

Diverse Image Synthesis from Semantic Layouts via Conditional IMLE

Ke Li*
UC Berkeley

ke.li@eecs.berkeley.edu

Tianhao Zhang*
Nanjing University

bryanzhang@smail.nju.edu.cn

Jitendra Malik
UC Berkeley

malik@eecs.berkeley.edu

Abstract

Most existing methods for conditional image synthesis are only able to generate a single plausible image for any given input, or at best a fixed number of plausible images. In this paper, we focus on the problem of generating images from semantic segmentation maps and present a simple new method that can generate an arbitrary number of images with diverse appearance for the same semantic layout. Unlike most existing approaches which adopt the GAN [11, 12] framework, our method is based on the recently introduced Implicit Maximum Likelihood Estimation (IMLE) [22] framework. Compared to the leading approach [3], our method is able to generate more diverse images while producing fewer artifacts despite using the same architecture. The learned latent space also has sensible structure despite the lack of supervision that encourages such behaviour. Videos and code are available at https://people.eecs.berkeley.edu/~ke.li/projects/imle/scene_layouts/.



Figure 1: Samples generated by our model. The 9 images are samples generated by our model conditioned on the same semantic layout as shown at the bottom-left corner.

1. Introduction

Conditional image synthesis is a problem of great importance in computer vision. In recent years, the commu-

nity has made great progress towards generating images of high visual fidelity on a variety of tasks. However, most proposed methods are only able to generate a single image given each input, even though most image synthesis problems are ill-posed, i.e.: there are multiple equally plausible images that are consistent with the same input. Ideally, we should aim to predict a distribution of all plausible images rather than just a single plausible image, which is a problem known as *multimodal image synthesis* [42]. This problem is hard for two reasons:

1. *Model*: Most state-of-the-art approaches for image synthesis use generative adversarial nets (GANs) [11, 12], which suffer from the well-documented issue of *mode collapse*. In the context of conditional image synthesis, this leads to a model that generates only a single plausible image for each given input regardless of the latent noise and fails to learn the distribution of plausible images.
2. *Data*: Multiple different ground truth images for the same input are not available in most datasets. Instead, only one ground truth image is given, and the model has to learn to generate other plausible images in an unsupervised fashion.

In this paper, we focus on the problem of multimodal image synthesis from semantic layouts, where the goal is to generate multiple *diverse* images for the same semantic layout. Existing methods are either only able to generate a fixed number of images [3] or are difficult to train [42] due to the need to balance the training of several different neural nets that serve opposing roles.

To sidestep these issues, unlike most image synthesis approaches, we step outside of the GAN framework and propose a method based on the recently introduced method of Implicit Maximum Likelihood Estimation (IMLE) [22]. Unlike GANs, IMLE by design avoids mode collapse and is able to train the same types of neural net architectures as generators in GANs, namely neural nets with random noise drawn from an analytic distribution as input.

This approach offers two advantages:

*Equal contribution.

1. Unlike [3], we can generate an arbitrary number of images for each input by simply sampling different noise vectors.
2. Unlike [42], which requires the simultaneous training of three neural nets that serve opposing roles, our model is much simpler: it only consists of a single neural net. Consequently, training is much more stable.

2. Related Work

2.1. Unimodal Prediction

Most modern image synthesis methods are based on generative adversarial nets (GANs) [11, 12]. Most of these methods are capable of producing only a single image for each given input, due to the problem of mode collapse. Various work has explored conditioning on different types of information. Various methods condition on a scalar that only contains little information, such as object category and attribute [25, 9, 5]. Other methods condition on richer labels, such as text description [28], surface normal maps [35], previous frames in a video [24, 33] and images [36, 14, 41]. Some methods only condition on inputs images in the generator, but not in the discriminator [27, 20, 40, 21]. [16, 28, 30] explore conditioning on attributes that can be modified manually by the user at test time; these methods are not true multimodal methods because they require manual changes to the input (rather than just sampling from a fixed distribution) to generate a different image.

Another common approach to image synthesis is to treat it as a simple regression problem. To ensure high perceptual quality, the loss is usually defined on some transformation of the raw pixels. This paradigm has been applied to super-resolution [1, 15], style transfer [15] and video frame prediction [32, 26, 8]. These methods are by design unimodal methods because neural nets are functions, and so can only produce point estimates.

Various methods have been developed for the problem of image synthesis from semantic layouts. For example, Karacan *et al.* [17] developed a conditional GAN-based model for generating images from semantic layouts and labelled image attributes. It is important to note that the method requires supervision on the image attributes and is therefore a unimodal method. Isola *et al.* [14] developed a conditional GAN that can generate images solely from semantic layout. However, it is only able to generate a single plausible image for each semantic layout, due to the problem of mode collapse in GANs. Wang *et al.* [34] further refined the approach of [14], focusing on the high-resolution setting. While these methods are able to generate images of high visual fidelity, they are all unimodal methods.

2.2. Fixed Number of Modes

A simple approach to generate a fixed number of different outputs for the same input is to use different branches or models for each desired output. For example, [13] proposed a model that outputs a fixed number of different predictions simultaneously, which was an approach adopted by Chen and Koltun [3] to generate different images for the same semantic layout. Unlike most approaches, [3] did not use the GAN framework; instead it uses a simple feedforward convolutional network. On the other hand, Ghosh *et al.* [10] uses a GAN framework, where multiple generators are introduced, each of which generates a different mode. The above methods all have two limitations: (1) they are only able to generate a fixed number of images for the same input, and (2) they cannot generate continuous changes.

2.3. Arbitrary Number of Modes

A number of GAN-based approaches propose adding learned regularizers that discourage mode collapse. BiGAN/ALI [6, 7] trains a model to reconstruct the latent code from the image; however, when applied to the conditional setting, significant mode collapse still occurs because the encoder is not trained until optimality and so cannot perfectly invert the generator. VAE-GAN [18] combines a GAN with a VAE, which does not suffer from mode collapse. However, image quality suffers because the generator is trained on latent code sampled from the encoder/approximate posterior, and is never trained on latent code sampled from the prior. At test time, only the prior is available, resulting in a mismatch between training and test conditions. Zhu *et al.* [42] proposed Bicycle-GAN, which combines both of the above approaches. While this alleviates the above issues, it is difficult to train, because it requires training three different neural nets simultaneously, namely the generator, the discriminator and the encoder. Because they serve opposing roles and effectively regularize one another, it is important to strike just the right balance, which makes it hard to train successfully in practice.

A number of methods for colourization [2, 38, 19] predict a discretized marginal distribution over colours of each individual pixel. While this approach is able to capture multimodality in the marginal distribution, ensuring global consistency between different parts of the image is not easy, since there are correlations between the colours of different pixels. This approach is not able to learn such correlations because it does not learn the joint distribution over the colours of *all* pixels.

3. Method

Most state-of-the-art approaches for conditional synthesis rely on the conditional GAN framework. Unfortunately, GANs suffer from the well-known problem of mode col-

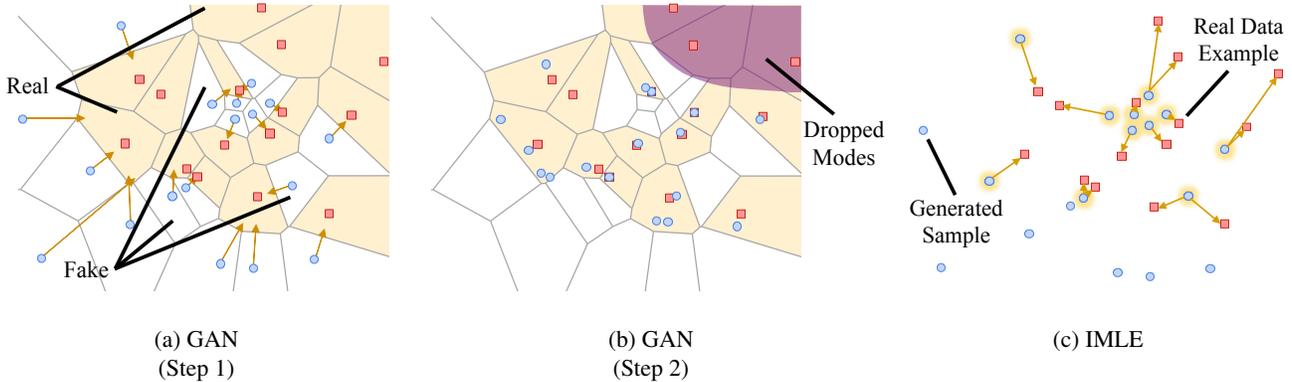


Figure 2: (a-b) How a (unconditional) GAN collapses modes (here we show a GAN with 1-nearest neighbour discriminator for simplicity). The blue circles represent generated images and the red squares represent real images. The yellow regions represent those classified as real by the discriminator, whereas the white regions represent those classified as fake. As shown, when training the generator, each *generated image* is essentially pushed towards the nearest *real image*. Some real images may not be selected by any generated image during training and therefore could be ignored by the trained generator – this is a manifestation of mode collapse. (c) An illustration of how Implicit Maximum Likelihood Estimation (IMLE) works. IMLE avoids mode collapse by reversing the direction in which generated images are matched to real images. Instead of pushing each generated image towards the nearest real image, for every real image, it pulls the nearest generated image towards it – this ensures that all real images are matched to some generated image, and no real images are ignored.

lapse, and in the context of conditional image synthesis, this causes the generator to ignore the latent input noise vector and always generate the same output image for the same input label, regardless of what the value of the latent noise vector. So, to generate different output images for the same input label, we must solve the underlying problem of mode collapse.

3.1. Why Mode Collapse Happens

We first consider the unconditional setting, where there is no input label. As shown in Figure 2(a-b), in a GAN, each generated image is made similar to *some* real image. Some images may not be selected by any real image. So after training, the generator will not be able to generate any image that is similar to the unselected real images, so it effectively ignores these images. In the language of probabilistic modelling, real images can be viewed as samples from some underlying true distribution of natural images, and the generator ignoring some of the real images means that the generative model assigns low probability density to these images. So, the modes (i.e.: the local maxima in the probability density) of the true distribution of natural images that represent the ignored images are not modelled by the generator; hence the name “mode collapse”. In the conditional setting, typically only one ground truth output image is available for every input label. As a result, mode collapse becomes more problematic, because the conditional distribution modelled by the generator will collapse to a single mode around the sole ground truth output image. This

means that the generator will not be able to output any other equally plausible output image.

3.2. IMLE

The method of Implicit Maximum Likelihood Estimation (IMLE) [22] solves mode collapse by reversing the direction in which generated images are matched to real images. In a GAN, each generated image is effectively matched to a real image. In IMLE, each real image is matched to a generated image. This ensures that all real images are matched, and no real images are left out. As shown in Figure 2(c), IMLE then tries to make each matched generated image similar to the real images they are matched to. Mathematically, it solves the optimization problem below. Here, \mathbf{z}_j 's denote randomly sampled latent input noise vectors, \mathbf{y}_i 's denote ground truth images, and T_θ denotes a neural net whose architecture is the same as the generator in GANs.

$$\min_{\theta} \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_m} \left[\frac{1}{n} \sum_{i=1}^n \min_{j=1, \dots, m} \|T_\theta(\mathbf{z}_j) - \mathbf{y}_i\|_2^2 \right]$$

3.3. Conditional IMLE

In the conditional setting, the goal is to model a family of conditional distributions, each of is conditioned on a different input label, i.e.: $\{p(\mathbf{y}|\mathbf{x} = \mathbf{x}_i)\}_{i=1, \dots, n}$, where \mathbf{x}_i 's denote ground truth input images, and \mathbf{y} denotes the generated output image. So, conditional IMLE [23] differs from standard IMLE in two ways: first, the input label is passed

into the neural net T_θ in addition to the latent input noise vector, and second, a ground truth output image can only be matched to an output image generated from its corresponding ground truth input label (i.e.: output images generated from an input label that is different from the current ground truth input label cannot be matched to the current ground truth output image). Concretely, conditional IMLE solves the following optimization problem, where $\mathbf{z}_{i,j}$'s denote randomly sampled latent input noise vectors and \mathbf{y}_i 's denote ground truth images:

$$\min_{\theta} \mathbb{E}_{\mathbf{z}_{1,1}, \dots, \mathbf{z}_{n,m}} \left[\frac{1}{n} \sum_{i=1}^n \min_{j=1, \dots, m} \|T_\theta(\mathbf{x}_i, \mathbf{z}_{i,j}) - \mathbf{y}_i\|_2^2 \right]$$

3.4. Probabilistic Interpretation

Image synthesis can be viewed as a probabilistic modelling problem. In unconditional image synthesis, the goal is the model the marginal distribution over images, i.e.: $p(\mathbf{y})$, whereas in conditional image synthesis, the goal is to model the conditional distribution $p(\mathbf{y}|\mathbf{x})$. In a conditional GAN, the probabilistic model is chosen to be an *implicit probabilistic model*. Unlike classical (also known as prescribed) probabilistic models like CRFs, implicit probabilistic models are not defined by a formula for the probability density, but rather a procedure for drawing samples from them. The probability distributions they define are the distributions over the samples, so even though the formula for the probability density of these distributions may not be in closed form, the distributions themselves are valid and well-defined. The generator in GANs is an example of an implicit probabilistic model. It is defined by the following sampling procedure, where T_θ is a neural net:

1. Draw $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2. Return $\mathbf{y} := T_\theta(\mathbf{x}, \mathbf{z})$ as a sample

In classical probabilistic models, learning, or in other words, parameter estimation, is performed by maximizing the log-likelihood of the ground truth images, either exactly or approximately. This is known as maximum likelihood estimation (MLE). However, in implicit probabilistic models, this is in general not feasible: because the formula for the probability density may not be in closed form, the log-likelihood function, which the sum of the log-densities of the model evaluated at each ground truth image, cannot be in general written down in closed form. The GAN can be viewed as an alternative way to estimate the parameters of the probabilistic model, but it has one critical issue of mode collapse. As a result, the learned model distribution could capture much less variation than what the data exhibits. On the other hand, MLE never suffers from this issue: because mode collapse entails assigning very low probability density to some ground truth images, this

would make the likelihood very low, because likelihood is the product of the densities evaluated at each ground truth image. So, maximizing likelihood will never lead to mode collapse. This implies that GANs cannot approximate maximum likelihood, and so the question is: is there some other algorithm that can? IMLE was designed with goal in mind, and can be shown to maximize a lower bound on the log-likelihood under mild conditions. Like GANs, IMLE does not need the formula for the probability density of the model to be known; unlike GANs, IMLE approximately maximizes likelihood, and so cannot collapse modes. Another added advantage comes from the fact that IMLE does not need a discriminator nor adversarial training. As a result, training is much more stable – there is no need to balance the capacity of the generator with that of the discriminator, and so much less hyperparameter tuning is required.

3.5. Formulation

For the task of image synthesis from semantic layouts, we take \mathbf{x} to be the input semantic segmentation map and \mathbf{y} to be the generated output image. (Details on the representation of \mathbf{x} are in the supplementary material.) The conditional probability distribution that would like to learn is $p(\mathbf{y}|\mathbf{x})$. A plausible image that is consistent with the input segmentation \mathbf{x} is a mode of this distribution; because there could be many plausible images that are consistent with the same segmentation, $p(\mathbf{y}|\mathbf{x})$ usually has multiple modes (and is therefore multimodal). A method that performs unimodal prediction can be seen as producing a point estimate of this distribution. To generate multiple plausible images, a point estimate is not enough; instead, we need to estimate the full distribution.

We generalize conditional IMLE by using a different distance metric $\mathcal{L}(\cdot, \cdot)$, namely a perceptual loss based on VGG-19 features [31], the details of which are in the supplementary material. The modified algorithm is presented in Algorithm 1.

3.6. Architecture

To allow for direct comparability to Cascaded Refinement Networks (CRN) [3], which is the leading method for multimodal image synthesis from semantic layouts, we use the same architecture as CRN, with minor modifications to convert CRN into an implicit probabilistic model.

The vanilla CRN synthesizes only one image for the same semantic layout input. To model an arbitrary number of modes, we add additional input channels to the architecture and feed random noise \mathbf{z} via these channels. Because the noise is random, the neural net can now be viewed as a (implicit) probabilistic model.

Noise Encoder Because the input segmentation maps are provided at high resolutions, the noise vector \mathbf{z} , which is

Algorithm 1 Conditional IMLE

Input Training semantic segmentation maps $\{\mathbf{x}_i\}_{i=1}^n$ and the corresponding ground truth images $\{\mathbf{y}_i\}_{i=1}^n$
Initialize parameters θ for neural net T_θ
for epoch = 1 **to** E **do**
 Pick a random batch $S \subseteq \{1, \dots, n\}$
 for $i \in S$ **do**
 Generate m i.i.d random vectors $\{\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,m}\}$
 for $j = 1$ **to** m **do**
 $\tilde{\mathbf{y}}_{i,j} \leftarrow T_\theta(\mathbf{x}_i, \mathbf{z}_{i,j})$
 end for
 $\sigma(i) \leftarrow \arg \min_{j \in \{1, \dots, m\}} \mathcal{L}(\mathbf{y}_i, \tilde{\mathbf{y}}_{i,j})$
 end for
 for $k = 1$ **to** K **do**
 Pick a random batch $\tilde{S} \subseteq S$
 $\theta \leftarrow \theta - \eta \nabla_\theta \left(\sum_{i \in \tilde{S}} \mathcal{L}(\mathbf{y}_i, \tilde{\mathbf{y}}_{i,\sigma(i)}) \right) / |\tilde{S}|$
 end for
end for

concatenated to the input channel-wise, could be very high-dimensional, which could hurt sample efficiency and therefore training speed. To solve this, we propose forcing the noise to lie on a low-dimensional manifold. To this end, we add a noise encoder module, which is a 3-layer convolutional net that takes the segmentation \mathbf{x} and a lower-dimensional noise vector $\tilde{\mathbf{z}}$ as input and outputs a noise vector \mathbf{z}' of the same size as \mathbf{z} . We replace \mathbf{z} with \mathbf{z}' and leave the rest of the architecture unchanged.

3.7. Dataset and Loss Rebalancing

In practice, we found datasets can be strongly biased towards objects with relatively common appearance. As a result, naïve training can result in limited diversity among the images generated by the trained model. To address this, we propose two strategies to rebalance the dataset and loss, the details of which are in the supplementary material.

4. Experiment

4.1. Dataset

The choice of dataset is very important for multimodal conditional image synthesis. The most common dataset in the unimodal setting is the Cityscapes dataset [4]. However, it is not suitable for the multimodal setting because most images in the dataset are taken under similar weather conditions and time of day and the amount of variation in object colours is limited. This lack of diversity limits what any multimodal method can do. On the other hand, the GTA-5 dataset [29], has much greater variation in terms of weather conditions and object appearance. To demonstrate this, we compare the colour distribution of both datasets and present the distribution of hues of both datasets in Figure 3. As

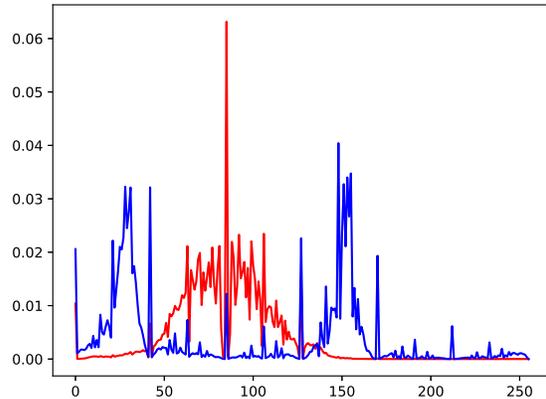


Figure 3: Comparison of histogram of hues between two datasets. Red is Cityscapes and blue is GTA-5.

shown, Cityscapes is concentrated around a single mode in terms of hue, whereas GTA-5 has much greater variation in hue. Additionally, the GTA-5 dataset includes more 20000 images and so is much larger than Cityscapes.

Furthermore, to show the generalizability of our approach and its applicability to real-world datasets, we train on the BDD100K [37] dataset and show results in Fig. 10.

4.2. Experimental Setting

We train our model on 12403 training images and evaluate on the validation set (6383 images). Due to computational resource limitations, we conduct experiments at the 256×512 resolution. We add 10 noise channels and set the hyperparameters shown in Algorithm 1 to the following values: $|S| = 400$, $m = 10$, $K = 10000$, $|\tilde{S}| = 1$ and $\eta = 1e - 5$.

The leading method for image synthesis from semantic layouts in the multimodal setting is the CRN [3] with diversity loss that generates nine different images for each semantic segmentation map and is the baseline that we compare to.

4.3. Quantitative Comparison

Quantitative comparison aims to quantitatively compare the diversity as well as quality of the images generated by our model and CRN.

Diversity Evaluation We measure the diversity of each method by generating 40 pairs of output images for each of 100 input semantic layouts from the test set. We then compute the average distance between each pair of output images for each given input semantic layout, which is then averaged over all input semantic layouts. The distance metric we use is LPIPS [39], which is designed to measure perceptual dissimilarity. The results are shown in Table 1. As



Figure 4: Comparison of generated images for the same semantic layout. The bottom-left image in (a) is the input semantic layout and we generate 9 samples for each model. See our website for more samples.

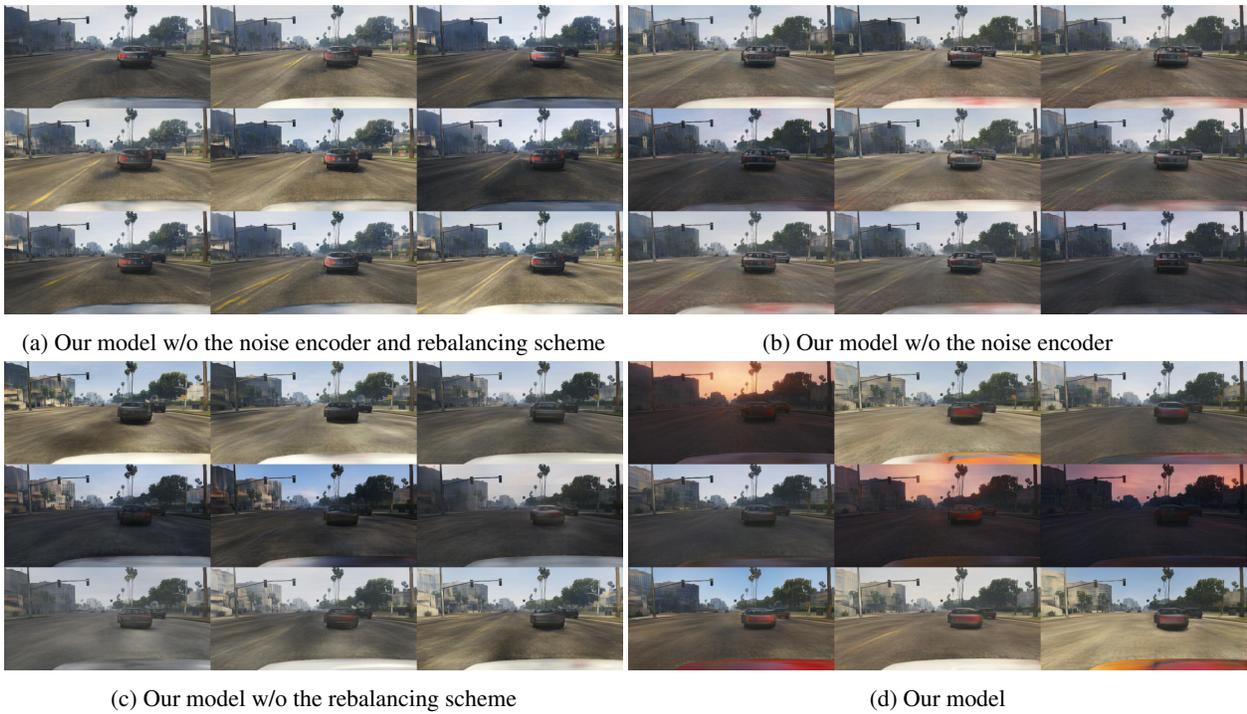


Figure 5: Ablation study using the same semantic layout as Fig. 4.



Figure 6: Images generated by interpolating between latent noise vectors. See our website for videos showing the effect of interpolations.

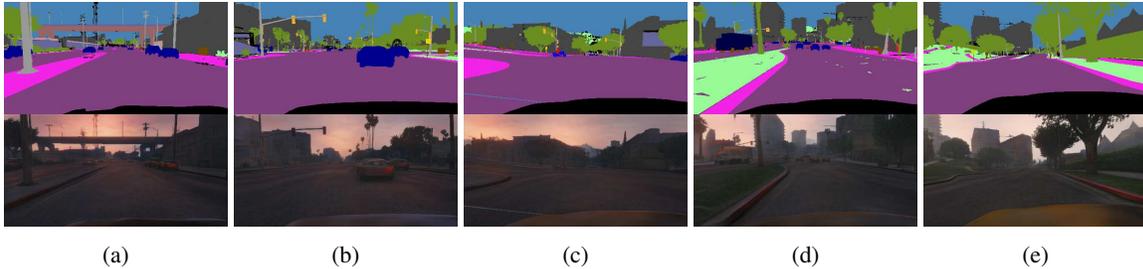


Figure 7: Style consistency with the same random vector. (a) is the original input-output pair. We use the same random vector used in (a) and apply it to (b),(c),(d) and (e). See our website for more examples.

Model	LPIPS score
CRN	0.11
CRN+noise	0.12
Ours w/o noise encoder	0.10
Ours w/o rebalancing scheme	0.17
Ours	0.19

Table 1: LPIPS score. We show the average perceptual distance of different models (including ablation study) and our proposed model gained the highest diversity.

shown, the proposed method outperforms the baselines by a large margin. We also perform an ablation study and find that the proposed method performs better than variants that remove the noise encoder or the rebalancing scheme, which demonstrates the value of each component of our method.

Image Quality Evaluation We now evaluate the generated image quality by human evaluation. Since it is difficult for humans to compare images with different styles, we selected the images that are closest to the ground truth image in ℓ_1 distance among the images generated by CRN and our method. We then asked 62 human subjects to evaluate the images generated for 20 semantic layouts. For each semantic layout, they were asked to compare the image generated by CRN to the image generated by our method and judge which image exhibited more obvious synthetic patterns. The result is shown in Table 2.

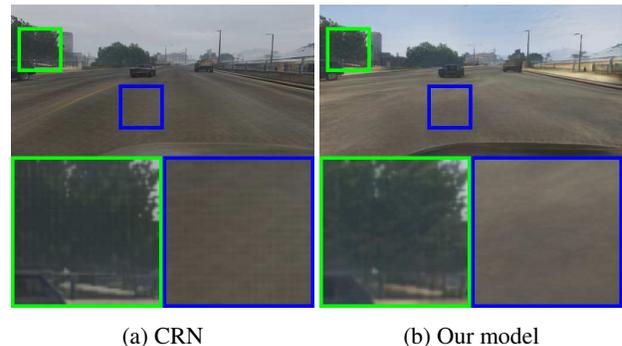


Figure 8: Comparison of artifacts in generated images.

% of Images Containing More Artifacts	
CRN	0.636 \pm 0.242
Our method	0.364 \pm 0.242

Table 2: Average percentage of images that are judged by humans to exhibit more obvious synthetic patterns. Lower is better.

4.4. Qualitative Evaluation

A qualitative comparison is shown in Fig. 4. We compare to three baselines, BicycleGAN [42], Pix2pix-HD with input noise [34] and CRN. As shown, Pix2pix-HD generates almost identical images, BicycleGAN generates images with heavy distortions and CRN generates images with

little diversity. In comparison, the images generated by our method are diverse and do not suffer from distortions. We also perform an ablation study in Fig. 5, which shows that each component of our method is important. In the supplementary material, we include results of the baselines with the proposed rebalancing scheme and demonstrate that, unlike our method, they cannot take advantage of it.

In addition, our method also generates fewer artifacts compared to CRN, which is especially interesting because the architecture and the distance metric are the same as CRN. As shown in Fig. 8, the images generated by CRN has grid-like artifacts which are not present in the images generated by our method. More examples generated by our model are shown in the supplementary material.

Interpolation We also perform linear interpolation of latent vectors to evaluate the semantic structure of the learned latent space. As shown in 6, by interpolating between the noise vectors corresponding to generated images during daytime and nighttime respectively, we obtain a smooth transition from daytime to nighttime. This suggests that the learned latent space is sensibly ordered and captures the full range of variations along the time-of-day axis. More examples are available in the supplementary material.

Scene Editing A successful method for image synthesis from semantic layouts enables users to manually edit the semantic map to synthesize desired imagery. One can do this simply by adding/deleting objects or changing the class label of a certain object. In Figure 9 we show several such changes. Note that all four inputs use the same random vector; as shown, the images are highly consistent in terms of style, which is quite useful because the style should remain the same after editing the layout. We further demonstrate this in Fig. 7 where we apply the random vector used in (a) to different segmentation maps in (b),(c),(d),(e) and the style is preserved across the different segmentation maps.

5. Conclusion

We presented a new method based on IMLE for multi-modal image synthesis from semantic layout. Unlike prior approaches, our method can generate arbitrarily many images for the same semantic layout and is easy to train. We demonstrated that our method can generate more diverse images with fewer artifacts compared to the leading approach [3], despite using the same architecture. In addition, our model is able to learn a sensible latent space of noise vectors without supervision. We showed that by taking the interpolations between noise vectors, our model can generate continuous changes. At the same time, using the same noise vector across different semantic layouts result in images of consistent style.

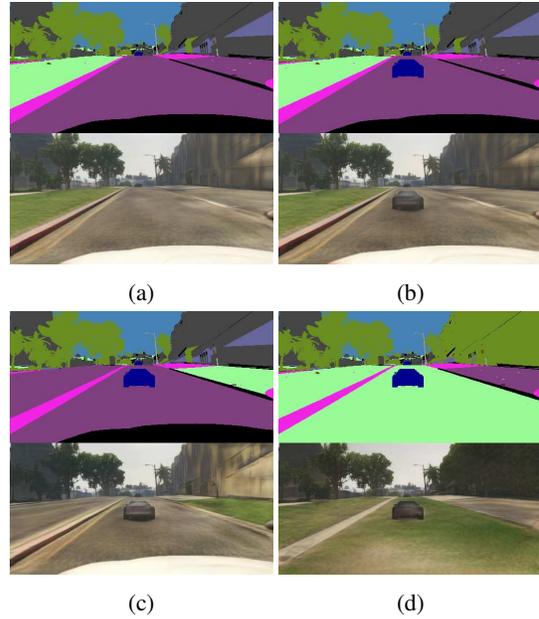


Figure 9: Scene editing. (a) is the original input semantic map and the generated output. (b) adds a car on the road. (c) changes the grass on the left to road and change the side walk on the right to grass. (d) deletes our own car, changes the building on the right to tree and changes all road to grass.

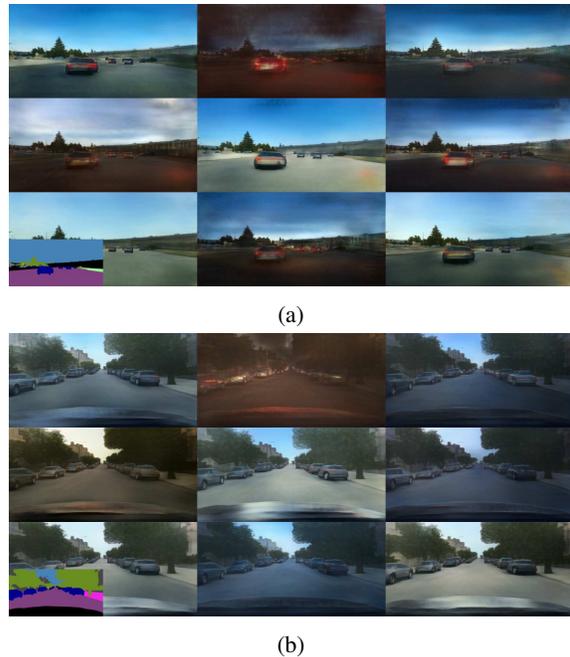


Figure 10: Images generated using our method on the BDD100K [37] dataset.

References

- [1] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. *arXiv preprint arXiv:1511.05666*, 2015. 2
- [2] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *European conference on computer vision*, pages 126–139. Springer, 2008. 2
- [3] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017. 1, 2, 4, 5, 8
- [4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- [5] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 2
- [6] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 2
- [7] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 2
- [8] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016. 2
- [9] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014. 2
- [10] Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. *arXiv preprint arXiv:1704.02906*, 1(4), 2017. 2
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2
- [12] Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *arXiv preprint arXiv:1407.4981*, 2014. 1, 2
- [13] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2012. 2
- [14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 2
- [15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 2
- [16] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7006–7015. IEEE, 2017. 2
- [17] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2
- [18] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 2
- [19] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016. 2
- [20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017. 2
- [21] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 2
- [22] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018. 1, 3
- [23] Ke Li, Shichong Peng, and Jitendra Malik. Super-resolution via conditional implicit maximum likelihood estimation. *arXiv preprint arXiv:1810.01406*, 2018. 3
- [24] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 2
- [25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2
- [26] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015. 2
- [27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2
- [28] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pages 217–225, 2016. 2
- [29] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision*

- (ECCV), volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. [5](#)
- [30] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. [2](#)
- [31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [4](#)
- [32] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015. [2](#)
- [33] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. [2](#)
- [34] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017. [2](#), [7](#)
- [35] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016. [2](#)
- [36] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision*, pages 517–532. Springer, 2016. [2](#)
- [37] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. [5](#), [8](#)
- [38] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016. [2](#)
- [39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint*, 2018. [5](#)
- [40] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016. [2](#)
- [41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. [2](#)
- [42] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017. [1](#), [2](#), [7](#)