Repurposing GANs for One-shot Semantic Part Segmentation

Nontawat Tritrong, et al., CVPR 2021(Oral)
2021/04/19 Vision Study

Contents

- Contributions & Meaning of this paper
- Methods
- Experiments
- Conclusion

Contributions & Meaning of this paper

- By utilizing a pre-trained GAN (StyleGAN-v2 here), semantic part segmentation on ***particular class*** can be achieved given *very few* images with part annotations.

One-shot Annotation

Generalization

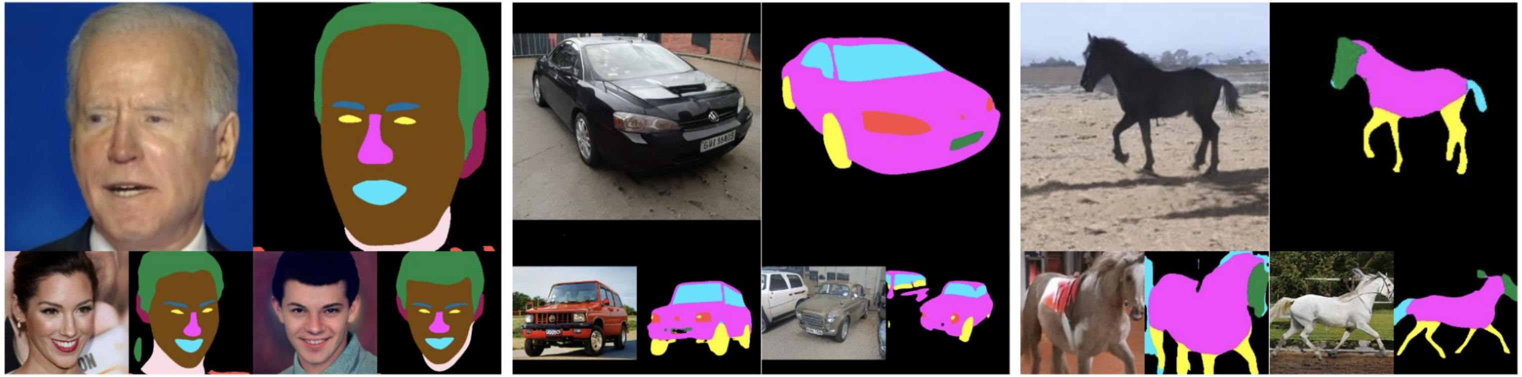


Figure 1: One-shot segmentation results. In each task, our segmentation network is given only one example of part labels.

Contributions & Meaning of this paper

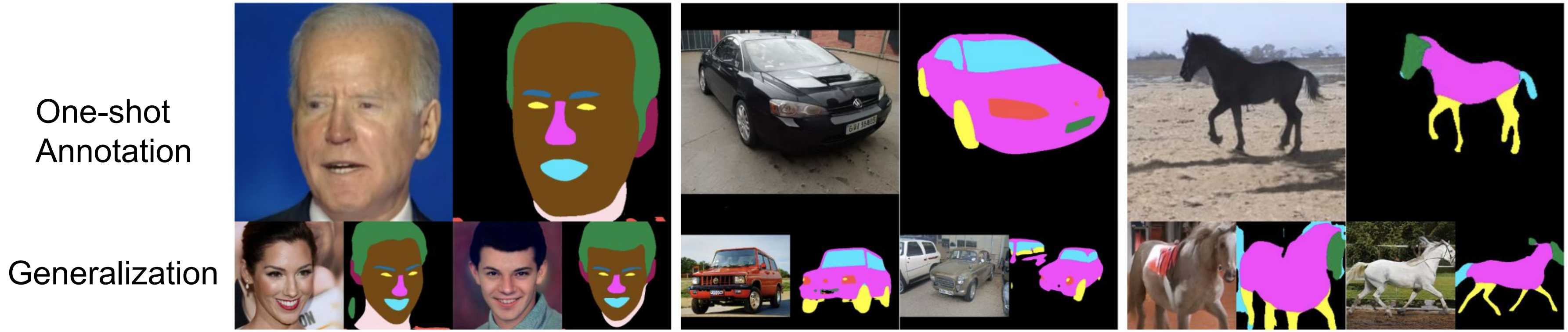
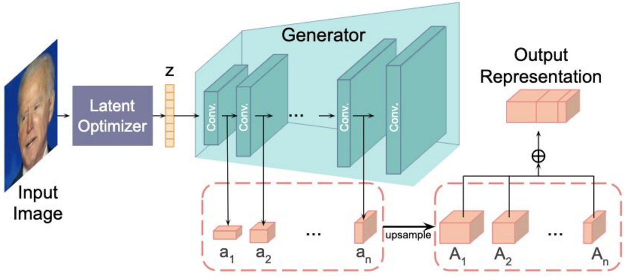
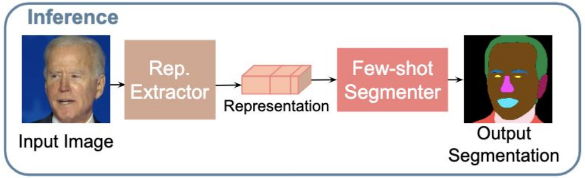


Figure 1: One-shot segmentation results. In each task, our segmentation network is given only one example of part labels.

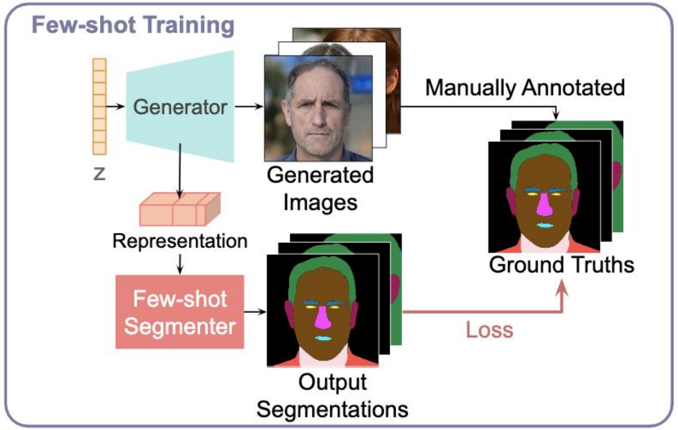
- Proposing a way to leverage the pre-trained GAN for “downstream” task by revealing GAN’s representations are readily discriminative.
- Extending a main idea to real-world scenarios.

Methods

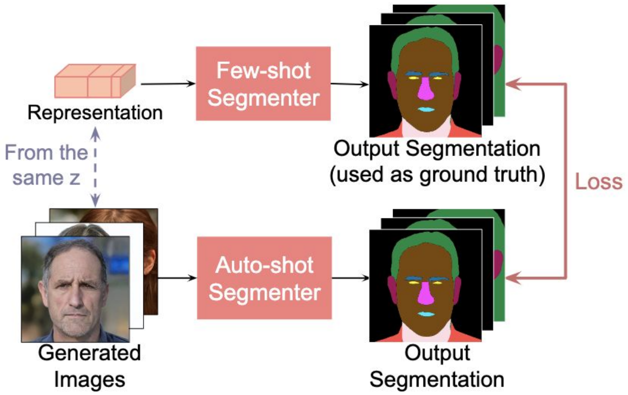




[2] Latent optimizer.
- Following same protocol as StyleGAN-v2.
- Optimization on *W*-space latent code.

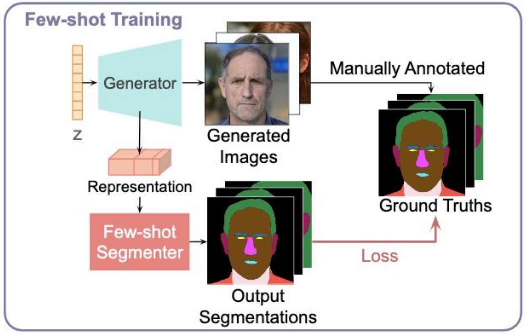


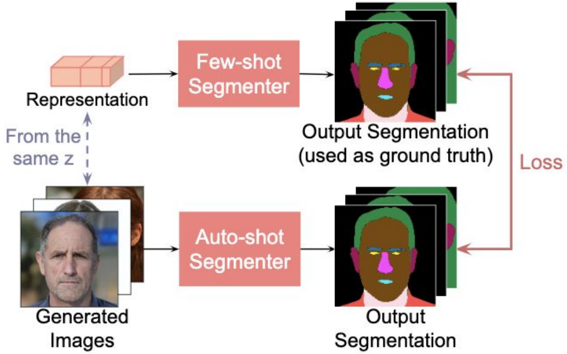
[1] Few-shot training.
- Training with *k*-shot images.
- Using Cross-entropy loss.



[3] Auto-shot Segmenter.
- Weakness; latent optimization is expensive.
- Training another *f*, which can receive an image directly.
- Training the networks using 5,000 samples.

Experiments - Experimental Setup





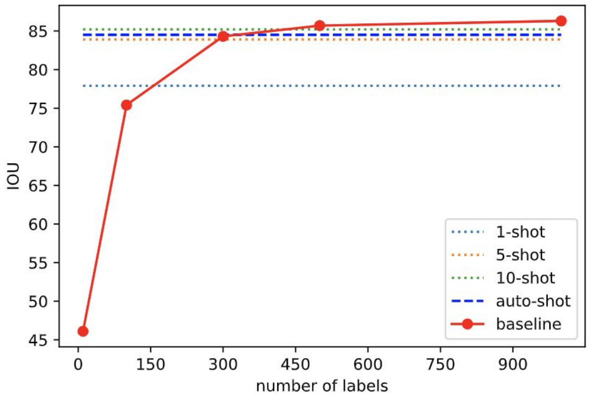
- Detailed training pipeline
  - GAN
    - Feature extraction layers resolutions- All layers except for last layers (4, 8, ...512)
  - Few-shot segmentation network
    - (1 + 8) layers CNN
    - 2-layer MLP
  - Auto-shot segmentation network
    - U-NET
- Datasets
  - Target objects classes: Human face (CelebAMask-HQ), Car, Horse (PASCAL-Part)
  - Few-shot segmentation network training: 5,000 generated images

Experiments - Quantitative Results

Table 1: Weighted IOU scores on few-shot human face segmentation.

Segmentation Network	Shots	3-class	10-class
CNN	1	71.7	77.9
	5	82.1	83.9
	10	**83.5**	**85.2**
MLP	1	75.3	74.1
	5	77.8	79.6
	10	77.2	77.2

<Variants ablation study>



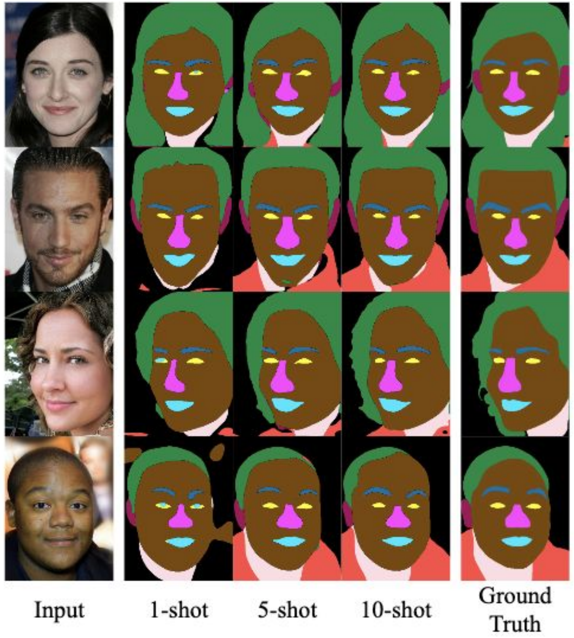
<Comparison with sup.>

Table 2: IOU scores of our 10-shot vs auto-shot segmenters on 10-class face segmentation. The auto-shot segmenter is trained with a dataset generated by the 10-shot segmenter. Both techniques have similar performance, which demonstrates the effectiveness of the dataset generation and auto-shot training process.

Network	Weighted IOU	Eyes	Mouth	Nose	Face	Clothes	Hair	Eyebrows	Ears	Neck	BG
10-shot segmenter	85.2	74.0	84.6	82.9	90.0	23.6	79.2	63.1	27.0	73.6	84.2
Auto-shot segmenter	84.5	75.4	86.5	84.6	90.0	15.5	84.0	68.2	37.3	72.8	84.7

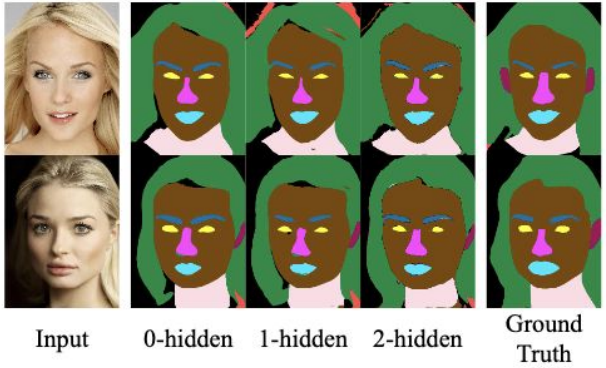
<IOU for each class>

Experiments - More Analysis with Human Face Results



<Qualitative results>

Model	Size	Weighted IOU
MLP	0 hidden layers	74.0
	1 hidden layer	72.2
	2 hidden layers	74.1
CNN	S	73.4
	M	75.2
	L	77.9



<Linear separability of extracted representations>

Experiments - Non-Human Objects Results

<Quantitative comparisons with Sup. baselines >

Table 4: IOU scores on PASCAL-Parts car segmentation.

Model	Body	Plate	Light	Wheel	Window	BG	Average
CNN[48]	73.4	41.7	42.2	66.3	61	67.4	58.7
CNN+CRF[48]	75.4	35.8	36.1	64.3	61.8	68.7	57
Ours (Auto-shot)	75.5	17.8	29.3	57.2	62.4	70.7	52.2
OMPS[62]	86.3	50.5	55.1	75.5	65.2	-	66.5
Ours (Auto-shot) w/o bg	76.4	17.5	29.3	52.5	64.1	-	47.9

<Quantitative comparisons with Sup. baselines >

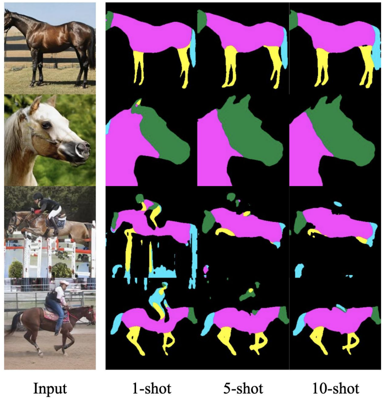
Table 5: IOU scores on PASCAL-Parts horse segmentation. "-" indicates no available result.

Model	Head	Neck	Torso	Neck+Torso	Legs	Tail	BG
Shape+Appearance[53]	47.2	-	-	66.7	38.2	-	-
CNN+CRF[48]	55.0	34.2	52.4	-	46.8	37.2	76.0
Ours (Auto-shot)	50.1	-	-	70.5	49.6	19.9	81.6

Part Annotation

One-shot Segmentation Results

<Auto-shot segmentation network results on *generated outputs*>



<Few-shot segmentation network results on *generated outputs*>

Experiments - How About Other Generative Models?

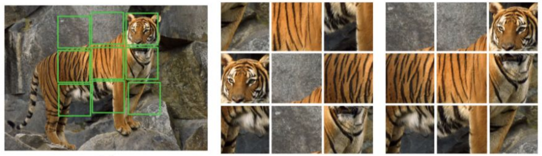
Table G: Weighted IOU scores on 10-shot face segmentation with representation learned from different tasks / networks.

Task / Network	3 class			10 class		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
GANs	71.7	82.1	83.5	77.9	83.9	85.2
VAE	55.1	69.7	72.8	51.6	58.4	65.5
Jigsaw Solving	23.3	46.3	60.0	41.6	54.9	60.4
Colorization	32.1	39.7	51.7	49.1	55.5	66.1
HED	38.2	48.5	60.9	48.9	67.2	70.3
Bilat Filtering	10.9	22.4	49.9	29.2	45.3	54.5

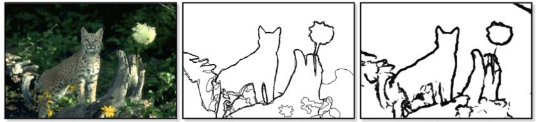
Fail to recognize the boundries as much as GANs do

Baselines

- (A) VAE
  - Enc -dec = ResNet - ResNet
  - Perceptual loss
- (B) Jigsaw



- (C) Image2Image translation
  - Backbone: Pix2Pix
  - (i) Holistically-nested edge detection (HED)

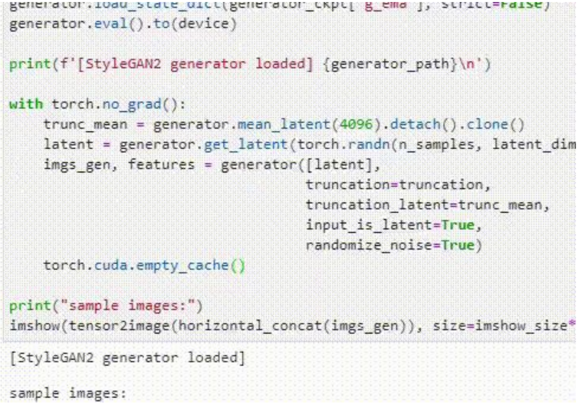


  - (ii) Color: in → out = L channel → AB channels
  - (ii) Bilateral filter
    - Noise-reducing smoothing filter
    - in → out = filtered image → original image

Take Home Messages

- *<u>Well-defined</u>* pretrained GANs provide richful representations for downstream tasks (*e.g.,* semantic part segmentations), meaning generative models bear a potential for practical use cases. (*Previous work focus on manipulating representations for another image generation.*)
- Training highly capable generative models are ***important*** to obtain robust representations.

<Video from official project>



<Unofficial annotation tool>