

A EXPERIMENT DETAILS AND MORE RESULTS

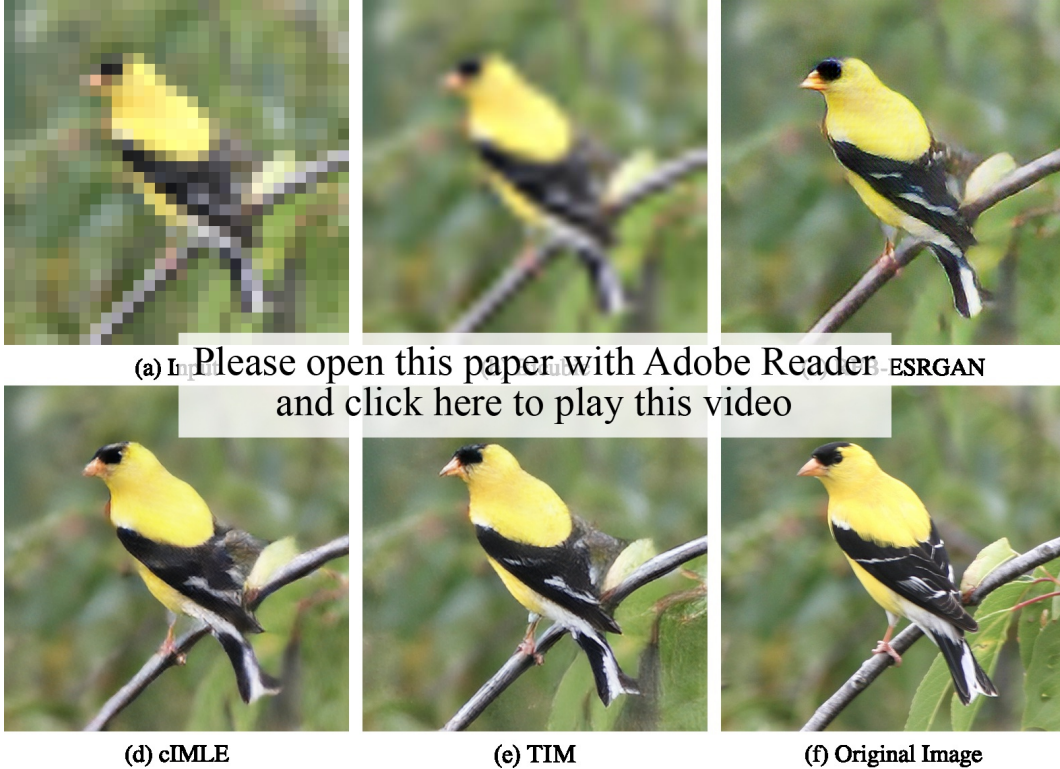
We include more details and results on $16\times$ super-resolution, colourization, image synthesis from scene layouts and image decompression in the following pages.

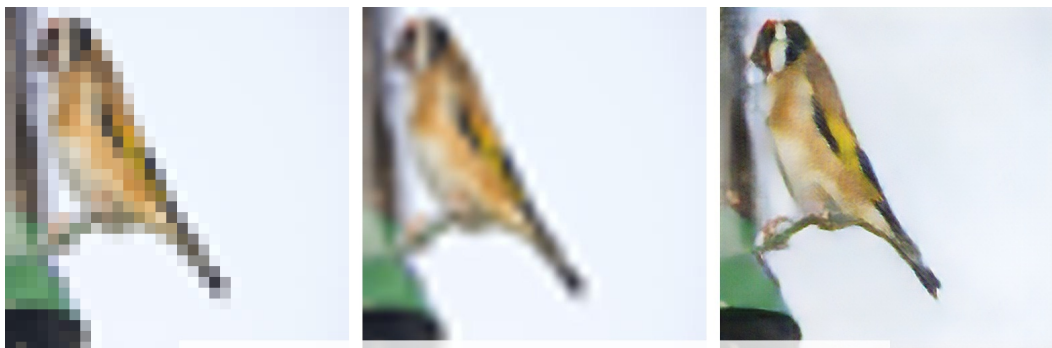
A.1 TRAINING DETAILS

We use a four-module design for all tasks. Input to each module is downsampled anisotropically from the full resolution input to fit the corresponding resolution, except for super-resolution, only the first module take in the low-resolution input. All models were trained for 150 epochs with mini-batch size of 1 using Adam optimizer with a learning rate of 1×10^{-4} on a NVIDIA V100 GPU.

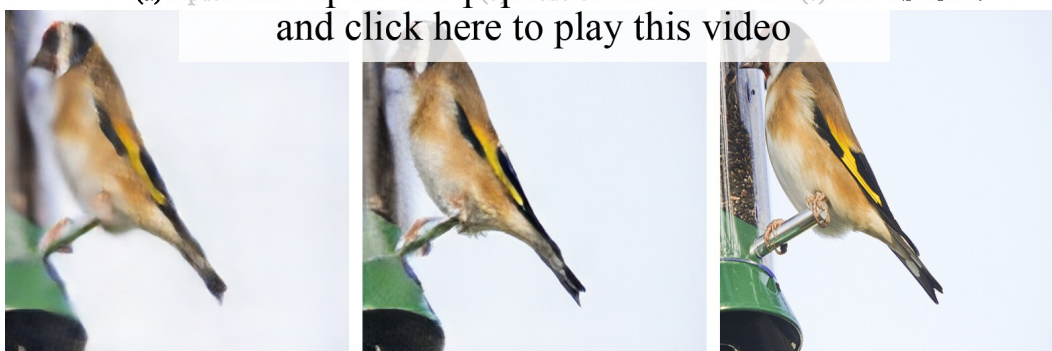
A.2 $16\times$ SUPER-RESOLUTION

Given a low-resolution input image, the goal is to generate different possible higher-resolution output images that are all geometrically consistent with the input. Applications include photo enhancement, remote sensing and medical imaging. Unlike most methods that consider relatively small upscaling factors of $2 - 4\times$, we apply our method to a much more challenging (Baker & Kanade, 2002) setting of $16\times$ upscaling. Our one-to-many method is a good fit for this setting since there could be multiple high-resolution images with perceptible differences in details that correspond to the same low-resolution image.





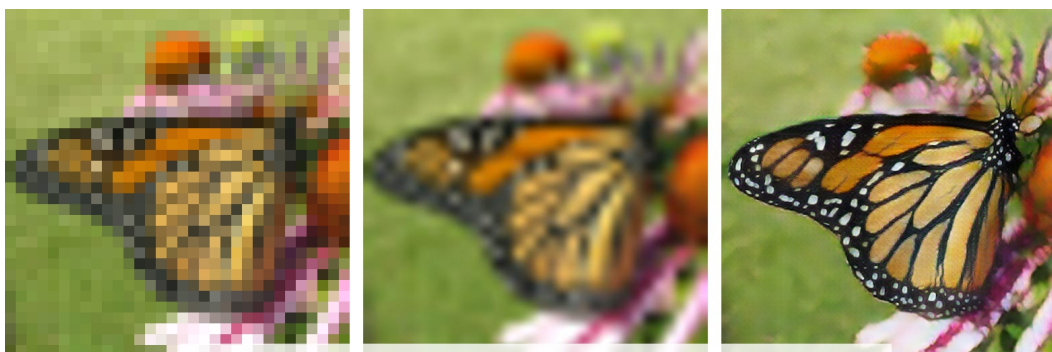
(a) Please open this paper with Adobe Reader and click here to play this video



(d) cIMLE

(e) TIM

(f) Original Image



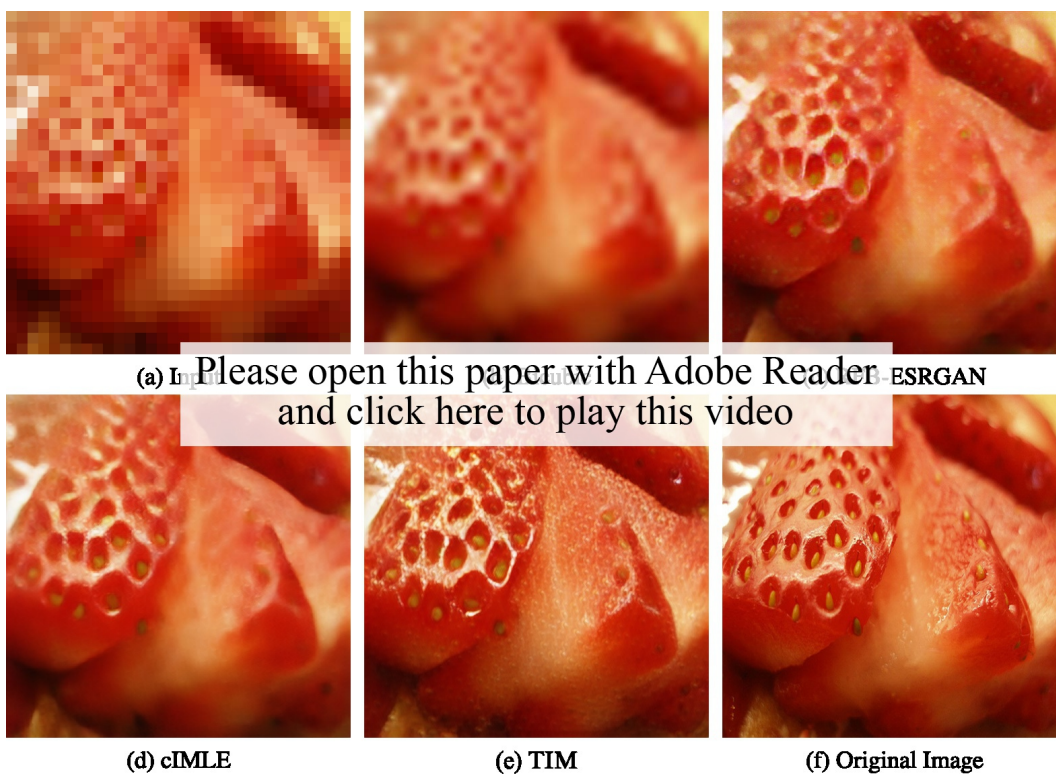
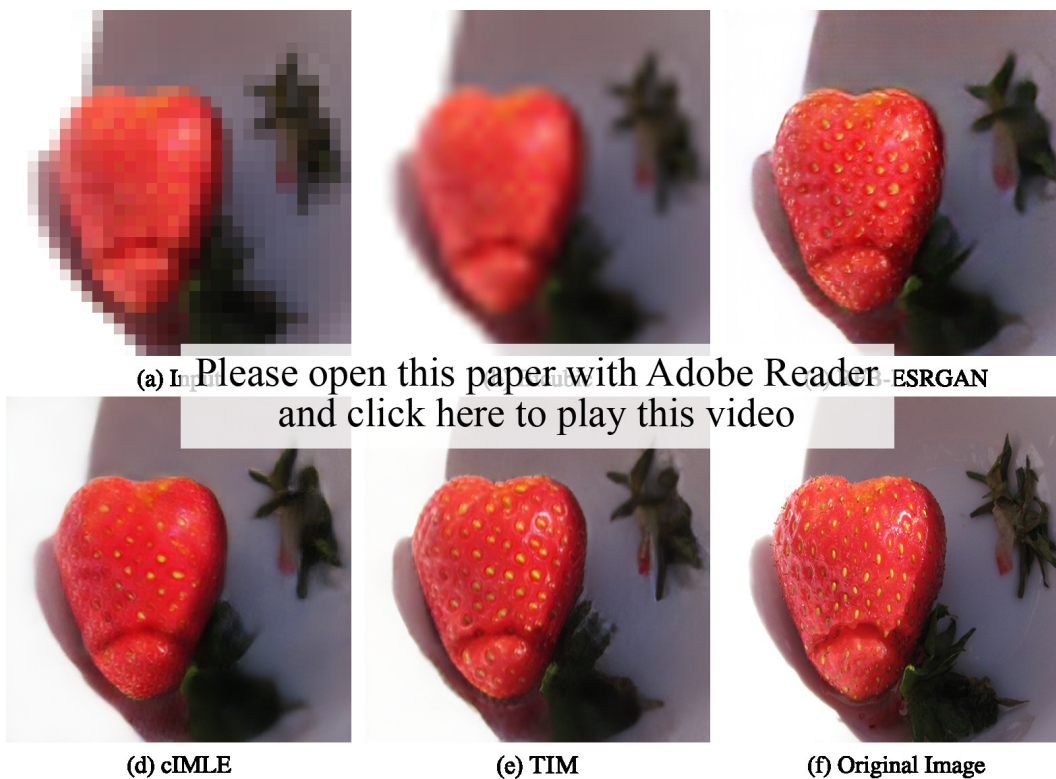
(a) Please open this paper with Adobe Reader and click here to play this video



(d) cIMLE

(e) TIM

(f) Original Image



A.3 IMAGE COLOURIZATION

Given a grayscale image, the goal is to generate colours for each pixel that are consistent with the content of the input image. Applications include the restoration of old photos, for example those captured by black-and-white cameras or those captured by colour cameras that has become washed out over time. Since there could be many plausible colourings of the same grayscale image, the goal of one-to-many colourization is to generate different plausible colourings, which would provide the user the ability to choose the preferred colouring among them.



(a) Please open this paper with Adobe Reader and click here to play this video



(d) Larson et al.



(e) cIMLE



(f) TIM



(a) Please open this paper with Adobe Reader and click here to play this video



(d) Larson et al.



(e) cIMLE



(f) TIM



(a) Please open this paper with Adobe Reader and click here to play this video



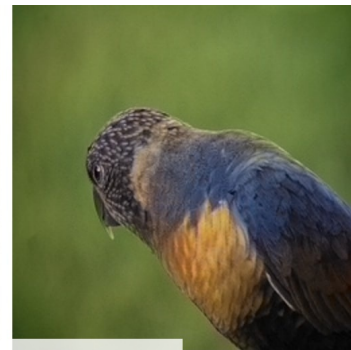
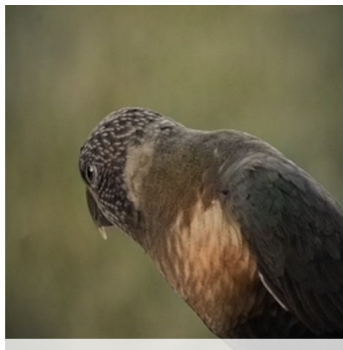
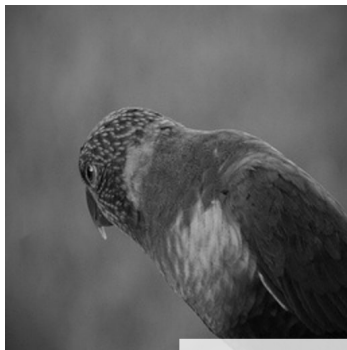
(d) Larson et al.



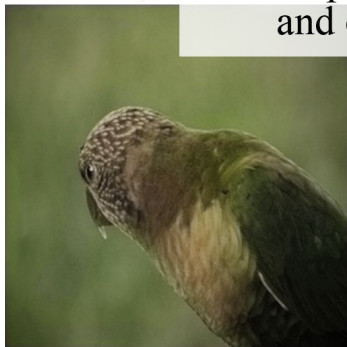
(e) cIMLE



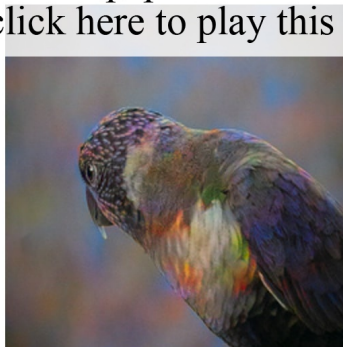
(f) TIM



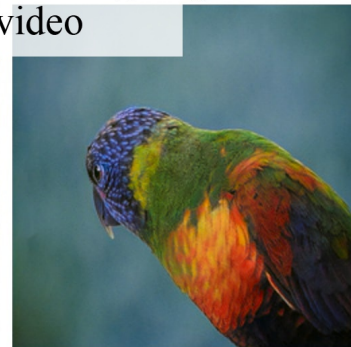
(a) Please open this paper with Adobe Reader and click here to play this video



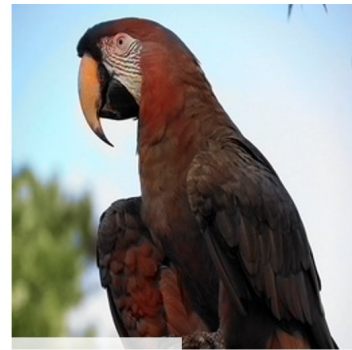
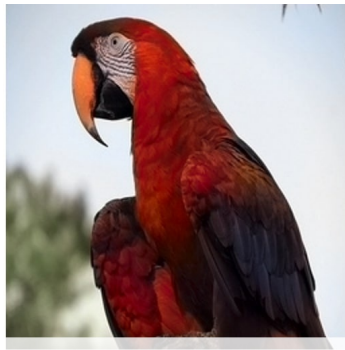
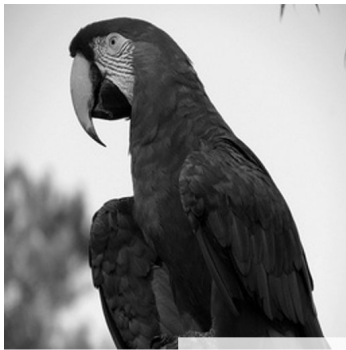
(d) Larson et al.



(e) cIMLE



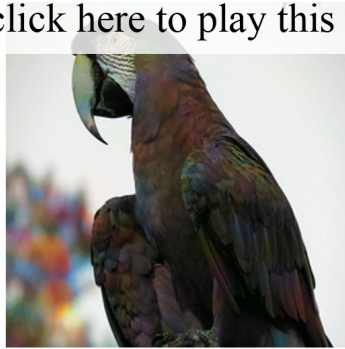
(f) TIM



Please open this paper with Adobe Reader
and click here to play this video



(d) Larson et al.



(e) cIMLE

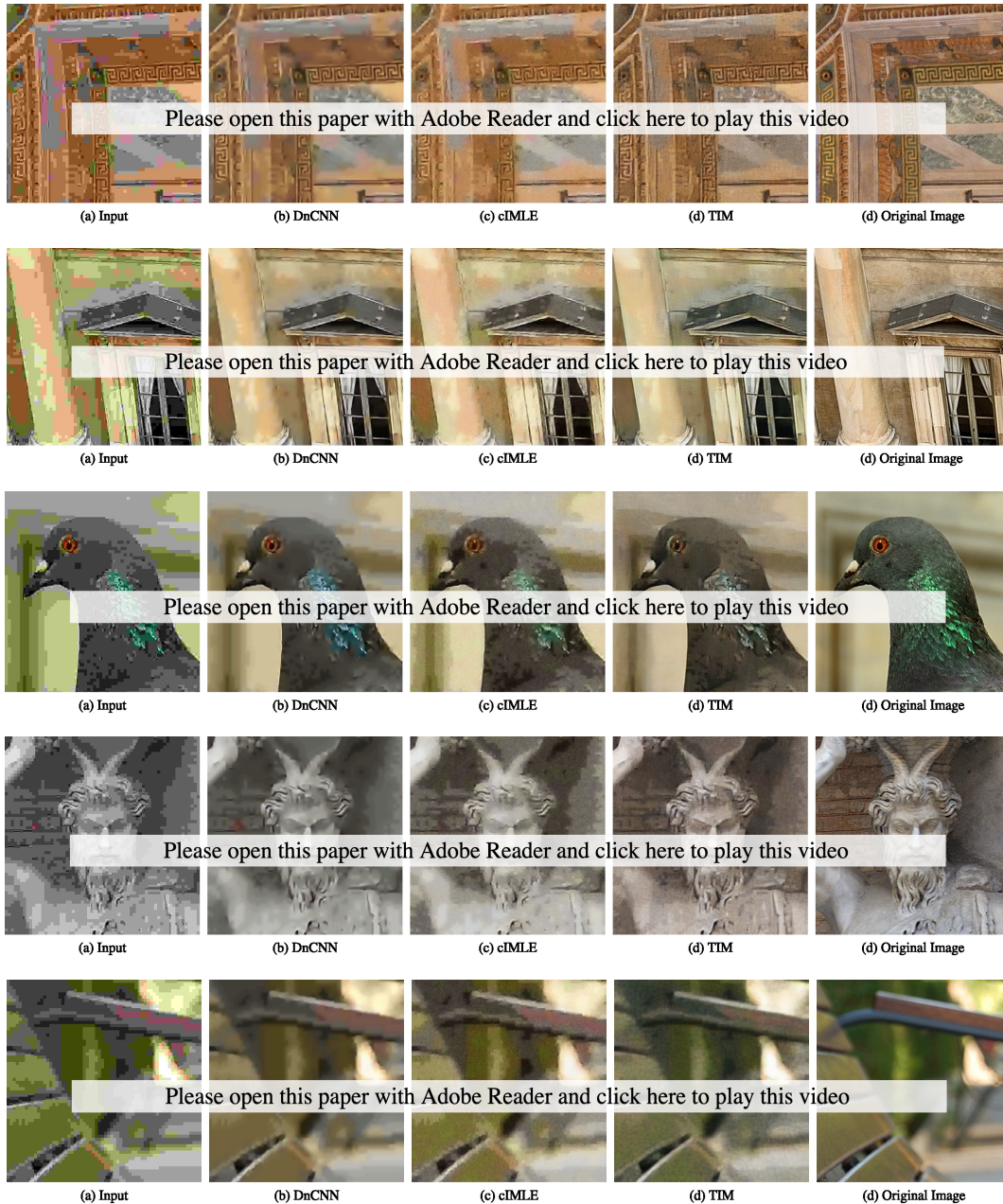


(f) TIM

A.4 IMAGE DECOMPRESSION

Given an image compressed aggressively with a standard lossy codec (e.g.: JPEG), the goal is to recover the original uncompressed image. Note that this task is different from image compression, where both the encoding and decoding model can be learned. Here, the encoding model is fixed (e.g.: JPEG), and only the decoding model is learned.

Image decompression is of practical interest since most images are saved in lossy compressed formats; noticeable artifacts may have been introduced during compression and the original uncompressed images have been lost. It would be nice to have the capability to recover an image free of artifacts. Because compression causes irreversible information loss, multiple artifact-free images are possible. With user guidance, the most preferred version can be selected and saved for future use.



	Scene Layout Synthesis	
	<i>TIM</i>	<i>cIMLE</i>
FID	59.71	61.62

Table 6: Comparison of perceptual quality, as measured by Fréchet Inception Distance (FID) between the observed images and the samples generated by our method (TIM) and the best-performing one-to-many method in terms of diversity, cIMLE. Lower values of FID are better.

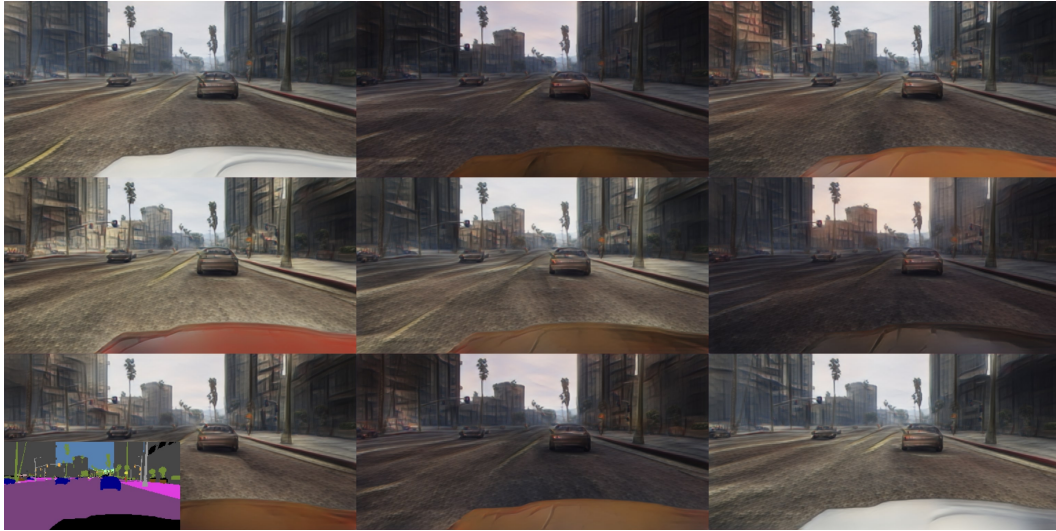
	Scene Layout Synthesis	
σ	<i>TIM</i>	<i>cIMLE</i>
0.3	.0254	0.0188
0.2	.000980	.000659
0.15	.000011	.000007

Table 7: Comparison of faithfulness weighted variance of the samples generated by our method (TIM) and the best-performing one-to-many method in terms of diversity, cIMLE. Higher value indicates more variation in the generated samples that are faithful to the original image. σ is the bandwidth parameter for the Gaussian kernel used to compute the faithfulness weights.

A.5 IMAGE SYNTHESIS FROM SCENE LAYOUTS

Given a semantic segmentation map, the goal is to synthesize realistic images from the segmentation map. This task is challenging as the category label and the shape of each segment are the only cues to the model; in particular, no information about the appearance is provided to the model. There are many scenes with the same layout that correspond to the same segmentation map. So, one-to-many prediction aims to generate multiple images with the same scene layout and different appearances.

Following prior work (Li* et al., 2020), we evaluate on the GTA-5 (Richter et al., 2016) dataset and apply dataset and loss rebalancing to compensate for the data imbalance of the datasets. As shown in Table 6 and 7, TIM outperforms the one-to-many baseline, cIMLE.



(a) TIM



(b) cIMLE



(a) TIM



(b) cIMLE

	<i>1 module</i>	<i>4 modules</i>
FID	94.84	72.75

Table 8: Comparison of Fréchet Inception Distance (FID) to the target of the samples generated by our method (TIM) using different number of modules on image decompression. Lower values of FID are better. Here it suggests that more modules produces higher quality samples.

σ	<i>1 module</i>	<i>4 modules</i>
0.3	.0513	.0495
0.2	.00301	.00380
0.15	.000110	.000223

Table 9: Comparison of faithfulness weighted variance of the samples generated by our method (TIM) using different number of modules on image decompression. Higher value is better. Here it suggests that more modules produces more diverse results while being faithful to the original image.

B EFFECT OF NUMBER OF MODULES

Here we show the quantitative results for using different number of modules of our method (TIM). As shown in Table 8 and 9, having more modules leads to better image quality and diversity which validates the need for having a multi-module design.

C EFFECT OF NUMBER SAMPLES PER IMAGE IN IMLE

Here we show the quantitative results for using different number of samples per image in IMLE of our method (TIM). As shown in Table 10, having more samples per image produces better results. Therefore, hierarchical sampling is essential to increase the effective number of samples that IMLE could search over while retaining computational efficiency.

D TRAINING STABILITY

In Figure 10, we visualize the output of TIM for a test input image over the course of training. As shown, the output quality improves steadily during training, thereby demonstrating training stability.

<i>Number Samples per Image</i>	<i>FID</i>
10	79.54
60	76.20
120	72.75

Table 10: Comparison of Fréchet Inception Distance (FID) to the target of the samples generated by our method (TIM) using different number of samples per image for IMLE on image decompression. Lower values of FID are better. Here it shows that more samples per image in IMLE produces higher quality samples.

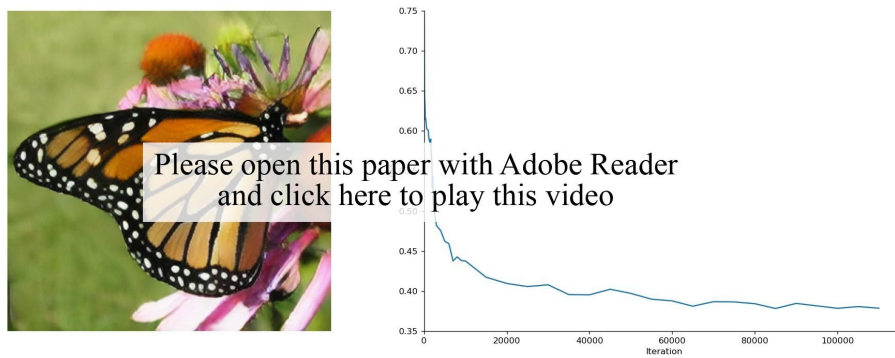


Figure 10: Click on the image to see output of model while it trains, demonstrating stable training. Video also available in supplementary materials.