Paper Reading

Xinwei Geng

2018-12-25

Visualizing and Understanding Generative Adversarial Networks

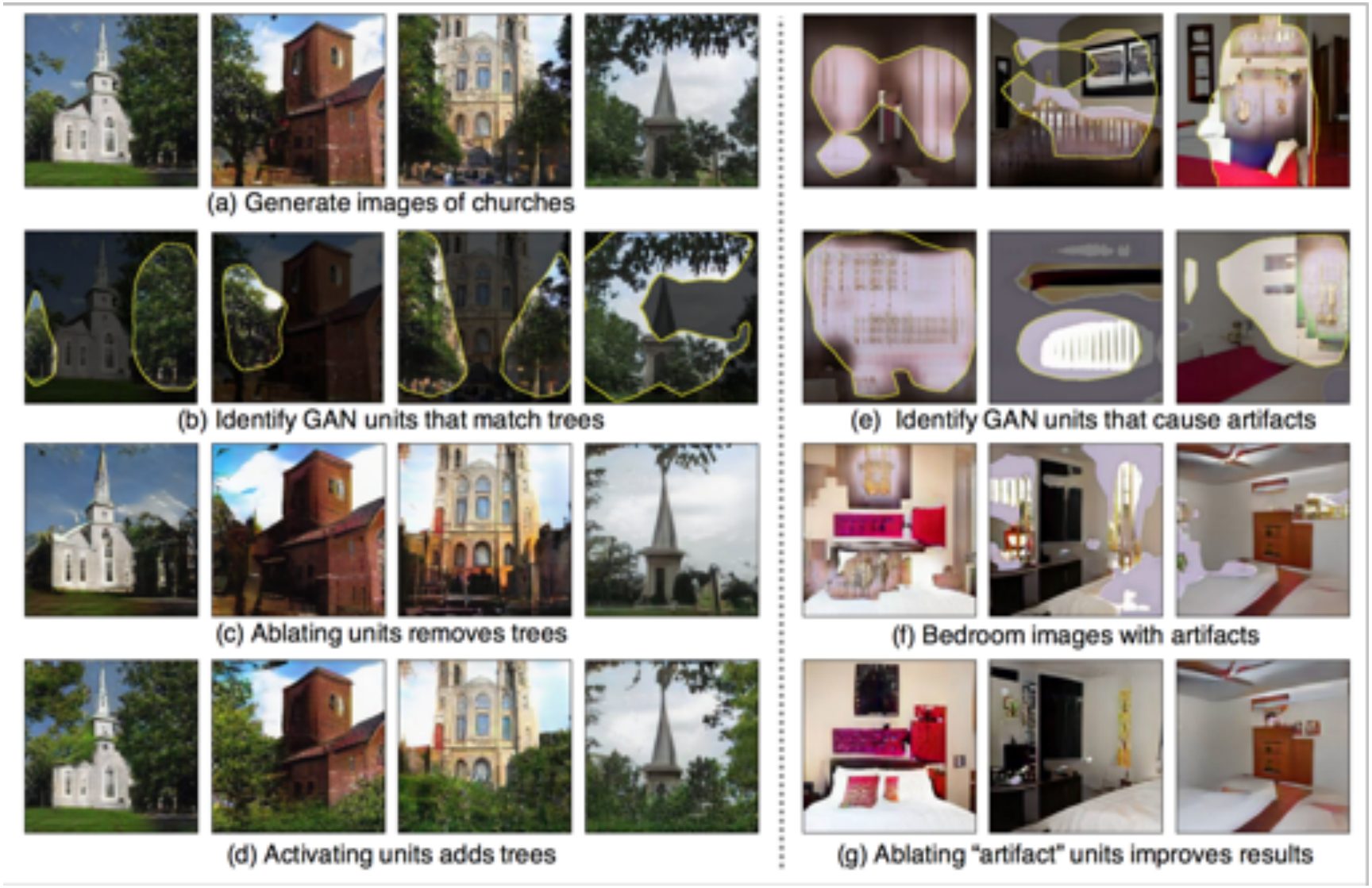
David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba

MIT & IBM

ICLR 2019
accepted as poster

Score: 7 7 8

Outlines


(a) Generate images of churches
(b) Identify GAN units that match trees
(c) Ablating units removes trees
(d) Activating units adds trees
(e) Identify GAN units that cause artifacts
(f) Bedroom images with artifacts
(g) Ablating "artifact" units improves results

how object information is encoded

- Use representation to describe the tensor *r* output from a particular layer of the generator G

$$\mathbf{x} = f(\mathbf{r}) = f(h(\mathbf{z})) = G(\mathbf{z}).$$

- For any concept *c*, it's possible to factor *r* at locations *P* into two components

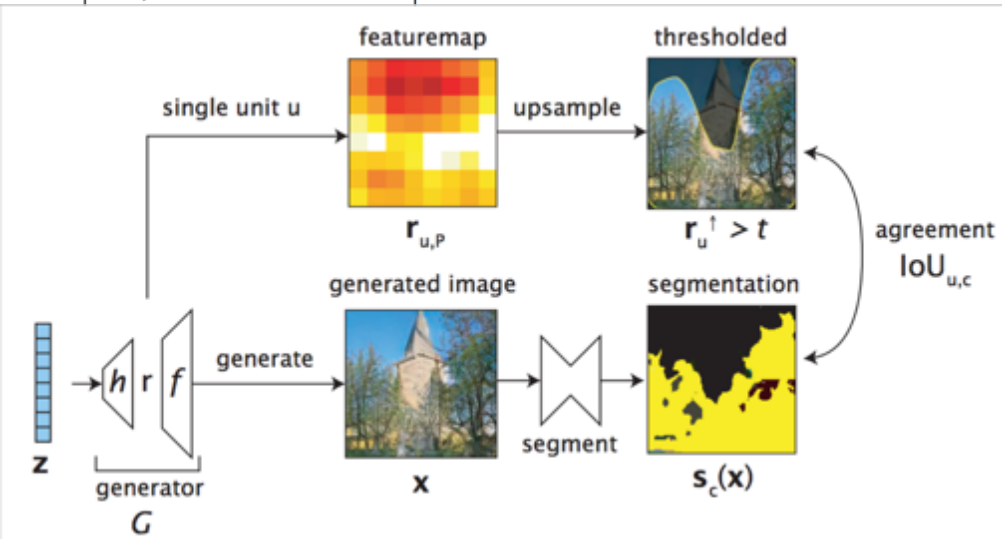
$$\mathbf{r}_{\mathbb{U},\mathrm{P}} = (\mathbf{r}_{\mathrm{U},\mathrm{P}}, \mathbf{r}_{\overline{\mathrm{U}},\mathrm{P}}),$$

- The structure of r in two phases
  - Dissection -> identify the classes that have an explicit representation in r by measuring the agreement between individual units of *r* and every class *c*
  - Intervention -> identify causal sets of units and measure causal effects between units and object classes by forcing sets of units on and off

Characterizing Units By Dissection

- Select a universe of concepts *c* for which we have a semantic segmentation $s_c(x)$ for each class ($s_c(x)$ is a binary mask where each pixel indicates the presence of class c in the generated image x)
- Quantify the spatial agreement between the unit *u*'s thresholded featuremap and a concept *c*'s segmentation [Bau et al. (2017)]
- Identified an object class that a set of units match closely

$$\text{IoU}_{u,c} \equiv \frac{\mathbb{E}_{\mathbf{z}}\left|(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \wedge \mathbf{s}_c(\mathbf{x})\right|}{\mathbb{E}_{\mathbf{z}}\left|(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \vee \mathbf{s}_c(\mathbf{x})\right|}, \text{where } t_{u,c} = \arg\max_t \frac{\text{I}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t; \mathbf{s}_c(\mathbf{x}))}{\text{H}(\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t, \mathbf{s}_c(\mathbf{x}))},$$



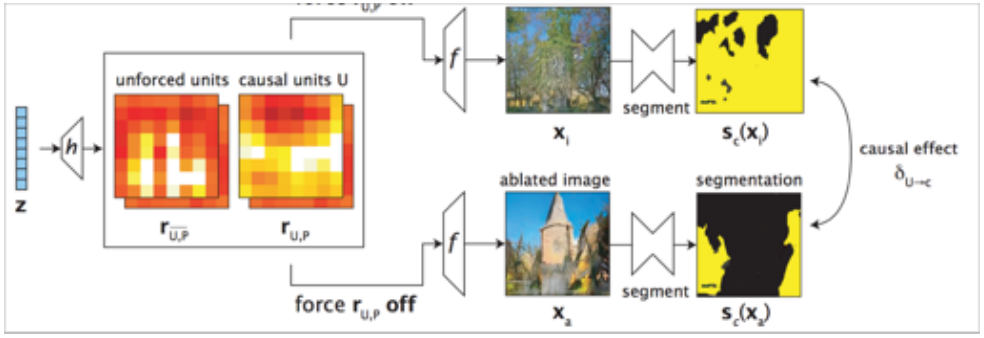
Measuring Causal Relationships Using Intervention

- The target is to identify combinations of units that cause an object
  - a unit that correlates highly with an output object might not actually cause that output
- Test whether a set of units *U* in *r* cause the generation of *c* by forcing the units of *U* on and off

Original image :	$\mathbf{x} = G(\mathbf{z}) \equiv f(\mathbf{r}) \equiv f(\mathbf{r}_{\mathrm{U,P}}, \mathbf{r}_{\overline{\mathrm{U,P}}})$
Image with U ablated at pixels P :	$\mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}})$
Image with U inserted at pixels P :	$\mathbf{x}_i = f(\mathbf{k}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

- This causality can be quantified by comparing the presence of trees in x_i and x_a and averaging effects over all locations and images.

$$\delta_{\mathrm{U}\to c} \equiv \mathbb{E}_{\mathbf{z},\mathrm{P}}[\mathbf{s}_c(\mathbf{x}_i)] - \mathbb{E}_{\mathbf{z},\mathrm{P}}[\mathbf{s}_c(\mathbf{x}_a)],$$



Finding sets of units with high ACE

- While these measures can be applied to a single unit, it's found that objects tend to depend on more than one unit
- Given a representation *r* with *d* units, exhaustively searching for a fixed-size set *U* with high $\delta_{U \to c}$ is prohibitive

Image with partial ablation at pixels P :	$\mathbf{x}'_a = f((\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\bar{\mathrm{P}}})$
Image with partial insertion at pixels P :	$\mathbf{x}'_i = f(\boldsymbol{\alpha} \odot \mathbf{k} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\bar{\mathrm{P}}})$
Objective :	$\delta_{\boldsymbol{\alpha} \to c} = \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_i)\right] - \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_a)\right],$

- We optimize α over the following loss with an L2 regularization

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}}(-\delta_{\boldsymbol{\alpha} \to c} + \lambda||\boldsymbol{\alpha}||_2),$$

Emergence of individual unit object detectors

Thresholding unit #65 layer 3 of a dining room generator matches 'table' segmentations with IoU=0.34.

Thresholding unit #37 layer 4 of a living room generator matches 'sofa' segmentations with IoU=0.29.

Interpretable units for different scene categories

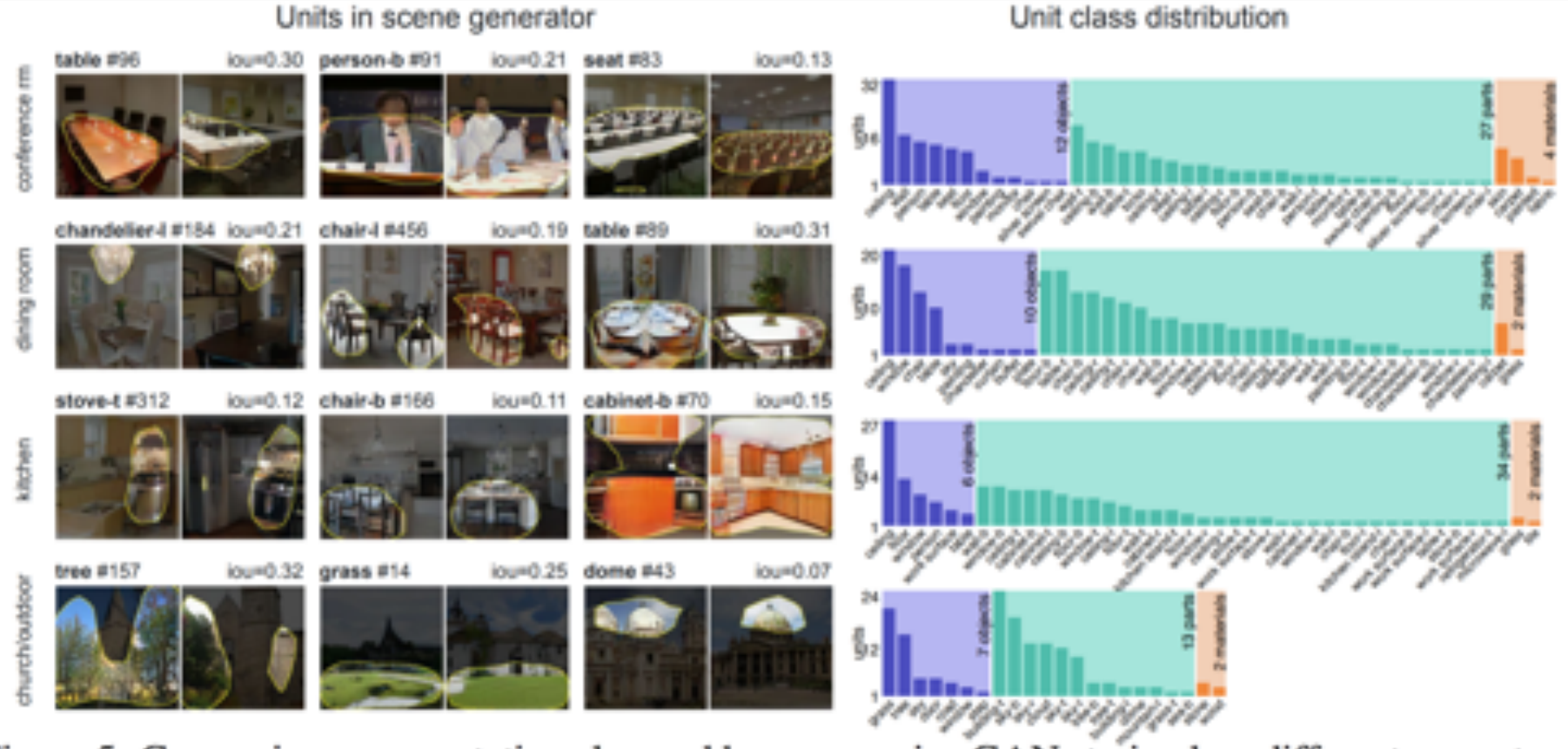


Figure 5: Comparing representations learned by progressive GANs trained on different scene types. The units that emerge match objects that commonly appear in the scene type: seats in conference rooms and stoves in kitchens. Units from `layer4` are shown. A unit is counted as a class predictor if it matches a supervised segmentation class with pixel accuracy > 0.75 and IoU > 0.05 when upsampled and thresholded. The distribution of units over classes is shown in the right column.

Interpretable units for different network layers

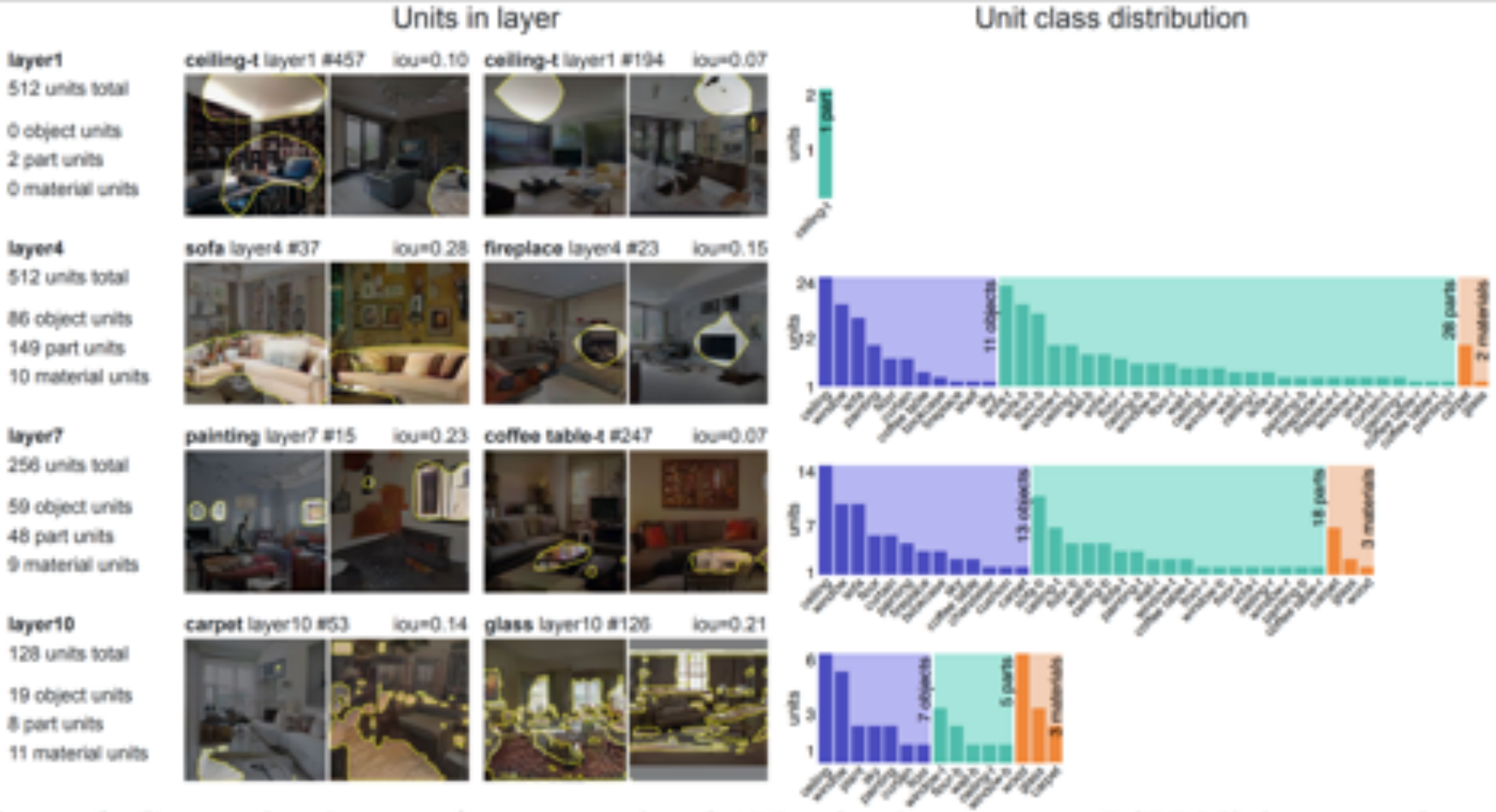


Figure 6: Comparing layers of a progressive GAN trained to generate LSUN living room images. The output of the first convolutional layer has almost no units that match semantic objects, but many objects emerge at layers 4-7. Later layers are dominated by low-level materials, edges and colors.

Interpretable units for different GAN models

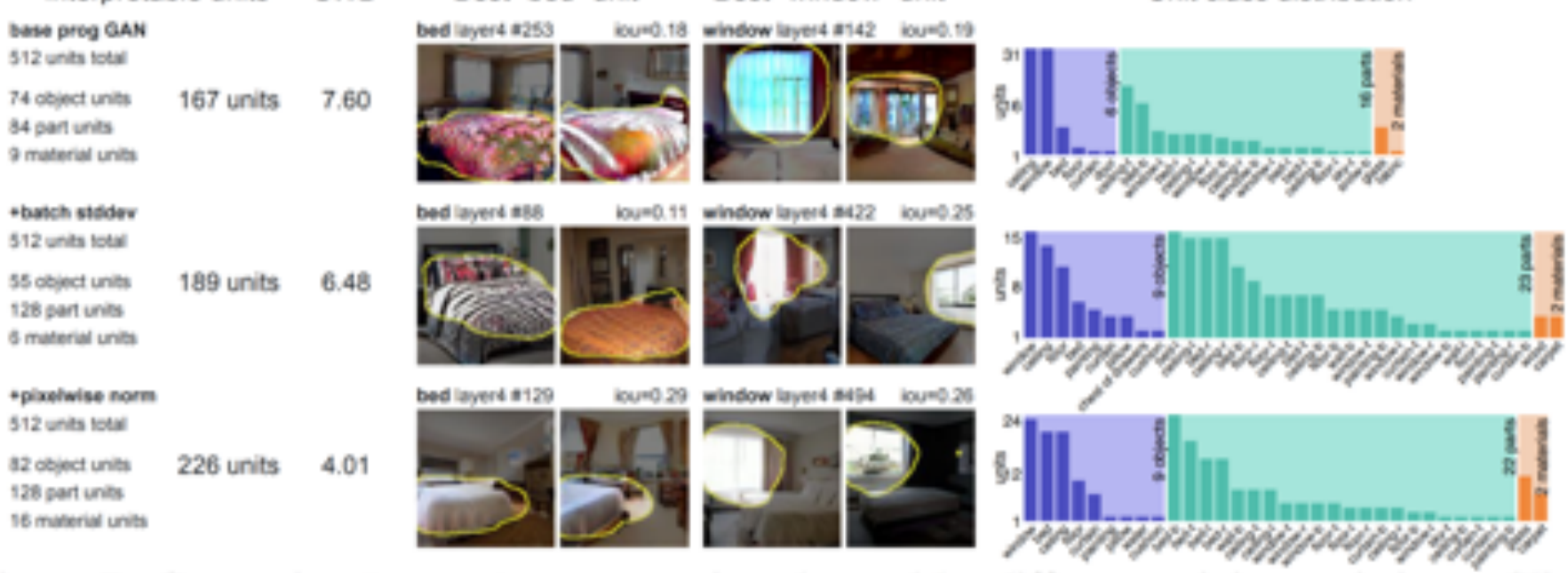


Figure 7: Comparing `layer4` representations learned by different training variations. Sliced Wasserstein Distance (SWD) is a GAN quality metric suggested by Karras et al. (2018): lower SWD indicates more realistic image statistics. Note that as the quality of the model improves, the number of interpretable units also rises. Progressive GANs apply several innovations including making the discriminator aware of minibatch statistics, and pixelwise normalization at each layer. We can see batch awareness increases the number of object classes matched by units, and pixel norm (applied in addition to batch stddev) increases the number of units matching objects.

Diagnosing And Improving GANs

Figure 8: (a) We show two example units that are responsible for visual artifacts in GAN results. There are 20 units in total. By ablating these units, we can fix the artifacts in (b) and significantly improve the visual quality as shown in (c).

Locating Causal Units With Ablation

conference room

church

kitchen

living room

bedroom

Figure 10: Comparing the effect of ablating 20 window-causal units in GANs trained on five scene categories. In each case, the 20 ablated units are specific to the class and the generator and independent of the image. In some scenes, windows are reduced in size or number rather than eliminated, or replaced by visually similar objects such as paintings.

Locating Causal Units With Ablation(Cont's)

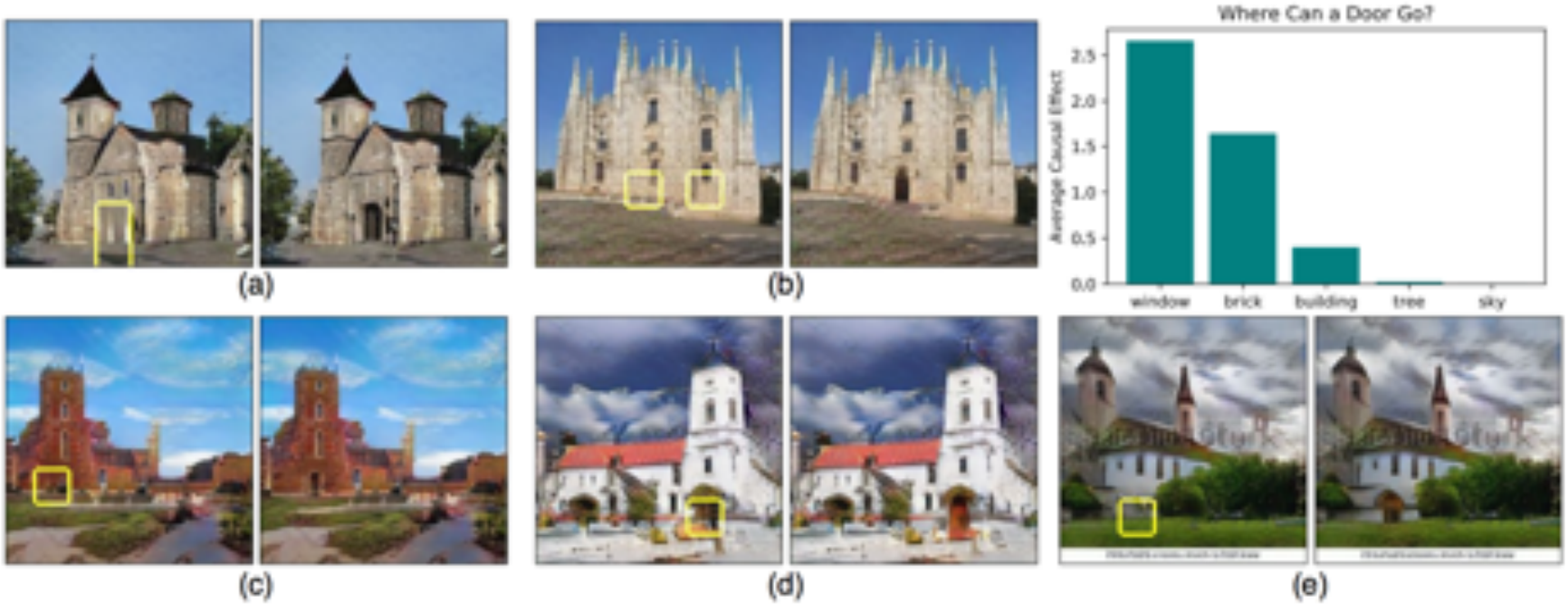


Figure 11: Inserting door units by setting 20 causal units to a fixed high value at one pixel in the representation. Whether the door units can cause the generation of doors is dependent on its local context: we highlight every location that is responsive to insertions of door units on top of the original image, including two separate locations in (b) (we intervene at left). The same units are inserted in every case, but the door that appears has a size, alignment, and color appropriate to the location. One way to add door pixels is to emphasize a door that is already present, resulting in a larger door (d). The chart summarizes the causal effect of inserting door units at one pixel with different contexts.

Conclusion

- Identity the causal units in deep generative models using intervention
- Adopt intervention into neural machine translation to interpret the internal units
- Similar to selecting subset of units with high ACE, assign a weight vector for hidden units and do optimization to choose the non-trivial units with SGD

Thanks & QA