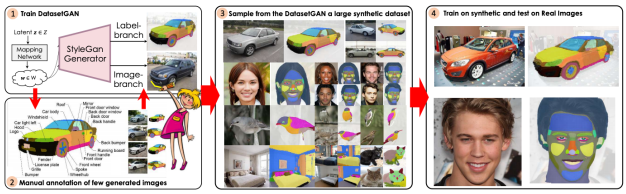

What: Synthesizes image datasets for various tasks using GAN's latent features.

Why: Training deep learning models requires large-scale datasets, but image annotation is laborious and time-consuming.

Annotating images with pixel-wise labels is a laborious task. Past methods try to reduce this annotation cost in semi-supervised or unsupervised manners, such as making pseudo labels, applying different augmentations, or using contrastive learning. This paper introduces a new method for creating image-annotation pairs using latent features of the pretrained StyleGAN.

How: Learn a "Style interpreter" network that outputs annotation labels from features learned by StyleGAN generator.

As demonstrated in the recent researches of image manipulations with GANs, GANs acquire disentangled semantic and geometric representations in their latent space. The key insight of this research is that we can use the latent feature to generate annotation labels for various tasks (*e.g.*, semantic segmentation or keypoint detection) corresponding to the generated images.

We build a model that generates image-annotation pairs from StyleGAN's latent space. Figure 1 shows the overall pipeline of the proposed method. First, a mapping network maps a random seed to StyleGAN's latent feature, and then the StyleGAN generates images from the latent feature. In the same time, the Style Interpreter network estimates the annotation labels of the generated image.

Fig. 2 shows the architecture of the "Style Interpreter" network. It is an ensemble MLP model that takes feature maps from intermediate layers of the StyleGAN's generator network. All feature maps are upsampled and concatenated so that their sizes align with the size of the generated images. The network estimates the annotation labels for each pixel from the stacked feature map. The weights are shared across pixels.

The Style Interpreter network is trained on the images generated with the StyleGAN and corresponding manual annotations. Because StyleGAN sometimes generates noisy images, we computed uncertainty of synthesized images with Jensen-Shannon divergence, and filtered out 10% most uncertain images.

Results: Outperforms transfer learning and semi-supervised learning based methods; Only dozens of manually annotated data suffice to train the Style Interpreter; The quality of synthesized data are important; Application to 3D model generation is shown.

Evaluated on two tasks: part segmentation and keypoint detection. For the part segmentation task, we evaluated the model on 5 categories (Car, face, bird, cat, bedroom). We used StyleGAN models trained on each category. For comparison, we used transfer learning based models and semi-supervised learning based models that are pretrained on other datasets. All of these models have the same architecture with our proposed model.

Figs. 4 and 5 show the synthesized images using the proposed method. Tab. 2 shows the quantitative results. The model trained on the synthesized dataset with our method achieves the best accuracy. Fig. 6 shows the relationship between the accuracy and the number of images used for training. It shows our methods is comparable to fully supervised methods with only 1% of annotation labels. Fig. 7 shows the qualitative results on the test set.

To investigate the effects of the quality of the synthesized images and annotations on the model performance, we performed experiments about sample selection. We tested three options: filtering synthesized images by annotators' confidence, simple active learning, and ensemble active learning methods. Tab. 6 shows that these three manipulations improves the accuracy, suggesting that the model performance actually depends on the quality of the synthesized dataset.

We also evaluated our method on the keypoint detection task. For that, we modified our model to output heatmaps of keypoint locations. As with the experiments of the segmentation task, we compared our model with transfer learning based methods. Tab. 2 shows the experimental results, in which our method achieves the best accuracy. Fig. 8 shows the qualitative results.

In addition, we applied the proposed model to generate realistic 3D models. From the style GAN's latent feature, we predict 3D shapes, textures, and keypoints using the Style Interpreter network. The network was finetuned with the rerendering loss. Fig. 9 shows the examples of generated 3D car models.

Thoughts: The StyleGAN needs to generate images that match the desired contexts, which would limit the application areas.

This method needs pretrained StyleGAN models that produce samples in expected senarios, which means the scope of application is limited to areas where large amounts of data are available. We will need powerfull transfer learing or semi-supervised methods to fine-tune the StyleGAN model on a small amount of images.