when number of annotated images is small.

2 RELATED WORK

2.1 GANS

Generative Adversarial Networks usually consist of two networks: Generator network and Discriminator network. Generator creates an image from the noise and Discriminator is trained to distinguish real and generated images. The Generator is trained to fool the Discriminator. After training procedure, the Generator should produce the images that are indistinguishable from real ones.

Since GANs were introduced in 2014 (Goodfellow et al., 2014) a lot of works improving the performance and quality of GANs appeared. This includes such works as WGAN-GP (Gulrajani et al.,

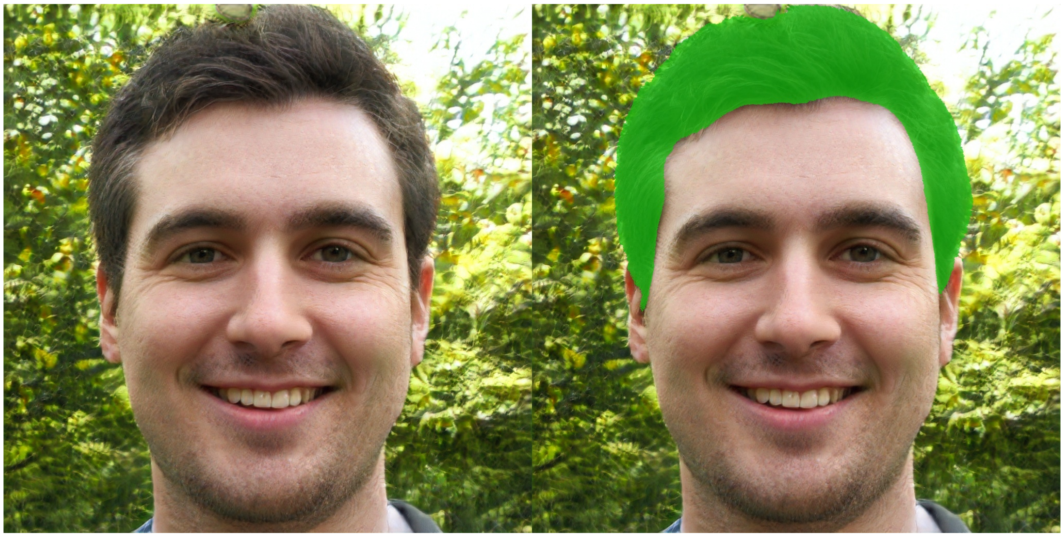


Figure 1: Example of generated image from StyleGAN-FFHQ and corresponding generated annotation for hair segmentation.

2017), spectral normalization (Miyato et al., 2018), ProGAN (Karras et al., 2017), StyleGAN (Karras et al., 2018) and others. Other studies propose new losses (Jolicoeur-Martineau, 2018; Mescheder et al., 2018), architectures and ways to incorporate the conditional information (Miyato & Koyama, 2018; Odena et al., 2017). Despite practical advances, the training dynamics of GANs are still not completely understood.

Pix2PixHD (Wang et al.) and SPADE (Park et al., 2019) achieve impressive results on creating high-resolution images from semantic segmentation masks. CycleGAN (Zhu et al., 2017b), MUNIT (Huang et al., 2018a), FUNIT (Liu et al., 2019) are image-to-image translation methods based on GANs.

There are several works that propose ways for studying and manipulating internal GAN features. For example, in GAN Dissection (Bau et al., 2018) authors present an analytic framework to visualize and understand GANs at the unit-, object-, and scene-level. The show that GANs lear internal neurons that match meaningful concepts. In Brock et al. (2016) authors introduce the Neural Photo Editor, the interface for exploring the learned latent space of generative models and making specific semantic changes to natural images.

2.2 Semantic segmentation

As semantic segmentation is a research topic of interest at this time, many methods were proposed in recent years. This includes Fully Convolutional Network (FCN) (Long et al., 2015), U-Net (Ronneberger et al., 2015), DeepLabs (Chen et al., 2014; 2017a;b; 2018) and others. DeepLabV3+ (Chen et al., 2018) achieves state-of-the-art results on popular benchmarks. Thus, we use this method for our baseline segmentation network.

3 Overview of the proposed method

The main idea of our method is addition of a light-weight decoder to already trained GAN. The decoder is trained to generate per-pixel annotation for the image which is generated by GAN. To train the decoder, several images are generated by GAN and manually annotated. Then decoder is trained with standard back-propagation. We show that only a few images are required in order to train the decoder due to its light-weight nature. Modified network is then used to generate a large dataset of images together with annotation. Separate segmentation network is then trained on this synhtetic dataset.

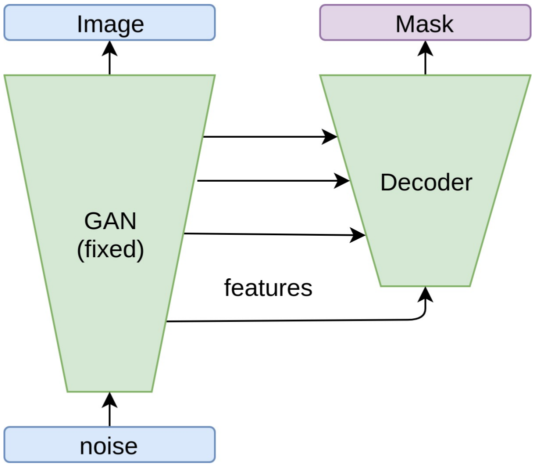


Figure 2: Schematic network architecture.

3.1 Baseline

We use StyleGAN (Karras et al., 2018) as our baseline image generation method and DeepLabV3+ (Chen et al., 2018) as baseline image segmentation method.

3.2 Training decoder

At this step we train Decoder for joint image and semantic segmentation masks synthesis (see Figure 1). Decoder accepts features from GAN and outputs segmentation mask. Decoder is trained in supervised manner on the pairs of input features and corresponding masks (see Figure 2). Such pairs can be collected simply by annotating generated images and storing corresponding intermediate features from GAN. Note that the original GAN remains fixed. We use cross-entropy loss to train Decoder. Interestingly, training takes a few minutes and Decoder successfully learns on a small number of training examples.

3.3 Training segmentation network on synthetic data

After Decoder is trained we can create a large dataset of pairs of GAN-generated images and corresponding masks predicted by Decoder. We train DeepLabV3+ (Chen et al., 2018) on this sythetic dataset. Our experiments show that such network successfully generalizes on real data.

3.4 Finding segmentation mask without training a segmentation network

We also experiment with Image2StyleGAN (Abdal et al., 2019) method which proposes a way to find embedding in StyleGAN for arbitrary real photo. In our case it means that we can find semantic segmentation mask for arbitrary photo without even training a separate segmentation network. Having a trained Decoder this can be done in two steps. Firstly, we find embedding in StyleGAN for specified photo and store intermediate features. Then we find segmentation mask from features using Decoder.

4 Experiments

4.1 Feasibility study

Firstly, we experiment with semantic segmentation of glasses for face photos. We annotate 11 generated by GAN face images with glasses masks. Then we apply our method to train DeepLabV3+ and finally test it on random photos from the web. We assess results visually. Results are shown in Figure 5.

Table 1: Evaluation results on LSUN-cars dataset with two classes (background and car)

Method	pre-trained backbone	accuracy	IoU
DeepLabV3+ (Chen et al., 2018)	-	0.8588	0.6983
Proposed method (ours)	-	0.9787	0.9408
DeepLabV3+ (Chen et al., 2018)	ImageNet	0.9641	0.9049
Proposed method (ours)	ImageNet	**0.9862**	**0.9609**

Table 2: Evaluation results on FFHQ dataset with two classes (background and hair)

Method	pre-trained backbone	accuracy	IoU
DeepLabV3+ (Chen et al., 2018)	ImageNet	0.8831	0.7549
Proposed method (ours)	ImageNet	**0.8967**	**0.8243**

We also demonstrate our model's ability to generalize to a specific part of the face, such as one specific tooth (see Figure 7).

4.2 Evaluation protocol

We test two variants of backbone for DeepLabV3+ (Chen et al., 2018): pretrained on ImageNet and not pretrained. We measure pixel accuracy and intersection-over-union averaged across the classes (mIoU).

4.3 LSUN-cars

We randomly sample subset of 100 images from validation part of LSUN-cars and annotate them with masks of cars. Then we randomly split dataset to train and test parts, 20 samples are used for training and 80 samples for testing. For baseline method, we use these 20 training samples to train DeepLabV3+ (Chen et al., 2018). For our proposed method, we also annotate 20 random images generated by StyleGAN and use them to train a Decoder. Then we generate 10000 synthetic samples and train DeepLabV3+ on them. Both methods are tested on 80 real samples. The results of evaluation are shown in Table 1. Examples of the results are shown in Figure 3.

4.4 FFHQ

We conduct same experiments on FFHQ dataset, but instead of car we use hair segmentation. The results are shown in Table 4. We also experiment with Image2StyleGAN (Abdal et al., 2019) for StyleGAN-FFHQ. An example of embedding and mask is shown in Figure 6.

4.5 LSUN-bedrooms

In this experiment we compare proposed method to baseline for a varying number of training samples to see the dynamics. As there is no semantic segmentation masks for LSUN-bedrooms and annotation is quite tedious, we use segmentation network from GluonCV package (He et al., 2018; Zhang et al., 2019) pretrained on ADE20K (Zhou et al., 2017; 2016) to create an annotation. We use only 13 classes out of 150 of ADE20K that correspond to bedrooms scenes. Results are shown in Figure 4

4.6 Implementation details

We use MXNet Gluon (Chen et al., 2015) for implementation of our algorithm. We convert StyleGAN models to Gluon. For training Decoder we use SGD with momentum 0.9 with the starting learning rate 0.01 and weight decay 1×10^{-3}. Our DeepLabV3+ network has ResNet-50 backbone. For training DeepLabV3+ we use SGD with momentum 0.9 with the starting learning rate 0.01 and weight decay 1×10^{-4}.

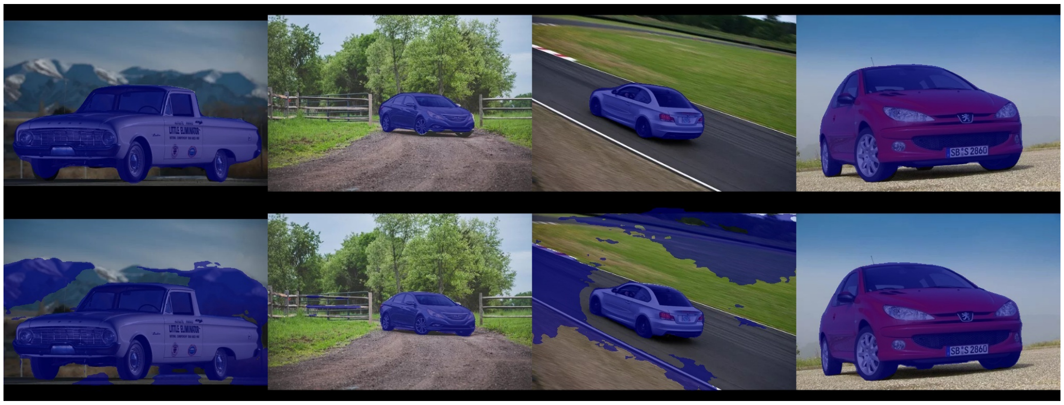
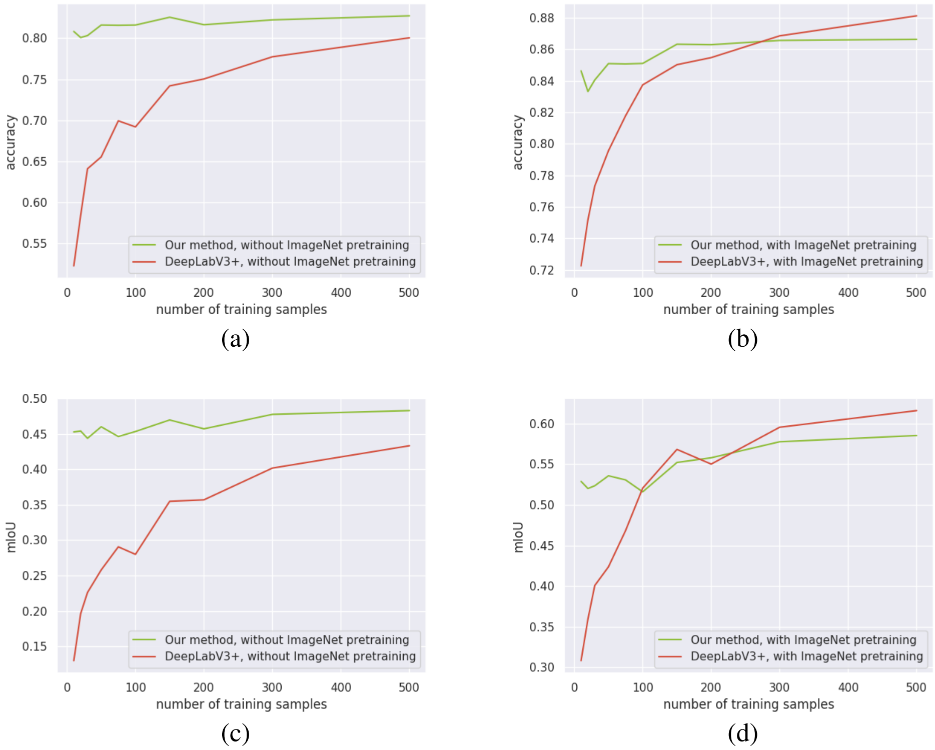


Figure 3: Examples of the results on LSUN-cars. Top line: proposed method. Bottom line: baseline. 20 images were used for training.



(a) (b)

(c) (d)

Figure 4: Comparison of proposed method to baseline on LSUN-bedrooms for a varying number of annotated samples. (a) - without ImageNet pretrained backbone, accuracy. (b) - with ImageNet pretrained backbone, accuracy. (c) - without ImageNet pretrained backbone, mean IoU. (d) - with ImageNet pretrained backbone, mean IoU.

4.7 Results and discussion.

Experiments show that proposed method works well when the number of training samples is small and outperforms regular supervised training by large margin in this case. However, when the number of training examples gets bigger the difference in accuracy decreases (Figure 4 (a), (c)). In the case when the pre-trained on ImageNet backbone is used proposed method begins to work worse (Figure 4 (b), (d)) after some point. This can be explained by the fact that GAN itself has limited

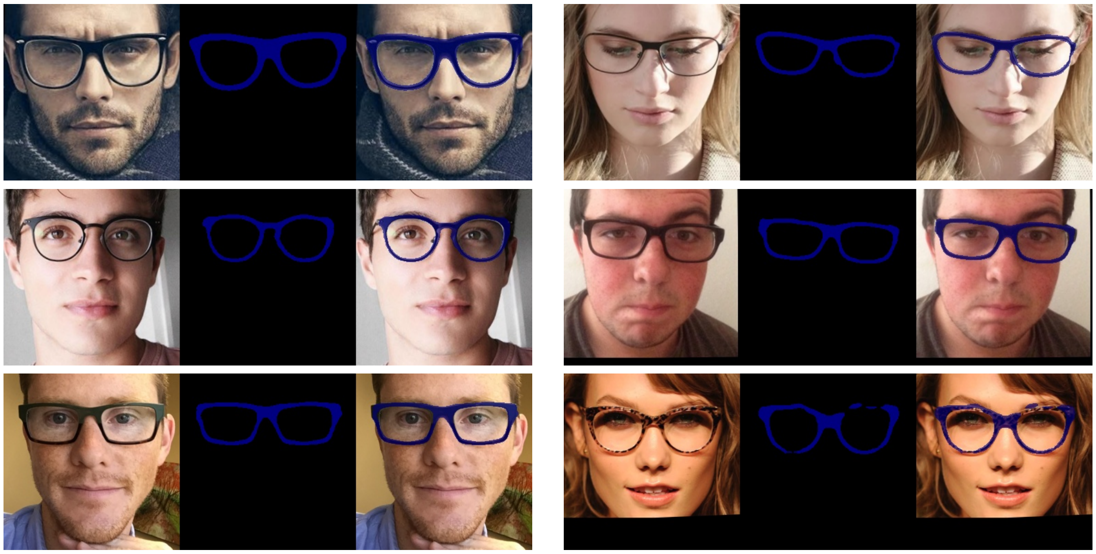


Figure 5: Examples of glasses segmentation on random images from the web. 11 images were annotated for training.

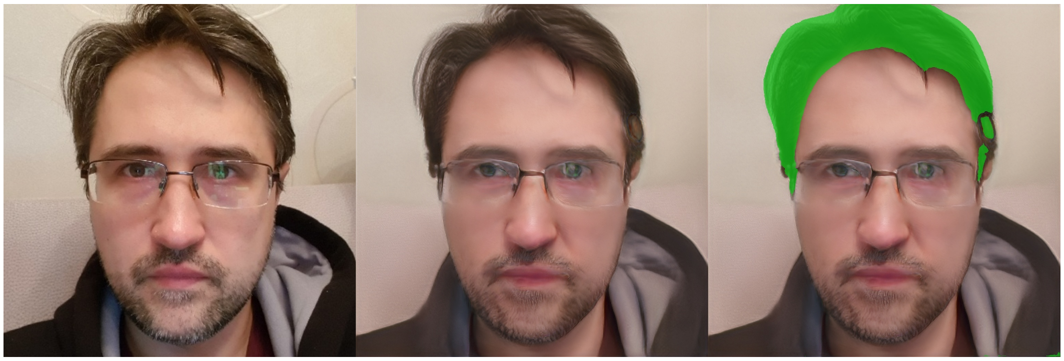


Figure 6: Example of embedding to StyleGAN and predicted mask.

capabilities: the quality of generated images is not perfect, and GAN is often unable to generate some rare objects. Therefore, these rare objects are missed in synthetic dataset. Additionally, the internal representation of GAN from which we project semantic masks may slightly differ from the real high-level representation. For example, the same features are probably used to represent a person's hair and beard. As a result, the quality of hair segmentation deteriorates.

5 Conclusion

We introduce a method for generating images along with semantic segmentation masks using pretrained GAN. It can be used for training separate segmentation network. The study shows that such segmentation network successfully generalizes on real data and performs well on various tasks.

The limitations of our method are associated with two factors. The first one is the lack of diversity of GANs. The second one is the imperfect internal representation of GANs. We assume that the gradual development of image generation algorithms will help to overcome current drawbacks of the proposed method, subsequently speeding up annotation process.

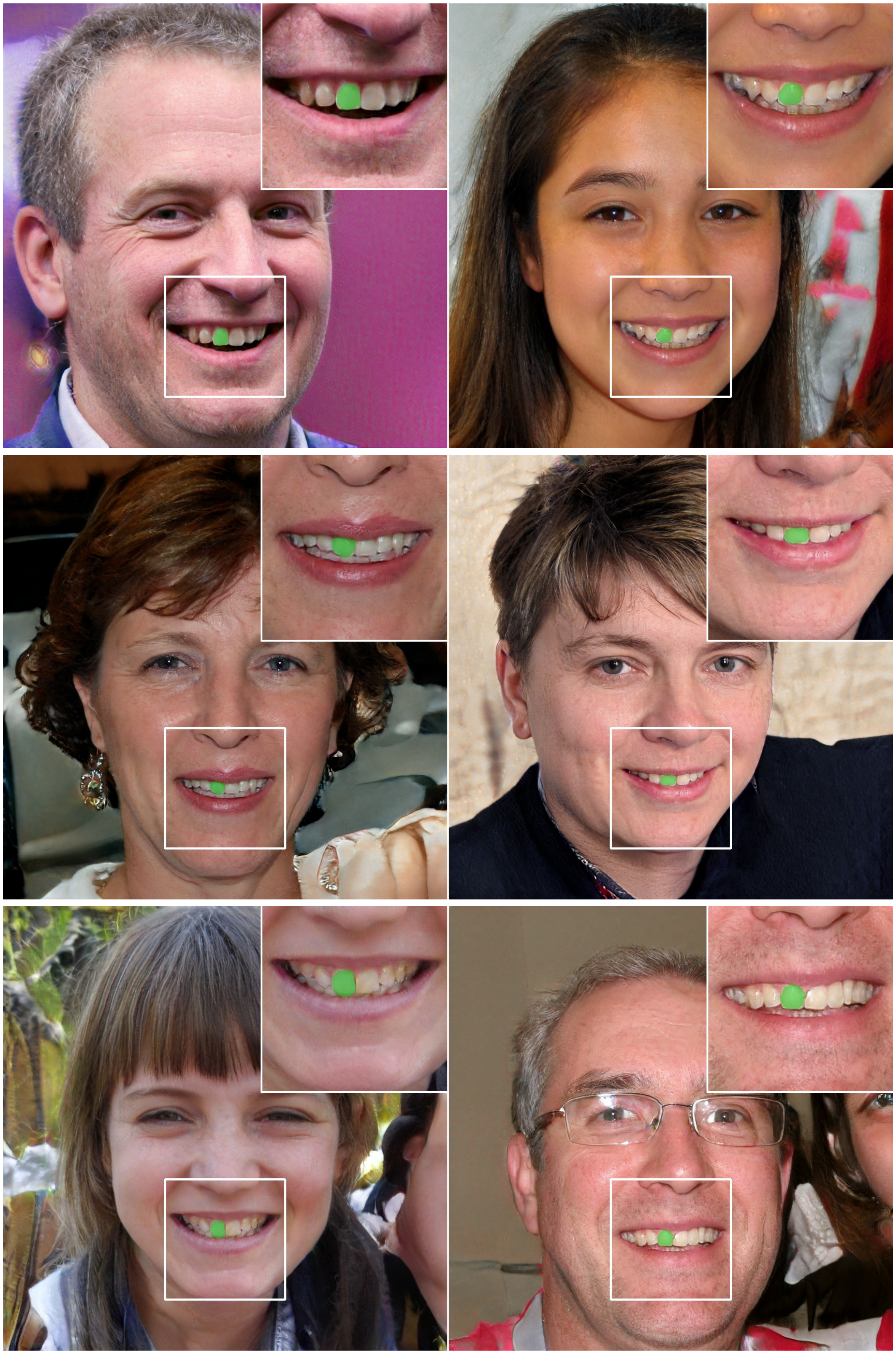


Figure 7: Example of generated image from StyleGAN-FFHQ and corresponding generated segmentation of the front left tooth. Only 5 images were annotated for training. Note that our model expectedly segmented only one of the many equally textured teeth.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? *arXiv preprint arXiv:1904.03189*, 2019.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017a.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017b.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*, 2018.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018a.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018b.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pp. 702–716. Springer, 2016.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *arXiv*, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *International Conference on Articulated Motion and Deformable Objects*, pp. 85–94. Springer, 2018.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2642–2651. JMLR. org, 2017.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *arXiv preprint arXiv:1903.07291*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. pix2pixhd: High-resolution image synthesis and semantic manipulation with conditional gans.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8456–8465, 2018.

Zhi Zhang, Tong He, Hang Zhang, Zhongyuan Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017a.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017b.