# MUSH: Multi-Scale Hierarchical Feature Extraction for Semantic Image Synthesis

Zicong Wang[1,2], Qiang Ren[1,2], Junli Wang[1,2], Chungang Yan[1,2], and
Changjun Jiang[1,2]

[1] Key Laboratory of Embedded System and Service Computing (Tongji University),
Ministry of Education, Shanghai 201804, China
[2] National (Province-Ministry Joint) Collaborative Innovation Center for Financial
Network Security, Tongji University, Shanghai 201804, China.
{wangzicong, rqfzpy, junliwang, yanchungang, cjjiang}@tongji.edu.cn

## A Implementation details

**Semantic feature extraction network architecture specifics.** The network adopts an encoder-decoder framework, which contains convolutional layers, downsampling layers and upsampling layers. After each convolutional layer, the feature map will be activated by a Leaky ReLU layer and be sent to a spectral norm layer and a synchronized batchnorm layer sequentially. The negative slope of Leaky ReLU is 0.2. We use six downsampling layers and six upsampling layers to obtain seven levels of features in total. For the semantic feature extraction network in generator, Each level contains two convolutional layers in encoder and two convolutional layers in decoder. We use 128 convolutional kernels for all convolutional layers. The network in discriminator has a similar structure, but differently, it has only one convolutional layer with 64 convolutional kernels in each level of both encoder and decoder.

**Overall framework specifics.** The generator contains seven MSFA residual blocks. We use an upsampling layer between every two MSFA residual blocks. Each MSFA residual block is also fed by two decoded feature maps from corresponding level of semantic feature extraction network. These feature maps are in the same size as the transformed input noise here. Output feature channel numbers of these seven MSFA residual blocks are: 1024, 1024, 1024, 512, 256, 128, 64 sequentially. We adopt a multi-scale architecture based on PatchGAN as the discriminator. Similar to the discriminator used in SPADE, it uses 4x4 convolutional layers with stride 2, spectral normalization, instance normalization and leaky ReLU layers to obtain discrimination results. However, it no longer takes the concatenation of the segmentation map and the image as input, but take them as input separately. Output feature channel numbers of 4x4 convolutional layers in discriminator are: 64, 128, 256, 512 sequentially and the negative slope of Leaky ReLU is 0.2.

**Additional training details.** We train our model for 200 epochs on Cityscapes and ADE20K, 100 epochs on COCO-Stuff. We use 512x256 image sizes and a batch size of 16 for Cityscapes, 256x256 image sizes and a batch size of 32 for ADE20K and COCO-Stuff.

## B Additional Results

Figure 1, 2 and 3 show additional qualitative comparison between MUSH and other models on ADE20K, Cityscapes and COCO-Stuff respectively. Figure 4 shows additional image synthesis results of GroupDNet-MUSH and INADE-MUSH (GroupDNet and INADE with our method applied to them).

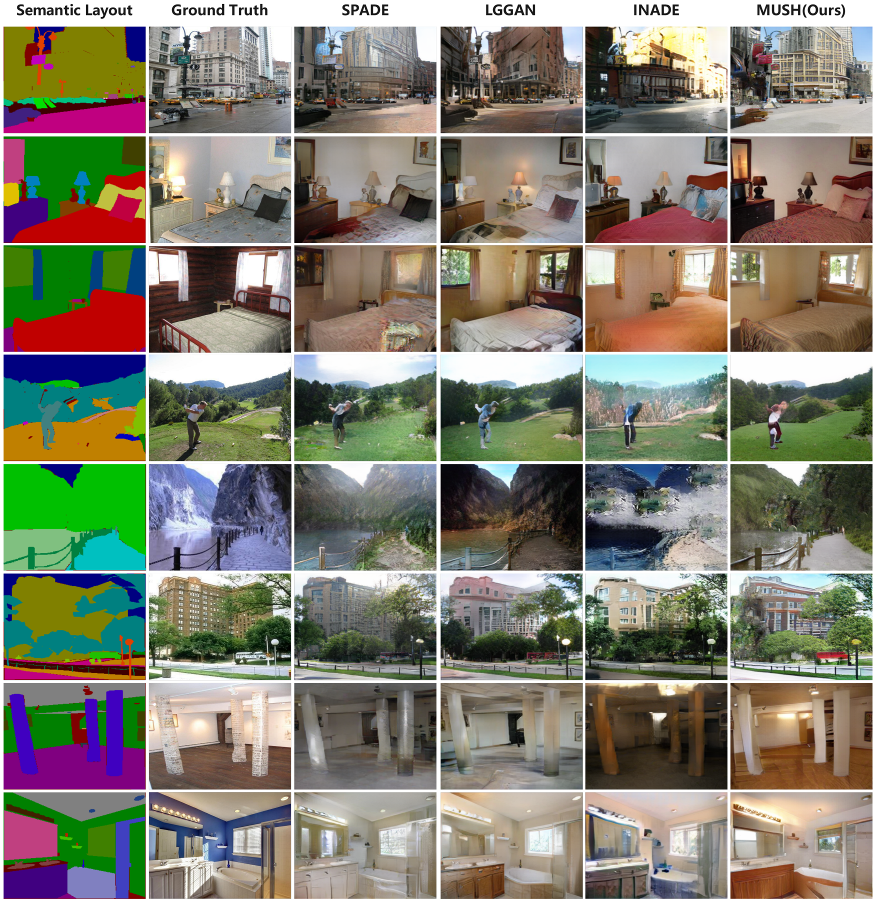

**Fig. 1.** Additional qualitative comparison between MUSH and other models on ADE20K.

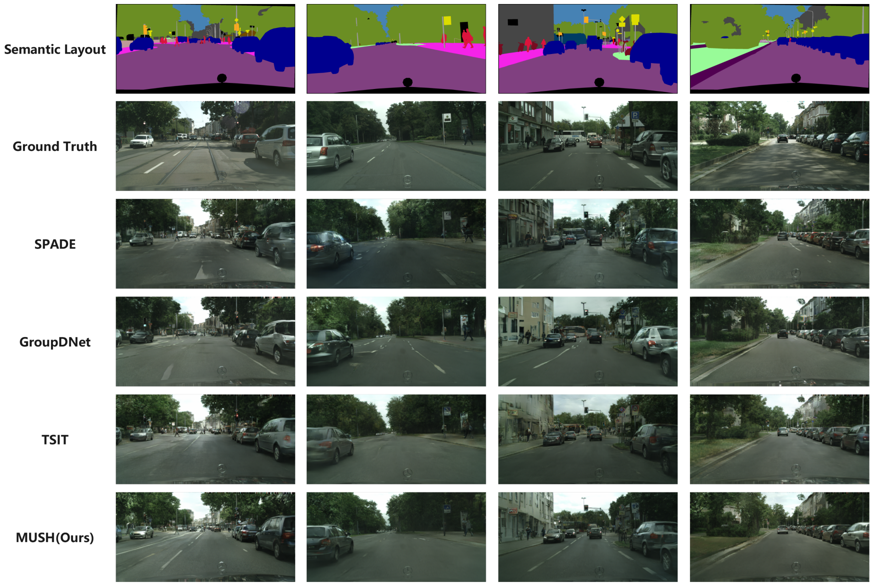

**Fig. 2.** Additional qualitative comparison between MUSH and other models on Cityscapes.

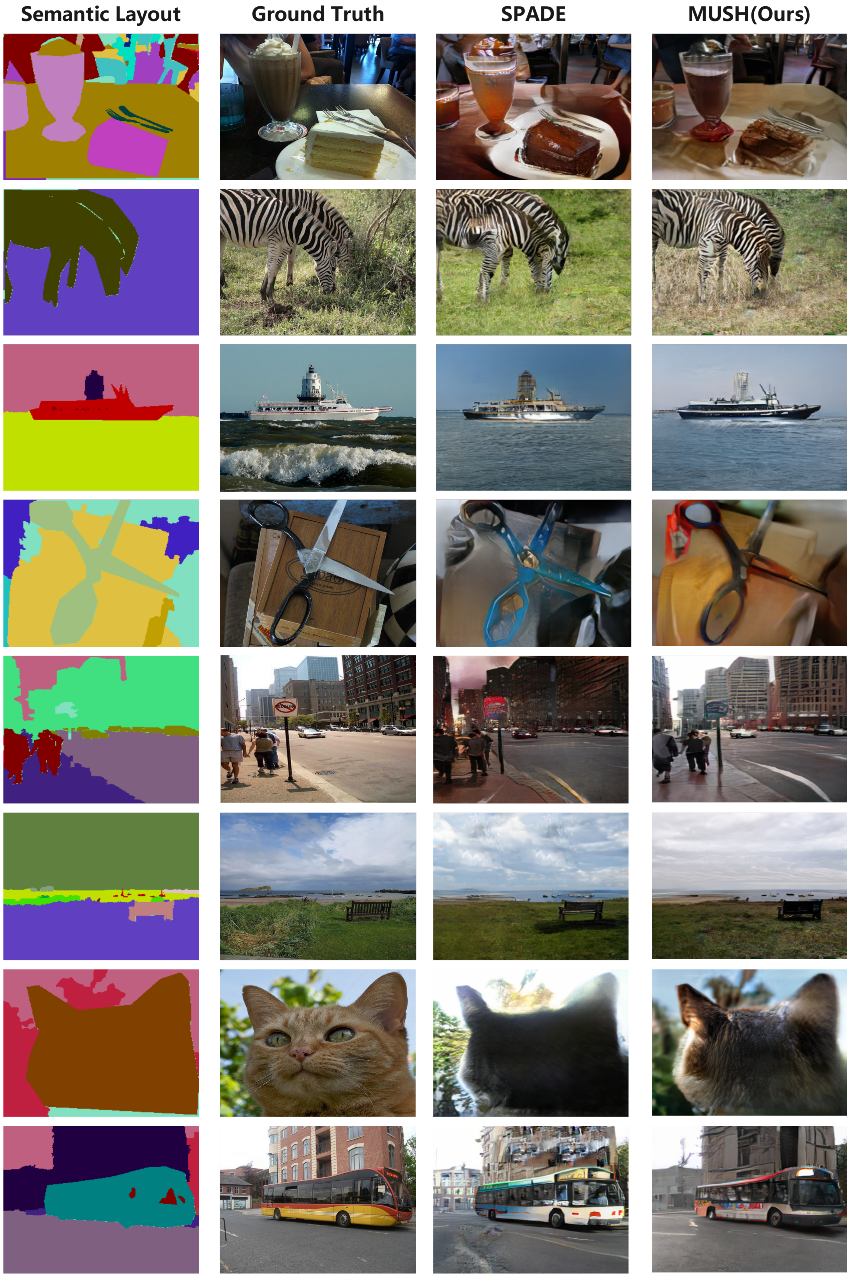

**Fig. 3.** Additional qualitative comparison between MUSH and other models on COCO-Stuff.

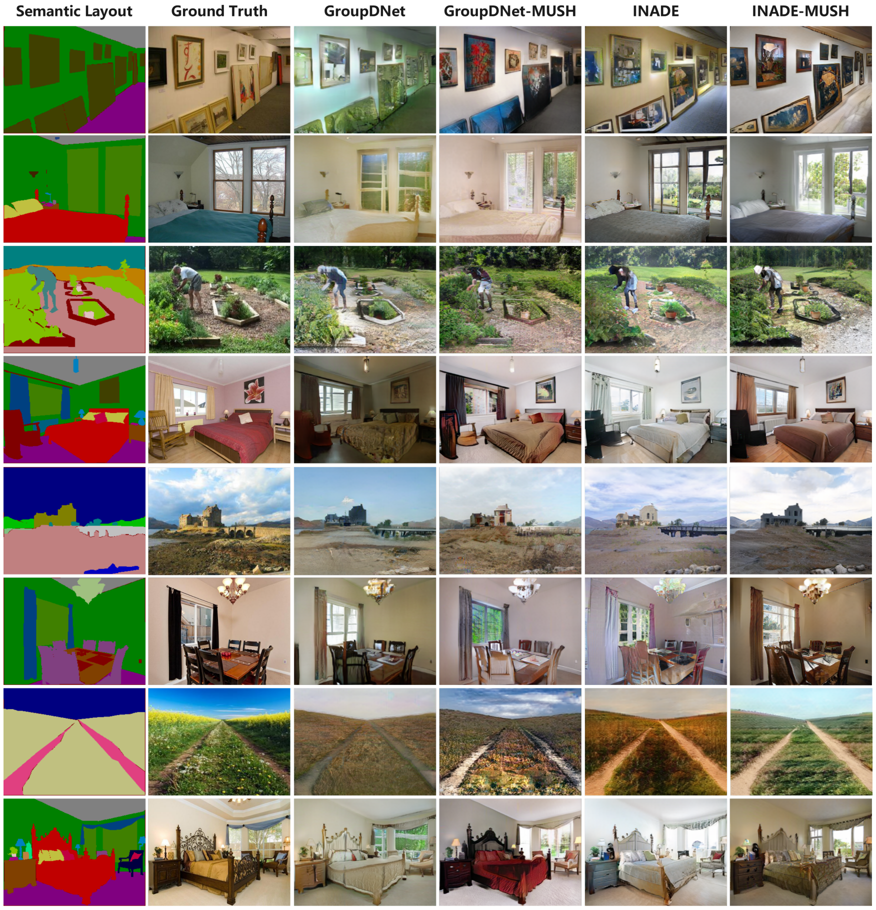

**Fig. 4.** Additional image synthesis results of GroupDNet-MUSH and INADE-MUSH.