Learning What and Where to Draw

Scott Reed\textsuperscript{1,3}, Zeynep Akata\textsuperscript{2}, Santosh Mohan\textsuperscript{1}, Samuel Tenka\textsuperscript{1}, Bernt Schiele\textsuperscript{2}, Honglak Lee\textsuperscript{1}
What object is meant to be drawn here? Can we control its location?
1. Bounding box

This bird is completely black.

1. Keypoints, e.g. 15 parts of a bird

This bird is bright blue.
Background:
Generative Adversarial Networks

The discriminator $D$ tries to distinguish real training data from synthetic images.

\[
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]
\]

The generator $G$ tries to fool $D$.

$$
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \\
\mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]
$$

Text-conditional GAN

$$\min_G \max_D V(D, G) = \mathbb{E}_{x,t \sim p_{data}(x,t)}[\log D(x, t)] + \mathbb{E}_{z \sim p_z(z), t \sim p_{data}(t)}[\log(1 - D(G(z, t)))]$$

• The discriminator $D$ tries to distinguish real (text, image) pairs from synthetic.
• The generator $G$ tries to fool $D$. 
Text-conditional GAN

This flower has small, round violet petals with a dark purple center

\[ z \sim \mathcal{N}(0, 1) \]
This flower has small, round violet petals with a dark purple center.
Text-conditional GAN

This flower has small, round violet petals with a dark purple center

\[ z \sim \mathcal{N}(0, 1) \]

This flower has small, round violet petals with a dark purple center

Generator Network
Text-conditional GAN

This flower has small, round violet petals with a dark purple center

$z \sim \mathcal{N}(0, 1)$

Generator Network

Discriminator Network

This flower has small, round violet petals with a dark purple center
Conditioning on bounding box

Generator Network
Conditioning on bounding box

Spatial replicate, crop to bbox

A red bird with a black fa

Generator Network
Conditioning on bounding box

- Spatial replicate, crop to bbox
- A red bird with a black face
- $z \sim \mathcal{N}(0, 1)$

Generator Network
Conditioning on bounding box

Spatial replicate, crop to bbox

A red bird with a black face

\( z \sim \mathcal{N}(0, 1) \)

Generator Network
Conditioning on bounding box
Conditioning on bounding box

Generator Network
Conditioning on bounding box

Generator Network

Discriminator Network
Conditioning on bounding box

Generator Network

Discriminator Network
Conditioning on bounding box

Generator Network

Discriminator Network
Conditioning on bounding box

Generator Network

Discriminator Network
Conditioning on bounding box

Generator Network

Discriminator Network

$z \sim \mathcal{N}(0, 1)$
Conditioning on keypoints

Generator Network
Conditioning on keypoints

Generator Network
Conditioning on keypoints

Generator Network
Conditioning on keypoints

Generator Network
Conditioning on keypoints

Generator Network
Conditioning on keypoints
Conditioning on keypoints

Generator Network
Conditioning on keypoints

Generator Network

Discriminator Network
Conditioning on keypoints

Generator Network

Discriminator Network
Conditioning on keypoints

Generator Network

Discriminator Network
Conditioning on keypoints

Generator Network

Discriminator Network
Conditioning on keypoints

Generator Network

Discriminator Network
Conditioning on keypoints
Moving the bird around with bounding box (noize z fixed)

Caption

This bird has a black head, a long orange beak and yellow body.
Moving the bird around with bounding box (noize z fixed)

Caption

This bird has a black head, a long orange beak and yellow body
Moving the bird around with bounding box (noize z fixed)

Caption
This bird has a black head, a long orange beak and yellow body

GT

Translation
Moving the bird around with bounding box (noize z fixed)

Caption
This bird has a black head, a long orange beak and yellow body

GT

Stretching
Moving the bird around with bounding box (noize z fixed)
Moving the bird around with key points (noize z fixed)

Caption
This bird has a black head, a long orange beak and yellow body.
Moving the bird around with key points (noize z fixed)

Caption
This bird has a black head, a long orange beak and yellow body

GT

Translation
Moving the bird around with key points (noize z fixed)

Caption
This bird has a black head, a long orange beak and yellow body

GT

Stretching
Moving the bird around with key points (noize $z$ fixed)
Generating birds with fixed text and fixed keypoints ($z$ changes)

This large black bird has a long neck and tail feathers.
Generating birds with fixed text and fixed keypoints (z changes)

GT This large black bird has a long neck and tail feathers.

GT This bird is mostly white with a thick black eyebrow, small and black beak and a long tail.

GT This is a small yellowish green bird with a pointy black beak, black eyes and gray wings.

GT This large white bird has an orange-tipped beak.

GT This bird has a bright red crown and black wings and beak.

GT This pale pink bird has a black eyebrow and a black pointy beak, gray wings and yellow underparts.
Generating birds with fixed text and *generated* keypoints

This white bird has gray wings, red webbed feet and a long, curved and yellow beak.
Generating birds with fixed text and *generated* keypoints

- **GT** This white bird has gray wings, red webbed feet and a long, curved and yellow beak.
- **GT** This bird is completely red with a red and cone-shaped beak, black face and a red nape.
- **GT** This water bird has a long white neck, black body, yellow beak and black head.
- **GT** This bird has a yellow head, black eyes, a gray pointy beak and orange lines on its breast.
- **GT** This bird is large, completely black, with a long pointy beak and black eyes.
- **GT** This small bird has a blue and gray head, pointy beak and a white belly.
Comparison to text-only conditional GAN:

- Ground-truth image and text caption
  - A small sized bird that has tones of brown and dark red with a short stout bill

- This bird has a yellow breast and a dark grey face

- The bird is solid black with white eyes and a black beak.
Comparison to text-only conditional GAN:

A small sized bird that has tones of brown and dark red with a short stout bill.

This bird has a yellow breast and a dark grey face.

The bird is solid black with white eyes and a black beak.

^ Text-only 64 x 64 GAN samples.
Comparison to text-only conditional GAN:

Ground-truth image and text caption

A small sized bird that has tones of brown and dark red with a short stout bill

This bird has a yellow breast and a dark grey face

The bird is solid black with white eyes and a black beak.

^ Text-only 128 x 128 GAN samples.
A small sized bird that has tones of brown and dark red with a short stout bill.

This bird has a yellow breast and a dark grey face.

The bird is solid black with white eyes and a black beak.

\(^ {\text{Ours, with fixed keypoints.}}\)
Comparison to text-only conditional GAN:

- Ground-truth image and text caption
  - A small sized bird that has tones of brown and dark red with a short stout bill
  - This bird has a yellow breast and a dark grey face
  - The bird is solid black with white eyes and a black beak.

^K^ Ours, with generated keypoints.
Comparison to text-only conditional GAN:

<table>
<thead>
<tr>
<th>Ground-truth image and text caption</th>
<th>A small sized bird that has tones of brown and dark red with a short stout bill</th>
<th>This bird has a yellow breast and a dark grey face</th>
<th>The bird is solid black with white eyes and a black beak</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN-INT-CLS (Reed et. al, 2016b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAWWN trained without key points</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAWWN Key points given</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAWWN Key points generated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Generating humans with fixed text and generated keypoints

<table>
<thead>
<tr>
<th>Caption</th>
<th>GT</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a woman in a yellow tank top is doing yoga.</td>
<td>![GT Image]</td>
<td>![Samples Images]</td>
</tr>
</tbody>
</table>
Generating humans with fixed text and generated keypoints

<table>
<thead>
<tr>
<th>Caption</th>
<th>GT</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a woman in a yellow tank top is doing yoga.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the man wearing the red shirt and white pants play golf on the green grass</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a man in a red sweater and grey pants swings a golf club with one hand.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a woman wearing goggles swimming through very murky water</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a woman in grey shirt is doing yoga.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a man in green shirt and white pants is swinging his golf club.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a man in an orange jacket, black pants and a black cap wearing sunglasses skiing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a man is skiing and competing for the olympics on the slopes.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Also works using PixelCNN instead of GAN:

\[\text{Structure: Class segmentation or keypoint map.}\]

This part is joint work with Aäron van den Oord, Nal Kalchbrenner, Victor Bapst, Matt Botvinick and Nando de Freitas.
Also works using PixelCNN instead of GAN:

**Structure:** Class segmentation or keypoint map.

**Text:** “a gray elephant standing next to a woman in a red dress.”
Also works using PixelCNN instead of GAN:

**Structure:** Class segmentation or keypoint map.

> Text: “a gray elephant standing next to a woman in a red dress.”
Also works using PixelCNN instead of GAN:

**Structure:** Class segmentation or keypoint map.

**Text:** “a gray elephant standing next to a woman in a red dress.”
A small sized bird with a yellow belly and black tipped head.

This bird has a long, narrow, sharp yellow beak ending in a black tip.

A yellow bird with grey wings and a short and small beak.
A man in a white shirt is holding a lacrosse racket.

A woman in an orange shirt is standing in front of a sink.

A man in a white shirt and black shorts holding a tennis racket and ball about to hit it on a tennis court.
A person carrying their surfboard while walking along a beach.

<table>
<thead>
<tr>
<th>Person</th>
<th>Surfboard</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Person" /></td>
<td><img src="image2.png" alt="Surfboard" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Person" /></td>
<td><img src="image4.png" alt="Surfboard" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Person" /></td>
<td><img src="image6.png" alt="Surfboard" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Person" /></td>
<td><img src="image8.png" alt="Surfboard" /></td>
</tr>
</tbody>
</table>
Conclusions

• Location conditioning is useful for image synthesis with GANs:
  – Adds additional layer of control and interpretability.
  – Works great for UCSD-CUB birds; more challenging for MHP humans dataset.
  – Fast enough for real-time interaction.

• PixelCNN can also do text- and structure-controllable density modeling:
  – Works well on CUB (using keypoints) and MS-COCO (using segmentation mask).
  – Not yet fast enough for real-time interaction.
Thank You!

Learning What and Where to Draw

Scott Reed\textsuperscript{1,3}, Zeynep Akata\textsuperscript{2}, Santosh Mohan\textsuperscript{1}, Samuel Tenka\textsuperscript{1}, Bernt Schiele\textsuperscript{2}, Honglak Lee\textsuperscript{1}
Questions?

Learning What and Where to Draw

Scott Reed\textsuperscript{1,3}, Zeynep Akata\textsuperscript{2}, Santosh Mohan\textsuperscript{1}, Samuel Tenka\textsuperscript{1}, Bernt Schiele\textsuperscript{2}, Honglak Lee\textsuperscript{1}
Extra Slides
Text-to-image samples (MS-COCO)

<table>
<thead>
<tr>
<th>GT</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>a group of people on skis stand on the snow.</td>
<td>a man in a wet suit riding a surfboard on a wave.</td>
</tr>
<tr>
<td>a table with many plates of food and drinks</td>
<td>two plates of food that include beans, guacamole and rice.</td>
</tr>
<tr>
<td>two giraffes standing next to each other in a forest.</td>
<td>a green plant that is growing out of the ground.</td>
</tr>
<tr>
<td>a large blue octopus kite flies above the people having fun at the beach.</td>
<td>there is only one horse in the grassy field.</td>
</tr>
<tr>
<td>a pitcher is about to throw the ball to the batter.</td>
<td>a picture of a very clean living room.</td>
</tr>
<tr>
<td>a sheep standing in an open grass field.</td>
<td>a toilet in a small room with a window and unfinished walls.</td>
</tr>
</tbody>
</table>
Looking more closely:

- No explicit notion of objects
- Maybe they are in the network *somewhere*...
- But how to access them and controllably re-render?
- No way for the model to “explain” its drawing choices.
- E.g. what is being depicted at *that* location?
Learning What and Where to Draw with PixelCNN

Joint work with
Aäron van den Oord, Nal Kalchbrenner, Victor Bapst, Matt Botvinick and Nando de Freitas.
MS-COCO

A person on snow skis with a backpack skiing down a mountain.

Caltech-UCSD Birds

A white body and head with a bright orange bill along with black coverts and rectrices.

MPII Human Pose

A young girl is wearing a black ballerina outfit and pink tights dancing.
Motivation

- GANs can do conditional image synthesis:
  - Text descriptions
  - keypoints, bounding box, surface normals, etc
Motivation

But they have drawbacks:

○ Difficult to train, pathological local minima, samples may lack diversity.
○ No likelihood, unclear how to compare models.
Motivation

Can PixelCNN solve these problems?
- Very simple and stable to train
- Tractable likelihood, state-of-the-art density model.
- Slower to sample (one pixel at a time)
This is a colorful bird with a blue and green body and orange eyes.

A black nape contrasts the white plumage of this bird, who is stretching its wingspan.

This small green bird with a thin curved bill has a blue patch by its eyes.
The head of the bird is red and the body is black and white.

The black bird looks like a crow has black beak, black wings and body.

A white body and head with a bright orange bill along with black coverts and rectrices.
Text: A gray elephant standing next to a woman in a red dress.
A man in a white shirt is holding a lacrosse racket.
A man in a blue and white "kuhl" biker outfit is riding a bike up a hill.
A swimmer is at the bottom on a pool taking off his swim gear.

A woman in an orange shirt is standing in front of a sink.
A man in a white button up shirt is standing behind an ironing board ironing a shirt.
A man wearing a blue t shirt and shorts is rock climbing.

A man in a white shirt and black shorts holding a tennis racket and ball about to hit it on a tennis court.
A man in a blue hat is holding a shovel in a dirt filled field.
A man practices his ice skating, wearing hockey boots, at an ice skating rink.
This bird is bright yellow.

This bird is completely black.

This bird is bright red.

This bird is completely green.

This bird is bright blue.

This bird is all white.
A person carrying their surfboard while walking along a beach.

A bathroom with a vanity mirror next to a white toilet.

A young man riding a skateboard down a ramp.
This laptop and monitor are surrounded by many wires.
The woman is riding her horse on the beach by the water.
A piece of cooked broccoli is on some cheese.
Three men wearing black and ties stand and smile at something.

A large cow walks over a fox in the grass.

Two women in English riding outfits on top of horses.