Computer Vision - Final Report

第 18 組：林學謙、林政旭、游鎮遠、謝崇志、王清翰

Learning to Predict Layout-to-image Conditional Convolutions for Semantic Image Synthesis: Ablation Track

Xihui Liu
The Chinese University of Hong Kong
xihuiliu@ee.cuhk.edu.hk

Guojun Yin
University of Science and Technology of China
gjyin91@gmail.com

Jing Shao
SenseTime Research
shaojing@sensetime.com

Xiaogang Wang
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

Hongsheng Li
The Chinese University of Hong Kong
hsli@ee.cuhk.edu.hk

1. Introduction

Generative Adversarial Networks (GANs) have excelled in generating realistic faces and simple objects but face challenges in creating photorealistic complex scenes. Semantic image synthesis, which generates images based on semantic layouts, is crucial for controlled image generation and interactive manipulation. At their core, GANs consist of two neural networks: a generator and a discriminator. The generator aims to produce data, such as images or text, while the discriminator's role is to distinguish between genuine data and artificially generated data. What makes GANs exceptionally powerful is their adversarial nature – the generator and discriminator engage in a continuous battle, with the generator striving to create data that is indistinguishable from real data, and the discriminator refining its ability to tell the two apart. This adversarial training process results in the generator becoming increasingly adept at generating realistic data. Existing GAN-based methods typically use label maps as inputs and employ encoder-decoder networks for image generation, but this may not preserve layout information effectively. SPADE enhances this process by using label maps to predict spatially-adaptive affine transformations for modulating activation in normalization layers. However, such feature modulation has limitations in representational power and flexibility. This perspective challenges the conventional use of convolutional layers in image synthesis.

In a generation network, traditional convolutional layers generate fine features in a uniform manner, using the same translation-invariant kernels for all samples and spatial locations. The proposed approach suggests employing distinct convolutional kernels based on the specific semantic labels and layout of each sample. This innovation seeks to enable more precise and context-aware image synthesis, potentially leading to more realistic and diverse results in complex scenes with different objects and elements. In response to the two challenges mentioned earlier, the paper introduces a method to predict spatially-adaptive convolution kernels based on input semantic layouts. This approach offers more explicit and effective control over image generation. To avoid overfitting and excessive GPU memory usage, the authors adopt the concept of depthwise separable convolution, breaking the operation into conditional depthwise and pointwise convolutions. The method provides predict conditional kernel weights using a global-context-aware network, allowing the semantic layout to fine-tune the generation process without significantly increasing network parameters or computational complexity. This innovation improves semantic layout control in image synthesis while remaining computationally efficient.

Current semantic image synthesis methods typically use a multi-scale PatchGAN discriminator, but it struggles to match the generator's capacity. This paper propose a more robust approach, focusing on critical image aspects: high-fidelity details and semantic alignment with the input layout map. The method employs multi-

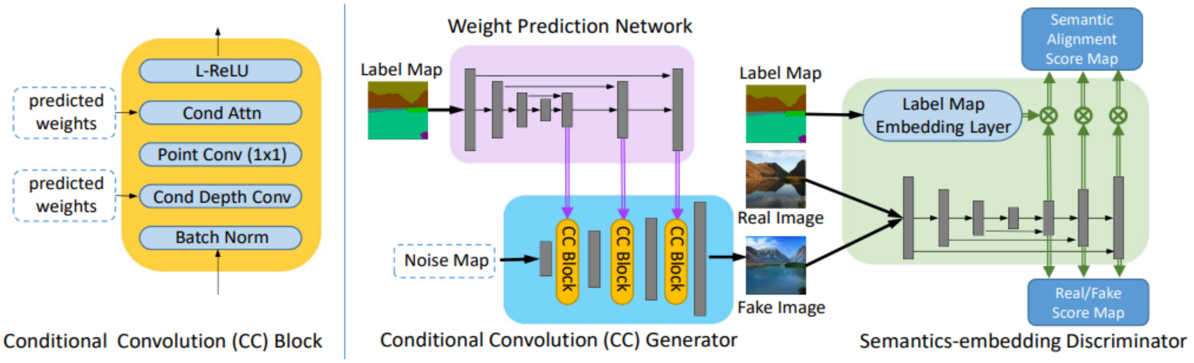

L-ReLU
predicted weights
Cond Attn
Point Conv (1x1)
predicted weights
Cond Depth Conv
Batch Norm
Conditional Convolution (CC) Block
Weight Prediction Network
Label Map
Noise Map
CC Block
CC Block
CC Block
Conditional Convolution (CC) Generator
Label Map
Real Image
Fake Image
Semantic Alignment Score Map
Label Map Embedding Layer
Real/Fake Score Map
Semantics-embedding Discriminator

Figure 1. Design of CC-FPSE

scale feature pyramids to enhance fine details like texture and edges and utilizes patch-based semantic embeddings to improve the spatial alignment between generated images and the input semantic layout. This dual-pronged strategy aims to create more realistic and semantically coherent images in the realm of semantic image synthesis.

This paper presents two key contributions: (1) Introducing a novel method for semantic image synthesis that predicts conditional convolution kernels based on semantic layouts, enabling adaptive generation control based on distinct labels. (2) Proposing a feature pyramid semantics-embedding discriminator, enhancing high-fidelity details and semantic alignment with input layouts.

2. Review

There are three methods about GAN, Pix2pixHD, spatially-adaptive denormalization(SPADE) and Conditional Convolution Feature Pyramid Semantics-Embedding (CC_FPSE).

2.1. Review previous work

The first method, Pix2PixHD, addresses the limitations of traditional conditional GANs, which often result in low resolution and unrealistic images. By introducing a novel adversarial loss and implementing new multi-scale generator and discriminator structures, this method achieves visually satisfying results at 2048 × 1024 resolution. The framework includes interactive visual manipulation features, integrating object instance segmentation information for flexible object manipulations. Additionally, it proposes a method to generate diverse results from the same input, allowing users to interactively edit object appearances. Human opinion studies indicate that this method significantly outperforms traditional approaches in terms of quality and resolution in deep image synthesis and editing. However, when the input is a single semantic label map, Pix2PixHD struggles to capture the distinctive features associated with that semantic, prompting the introduction of a new method, SPADE, to address this issue.

SPADE introduces spatially-adaptive normalization as a simple yet effective layer for synthesizing photorealistic images based on input semantic layouts. Unlike previous methods that directly input semantic layouts into deep networks, SPADE adjusts activations in normalization layers through spatially-adaptive, learned transformations, preserving semantic information more effectively. Experimental results on challenging datasets demonstrate the method's superiority over existing approaches in terms of visual fidelity and alignment with input layouts. Importantly, the model allows users to control both semantic and stylistic aspects. While SPADE resolves the issue of controlling single semantics, its detailed output is suboptimal due to the use of a 3x3 convolutional neural network. It struggles to generate finer details when dealing with larger label maps of the same semantic. To overcome this limitation, a different approach has been proposed to enhance the realism of generated details.

2.2. Summarize key point

CC_FPSE improves the process of synthesizing photorealistic images from semantic layouts. The whole model design is shown in Figure 1. By conditioning convolutional kernels in the generator on different-sized unique semantic labels, the method aims to better exploit the semantic layout for image generation. Additionally, it introduces a feature pyramid semantics-embedding discriminator, more effectively enhancing details and semantic alignment between generated images and input semantic layouts compared to the Pix2PixHD discriminator. The method achieves state-of-the-art results

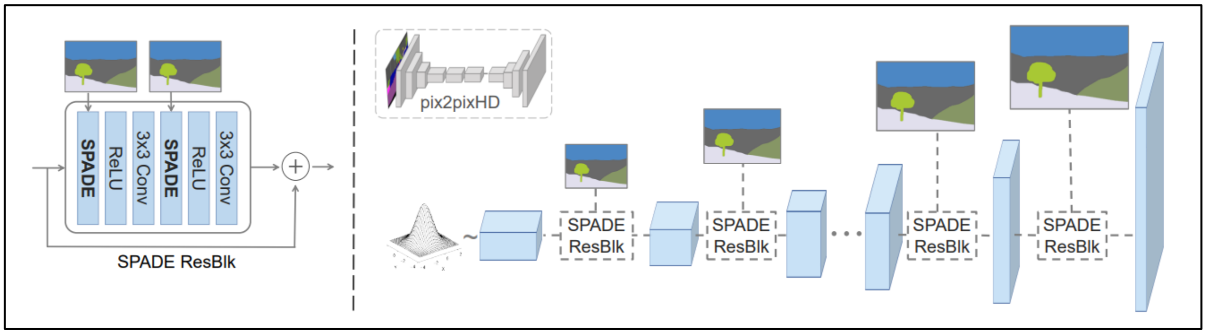


Figure 2. Diagram of SPADE

on various semantic segmentation datasets, showcasing its effectiveness through both quantitative metrics and subjective evaluations.

3. Technical part

3.1. Technical part: Summary of technical solutions

Our project target is ablation of the CC_FPSE, we will ablate SPADE and CC_FPSE with its different part. We have four combinations of SPADE and CC_FPSE. The first one is SPADE generator with discriminator of CC_FPSE. The second is SPADE generator with discriminator of CC-FPSE without Semantics embedding. The last two examples are CC_FPSE with the feature pyramid Semantics embedding discriminator and only feature pyramid part.

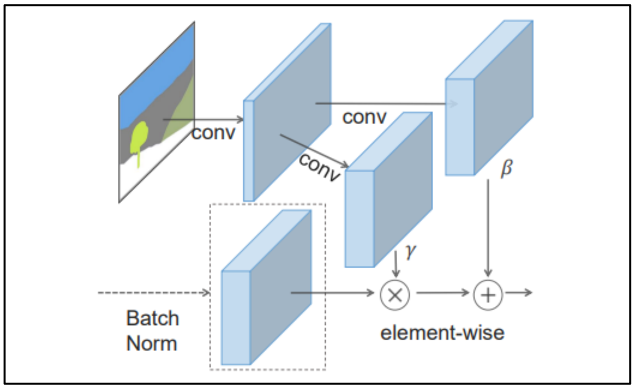


Figure 3. Diagram of SPADE

3.2. Technical part: Details of the technical solution

3.2.1 Generator

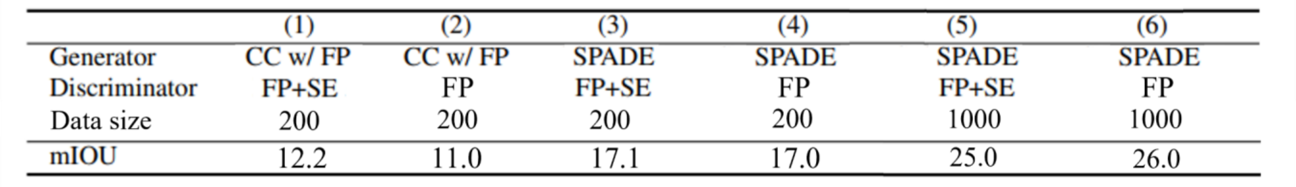
3.2.1.1 SPADE

The SPADE (Spatially Adaptive Denormalization) technique eliminates the need to input the segmentation map into the first layer of the generator because the learned modulation parameters already contain sufficient information about the label layout. As a result, we simplified the generator by discarding the encoder part commonly used in recent architectures, making the network more lightweight. This simplified generator can accept a random vector as input, enabling straightforward multi-modal synthesis. Figure 2 and 3 illustrates our generator detail architecture, utilizing ResNet blocks and upsampling layers. The modulation parameters for all normalization layers are learned through SPADE. To address different scales in residual blocks, we downsample the semantic mask to match spatial resolution. Training the generator involves a multi-scale discriminator and a loss function similar to pix2pixHD, with the least squared loss term replaced by the hinge loss term. We experimented with various ResNet-based discriminators, yielding similar results with a higher GPU memory requirement. Incorporating SPADE into the discriminator also showed comparable performance. Regarding the loss function, we observed that removing any loss term in the pix2pixHD loss function leads to degraded generation results.

3.2.1.2 CC_FPSE

Conditional convolution from CC_FPSE, takes a low-resolution noise map as input. It uses conditional convolution blocks and upsampling layers alternatively to gradually refine intermediate feature maps and ultimately generate the output image. Unlike traditional convolution layers that universally apply the same kernels to all samples and spatial locations, our approach seeks to enhance flexibility for semantic image synthesis. We argue for the need to adapt convolutional operations to distinct semantic layouts. To achieve this, we propose predicting convolutional kernel weights based on the semantic layout,

Table 1 The result from our experiment



	(1)	(2)	(3)	(4)	(5)	(6)
Generator	CC w/ FP	CC w/ FP	SPADE	SPADE	SPADE	SPADE
Discriminator	FP+SE	FP	FP+SE	FP	FP+SE	FP
Data size	200	200	200	200	1000	1000
mIOU	12.2	11.0	17.1	17.0	25.0	26.0

allowing the convolution layer to be aware of unique semantic labels at target locations. To integrate layout information into the image generation process, we introduce a weight prediction network that takes the semantic label map as input and outputs predicted convolutional kernel weights. To manage computational costs and GPU memory usage, we decompose the convolutional layer into depthwise convolution and pointwise convolution, predicting only the weights of lightweight depthwise convolutions.

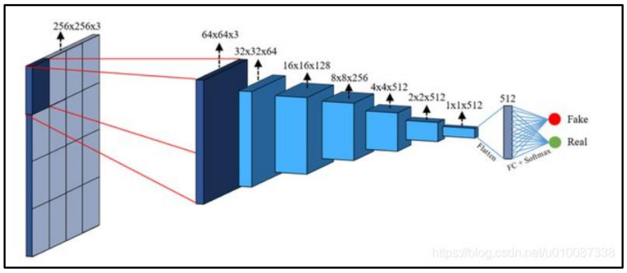


Figure 4. Design of PatchGAN

3.2.2 Discriminator

3.2.2.1 SPADE

In the SPADE, it adopts the PatchGAN, as shown in Figure 4, discriminator architecture for both PatchGAN and SPADE discriminator. The discriminator is designed to model high-frequency structures while relying on an L1 term for low-frequency correctness. It classifies each N × N patch in an image as real or fake, operating convolutionally across the image and averaging responses for the final output.

The SPADE discriminator incorporates modifications to the PatchGAN by using a 70 × 70 PatchGAN configuration. Each layer, denoted as Ck, represents a 4 × 4 ConvolutionInstanceNorm-LeakyReLU layer with k filters and a stride of 2. A convolution is applied after the last layer to produce a 1-dimensional output. Notably, the first C64 layer does not use InstanceNorm. Leaky ReLUs with a slope of 0.2 are employed throughout.

This combined SPADE discriminator architecture efficiently captures both high and low-frequency information in the image generation process, balancing accuracy and computational efficiency.

3.2.2.2 CC_FPSE

The CC_FPSE discriminator is designed to enhance semantic image synthesis by focusing on high-fidelity details and semantic alignment. Instead of the common multi-scale PatchGAN discriminator, we introduce a novel approach that overcomes limitations in discriminating fine details and ensuring spatial semantic alignment.

Motivated by the need for a more effective discriminator, our design incorporates a feature pyramid structure to emphasize low-level details and utilizes a semantics-embedding mechanism to enforce spatial alignment between generated images and input semantic layouts.

Traditional image generation often results in images with blurry edges and artifacts, indicating a need for increased attention to low-level details. Our feature pyramid discriminator addresses this by creating a single feature pyramid, generating a multi-scale feature representation with both global semantics and fine low-level texture and edge information. This architecture combines high-level semantic feature maps with low-level feature maps, resulting in stronger constraints on both semantic information and fine details.

In contrast to conventional discriminators, which concatenate images and semantic label maps, the CC_FPSE discriminator introduces a patch-based semantics embedding approach. Adapting the concept of a projection discriminator, our method computes the inner product between feature vectors and embedded semantic labels. This encourages semantic alignment between generated images and conditional semantic layouts, enhancing both high-fidelity and semantic coherence.

4. Experiment

Table 1 is our experiment from four combinations of CC_FPSE and SPADE. In addition, we use two data size to train our model, and there is still one model cannot be trained on time. That is the reason why we only put SPADE comparison.

5. Conclusions

After conducting the experiment, we have observed that the Feature Pyramid Network (FPN) contributes to making the details more reasonable. However, the Cascade Context Feature Pyramid Network (CC_FPSE) requires a larger amount of training data to achieve the optimal mean Intersection over Union (mIOU). Furthermore, during the ablation study comparing Spatially Adaptive Normalization (SPADE) and CC_FPSE, the results in Table 1 indicate that a reduced amount of data leads to superior performance for the simpler model. This finding aligns with the principles we have learned in our machine learning course, where simpler models tend to converge faster than complex ones. This observation provides a reasonable explanation for the results presented in the table.

References

[1] GauGAN: Changing Sketches into Photorealistic Masterpieces, https://www.youtube.com/watch?v=p5U4NgVGAwg

[2] Wang, Ting-Chun, et al. "High-resolution image synthesis and semantic manipulation with conditional gans." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[4] Liu, Xihui, et al. "Learning to predict layout-to-image conditional convolutions for semantic image synthesis." Advances in Neural Information Processing Systems 32 (2019).

[5] https://www.researchgate.net/figure/The-PatchGAN-structure-in-the-discriminator-architecture_fig5_339832261

[6] https://iq.opengenus.org/downsampling-and-upsampling-in-cnn/

[7] https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

Appendix

The below picture show the one of the test image from the coco dataset.

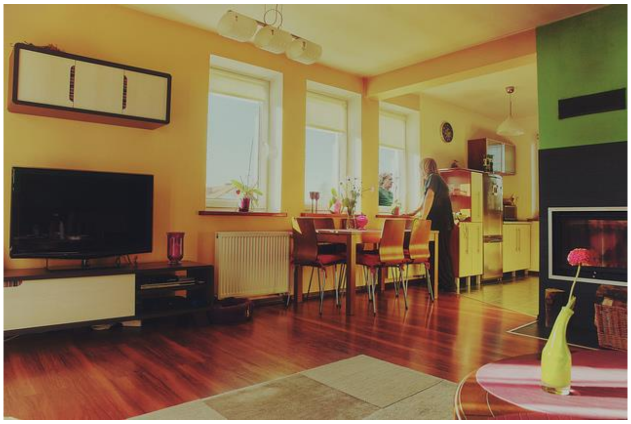


Figure 5. Original Image

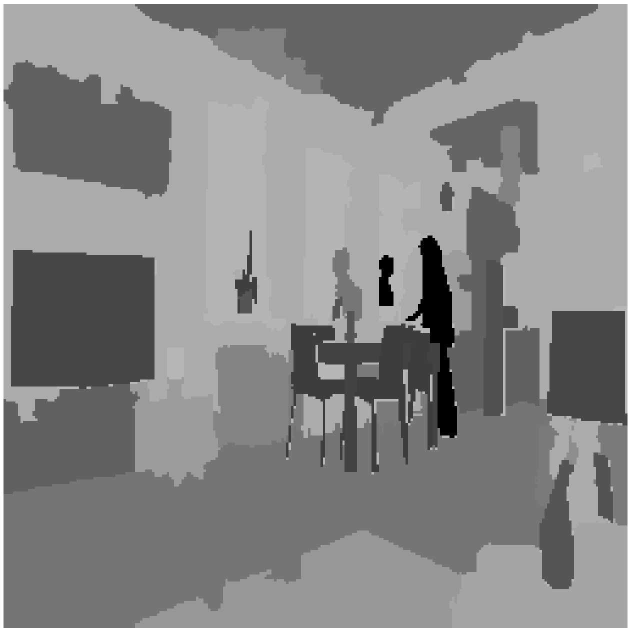


Figure 6. Label Image

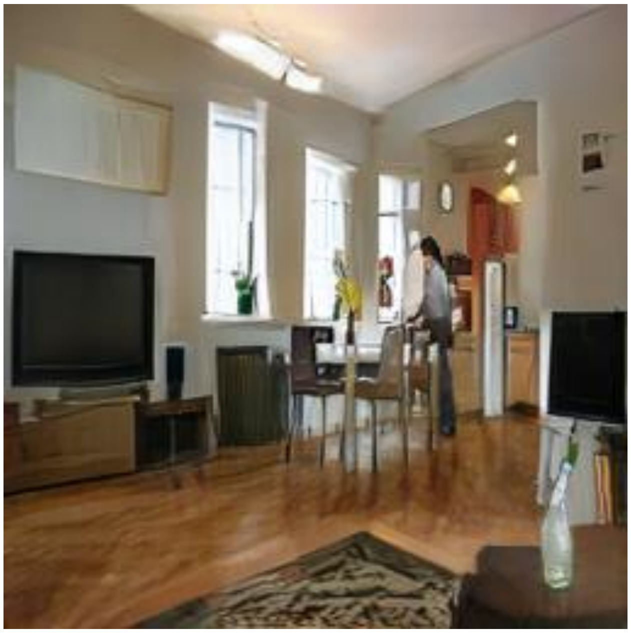


Figure 7. Result of using the paper's pretrain model

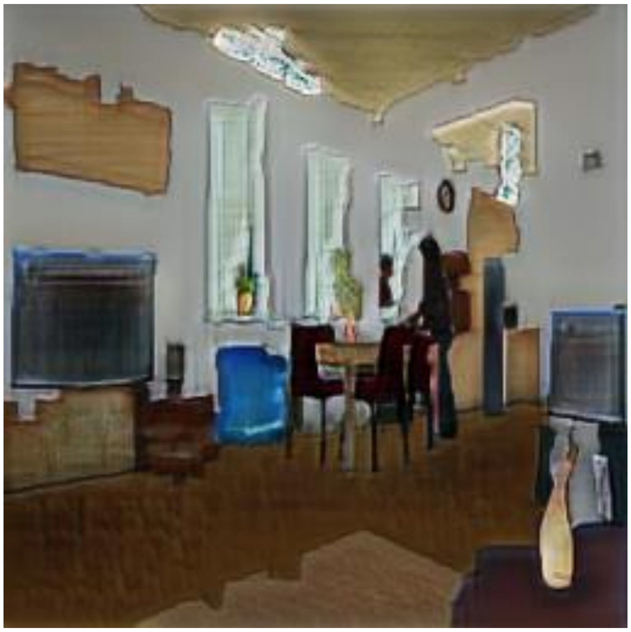


Figure 8. Result of CC-FPSE-200

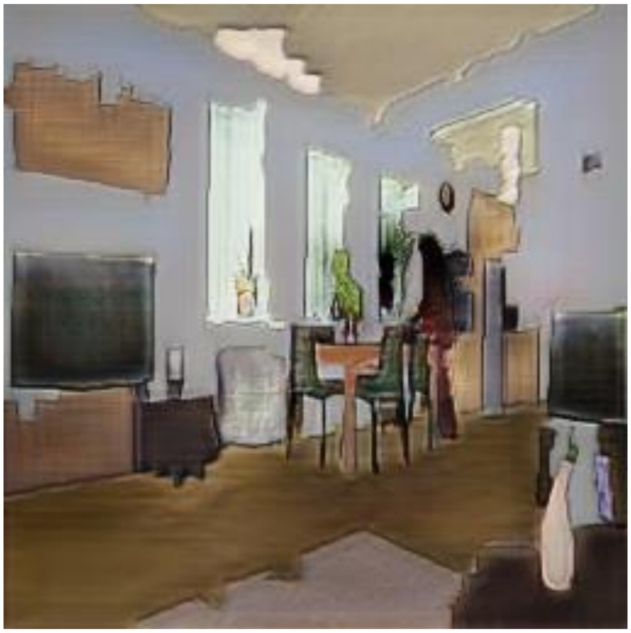


Figure 9. Result of CC-FP-200

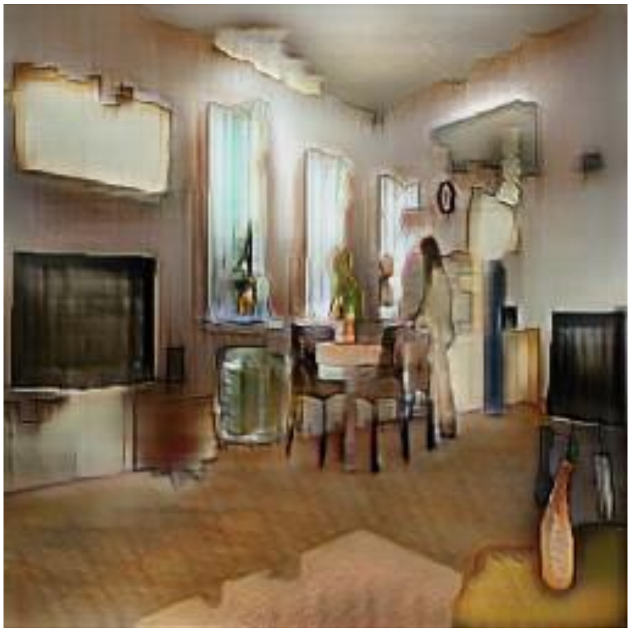


Figure 10. Result of SPADE-FPSE-200

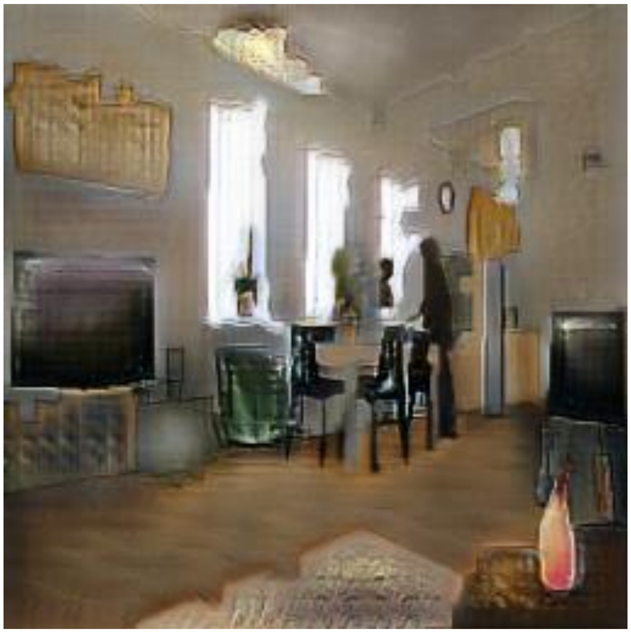


Figure 11. Result of SPADE-FP-200

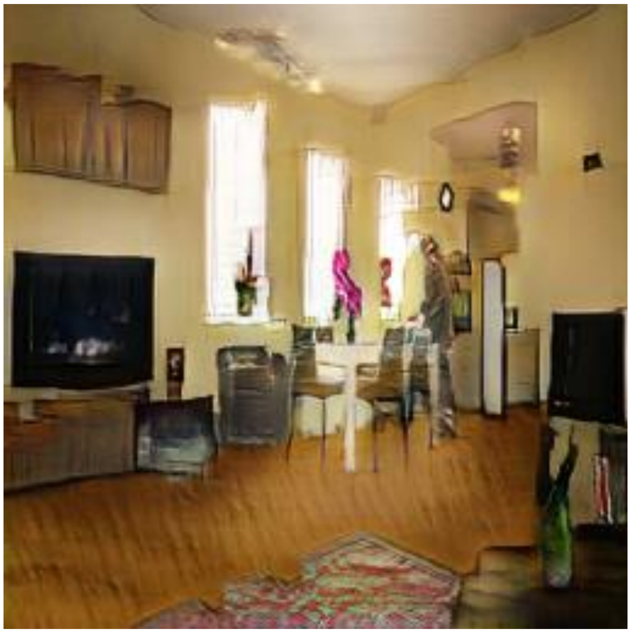


Figure 13. Result of SPADE-FP-1000

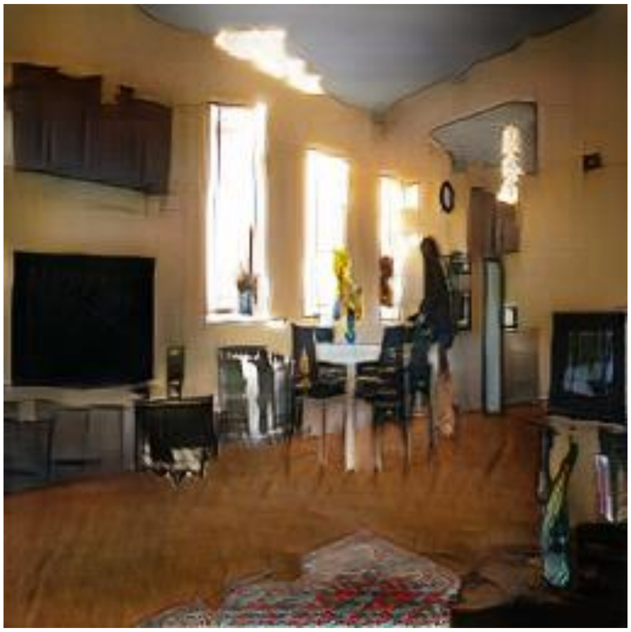


Figure 12. Result of SPADE-FPSE-1000