Computer Vision - Project Proposal

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-

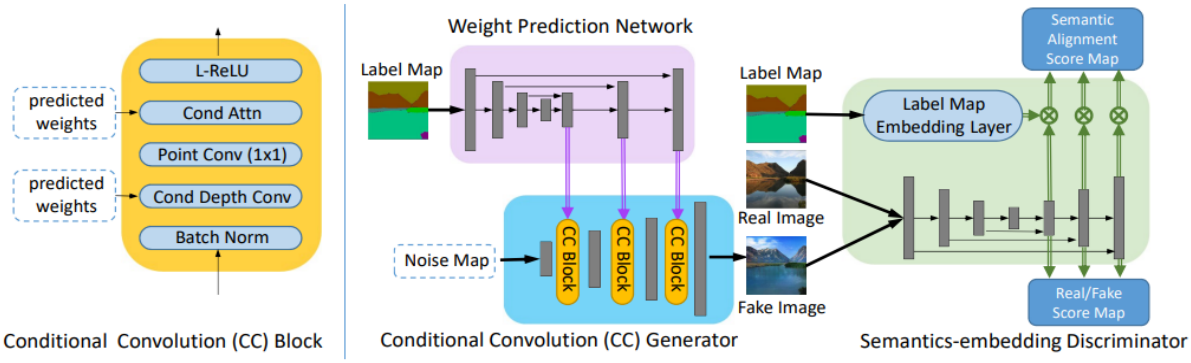

Figure 1: (Left) The structure of a Conditional Convolution Block. (Right) The overall framework of the proposed CC-FPSE.

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. Method

In the approach, it introduces an innovative GAN network architecture consisting of a Conditional Convolution Generator and a Semantics-embedding Discriminator from the paper, as depicted in Figure 1. The primary function of the Conditional Convolution Generator is to synthesize a virtual image that closely resembles a real photograph. In contrast, the Semantics-embedding Discriminator's role is to recognize the differences between virtual and real images. Based on the discriminator's evaluations, the Conditional Convolution Generator refines its output. Simultaneously, the discriminator continuously enhances its ability to distinguish between the images for more precise discrimination.

2.1. Conditional Convolution Generator

The Conditional Convolution Generator takes a low-resolution map as its input and generates a synthetic image through Conditional Convolution Blocks and the process of upsampling. The convolution kernels within the CC Block are primarily derived from a weight prediction network.

However, when employing a similar method to SPADE, challenges may arise in capturing critical information when there is a substantial area with the same semantic label. To address this, the paper proposes a global-context-aware weight prediction network with a feature pyramid structure. This approach involves using varying kernel sizes and weights to preserve specific features of the label map. Furthermore, in an effort to reduce the computational demands associated with conventional convolution, it incorporates depthwise separable convolution within the CC Block, focusing exclusively on the prediction of depthwise convolutional kernel weights. This significantly reduces the number of parameters requiring prediction.

Additionally, the author of the paper introduce a conditional attention layer within the CC Block to regulate the flow of information to the subsequent layer, thereby optimizing the generation of synthetic images.

2.2. Semantics-embedding Discriminator

A good discriminator is its capability to capture high-fidelity details while maintaining semantic alignment. To achieve this, they construct multi-scale feature pyramids, which enhance high-fidelity details such as texture and edges. Furthermore, they employ a semantics-embedding discriminator to ensure spatial semantic alignment between the generated images and the input semantic layout. As illustrated in Figure 1 (right panel), they use upsampleing feature-rich maps and combine them to generate multi-scale features with finer details and semantic information, thereby enforcing more robust constraints.

On the other hand, to promote semantic consistency between the generated images and the label map, the paper introduces a patch-based semantics embedding discriminator. This entails computing the dot product of the label map and multi-scale feature pyramids to produce a

semantic matching score map, which is then incorporated into the real/fake score, finally yielding the discriminative score. This innovative strategy serves to actualize high-detail images while ensuring semantic alignment.

3. Milestone

10/28: paper survey finished
10/30: project pitch
11/20: midterm-project due
12/8: code accomplished (include training & testing)
12/15: validation data collected
12/18 or 12/25: final project presentation
1/5: adjust project & final report finished
1/8: final project report due

References

[1] Xihui. Liu, Guojun. Yin, Jing. Shao, Xiaogang. Wang and Hongsheng. Li, "Learning to Predict Layout-to-image Conditional Convolutions for Semantic Image Synthesis," *arXiv*, 10 Jan, 2020.