

# Emergence of Object Segmentation in Perturbed Generative Models

Adam Bielski, Paolo Favaro  
University of Bern

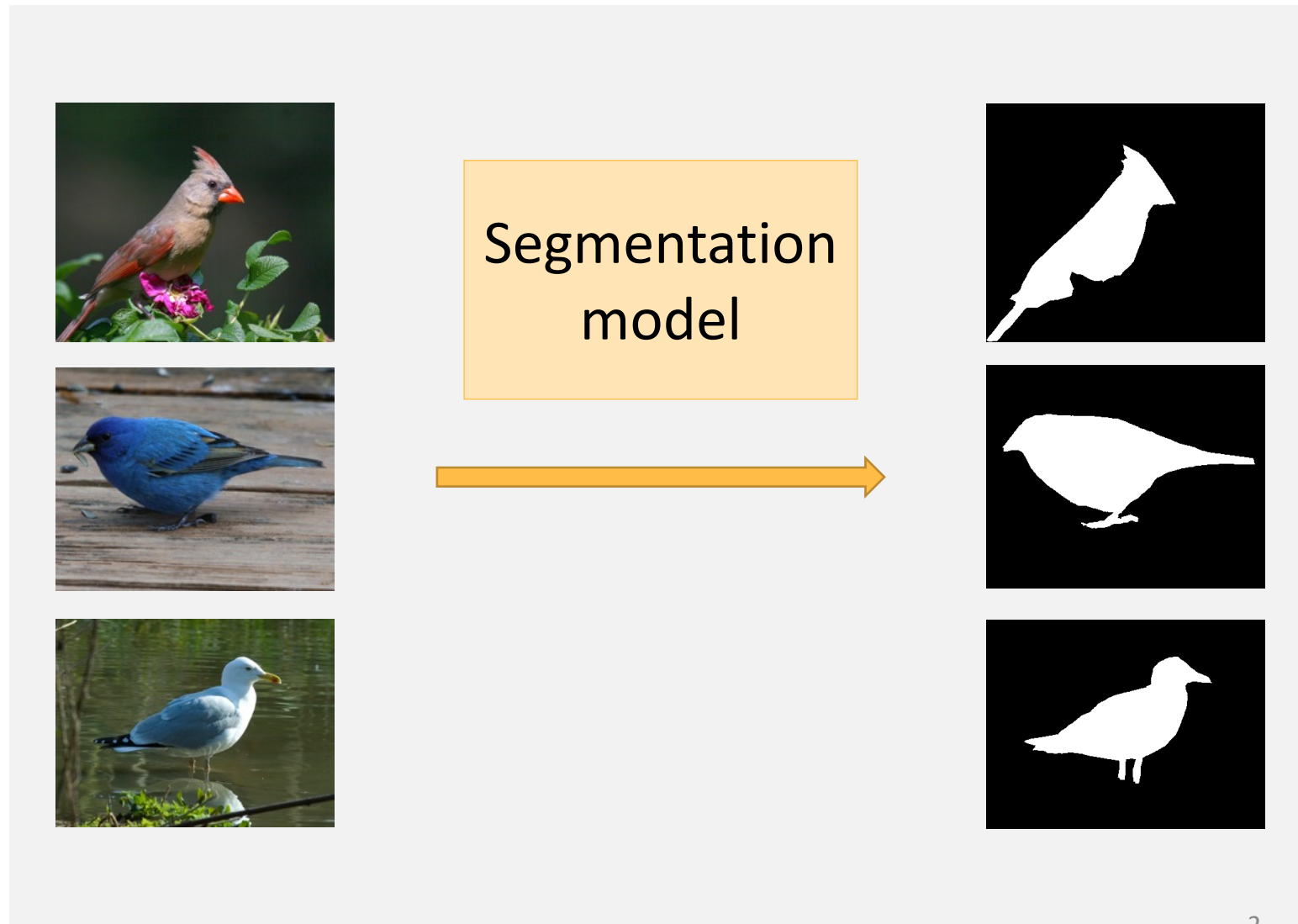
*u*<sup>b</sup>

---

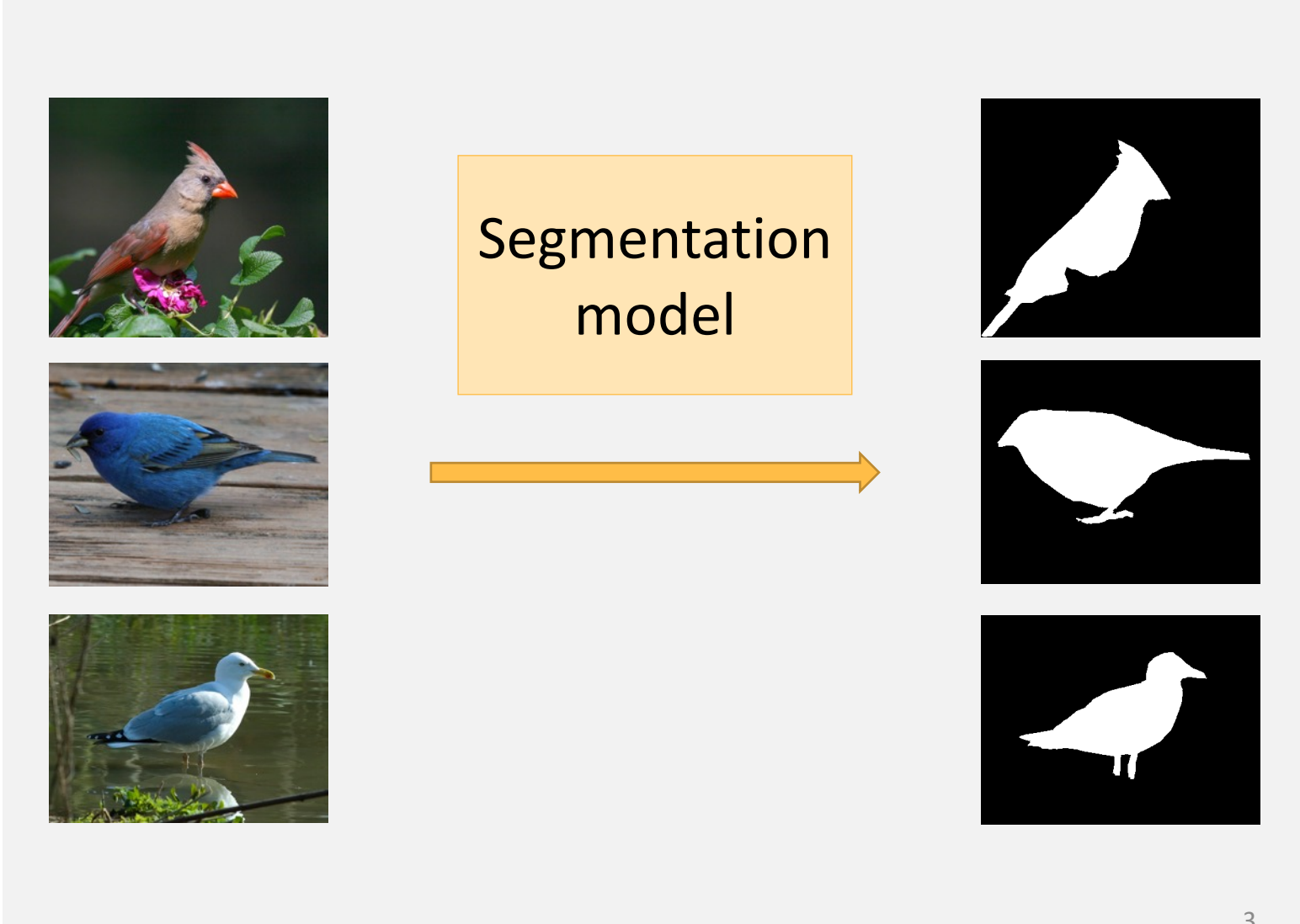
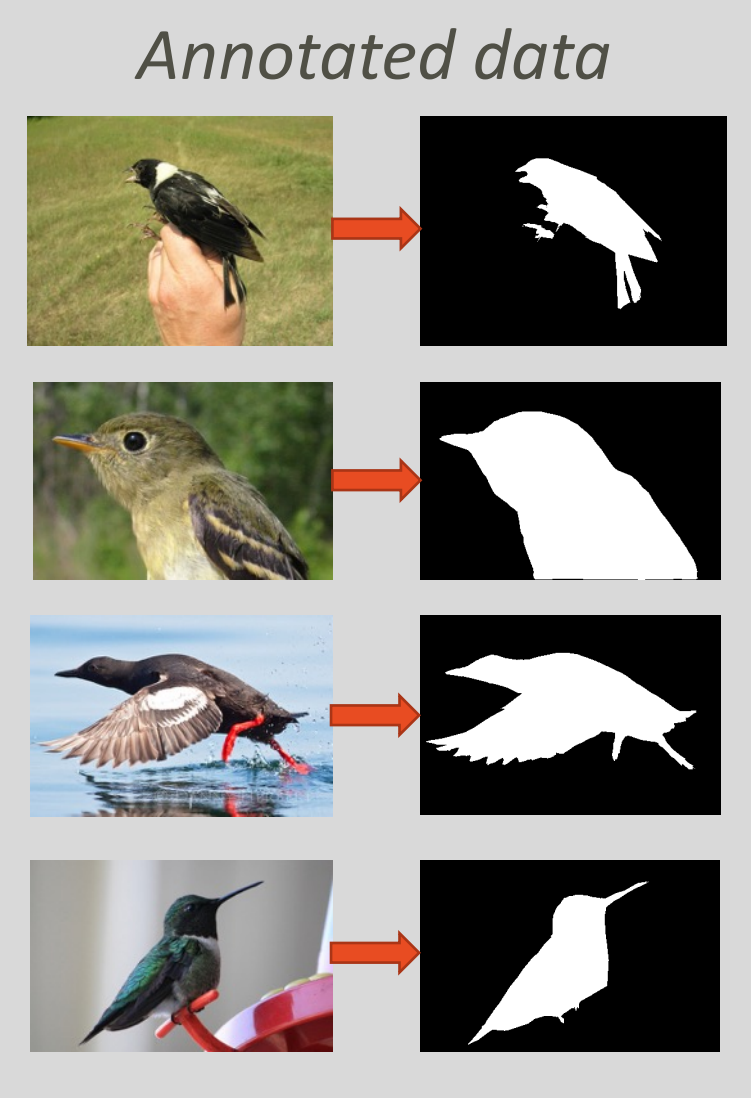
<sup>b</sup>  
**UNIVERSITÄT  
BERN**



# Object segmentation

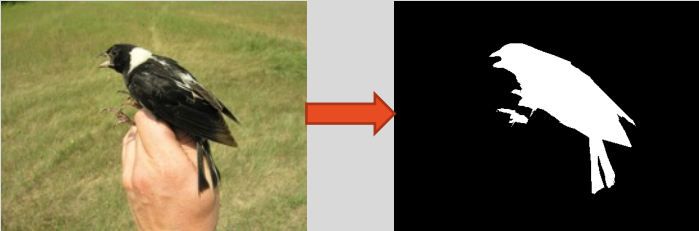


# Object segmentation

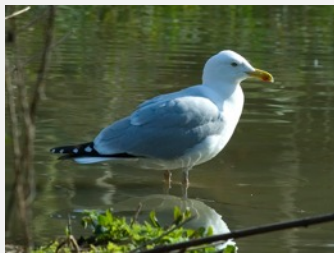


# Object segmentation

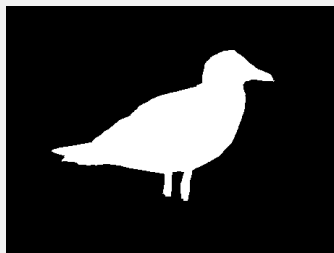
*Annotated data*



- *Human annotated*
- *Expensive*
- *Expertise knowledge*



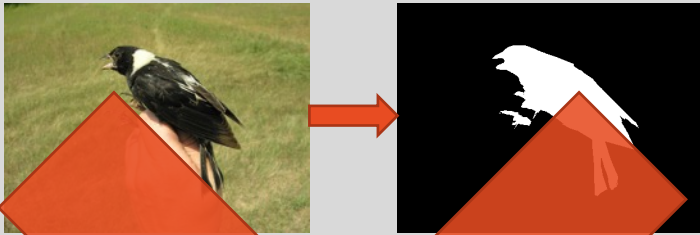
Segmentation model





# Unsupervised object segmentation

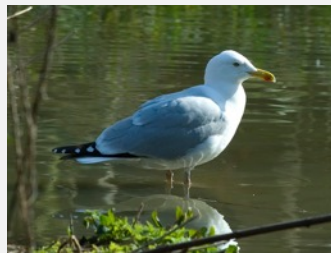
*Annotated data*



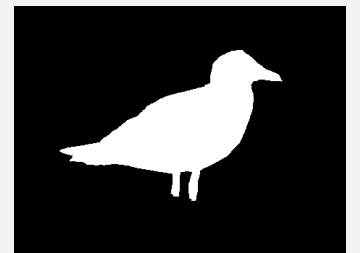
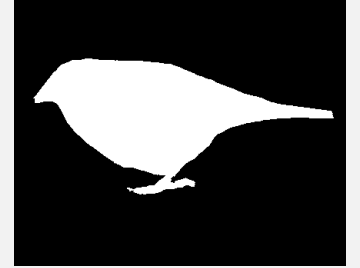
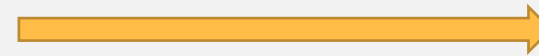
- *Human annotated*
- *Expensive*
- *Expertise knowledge*



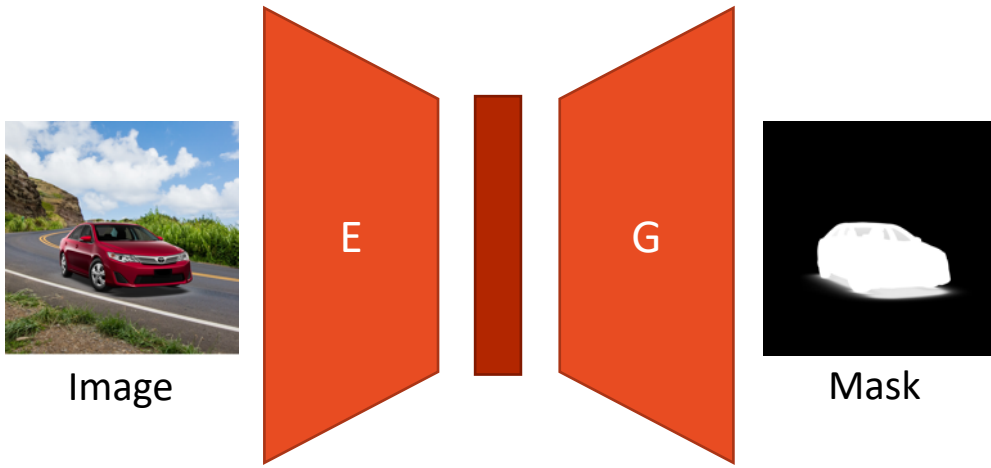
*Can we do it without labels?*



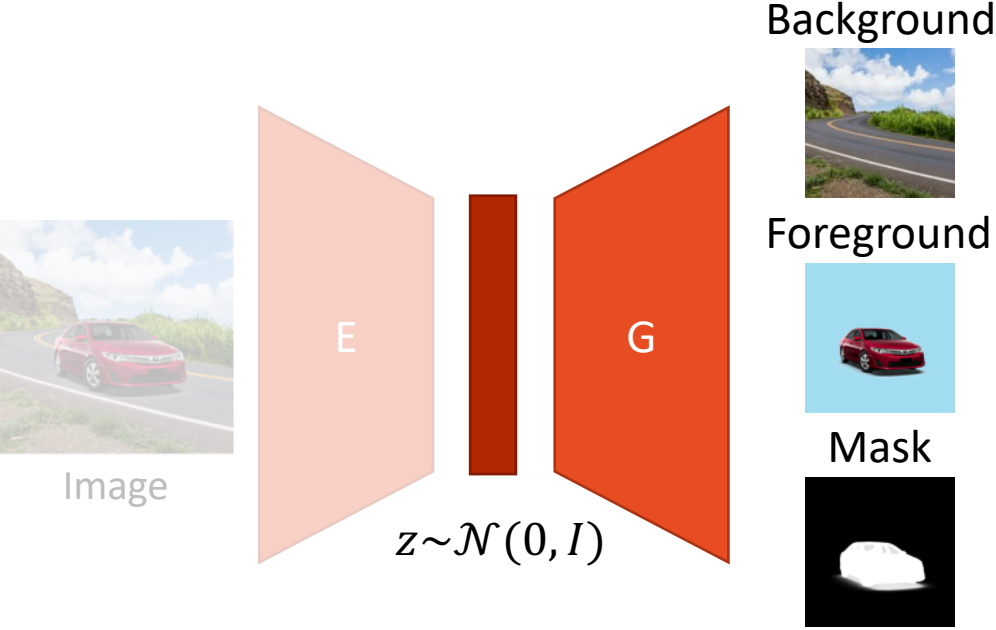
Segmentation  
model



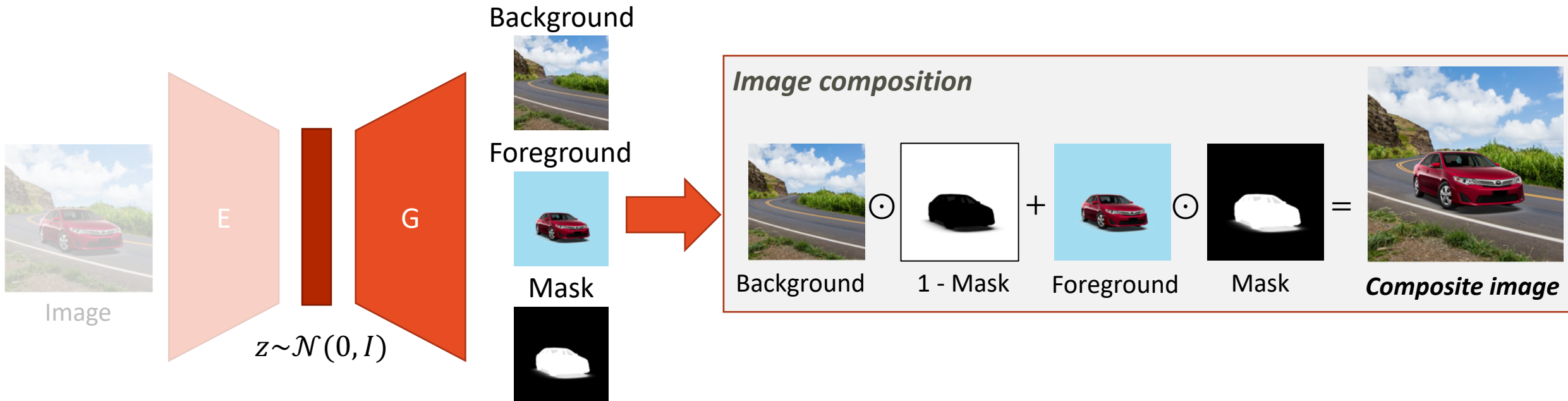
# Unsupervised object segmentation



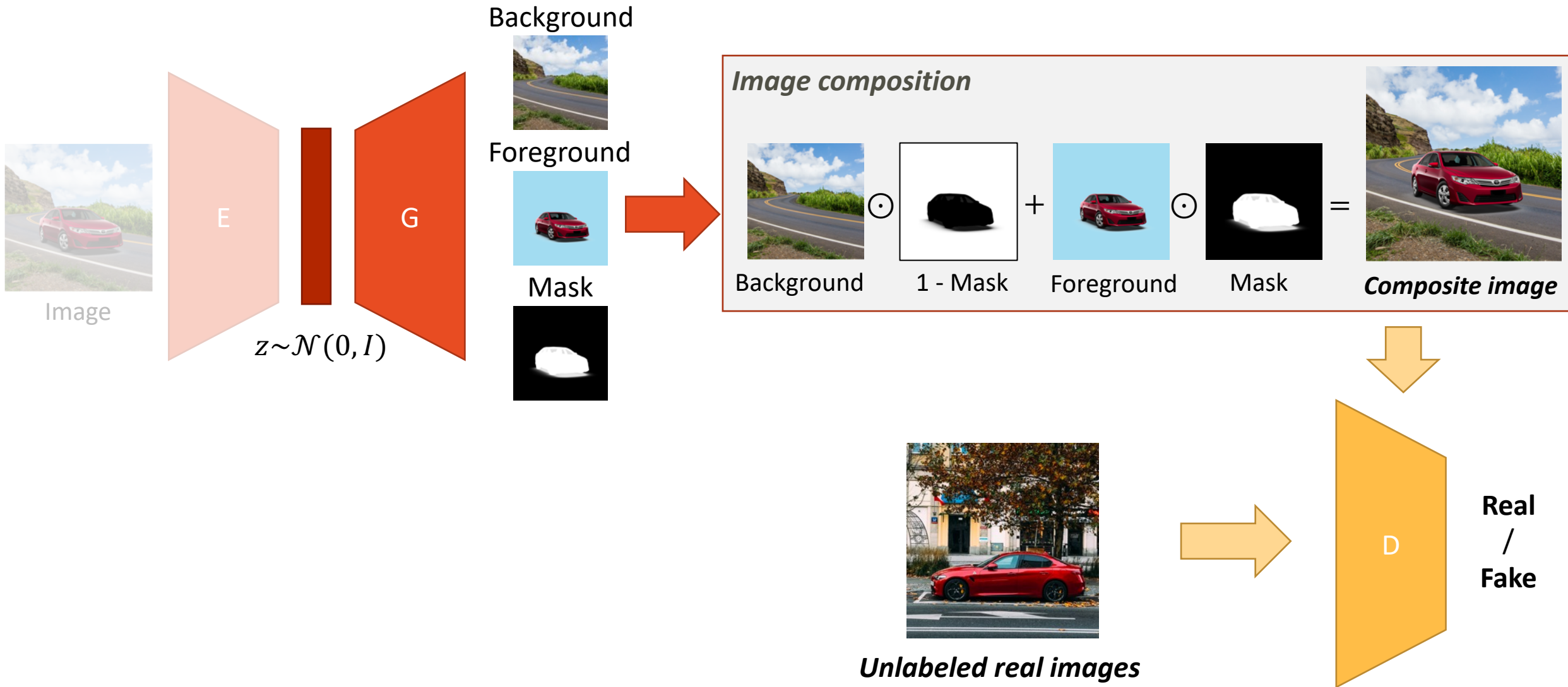
# Learning to generate layered scenes with GAN



# Learning to generate layered scenes with GAN



# Learning to generate layered scenes with GAN



# Degenerate solutions

Background



Foreground



Mask



Composite image



# Degenerate solutions

Background



Foreground



Mask



Composite image





# Degenerate solutions

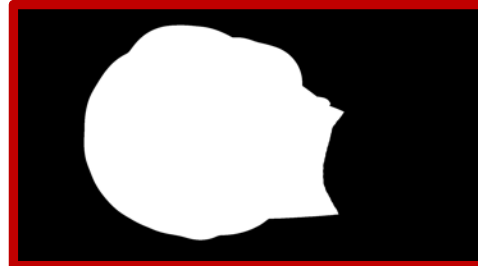
Background



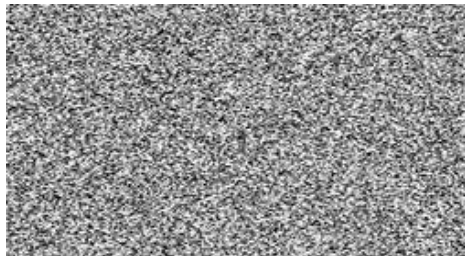
Foreground



Mask



Composite image



# Degenerate solutions

Background



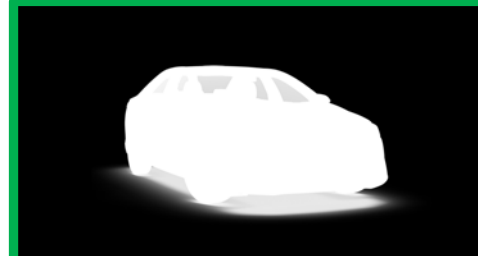
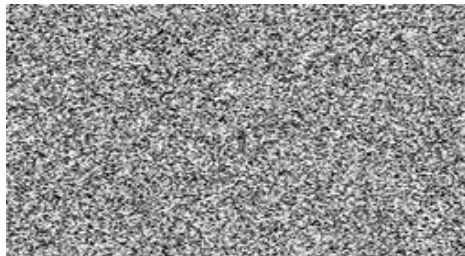
Foreground



Mask



Composite image







What is a correct partition?



What is a correct partition?





Random shifts





Random shifts







Random shifts







Random shifts





Invalid partition  $\Rightarrow$  invalid scene after a small shift

Valid partition  $\Rightarrow$  valid scene after a small shift

# Random shifts





# Avoiding degenerate solutions

Background



Foreground



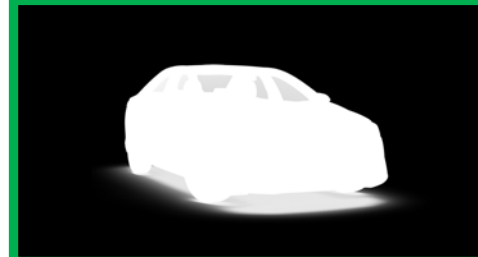
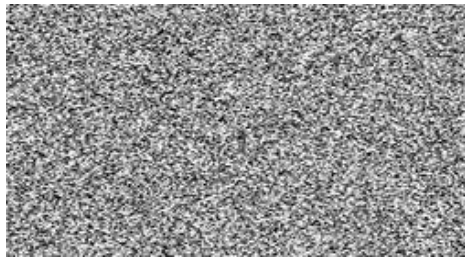
Mask



Composite image



Composite image w/shift



# Avoiding degenerate solutions

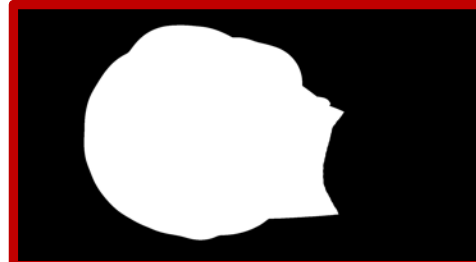
Background



Foreground



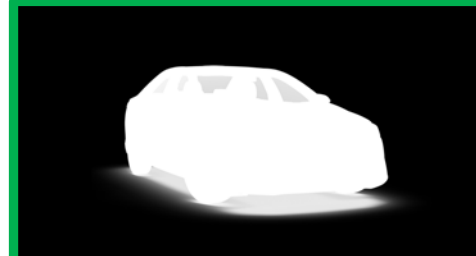
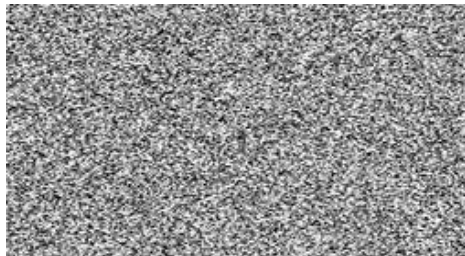
Mask



Composite image



Composite image w/shift





# Avoiding degenerate solutions

Background



Foreground



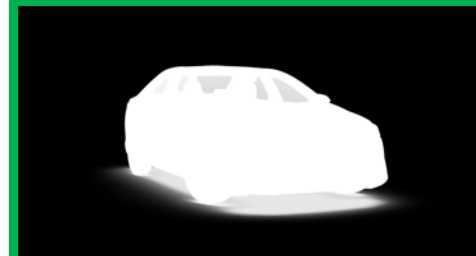
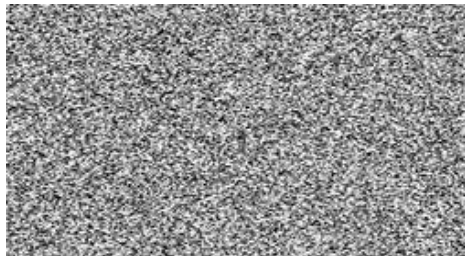
Mask



Composite image



Composite image w/shift





# Avoiding degenerate solutions

Background



Foreground



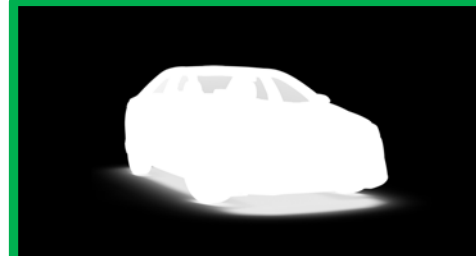
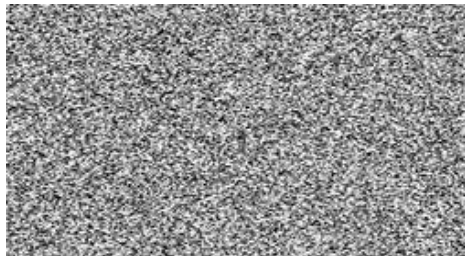
Mask



Composite image



Composite image w/shift





# Avoiding degenerate solutions

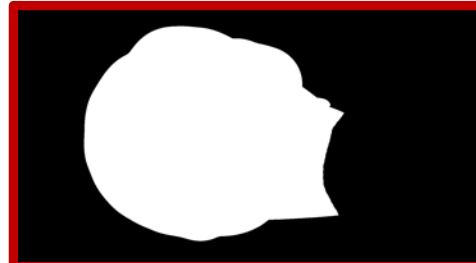
Background



Foreground



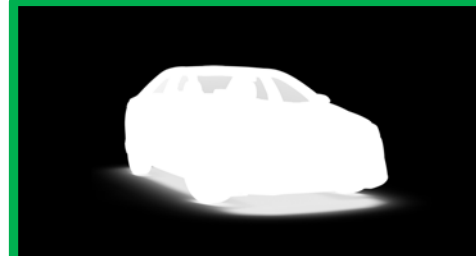
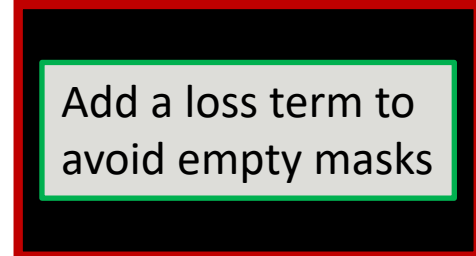
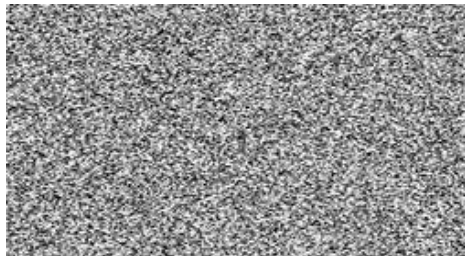
Mask



Composite image



Composite image w/shift





# Avoiding degenerate solutions

Background



Foreground



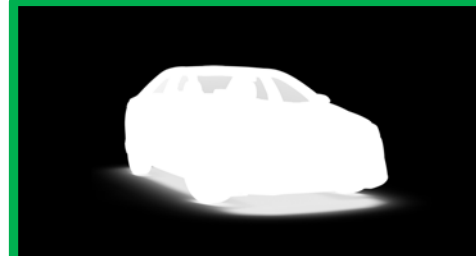
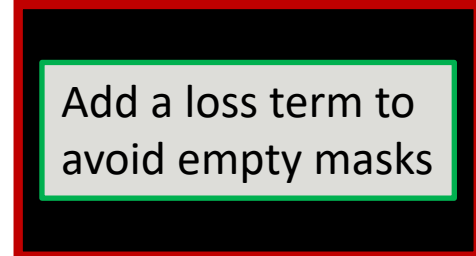
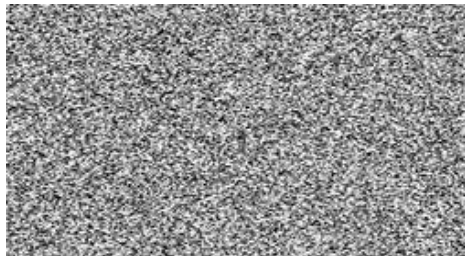
Mask



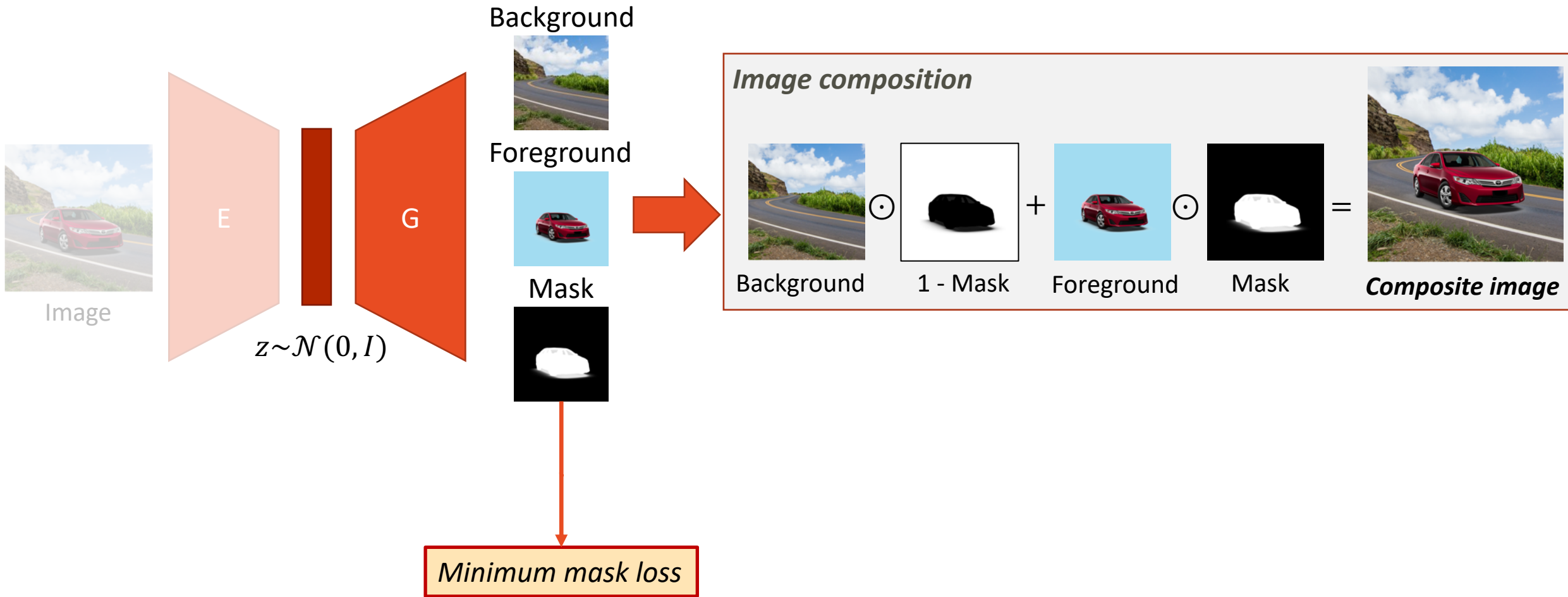
Composite image



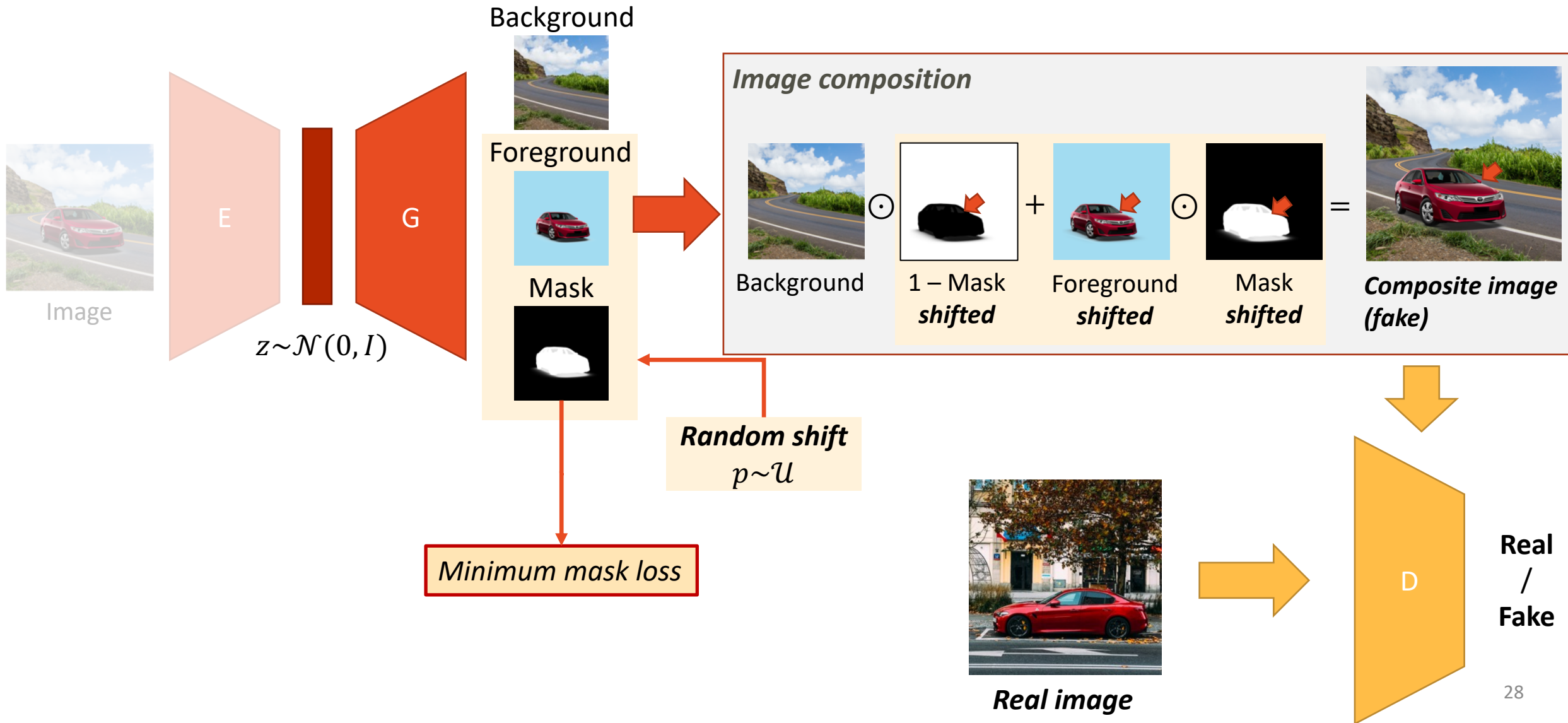
Composite image w/shift



# Learning with perturbations

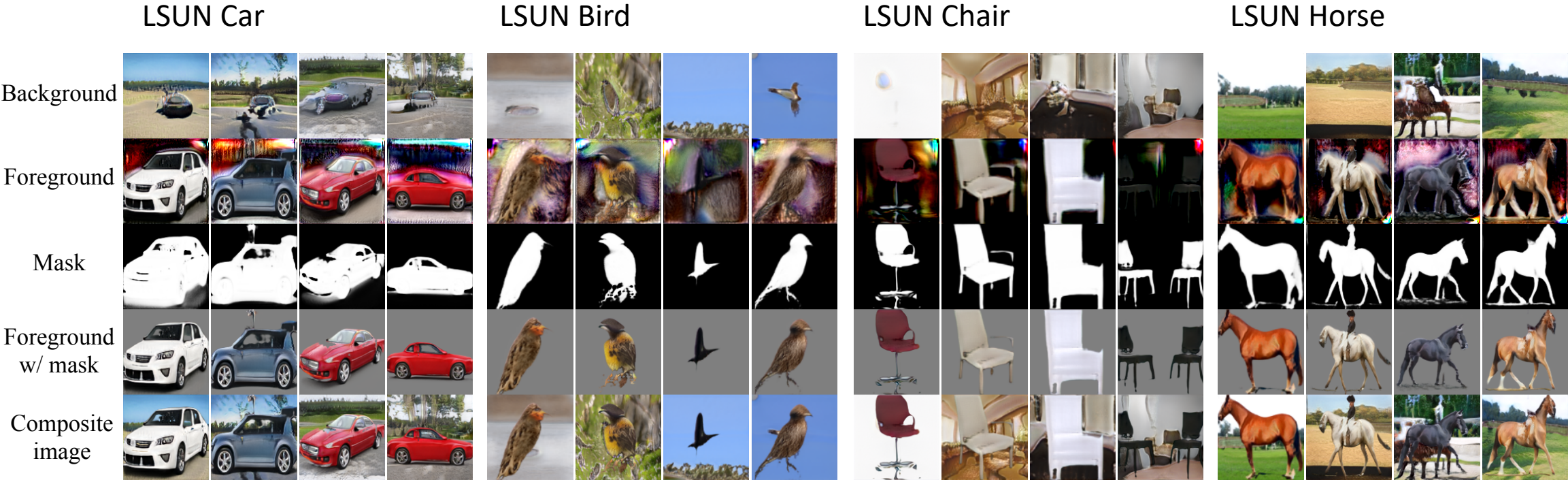


# Learning with perturbations

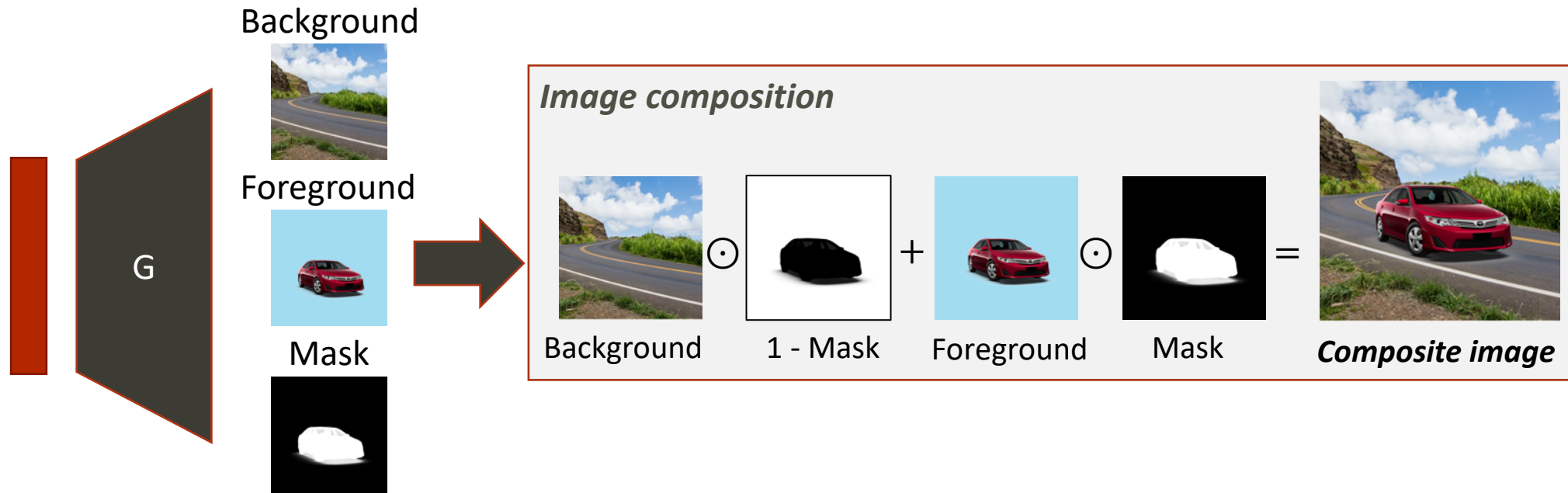




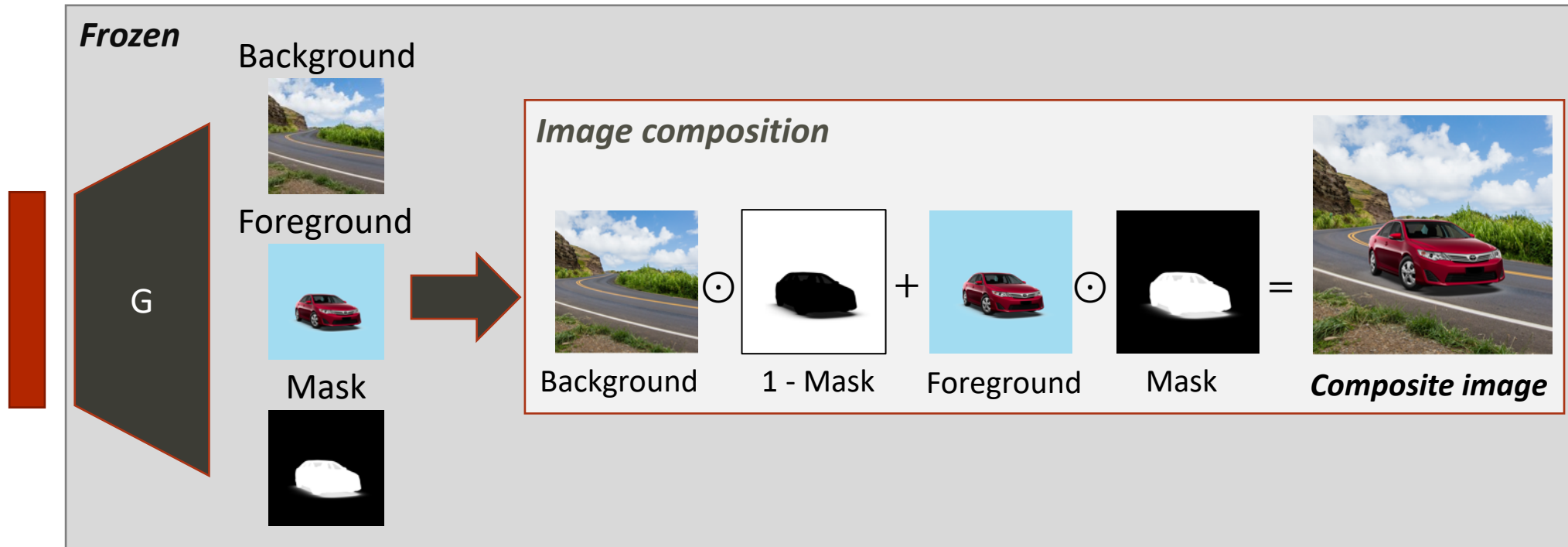
# Generation results



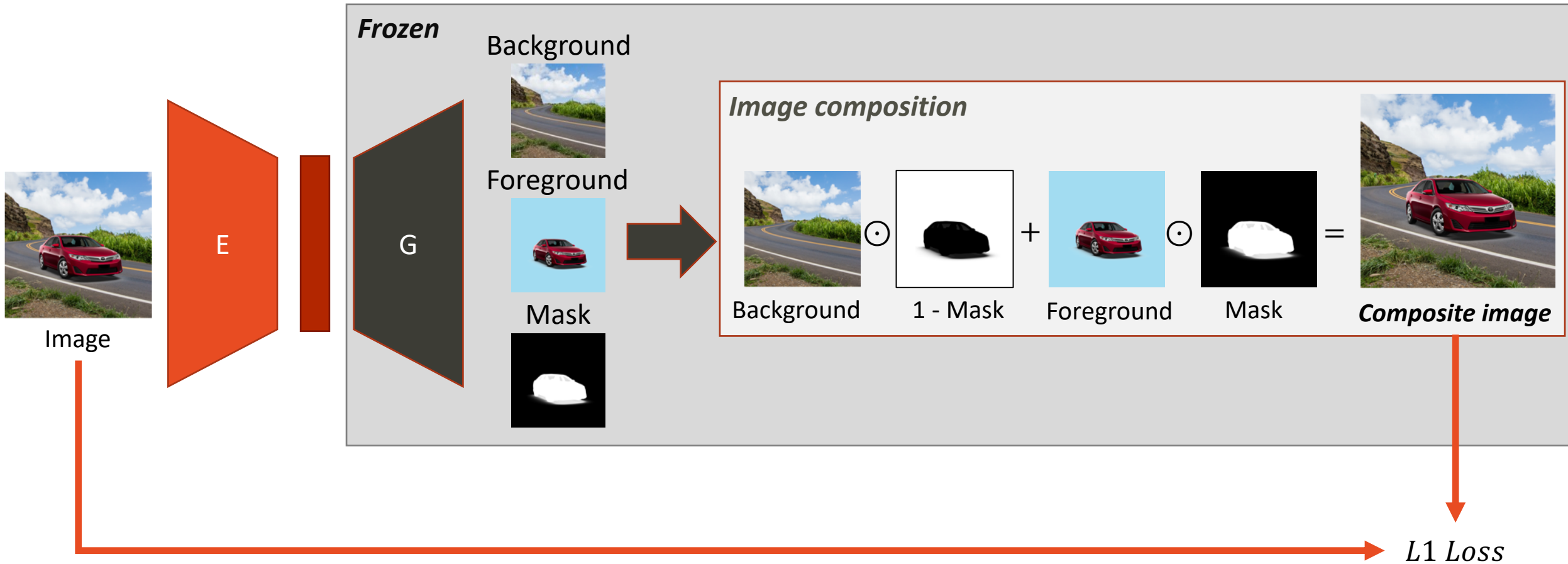
# Learning to segment



# Learning to segment

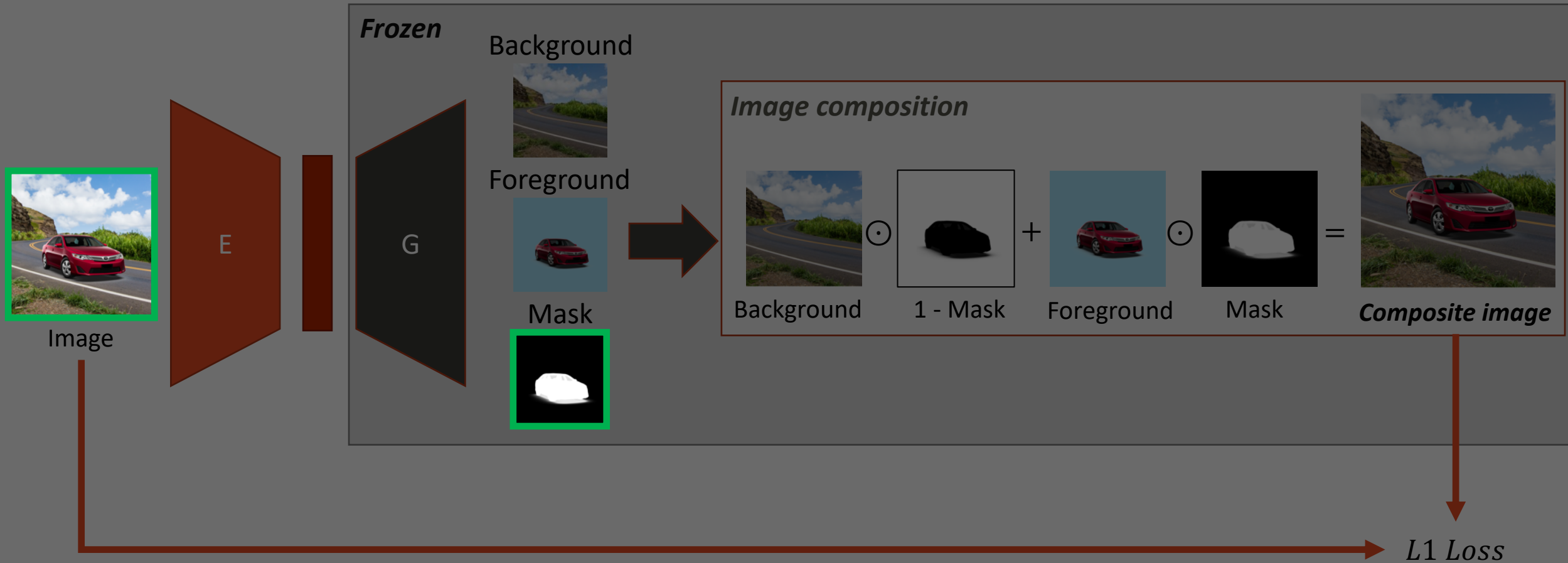


# Learning to segment





# Learning to segment







### Learning to generate layered scenes

We train a generative model that produces a layered image representation: **background, foreground and mask**.

We render a full image through alpha compositing.



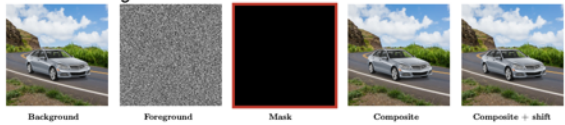
We notice that if the segmentation is valid, we can apply a small random shift to the foreground and still get a valid composite image

Example: If background and foreground images are the same, any mask produces a valid composite image



Random shift exposes invalid segmentation

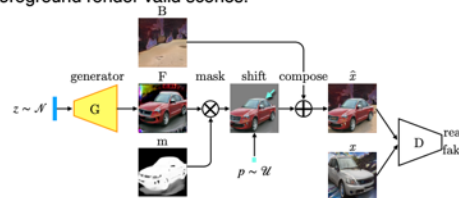
If the generated mask is empty, the composite image is valid even when we shift the foreground



We add a loss term to avoid empty masks

### Implementation

We train a StyleGAN with two generators, separate for a background and a foreground with a mask. It is trained so that the composite images with a shifted foreground render valid scenes.



We define two loss terms on generated masks to encourage binarization and assert minimum mask coverage and add them to WGAN-GP generator loss.

$$\mathcal{L}_{size} = \gamma_1 \mathbb{E}_{z \sim \mathcal{N}(0, I_d)} [\max\{0, \eta - 1/MN|G_m(z)|_1\}]$$

$$\mathcal{L}_{binary} = \gamma_2 \mathbb{E}_{z \sim \mathcal{N}(0, I_d)} [\min\{G_m(z), 1 - G_m(z)\}]$$

### Generator results



StyleGAN trained on 100k images for 4 categories from LSUN object dataset: Car, Horse, Chair, Bird.

Minimum mask coverage set to 25%, 20%, 15%, 15% respectively.

### Ablation study

We validate the importance of random shift and other parameters on LSUN Car.

Setting	64 × 64			128 × 128		
	mIoU	reference mIoU	detected cars	mIoU	reference mIoU	detected cars
(a) Default parameters	0.685	0.440	6295	0.533	0.432	7090
(b) No shift $\beta = 0$	0.039	0.428	6738	0.025	0.419	7978
(c) 25% shift $\beta = 0.25$ -best	0.144	0.434	6493	0.064	0.426	7259
(d) Bg contrast jitter = (0.7, 1.3)	<b>0.765</b>	0.454	6089	<b>0.673</b>	0.436	7046
(e) No random crops	0.264	0.374	6339	0.136	0.365	7520
(f) Mask size $\gamma = 10.1$	0.737	0.443	6245	0.643	0.430	7241
(g) Min. mask size $\eta = 5\%$	0.693	0.458	6202	0.552	0.430	7256
(h) Single generator	0.550	0.446	<b>6903</b>	0.464	0.435	7544

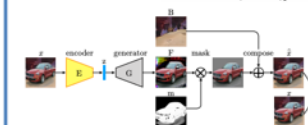
We train the generator on a dataset with two object categories



### Segmentation

We train encoders with a fixed generator to get segmentation for real images

$$\mathcal{L}_{auto} = \mathbb{E}_{x \sim p_x} \|x_E - x\|_1 + \mathbb{E}_{x \sim p_x} \|D_{real}(x_E) - D_{real}(x)\|_2^2$$



Setting	LSUN Car		CUB-200	
	mIoU	mIoU	mIoU	mIoU
Our method	0.540	0.380		
GrabCut	0.559	0.453		
Full image mask	0.402	0.132		



Poster #60

East Exhibition Hall B + C