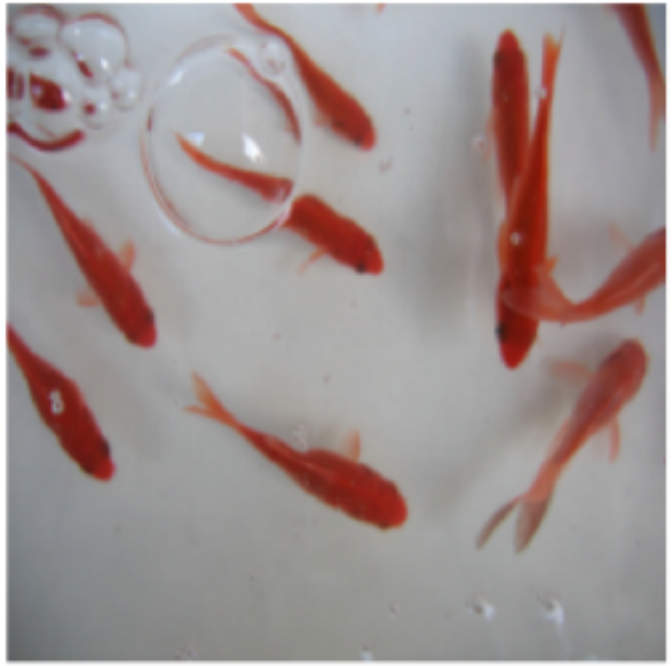
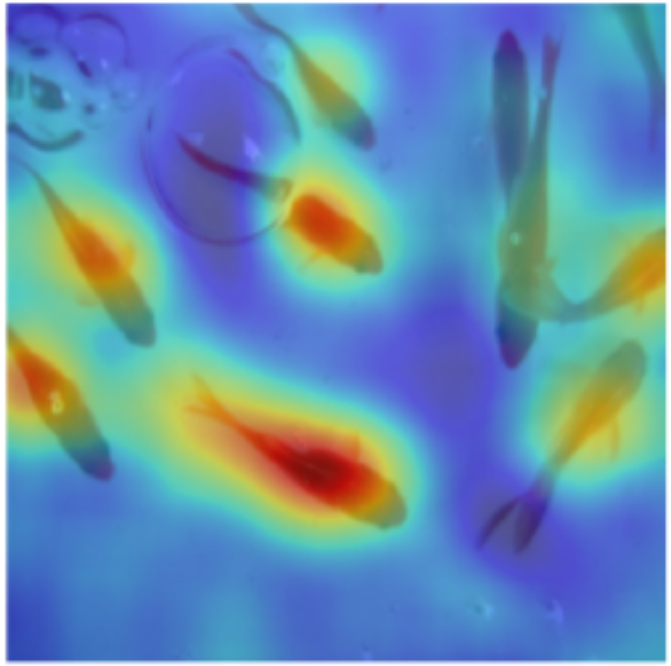
Understanding units in a deep neural network

Jaehyoung Hong

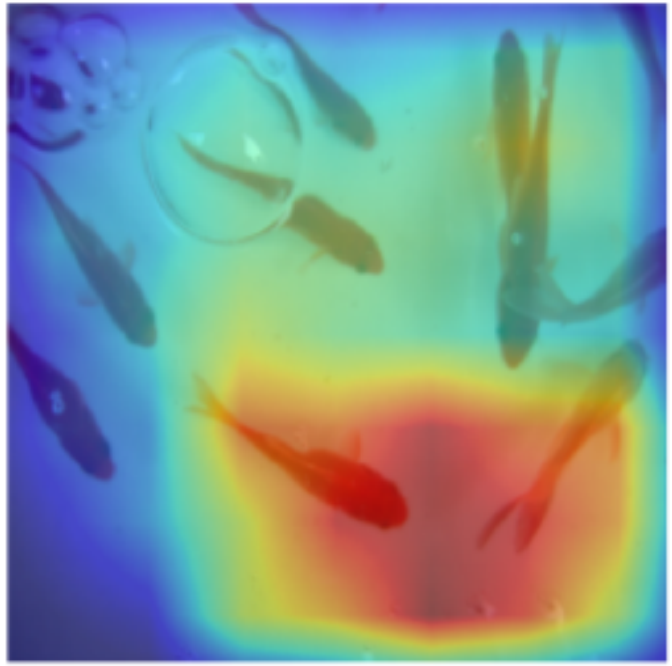
There are three major classes of explainable AI (X-AI)



(a) Input



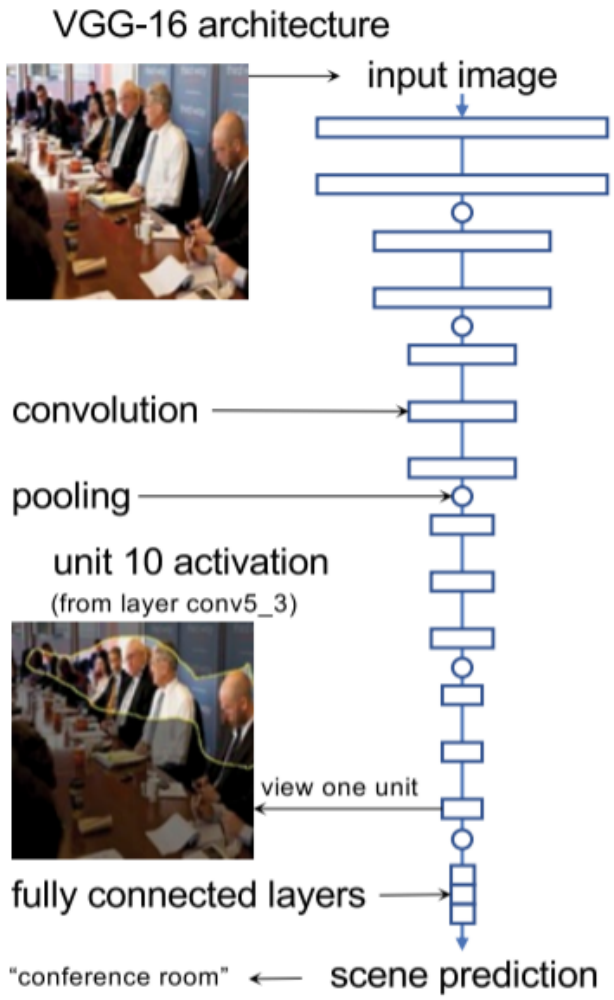
(b) RISE (ours)



(c) GradCAM

Which parts of input give such result?

Network dissection focuses on what a network is looking for and why



- Which layer is important to classify "conference room"?
- Which unit is important to classify "conference room"?

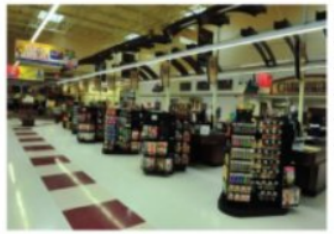
Similar to mask based X-AI, delete some units or layers
(Previously, using auxiliary model)

Directly interpret the internal computation of the network itself

First, identify individual units of network that is trained for scene classification

GT: cafeteria
top-1: cafeteria (0.179)
top-2: restaurant (0.167)
top-3: dining_hall (0.091)
top-4: coffee_shop (0.086)
top-5: restaurant_patio (0.080)

GT: classroom
top-1: locker_room (0.585)
top-2: lecture_room (0.135)
top-3: conference_center (0.061)
top-4: classroom (0.033)
top-5: elevator door (0.025)



GT: drugstore
top-1: supermarket (0.286)
top-2: hardware_store (0.248)
top-3: drugstore (0.120)
top-4: department_store (0.087)
top-5: pharmacy (0.052)

GT: natural canal
top-1: swamp (0.529)
top-2: marsh (0.232)
top-3: natural canal (0.063)
top-4: lagoon (0.047)
top-5: rainforest (0.029)

GT: creek
top-1: forest broadleaf (0.307)
top-2: forest_path (0.208)
top-3: creek (0.086)
top-4: rainforest (0.076)
top-5: cemetery (0.049)

GT: greenhouse indoor
top-1: greenhouse indoor (0.479)
top-2: greenhouse outdoor (0.055)
top-3: botanical_garden (0.044)
top-4: assembly_line (0.025)
top-5: vegetable_garden (0.022)

GT: chalet
top-1: ski_resort (0.141)
top-2: ice_floe (0.129)
top-3: igloo (0.114)
top-4: balcony exterior (0.103)
top-5: courtyard (0.083)

GT: crosswalk
top-1: crosswalk (0.720)
top-2: plaza (0.060)
top-3: street (0.055)
top-4: shopping_mall indoor (0.039)
top-5: bazaar outdoor (0.021)

GT: market outdoor
top-1: promenade (0.569)
top-2: bazaar outdoor (0.137)
top-3: boardwalk (0.118)
top-4: market outdoor (0.074)
top-5: flea_market indoor (0.029)

- Task : scene classification
- Data : Place365, images each labeled with one of 365 scene classes (Above)
- Network : VGG-16 classifier trained for classifying Place365

Intersection over union (IoU) ratio quantifies the agreement between concept c (1825 classes) and unit u

- $a_u(x, p)$: Calculated output of unit u given image position p of image x

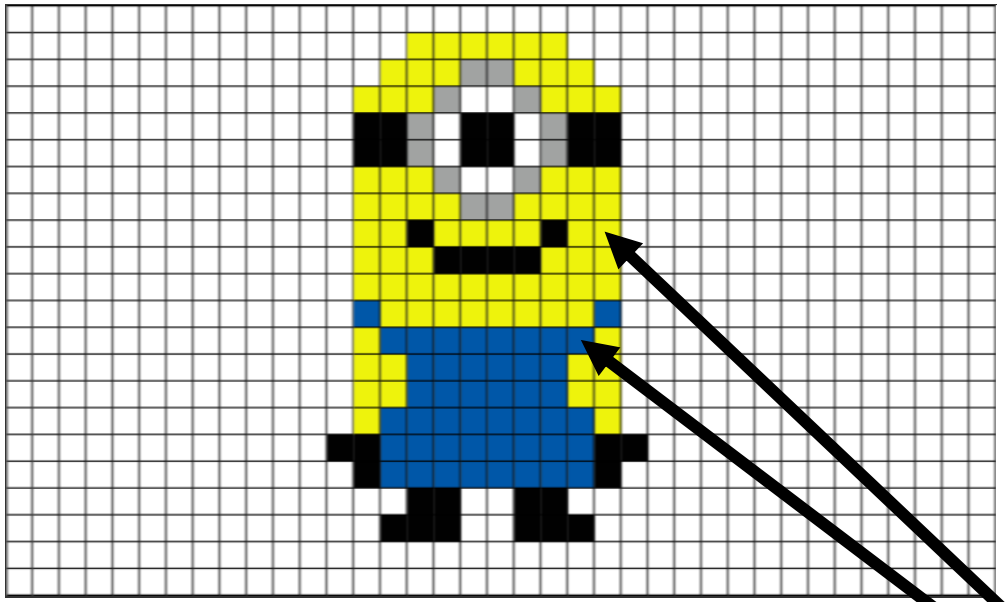
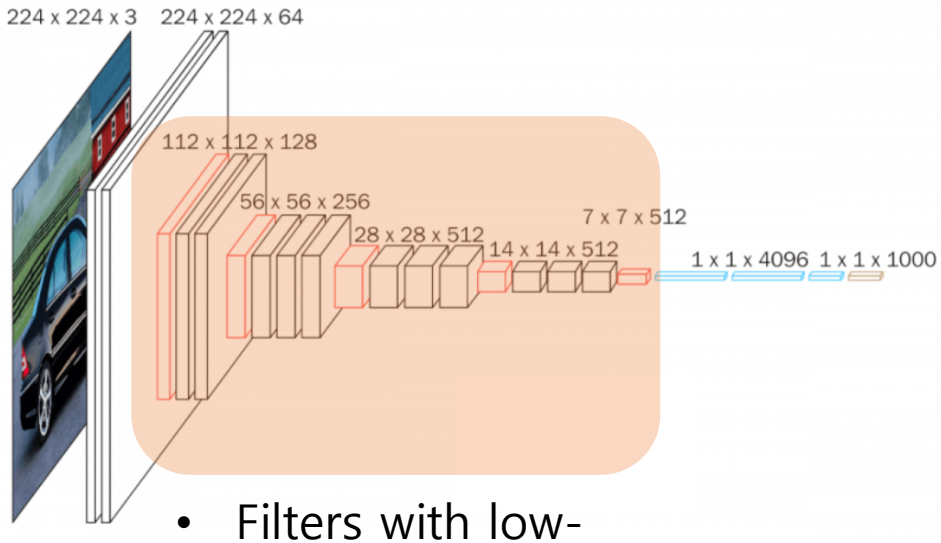


Image x

Position p

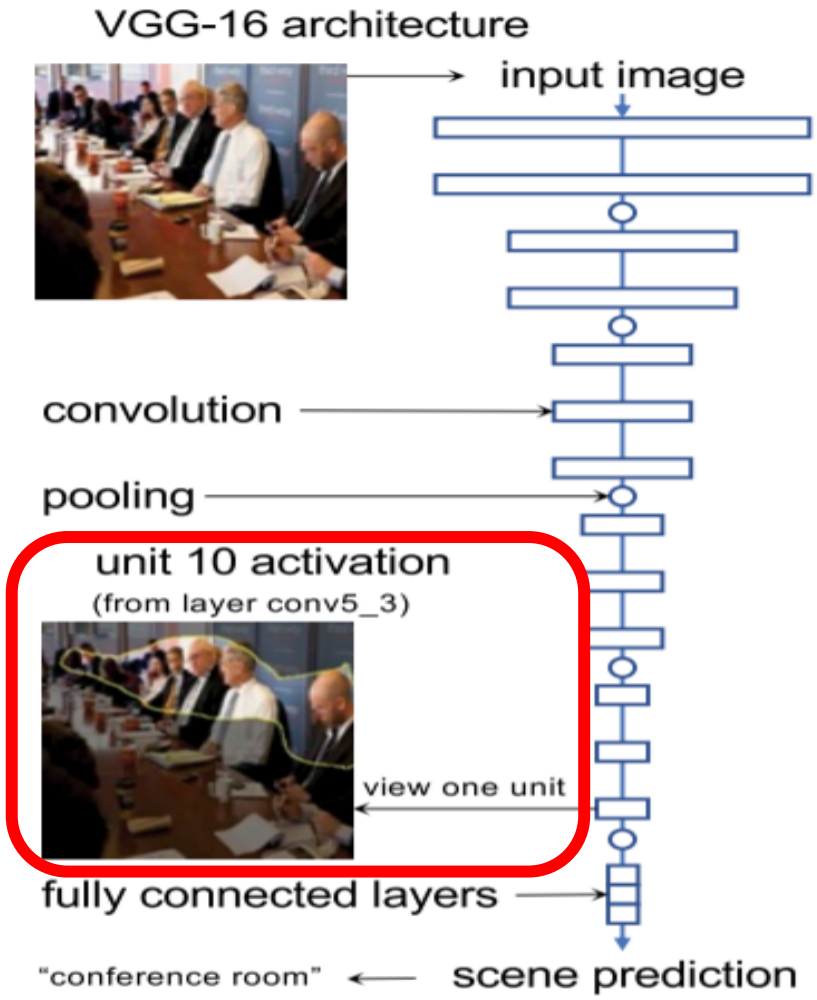


- Filters with low-resolution output

→ Bilinear up sampling

Intersection over union (IoU) ratio quantifies the agreement between concept c (1825 classes) and unit u

- $t_u = max_t P_{x,p}[a_u(x,p) > t]$: The top 1% quantile level for a_u



✓ Highlight the activation region : $\{p|a_u(x,p) > t_u\}$

✓ Unit 10 finds the **"heads"** in the image

How about matching such activated part with **concept**

Intersection over union (IoU) ratio quantifies the agreement between concept c (1825 classes) and unit u

<Previously trained Computer vision segmentation model>

- $s_c : (x, p) \rightarrow \{0,1\}$
- Trained to predict the presence image of the visual concept c (1825 classes) within image x at position p

$$\mathrm{IoU}_{u,c} = \frac{\mathbb{P}_{x,p}\left[s_c(x,p) \wedge \left(a_u(x,p) > t_u\right)\right]}{\mathbb{P}_{x,p}\left[s_c(x,p) \vee \left(a_u(x,p) > t_u\right)\right]}.$$

➡ Agreement between activated region and most important concept of that region

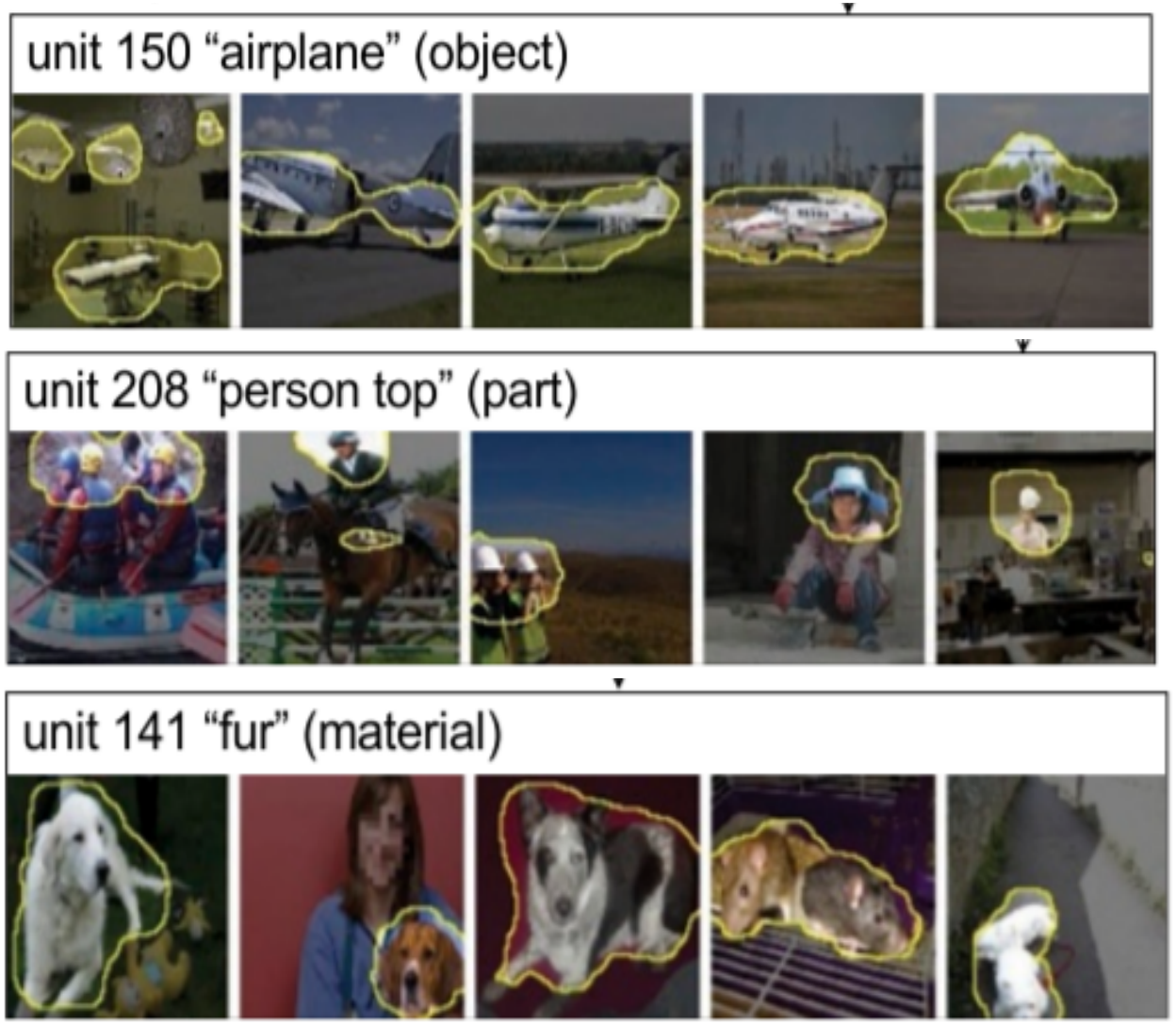
unit 10 activation
(from layer conv5_3)

view one unit

"Head"

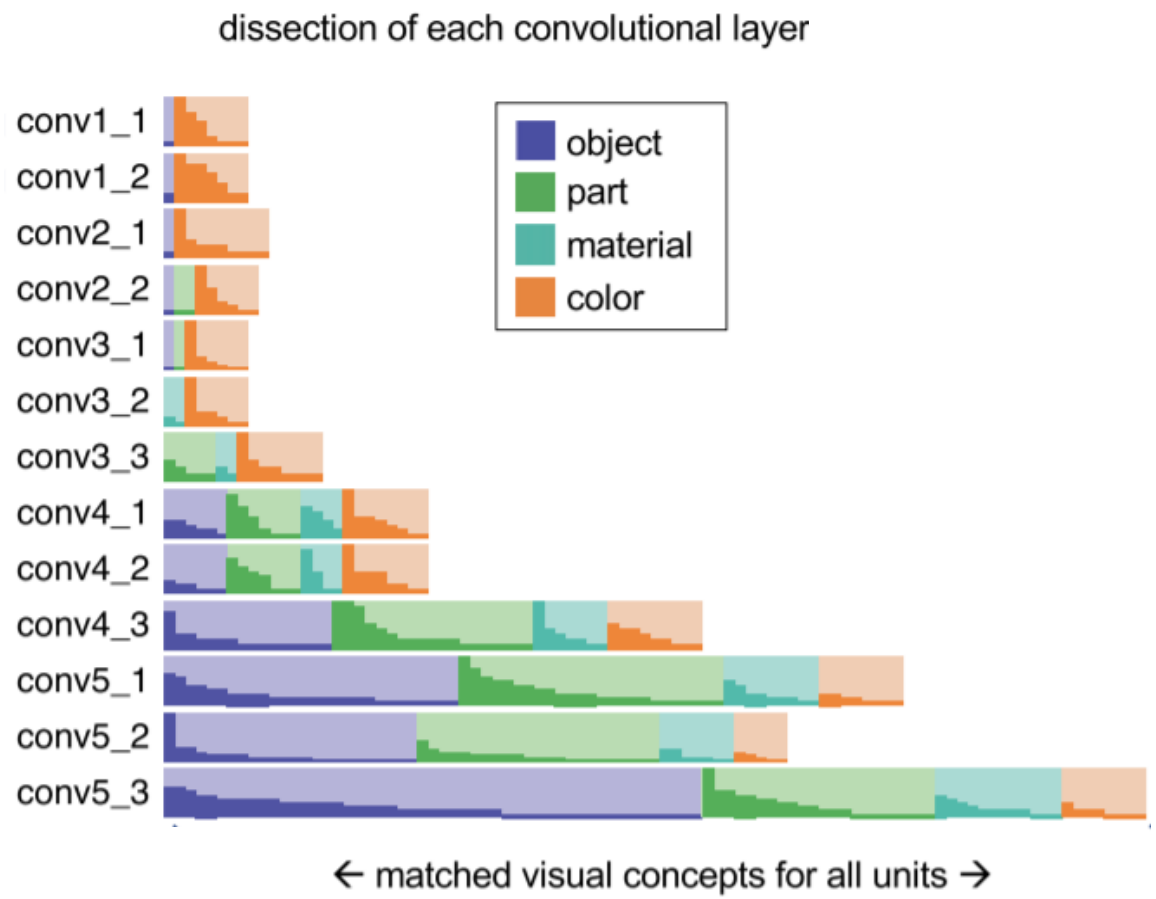
✓ **High IOU : The unit finds such concept**

Dissection (Finding matched concept of each layer and unit) shows difference between each units



✓ Units in same layer

Dissection shows difference between each layers



- ✓ **Object :** Airplane, House…
- ✓ **Part :** Person-top, leg….
- ✓ **Material :** Fur, Skin….
- ✓ **Color :** Red, Blue…

- ✓ Conv5_1 detected the largest number of the ‘part’
- ✓ Last layer detected the largest number of the ‘object’

Dissection shows difference between each label context

- Stylized image changes IoU of some units while IoU of some units are unchanged

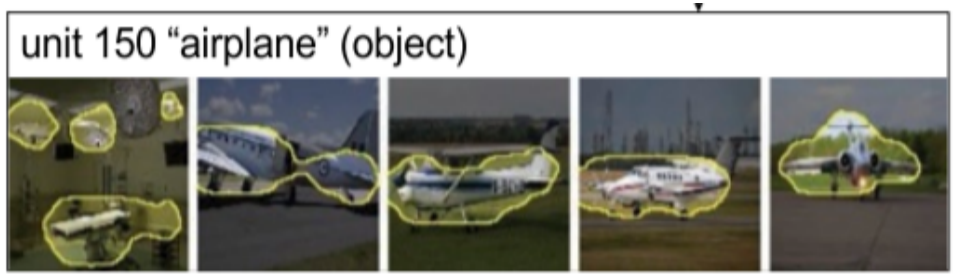
<Shape-sensitive unit>

<Texture-sensitive unit>

✓ IoU(s,u) : Difference of IoU for original and stylized image

Interestingly, object detectors emerge despite the absence of object labels in the training task

- Place365 : Only scene label exists ("Airport terminal", "Airfield" rather than "Airplane")

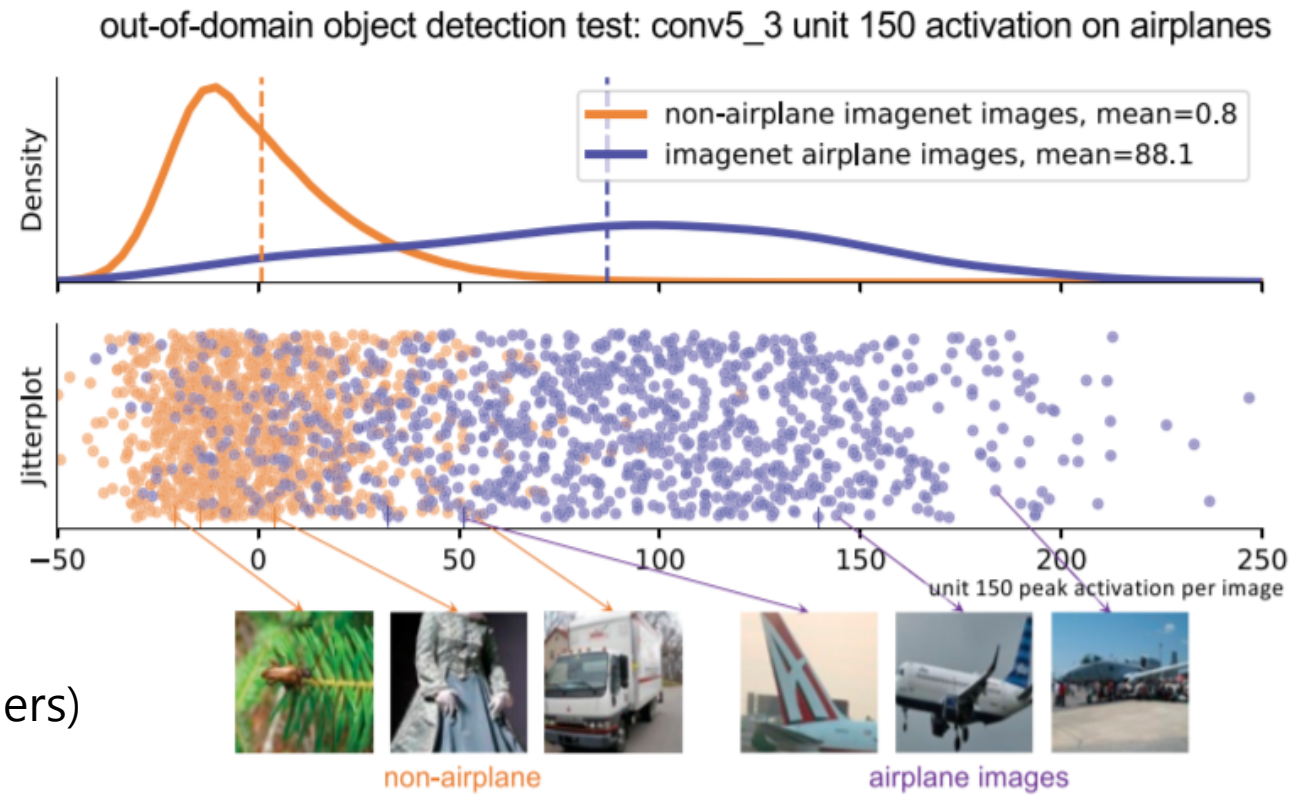


$\boldsymbol{IoU = 9\%}$

Another scene image dataset
: $\max_p a_u(x, p)$ for all x

Simple model : Peak activation > 23.4

85.6% accuracy ("Airliner" or "Warplane" vs Others)

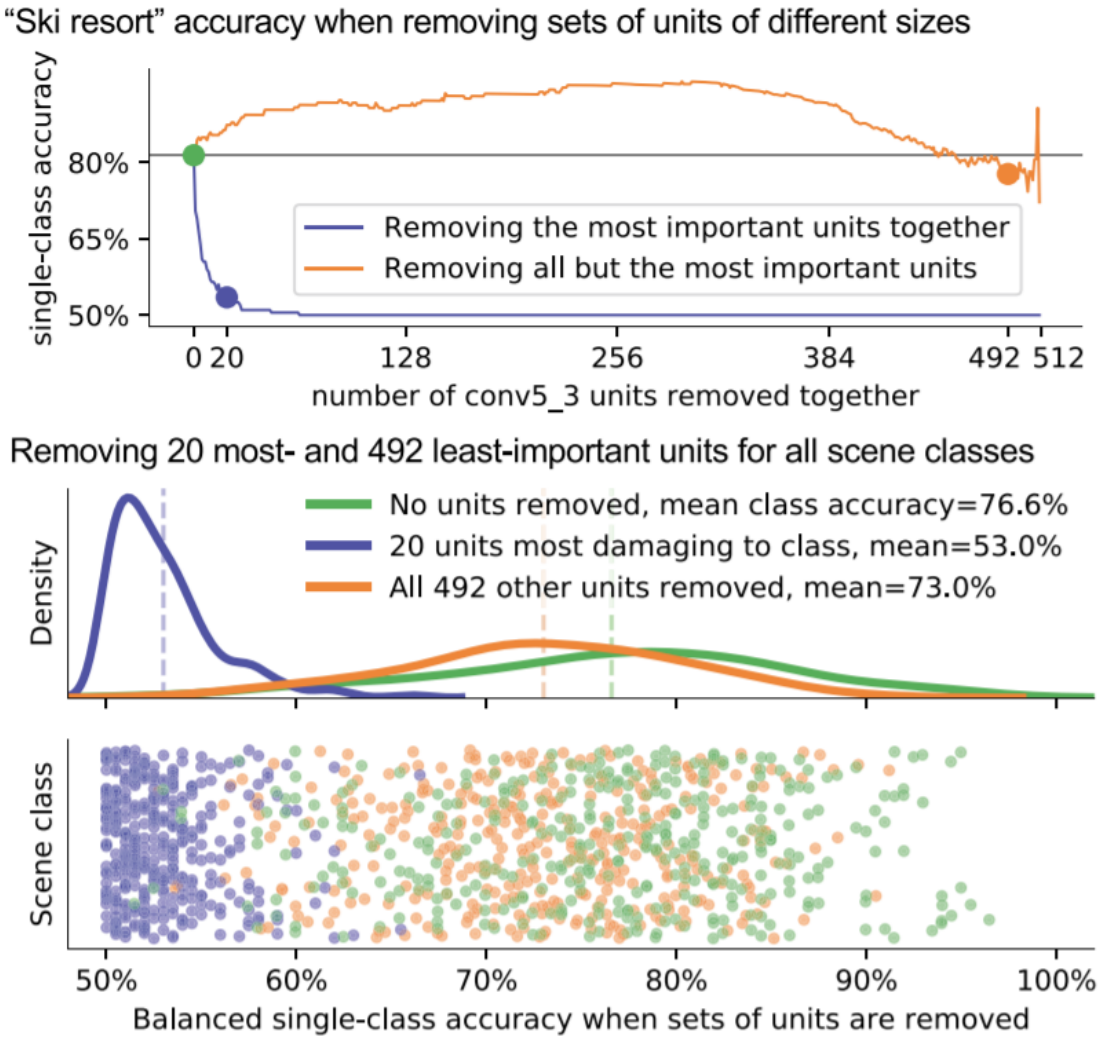


Removing top-important object detector units for scene classification decreases specific accuracy while overall accuracy is unchanged

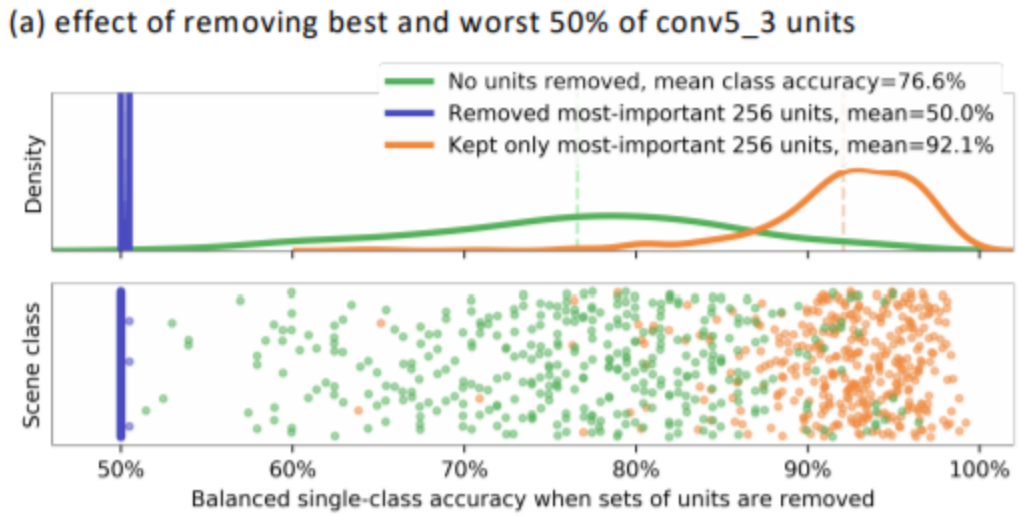
- Removing "each unit" by forcing its output 0
  → Find most important unit (Causes most decrement of accuracy for two-way classification)

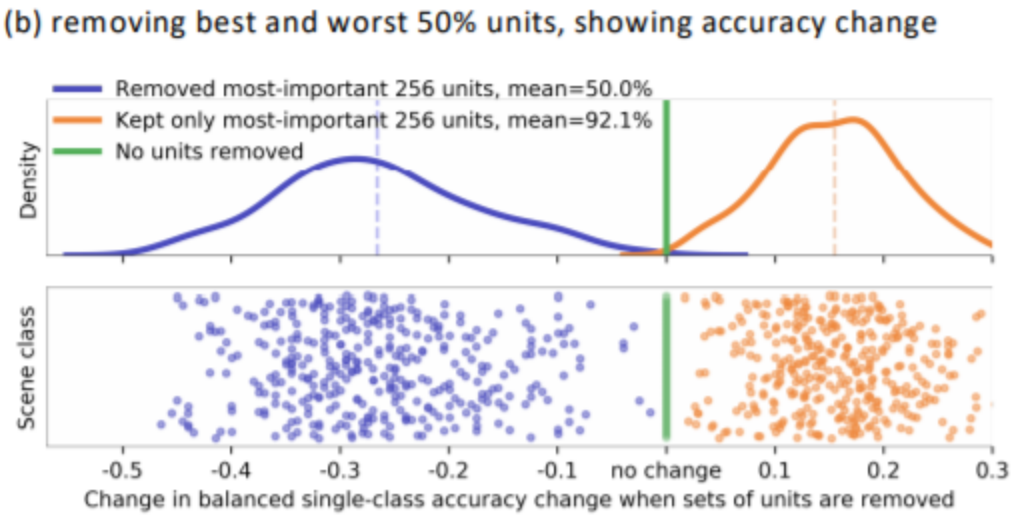
	Balanced single-class 'ski resort' accuracy	All-class accuracy
Unchanged vgg-16:	81.4%	53.3%
4 most important units removed:	64.0%	53.2%
20 most important units removed:	53.5%	52.6%
492 least important units removed:	77.7%	2.1%
Chance level	50.0%	0.27%

Change by removing important units are larger than that by removing all other units



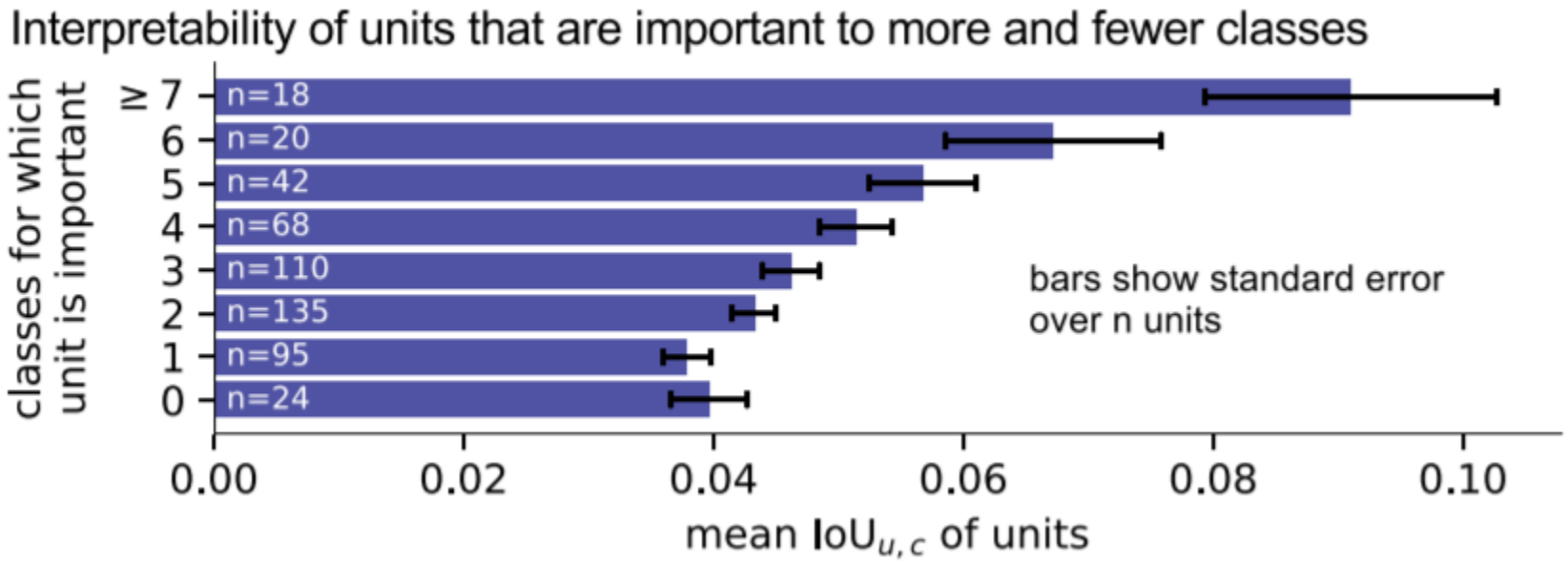
Change by removing important units are larger than that by removing all other units





✓ Removing useless unit makes better single-class accuracy

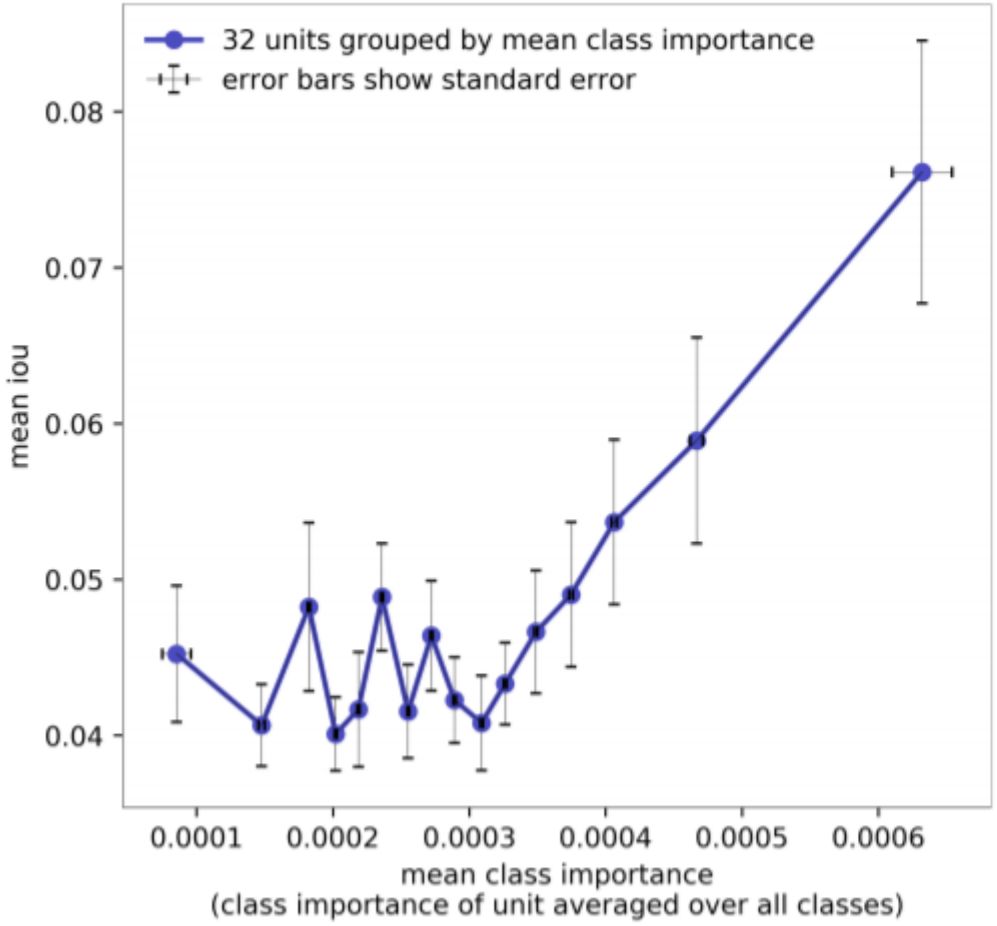
Most interpretable units are those that are important to many different classes



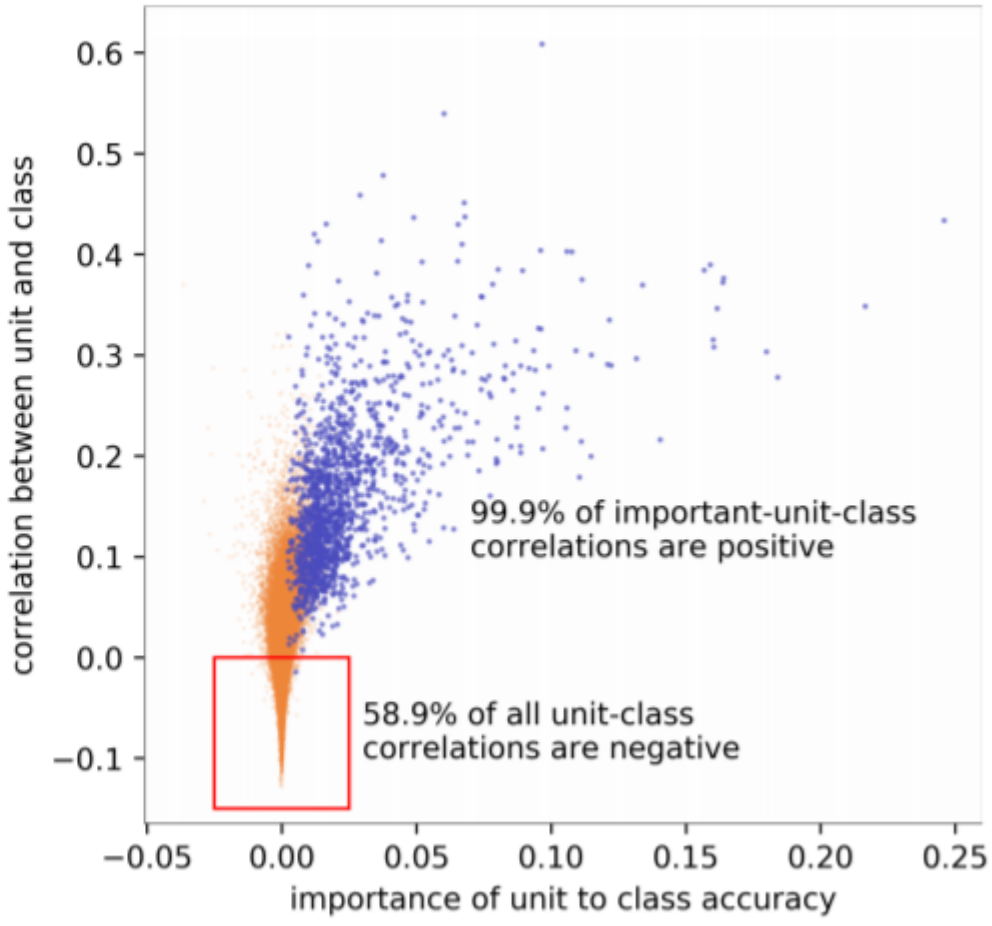
- ✓ Units which are important for one units have less mean IoU
- ✓ More clear interpretation can uses for many classes

Most interpretable units are those that are important to many different classes

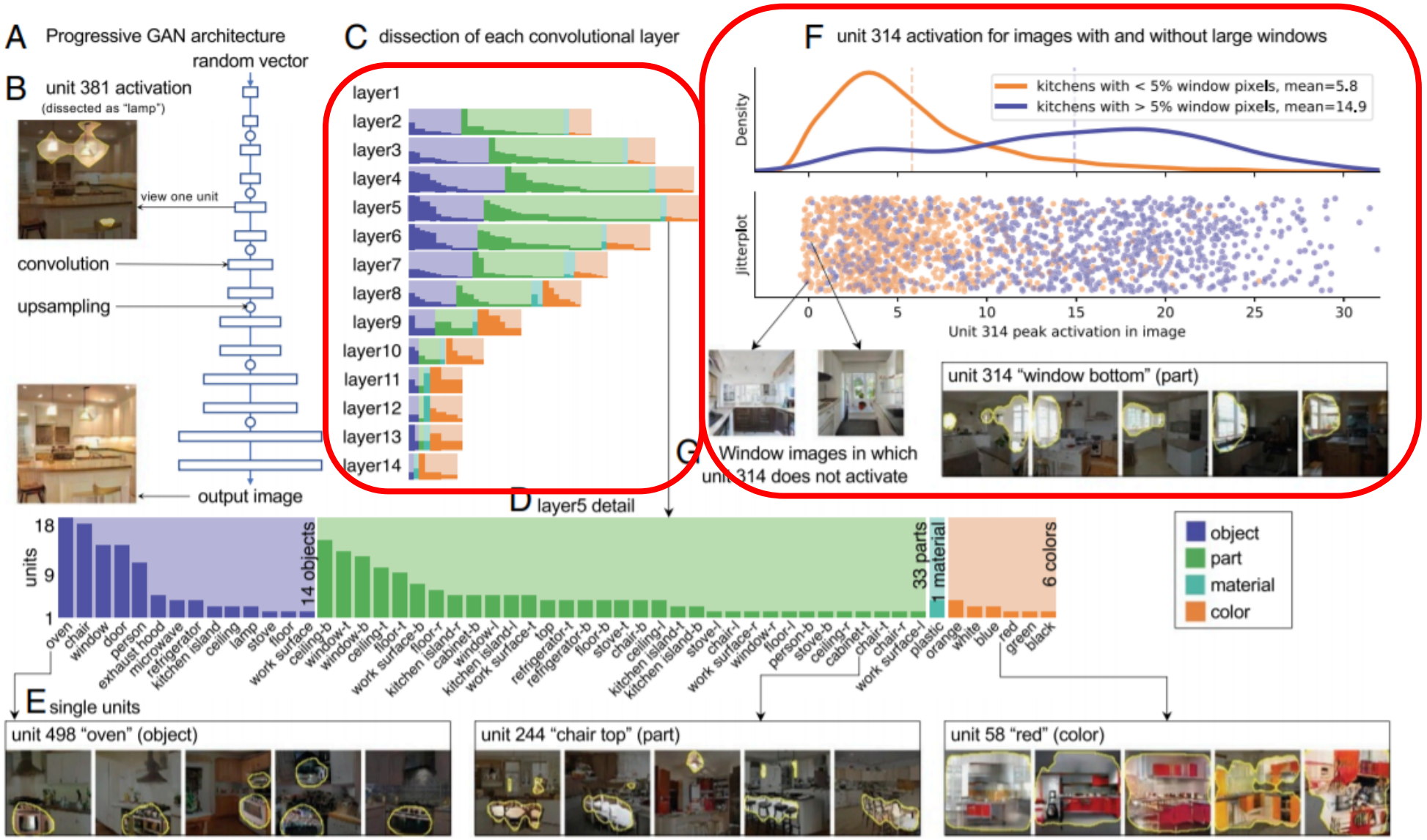
(g) interpretability vs mean class importance



(h) unit-class correlation for top-4 important units vs other units



Similar application to GAN provides similar results

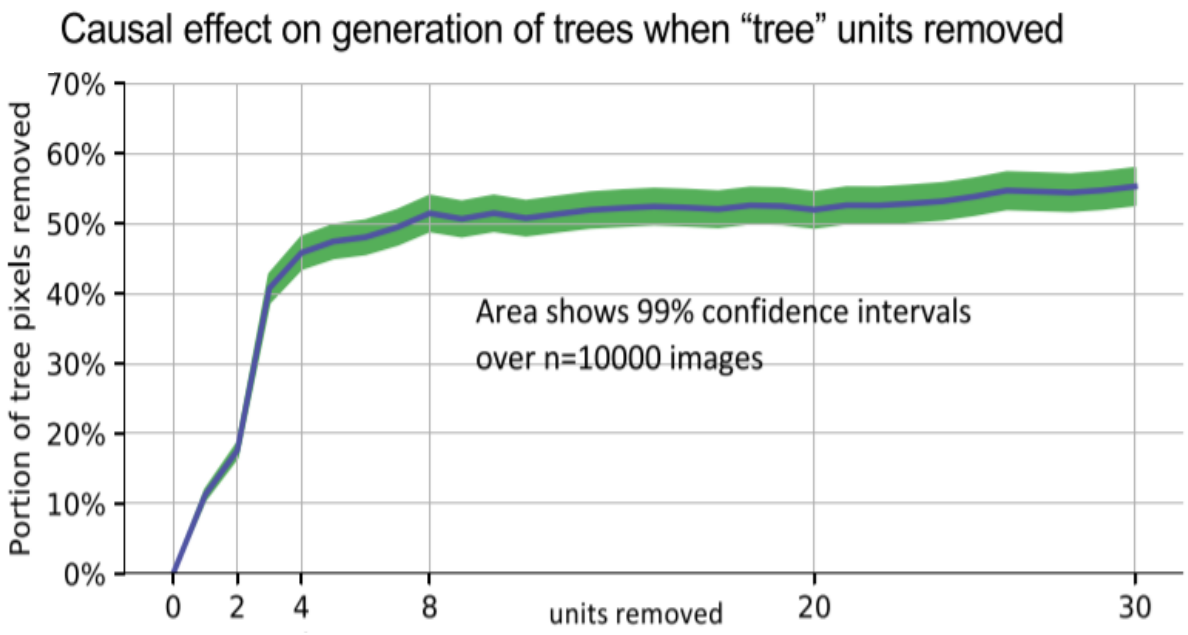


✓ Unit 314 is activated more when make large window

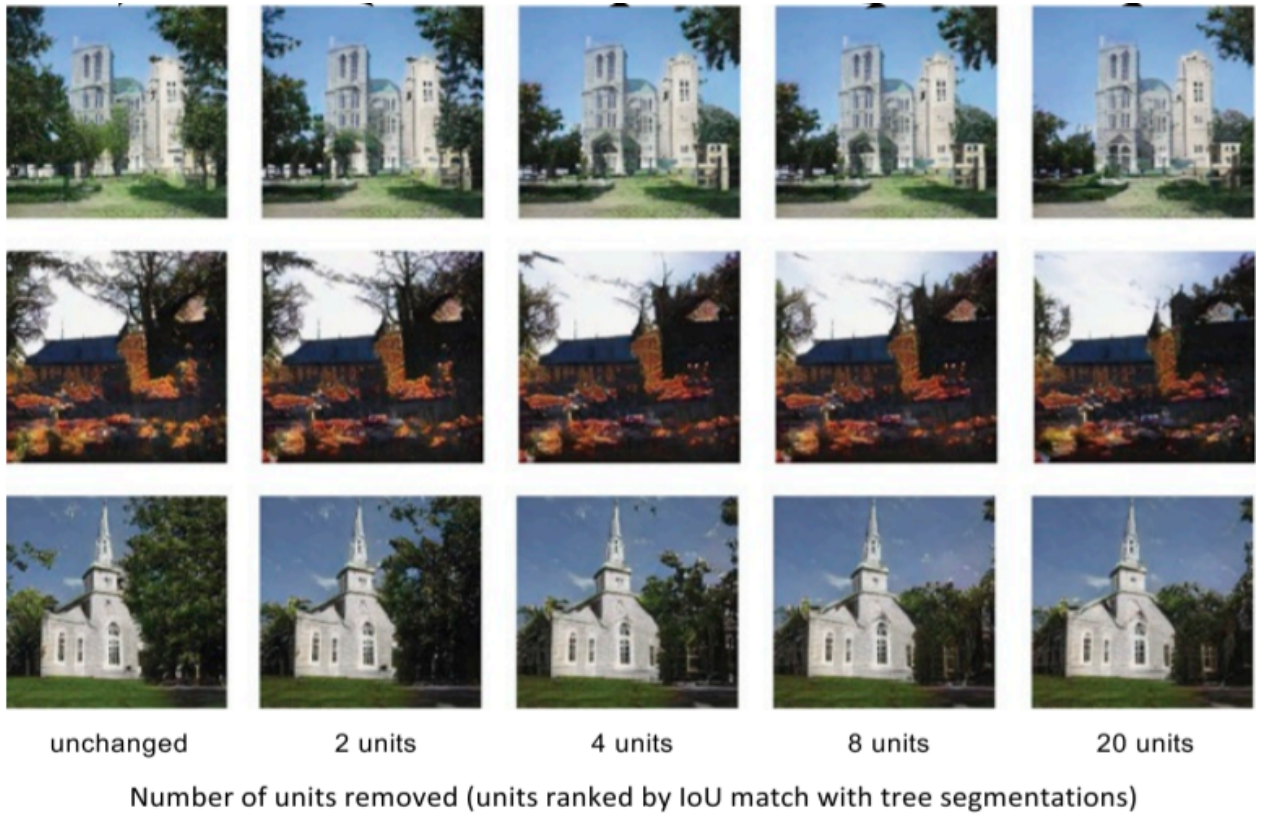
✓ GAN trained to imitate Kitchen image

Removing top important units for ‘tree’ causes less tree in the generated image

- Top important unit : According to $IoU_{u,tree}$ (Different with previous criteria for image classification)



- **Not completely removed**



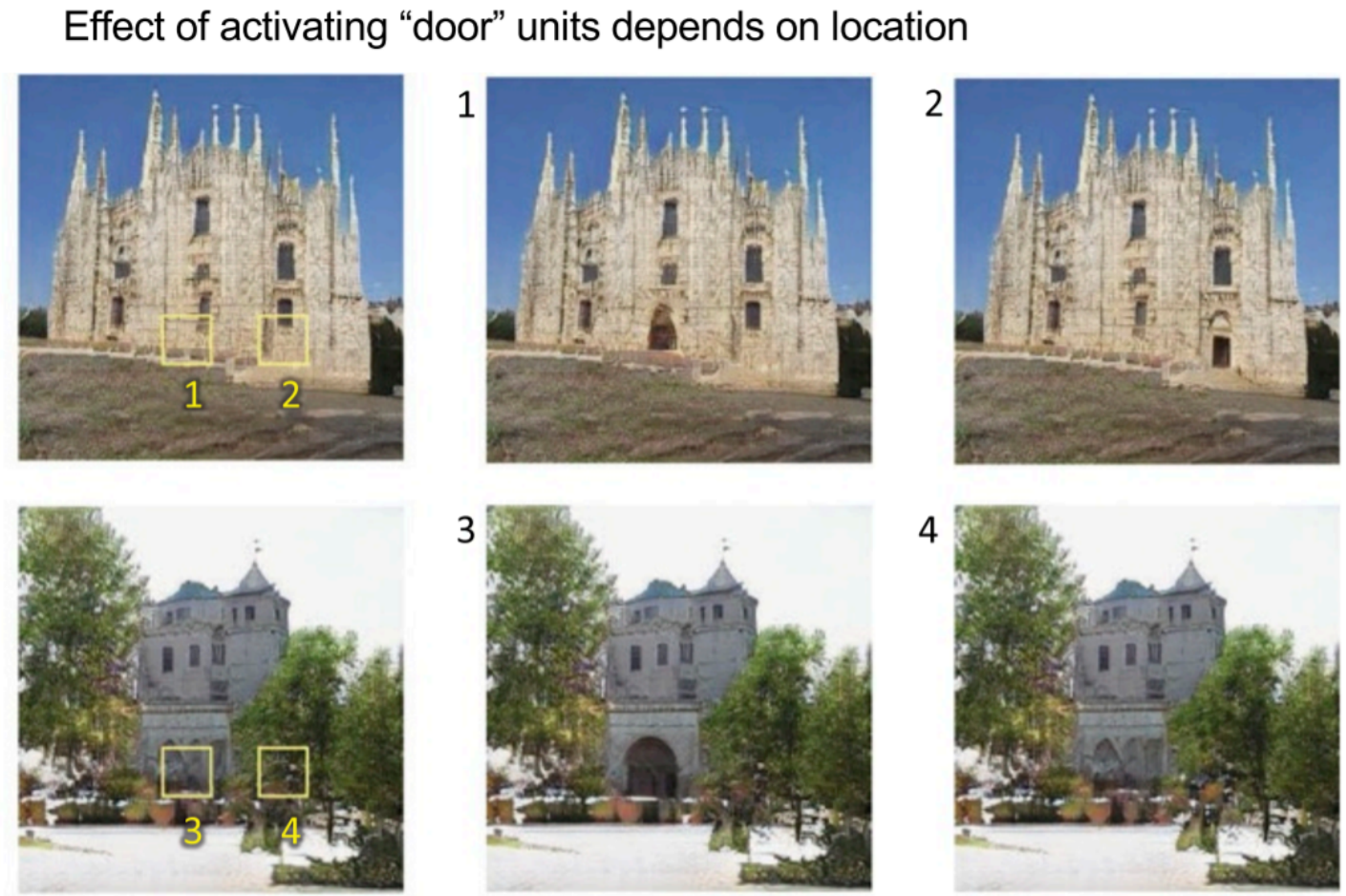
GAN learns more than flat summarization of visible pixel patterns

Number of units removed (units ranked by IoU match with tree segmentations)

- When tree is removed, GAN generates the detail of the building, rather than make it blank space

Activating specific units increase their concept in generated image while it depends on location

- For specific location $\boldsymbol{p}$, increase activation of 'door' unit



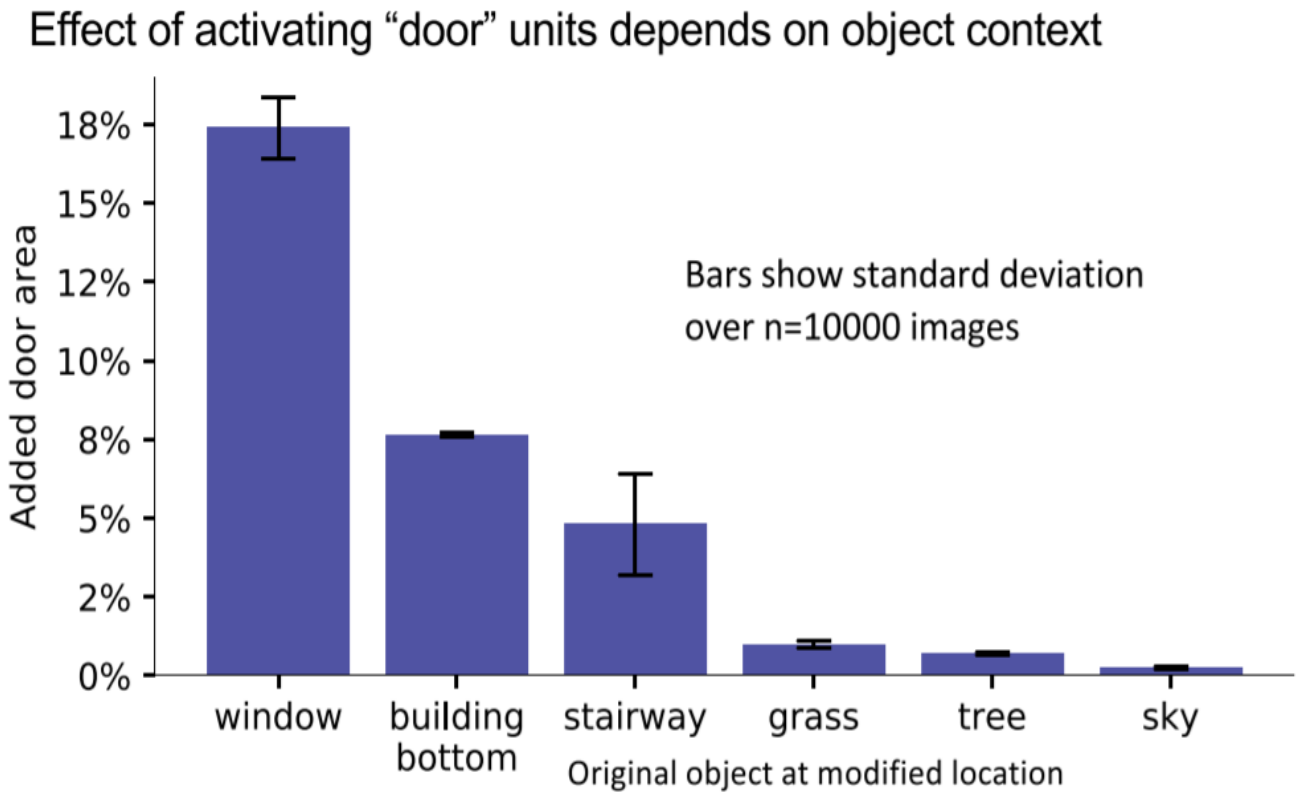
✓ Activates on 1, 3 makes door

✓ Activates on 2 makes door but not appropriate door (delete original door)

✓ Activates on 4 do not makes door

Difference?

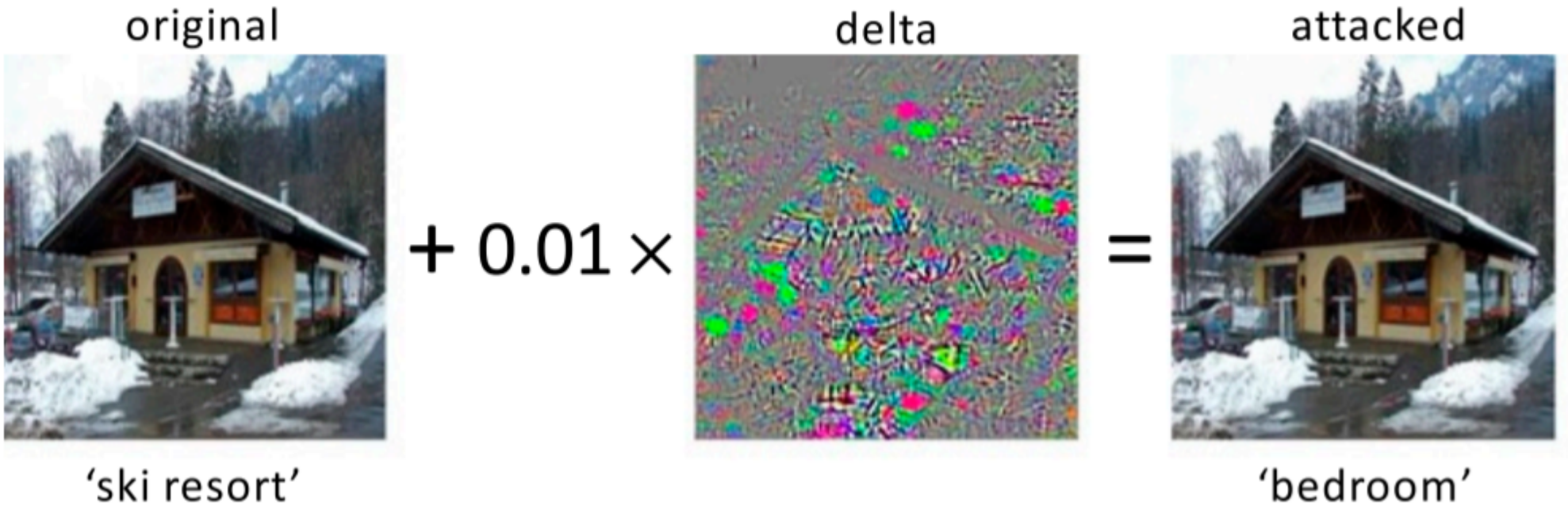
Activating specific units increase their concept in generated image while it depends on object context



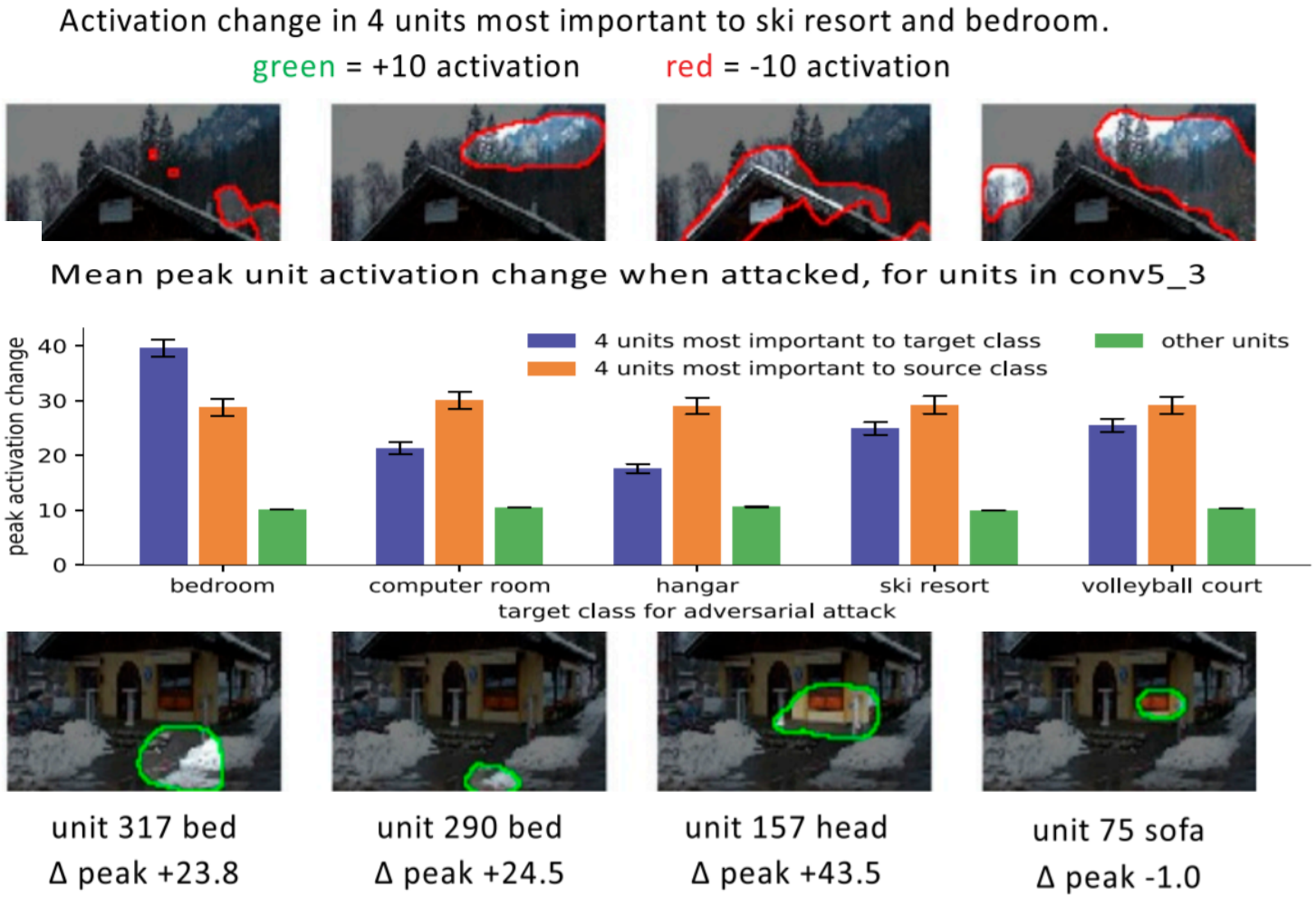
✓ GAN learns more from unsupervised learning

Application on adversarial attack provides hints for how adversarial attack fools the classifier

Adversarial attack changes an image imperceptibly to fool the classifier

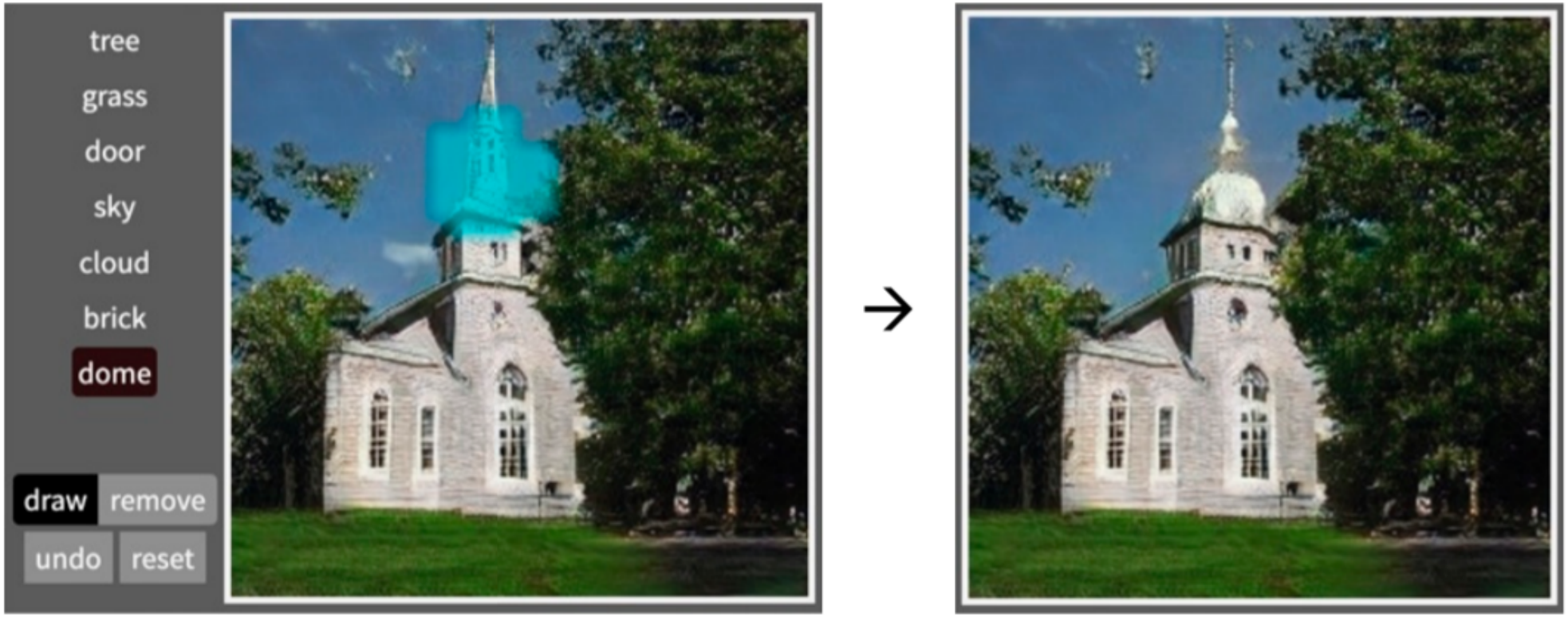


Application on adversarial attack provides hints for how adversarial attack fools the classifier



- ✓ Adversarial attack changes activation of the small amount of units (Imperceptible change)
- ✓ Adversarial attack changes activation of the top important units for classification (Fools classifier)
- How about defense adversarial attack by using randomly deactivate top important unit?

Application on GAN make semantic painting available



✓ Activates / Deactivates top important units

Explainable AI will need for judgment AI, medical AI or fake detection