Yandex Research

Big GANs Are Watching You: Towards Unsupervised Object Segmentation with Off-the-Shelf Generative Models

Andrey Voynov   Stanislav Morozov   Artem Babenko

Problem setup and prior works

Problem setup:
- Pixel-level labeling is expensive
- Fully unsupervised training
- Off-the-shelf generative models

Prior works:
- Train generative models to predict object segmentation
- Use pretrained supervised models in their training protocol

Image credit: https://neurohive.io/en/news/u-net-u-squared-net-a-new-neural-network-for-salient-object-detection/

Interpretable Directions in the GAN Latent Space

Andrey Voynov and Artem Babenko, “Unsupervised discovery of interpretable directions in the GAN latent space”, ICML 2020:

- Unsupervised model-agnostic method to identify interpretable directions in the latent space of a pretrained GAN model
- “Background removal” directions was discovered only for BigGAN that was trained under the supervision from image class labels

In a nutshell, it seeks to learn K directions in the latent space $h_1, \dots, h_K$ such that the sets of pairs $\{G(z), G(z + h_i) | z \sim \mathcal{N}(0, \mathbb{I})\}$ with different $i = 1, \dots, K$ should be easy to distinguish from each other by a CNN classifier, which is trained jointly with $h_1, \dots, h_K$

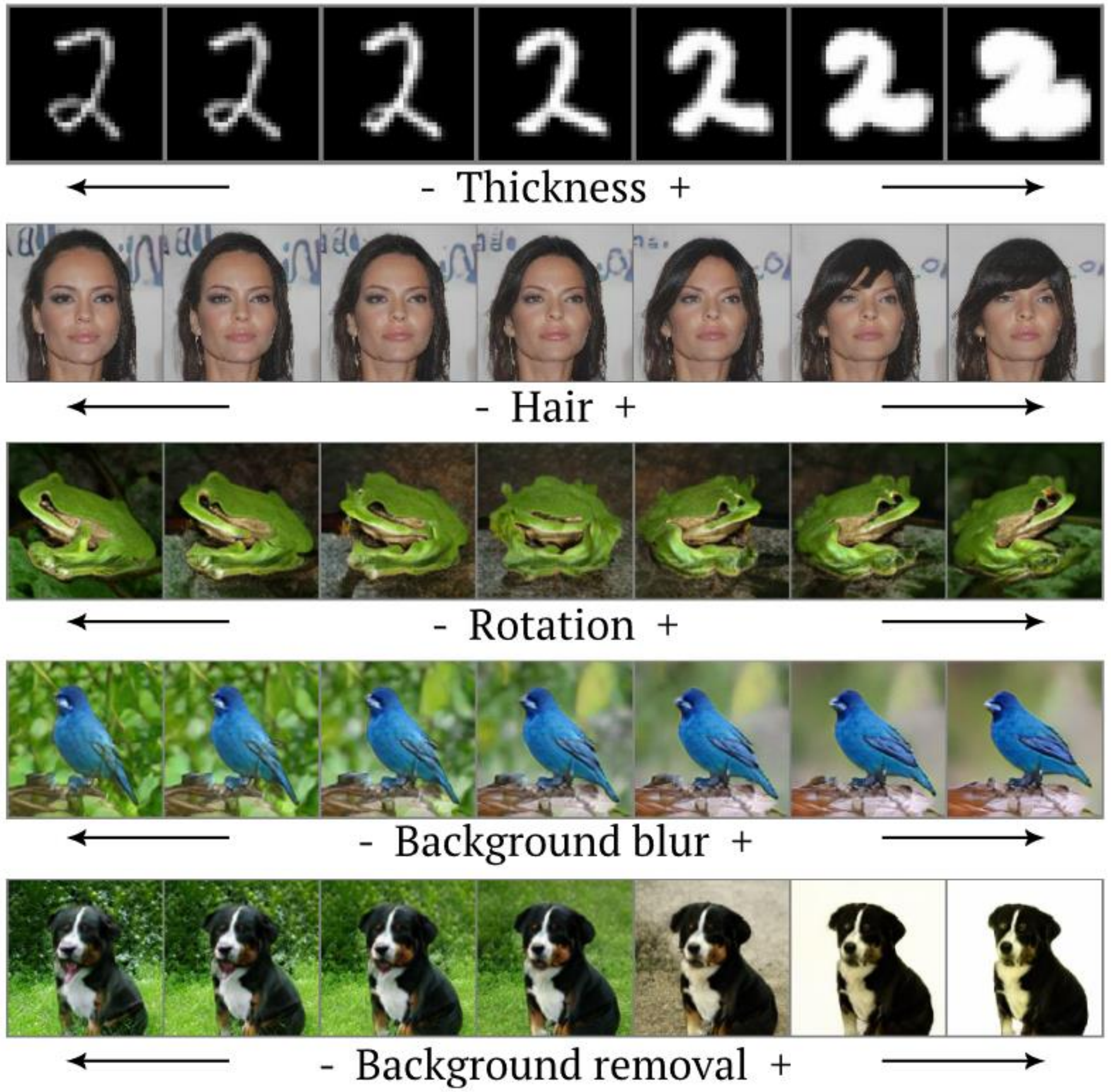
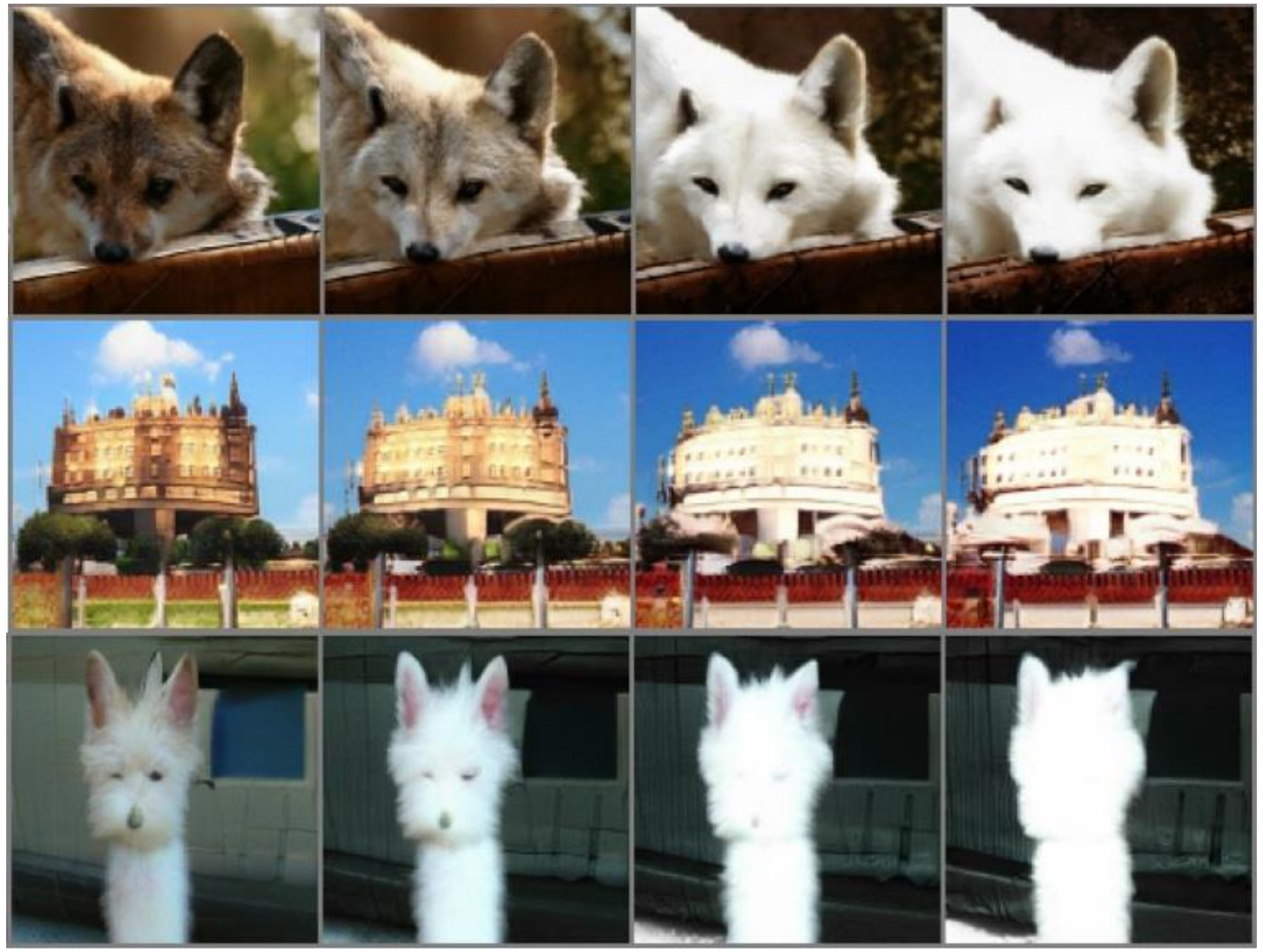


Image credit: Andrey Voynov and Artem Babenko, ICML 2020

Exploring the latent spaces of unsupervised GANs

BigBiGAN



StyleGAN2

BigBiGAN and StyleGAN2 do not possess any directions that have clear “background removal” effect, however, they both possess directions that have different effect on the object and background pixels

Exploring the latent spaces of unsupervised GANs

We produce a binary mask M for an image $G(z)$ by comparing its intensity with the shifted image $M = [G(z + h_{bg}) > G(z)]$ after greyscale conversion

Additional heuristics

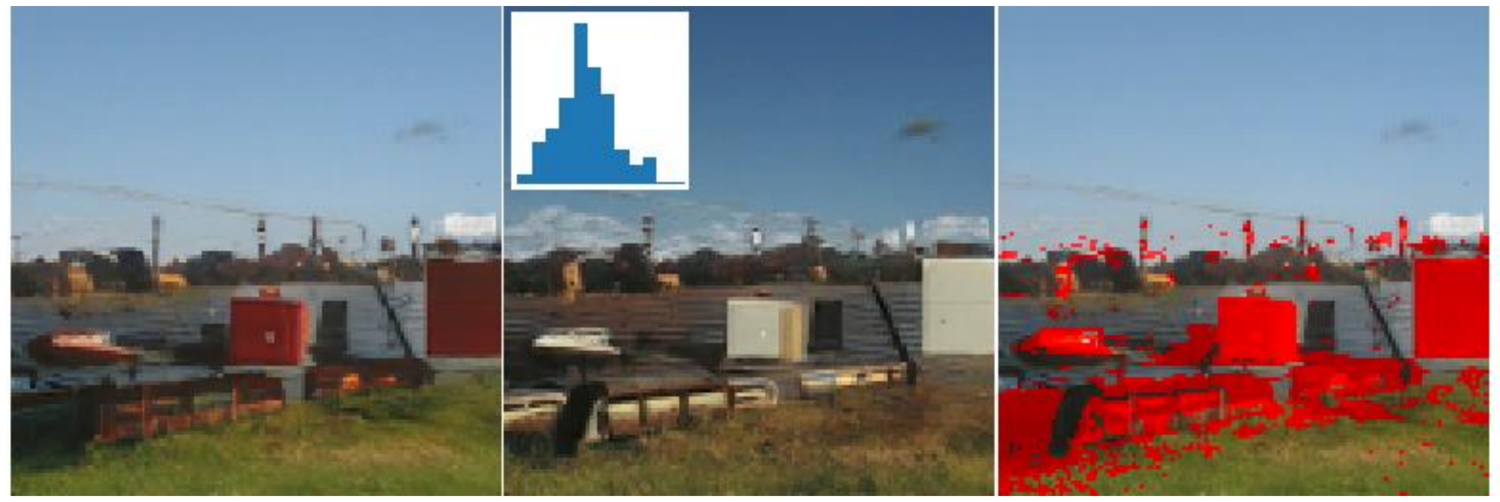
Adaptation to the particular segmentation task

- BigBiGAN trained on the Imagenet samples the latent codes from the standard Gaussian distribution $z \sim \mathcal{N}(0, \mathbb{I})$
- The Imagenet distribution can be suboptimal for the particular segmentation task
- BigBiGAN is equipped with an encoder that maps images to the latent space
- To make the distribution closer to the particular dataset $I = \{I_1, \dots, I_N\}$ we sample z from the latent space regions that are close to the latent codes of I: $\{E(I_i) + \alpha\xi \mid i \sim U\{1, N\}, \xi \sim \mathcal{N}(0, \mathbb{I})\}$

Improving saliency masks

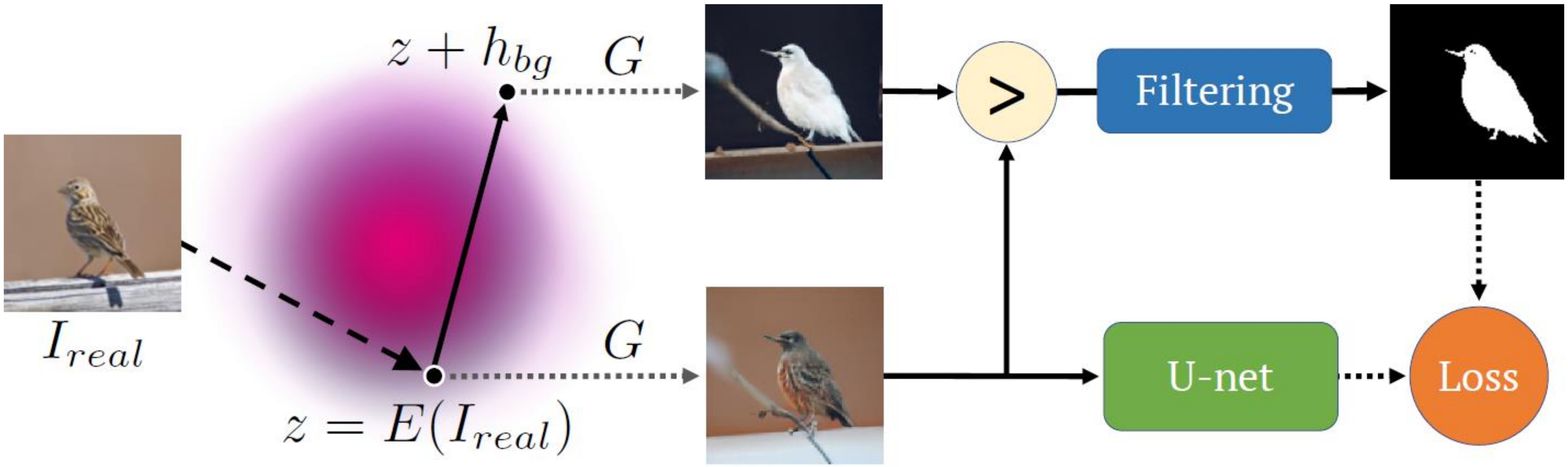
Mask size filtering

Histogram filtering



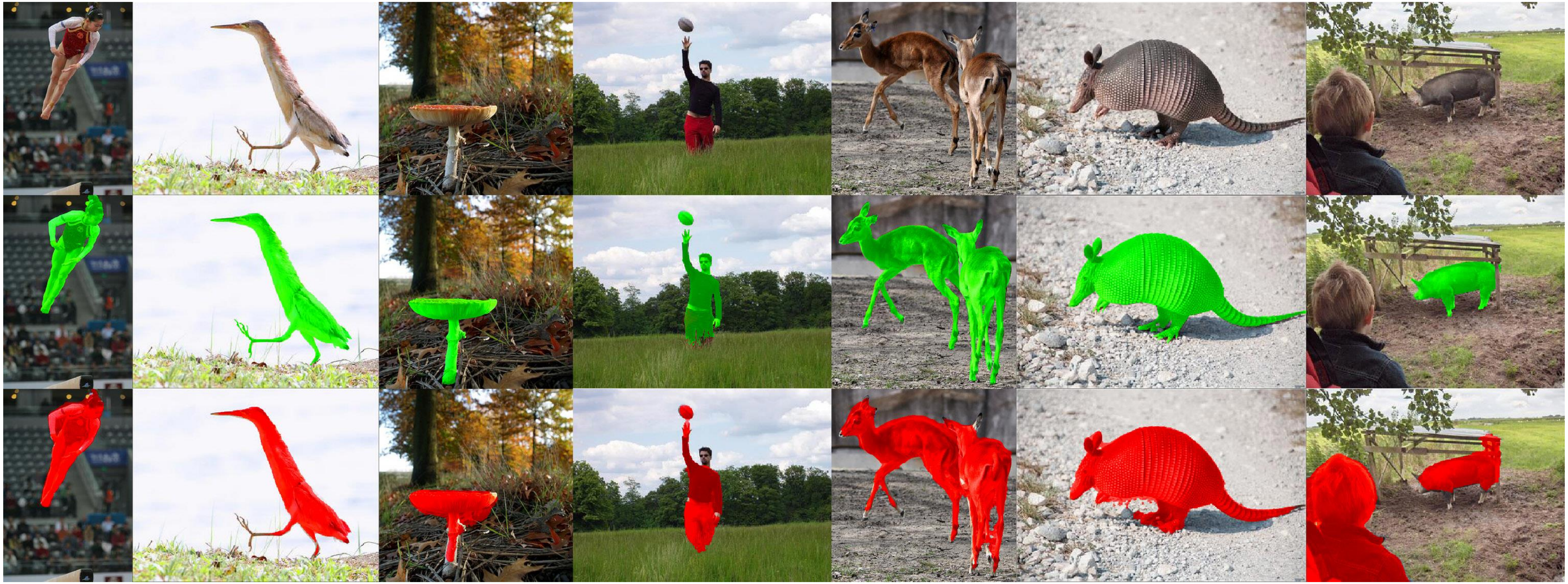
Connected components filtering

Putting all together



- Fully unsupervised
- Does not require GAN training
- The only hyperparameters to tune are batch size, learning rate and a number of optimizer steps
- Training takes approximately seven hours on two Nvidia 1080Ti cards

Qualitative results



Top: Images from the DUTS-test dataset. *Middle:* Groundtruth masks. *Bottom:* Masks produced by the E-BigBiGAN method

Quantitative results

max F_β

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall},$$

$$Precision = \frac{TP}{TP+FP},$$

$$Recall = \frac{TP}{TP+FN}.$$

We compute F-measure for 255 uniformly distributed binarization thresholds and report its maximum value. $\beta = 0.3$

IoU

$$IoU = \frac{\mu(s \cap m)}{\mu(s \cup m)}$$

After the binarization with threshold 0.5

Accuracy

The proportion of pixels that have been correctly assigned to the object/background after the binarization with threshold 0.5

Method	CUB-200-2011			Flowers		
	maxF_β	IoU	Accuracy	maxF_β	IoU	Accuracy
PerturbGAN	—	0.380	—	—	—	—
ReDO	—	0.426	0.845	—	0.764	0.879
OneGAN	—	0.555	—	—	—	—
BigBiGAN	0.794	0.683	0.930	0.760	0.540	0.765
E-BigBiGAN (w/o z-noising)	0.750	0.619	0.918	0.814	0.689	0.874
E-BigBiGAN (with z-noising)	**0.834**	**0.710**	**0.940**	**0.878**	**0.804**	**0.904**
std	0.005	0.007	0.002	0.001	<0.001	<0.001

Salient object detection

Method	ECSSD			DUTS			DUT-OMRON		
	maxF_β	IoU	Accuracy	maxF_β	IoU	Accuracy	maxF_β	IoU	Accuracy
HS	0.673	0.508	0.847	0.504	0.369	0.826	0.561	0.433	0.843
wCtr	0.684	0.517	0.862	0.522	0.392	0.835	0.541	0.416	0.838
WSC	0.683	0.498	0.852	0.528	0.384	0.862	0.523	0.387	**0.865**
DeepUSPS	0.584	0.440	0.795	0.425	0.305	0.773	0.414	0.305	0.779
BigBiGAN	0.782	0.672	0.899	0.608	0.498	0.878	0.549	0.453	0.856
E-BigBiGAN	**0.797**	**0.684**	**0.906**	**0.624**	**0.511**	**0.882**	**0.563**	**0.464**	0.860

The comparison of unsupervised saliency detection methods

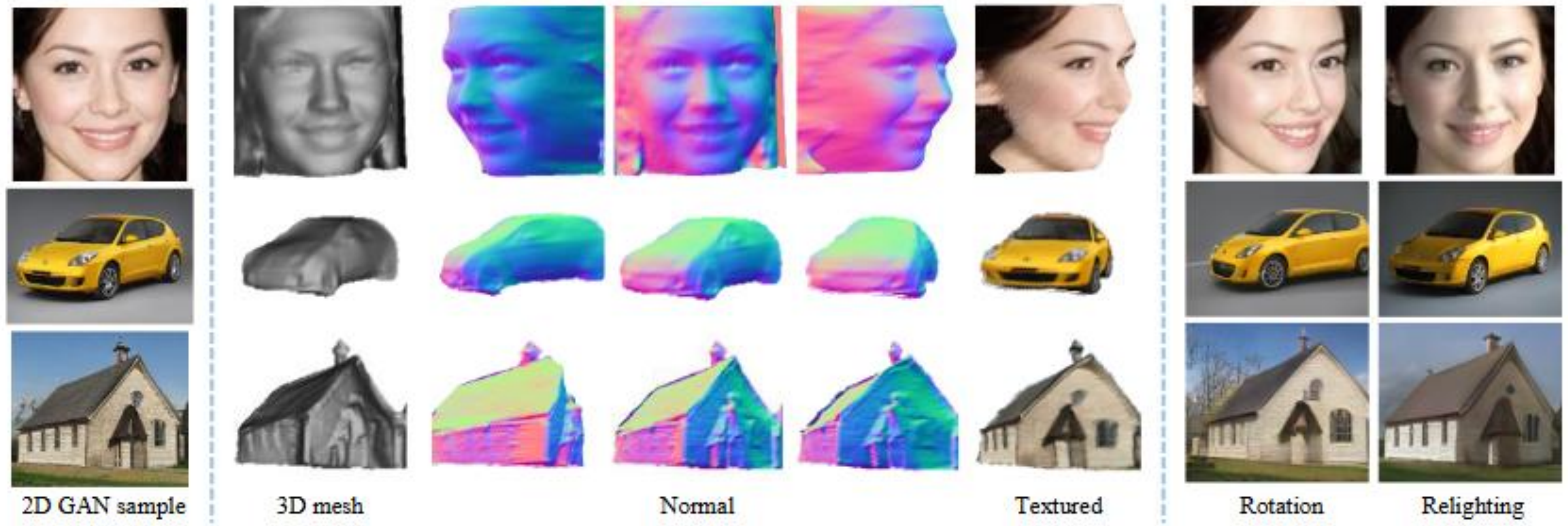
Ablation

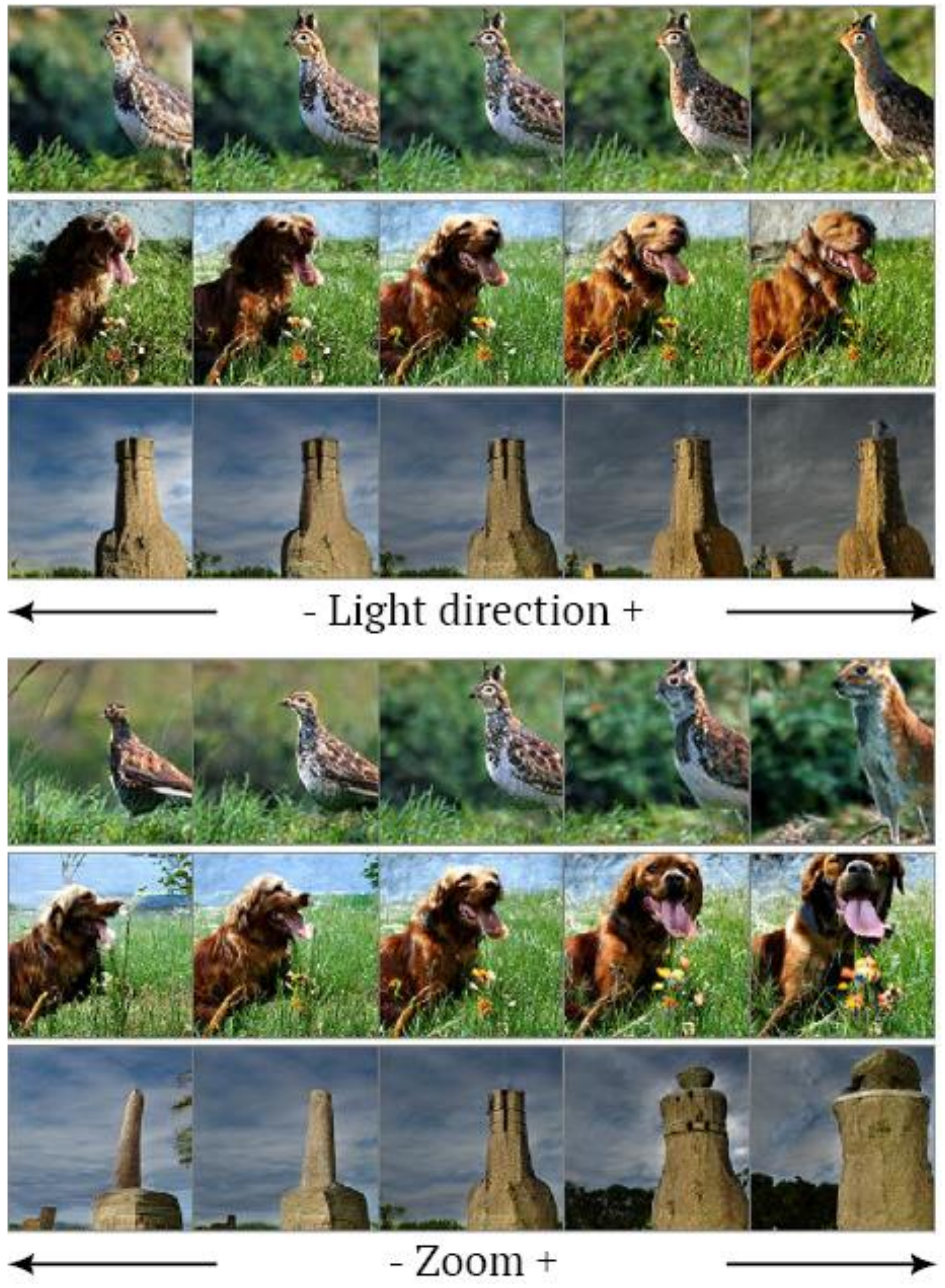
Method	ECSSD			DUTS			DUT-OMRON		
	maxF_β	IoU	Accuracy	maxF_β	IoU	Accuracy	maxF_β	IoU	Accuracy
Base	0.737	0.626	0.859	0.575	0.454	0.817	0.498	0.389	0.758
+Imagenet embeddings	0.773	0.657	0.874	0.616	0.483	0.832	0.533	0.413	0.772
+Size filter	0.781	0.670	0.900	0.62	0.499	0.871	0.552	0.443	0.842
+Histogram	0.779	0.670	0.900	0.621	0.503	0.875	0.555	0.450	0.850
+Connected components	**0.797**	**0.684**	**0.906**	**0.624**	**0.511**	**0.882**	**0.563**	**0.464**	**0.860**

Impact of different components in the E-BigBiGAN pipeline

Research direction

- GANs for generating training sets
- General segmentation
- 3D reconstruction (Xingang Pan et al, “Do 2D GANs Know 3D Shape? Unsupervised 3D shape reconstruction from 2D image GANs”, 2020)
- Object localization
- and much more…





Thank you!