Table 1: More ablation studies.

	Cityscapes					COCO-Stuff		
Generator	SPADE	SPADE	SPADE	SPADE w/ FP	CondGen w/FP	SPADE	SPADE w/ FP	CondGen w/FP
Discriminator	MsPatch	FP	FP+SE	FP+SE	FP+SE	MsPatch	FP+SE	FP+SE
mIOU	62.3 (baseline)	62.9 **(R1)**	63.11 **(R1)**	64.17 **(R1&R3)**	64.9 (ours)	37.4 (baseline)	38.79 **(R1&R3)**	40.1 (ours)

Figure 1: A sequence of generated images from Cityscapes.

To reviewer #1

Q1: (1) Motivation behind the proposed method? (2) V layer is only smaller than a convolution layer by a factor D (which is 3?). (3) Make more sense to predict $D \times C \times K \times K \times L$ kernels.

A1: (1) The key idea is to generate spatially-varying convolution kernels at different spatial locations according to the input layouts, so that the image synthesis process can be better controlled by the semantic layouts. (2) V layer is smaller by a factor of D. However, D is the number of input channels and generally ranges from 64 to 1024 in different layers. (3) The reviewer's suggestion on predicting $D \times C \times K \times K \times L$ kernels has two limitations: (3.1) it takes more parameters than our design. If we predict $D \times C \times K \times K \times L$ kernels, the last layer of the weight prediction network would be a fully-connected layer with $D \times C \times K \times K \times L$ output dimensions. While in our design, the last layer of the weight prediction network is a convolution layer with $C \times K \times K$ output dimensions. For instance, $L = 182$ for COCO dataset, $K = 3$, and C and D ranges from 64 to 1024.(3.2) The kernels are fixed for all spatial locations with the same label, regardless of their distinct contexts. For instance, a pixel in the top and bottom regions of a "person" would share convolutional kernels, which cannot effectively achieve appearance variations for pixels with the same label.

Q2: Contribition of the discriminator design.

A2: The contribution of our discriminator design is twofold. (1) This is the first time to adopt multi-scale feature pyramids in the discriminator to promote high-fidelity details such as textures and edges. (2) The patch-based semantic embeddings are adopted to enhance the spatial semantic alignment between the generated images and input semantic layout. The idea is inspired by projection discriminator which computes the dot product between the class label and image feature vector, but we are the first to adapt this idea to the spatial label map and make it work on semantic image synthesis. We will add more explanations in the final version.

Q3: Number of parameters and more ablation studies comparing with SPADE.

A3: The parameter size of our proposed generater is 107.4 million, which is similar to that of SPADE (96 million). Although predicting convolutional kernels requires more parameters than predicting scale and bias vectors in SPADE, we predict only 1/3 of the convolutional layer weights (as shown in CC Block in original Fig. 2), while SPADE predicts all BN parameters. Following the reviewer's suggestions, we did more ablation studies as shown in Table 1.

To Reviewer #2

Q4: How impacted is this network by small changes in the input map?

A4: We conduct an experiment on generating videos. We extract segmentation maps from a video sequence in Cityscapes by a semantic segmentation model, and use those segmentation maps and the same noise vector to generate a sequence of images. As shown in Figure 1, the generated results for adjacent frames are smooth and consistent.

Q5: Human perceptual evaluation details.

A5: Each pair of images is annotated by 5 workers independently, and in total 20 workers involved in the human evaluation. The 5 workers reach an agreement for 70% of the cases.

Q6: More ablation studies and code.

A6: More ablation studies are shown in Table 1. We will release code upon paper acceptance.

To Reviewer #3

Q7: Compare the SPADE-like method (learned scaling parameters) to the proposed method (predicting conditional convolution) with the same discriminator and weight prediction network (SPADE w/ FP + FP+SE).

A7: The comparisons on COCO and Cityscapes are shown in Table 1, which demonstrates the effectiveness of predicting conditional convolution (ours) over predicting scaling parameters for BN (SPADE).