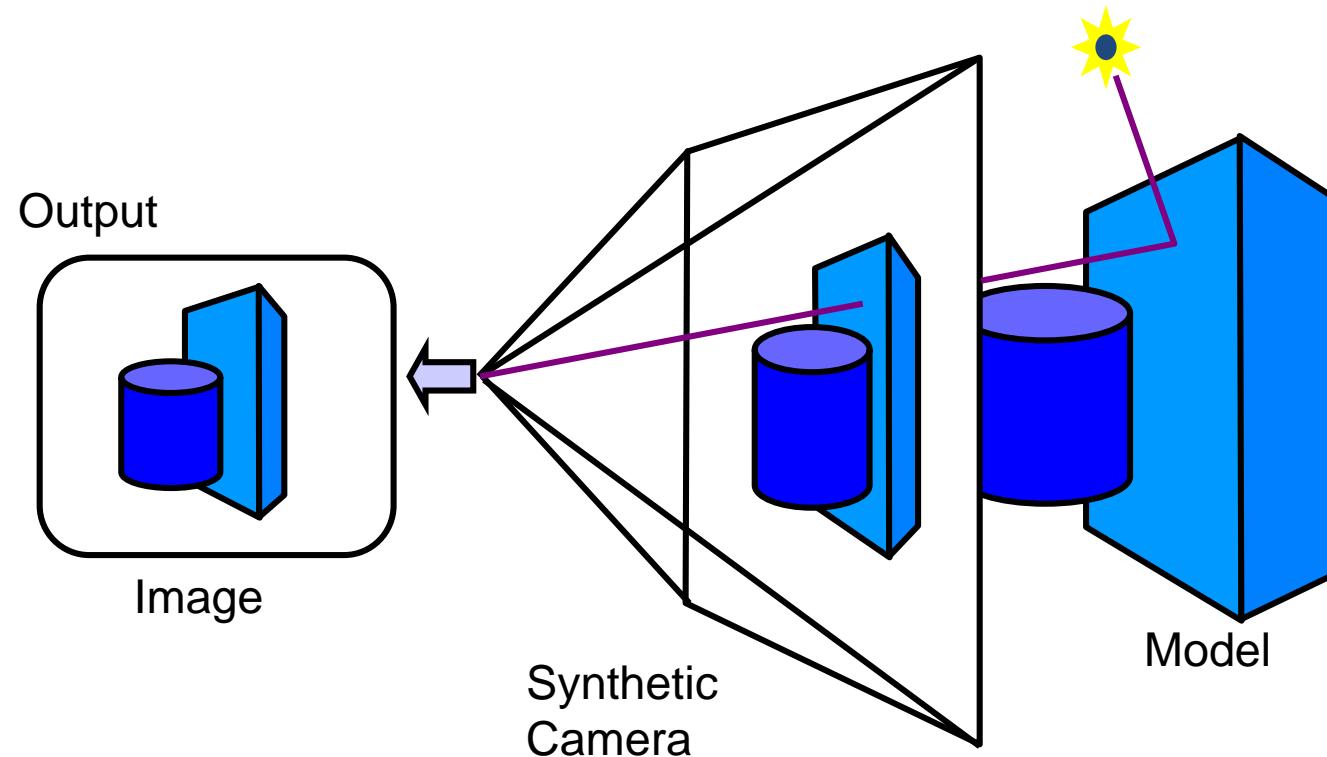


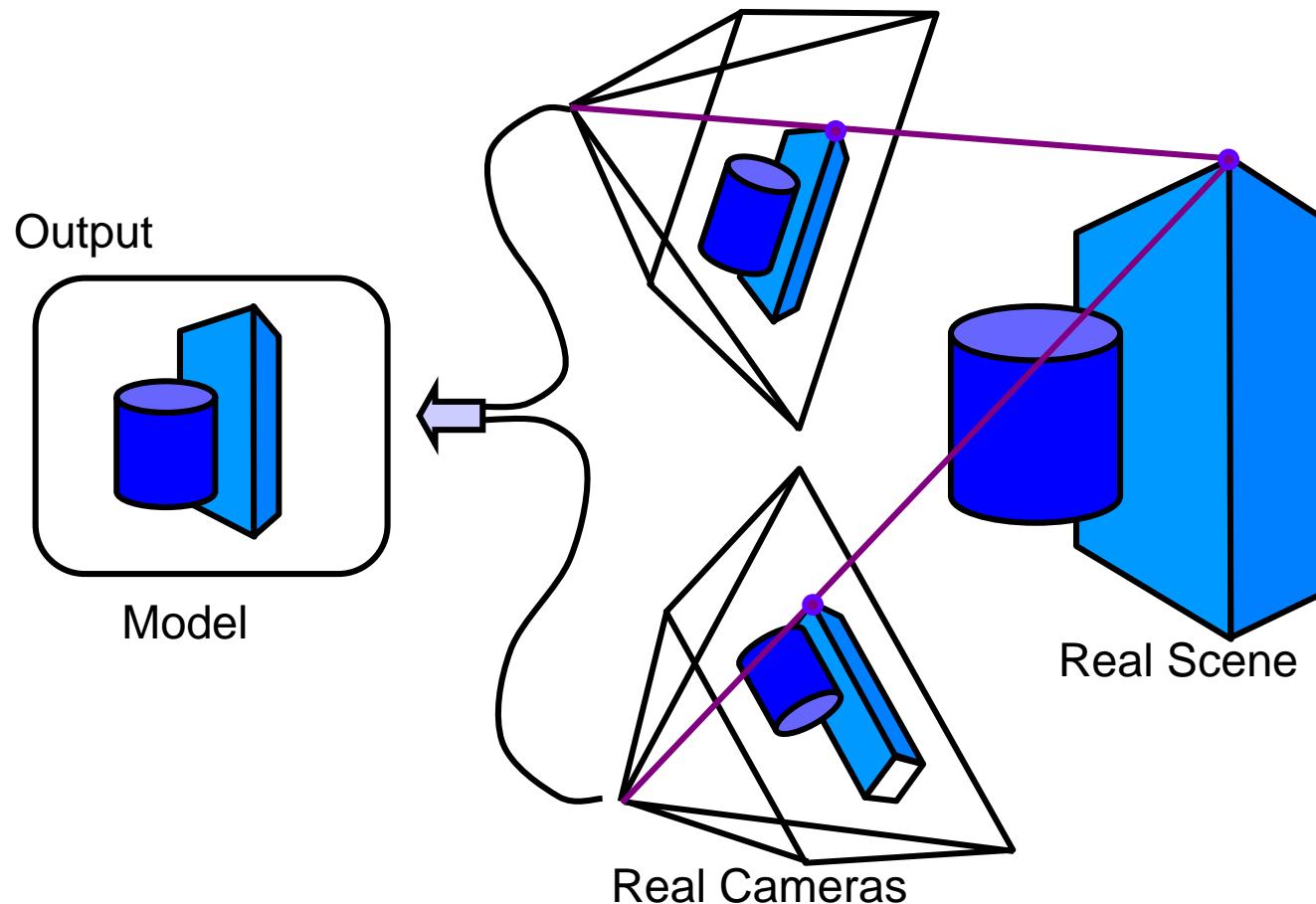
Machine Learning in Computer Vision and Graphics

6.036 Introduction to Machine Learning

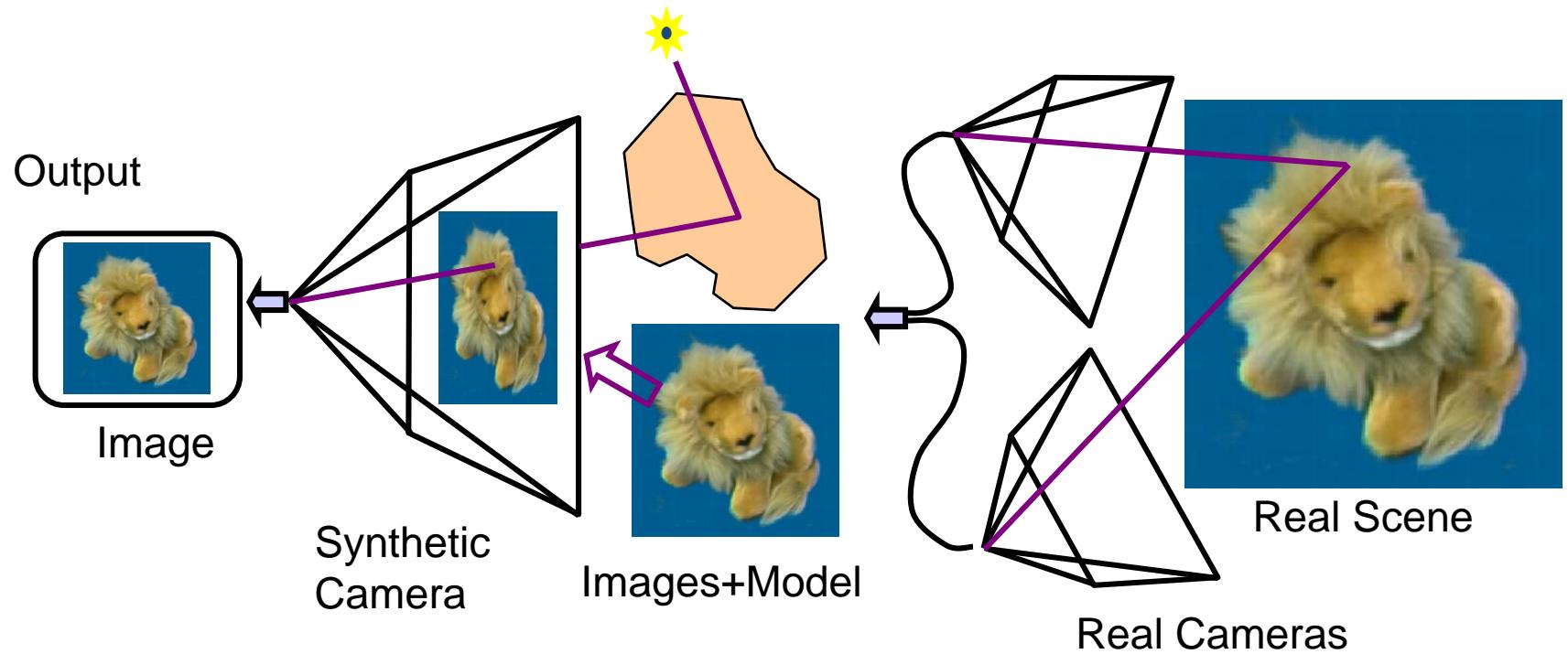
Computer Graphics – Synthesis



Computer Vision - Analysis



Often Together



Machine learning is a core component of many vision and graphics algorithms

Let's start with something familiar...

Project 3

Image Segmentation

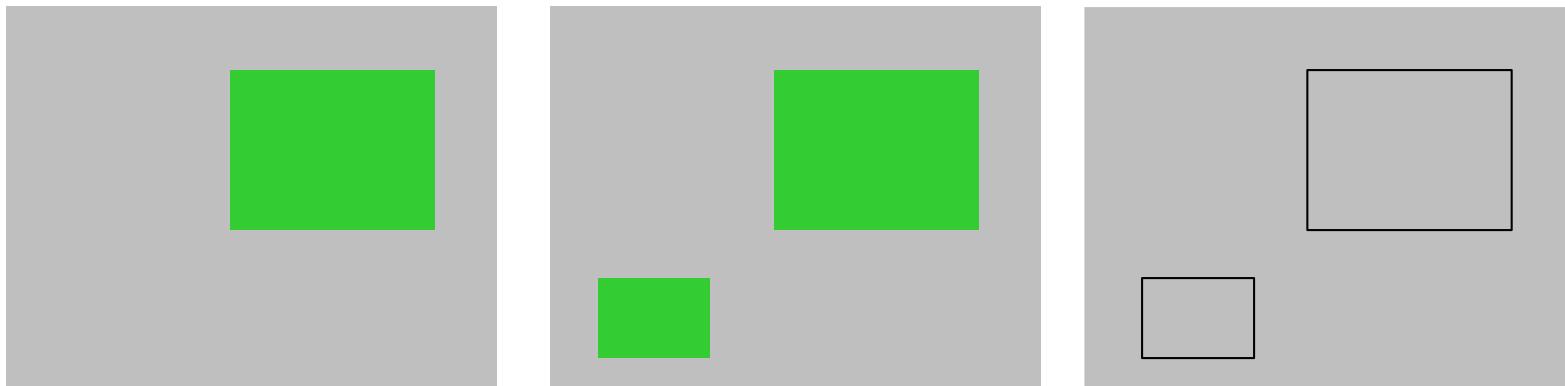
- Split an image into non-overlapping regions
- Group together similar-looking pixels



“superpixels”

Issues

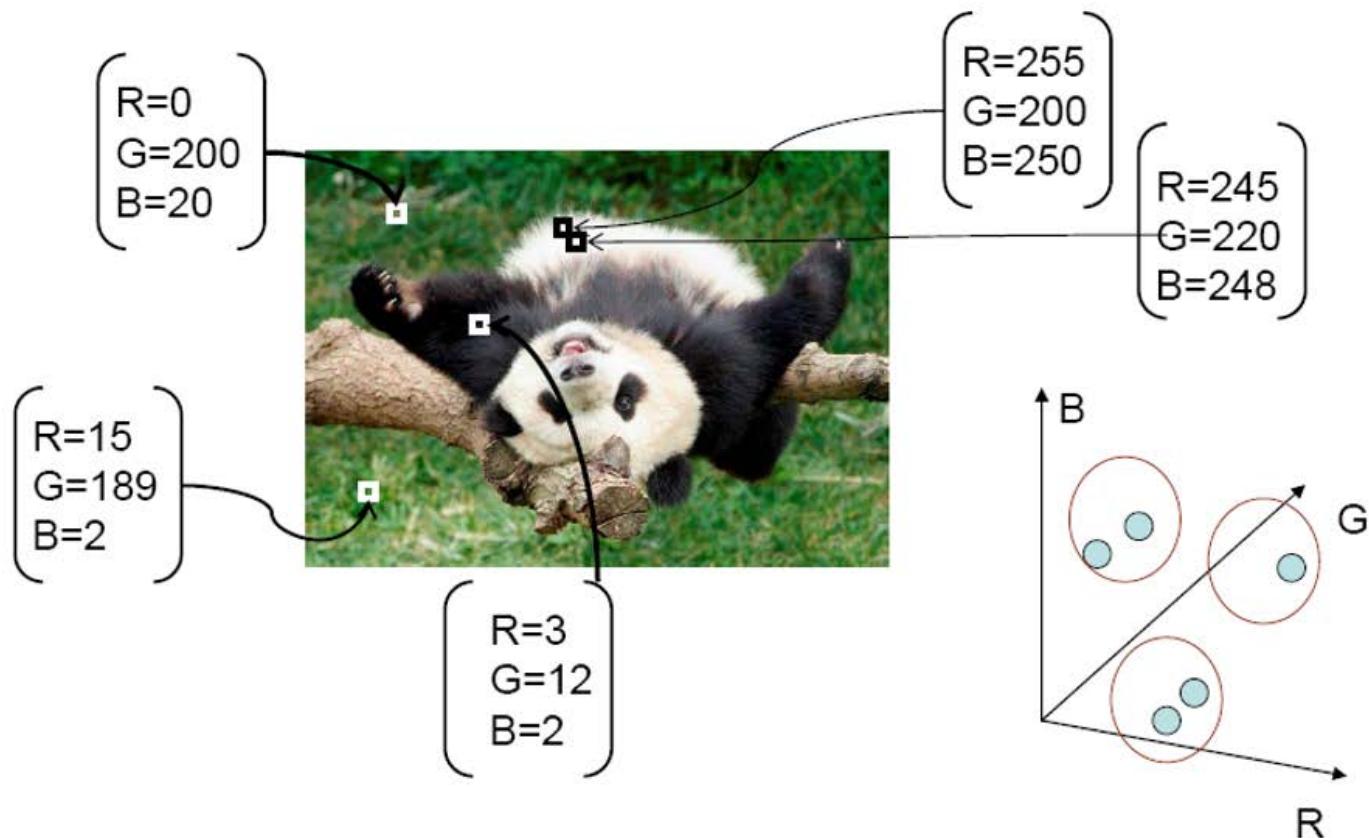
- How do we decide that two pixels are likely to belong to the same region?



- How many regions are there?

Segmentation: Clustering

- Cluster similar pixels together



A Simple Segmentation Algorithm

- Each pixel is described by a vector
 - e.g., $z = [r, g, b]$
- Run a clustering algorithm (e.g. k-means) using some distance between pixels:

$$D(\text{pixel}_i, \text{pixel}_j) = || z_i - z_j ||^2$$

K-Means Clustering

- Given k , the k -means algorithm consists of four steps:

- Select initial centroids at random.
- Assign each object to the cluster with the nearest centroid.
- Compute each centroid as the mean of the objects assigned to it.
- Repeat previous 2 steps until no change.

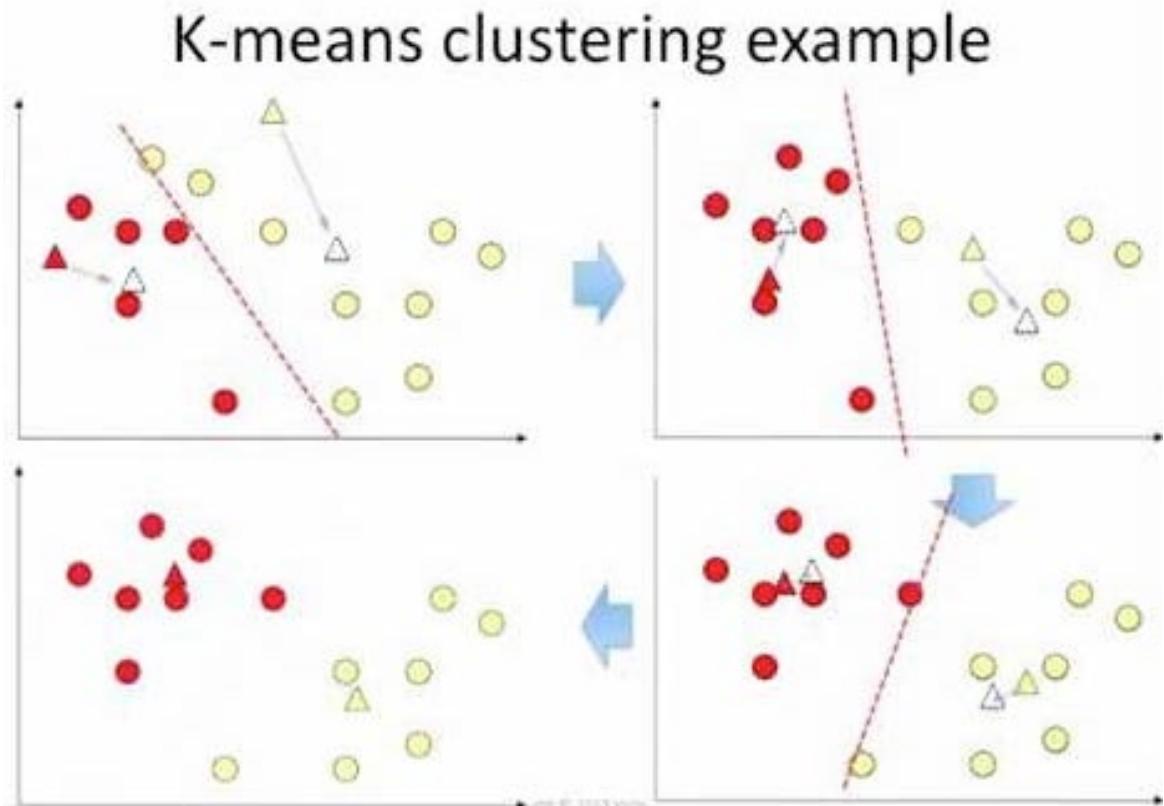
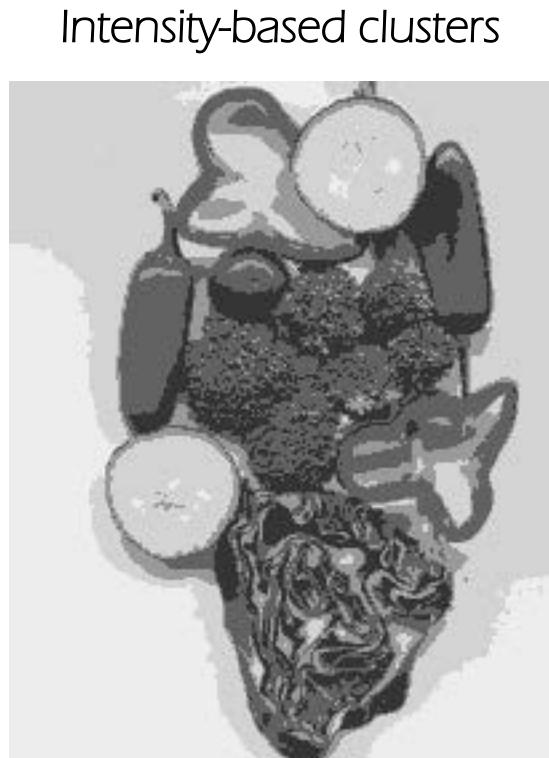


Image Segmentation: K-Means

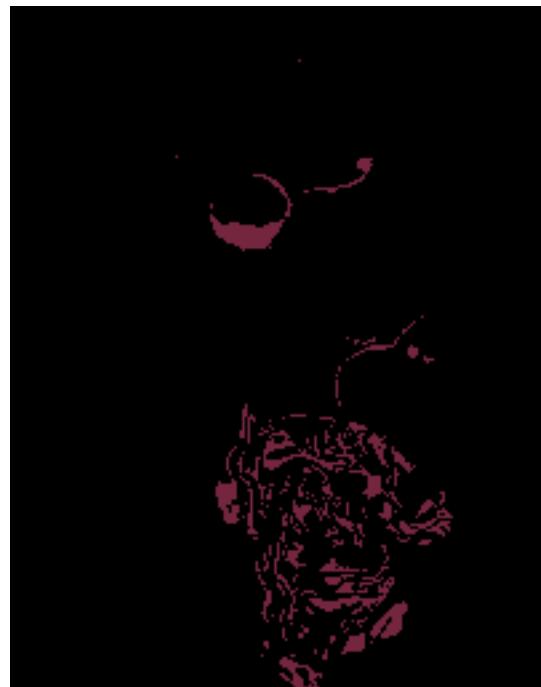
- K-means ($k=5$) clustering based on intensity (middle) or color (right) is essentially vector quantization of the image attributes



each pixel is replaced with the mean value of its cluster



K-means using
color alone
($k=11$ clusters)
Showing 4 of the
segments, (not
necessarily connected)
Some are good, some
meaningless



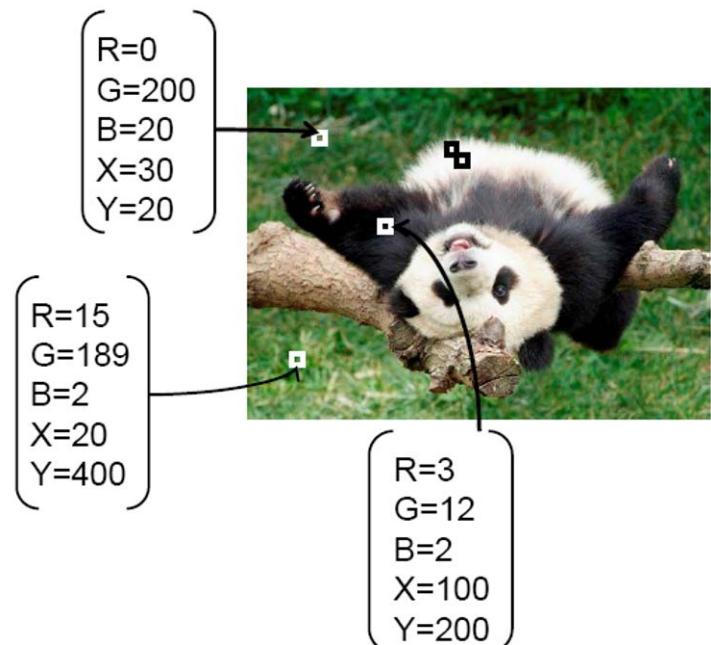
Including Spatial Relationships

Augment data to be clustered with spatial coordinates.

$$z = \begin{pmatrix} r \\ g \\ b \\ x \\ y \end{pmatrix}$$

color coordinates
(or r,g,b)

spatial coordinates



K-Means based on (r,g,b,x,y)

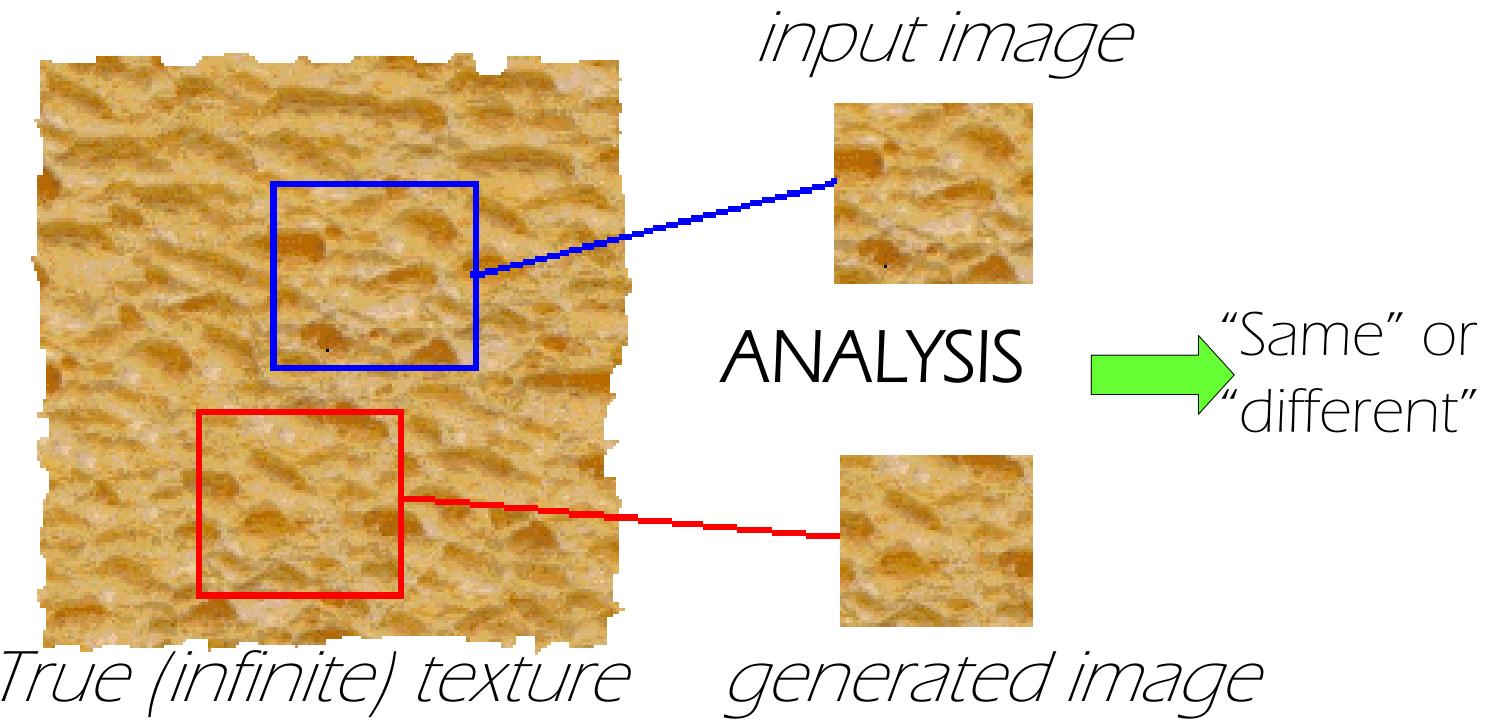


K-means using color and position, 20 segments



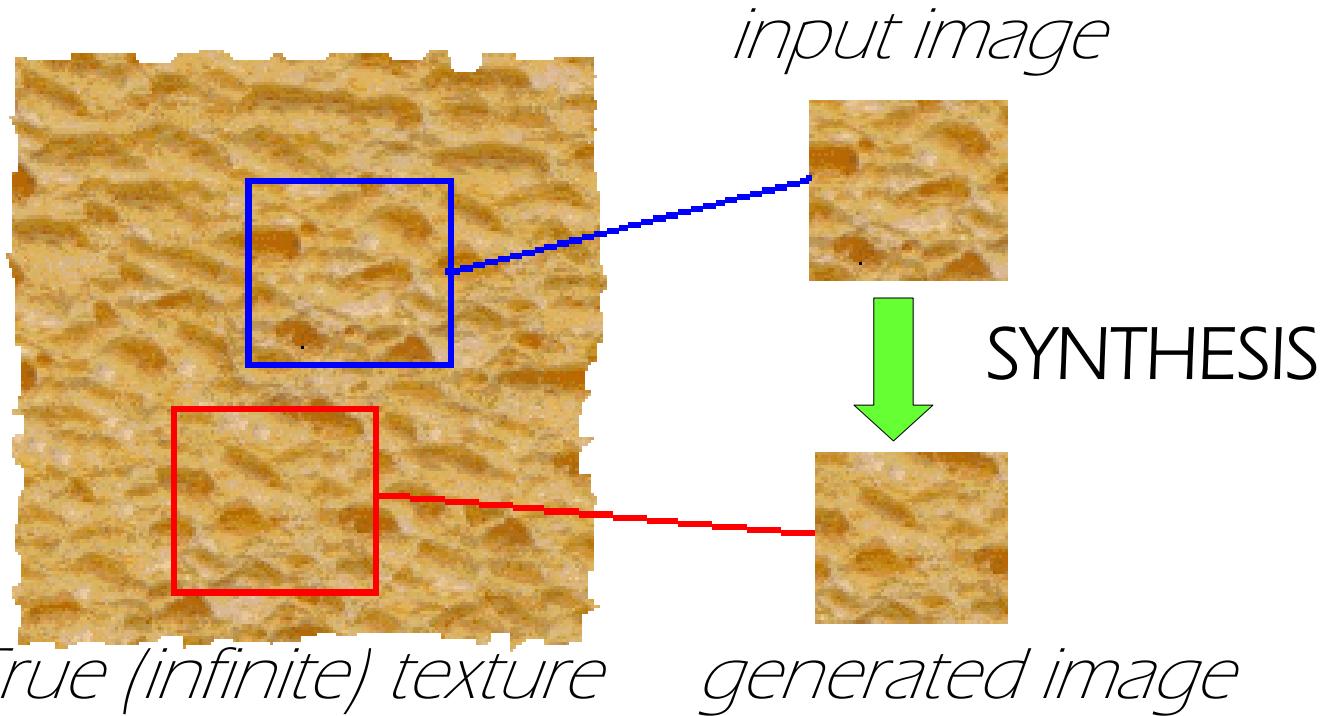
Let's look at other classic
problems

Texture Analysis



Compare textures and decide if they're made of the same “stuff”.

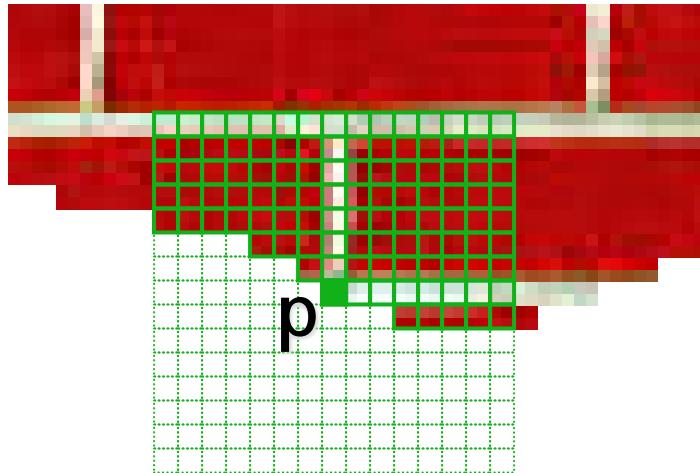
Texture Synthesis



Given a finite sample of some texture, the goal is to synthesize other samples from that same texture

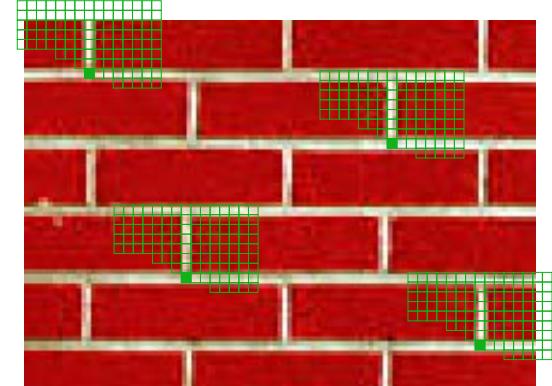
- The sample needs to be "large enough"

Textures Synthesis



Synthesizing a pixel

non-parametric
sampling

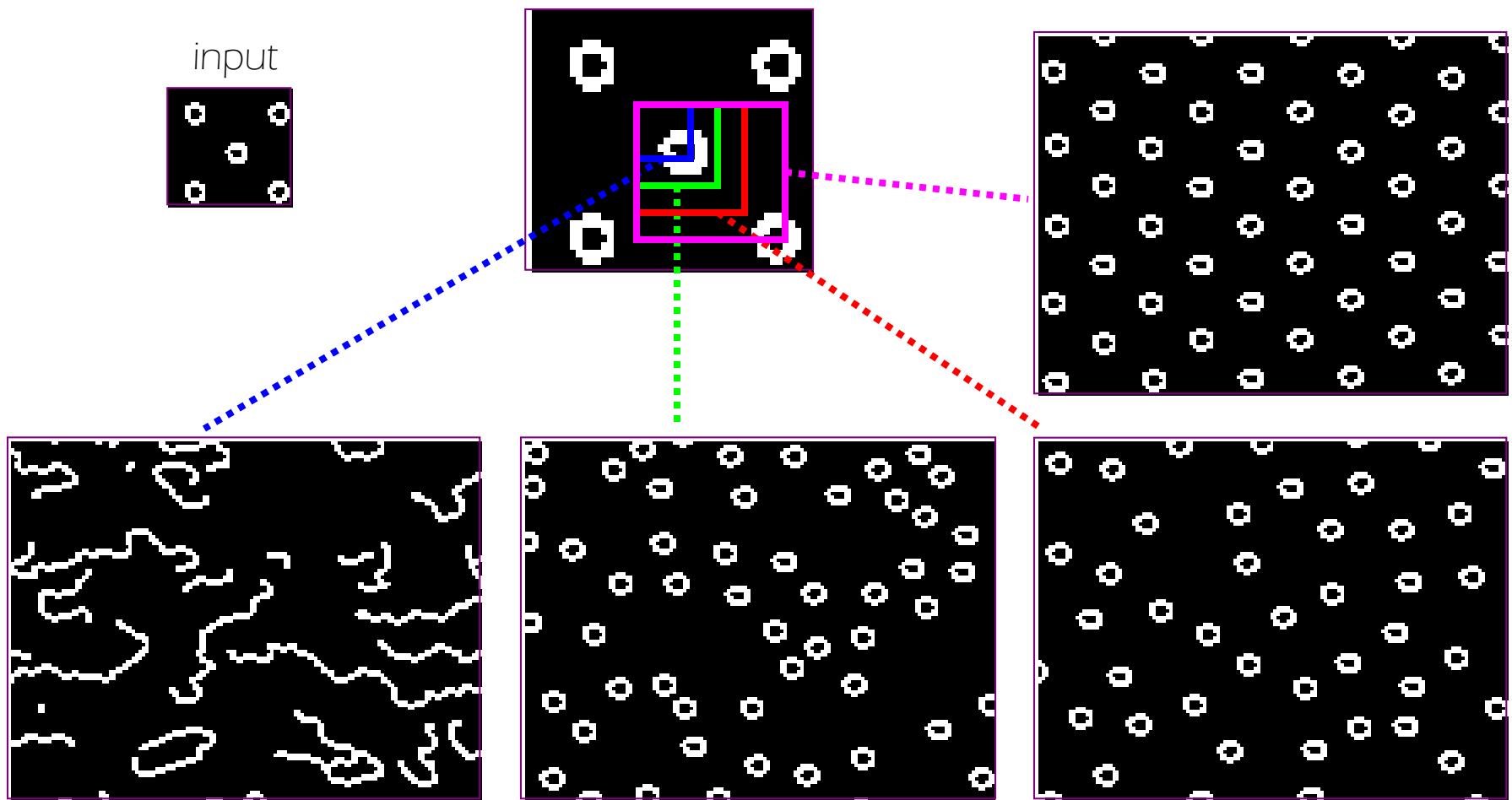


Input image

Assuming Markov property, compute $P(p | N(p))$

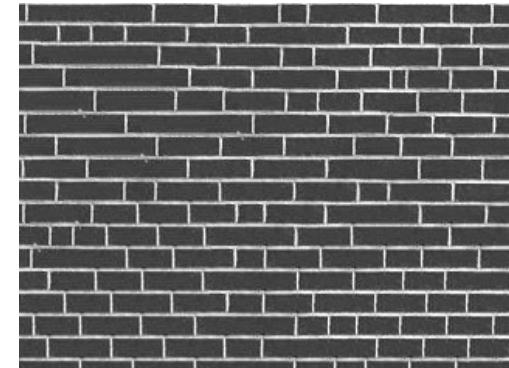
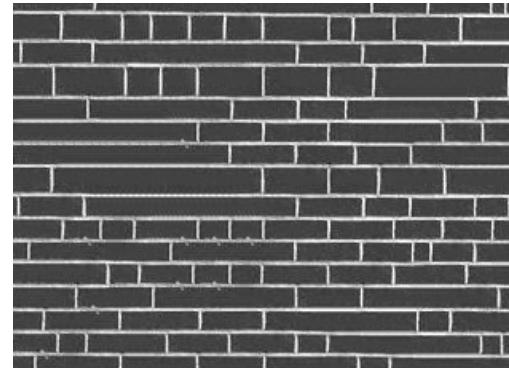
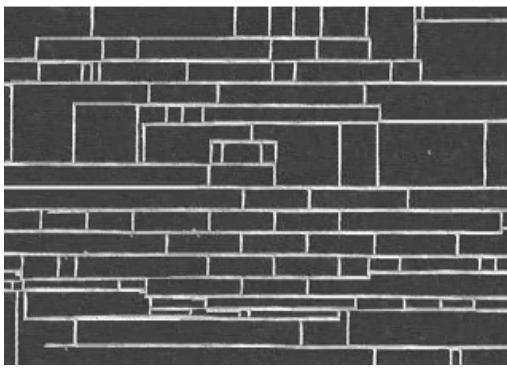
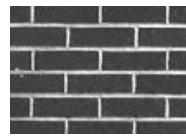
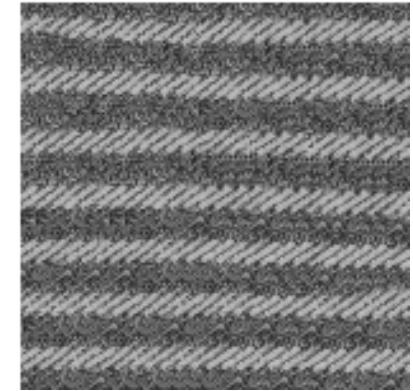
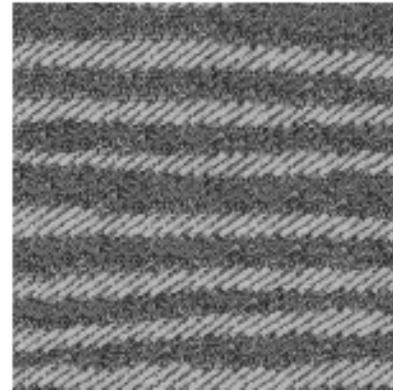
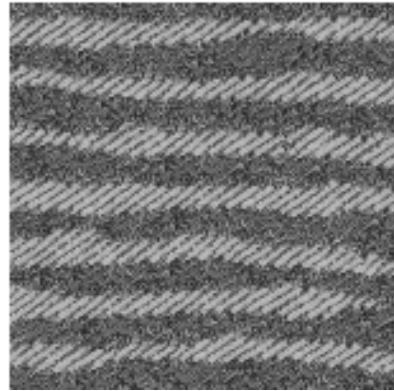
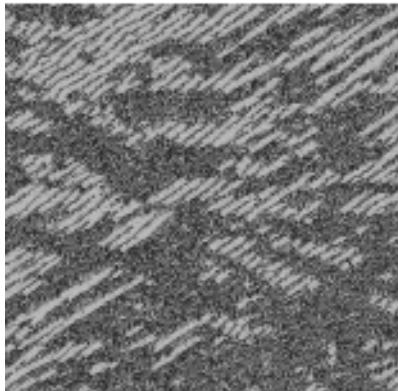
- Building explicit probability tables infeasible
- Instead, we *search the input image* for all similar neighbourhoods – that's our pdf for p
- To sample from this pdf, just pick one match at random

Neighborhood Window



Source: Efros

Varying Window Size



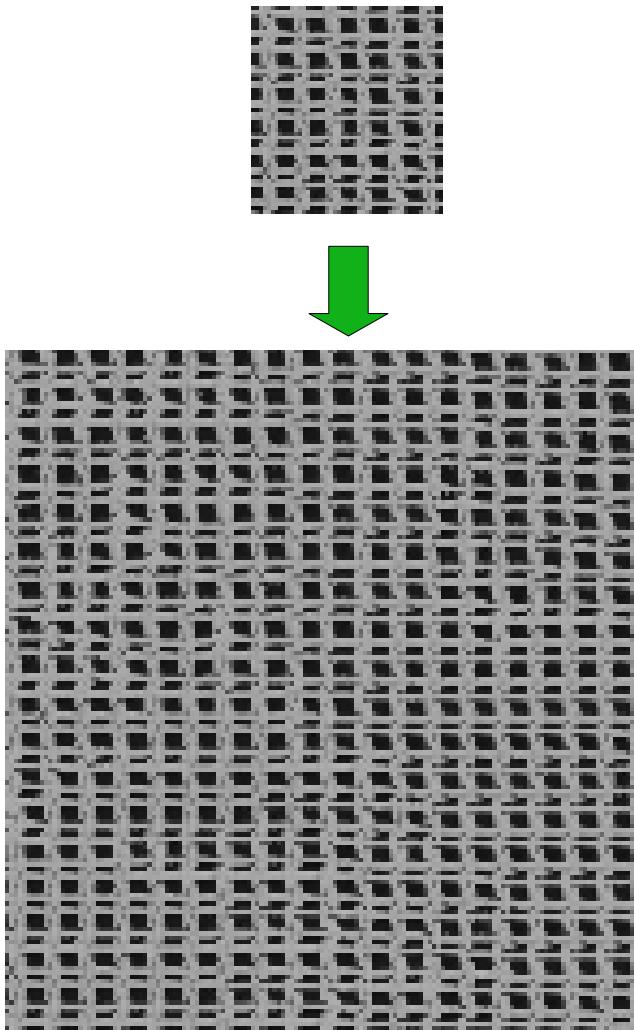
Increasing window size



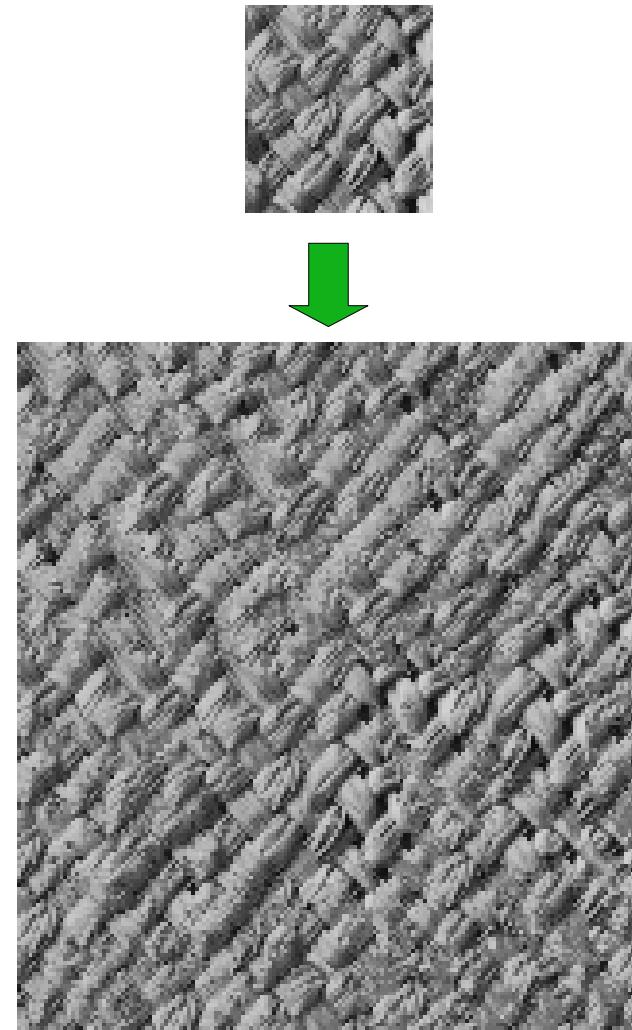
Source: Efros

Synthesis Results

French canvas



rafia weave



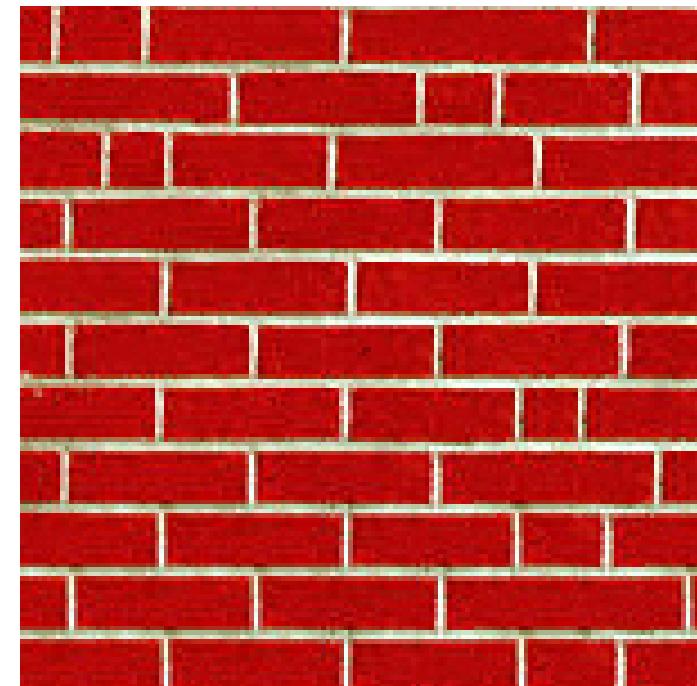
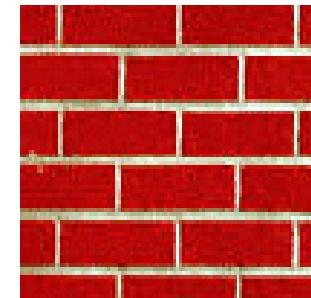
Source: Efros

More Results

white bread



brick wall



Source: Efros

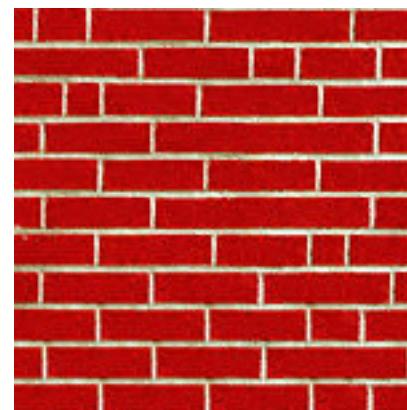
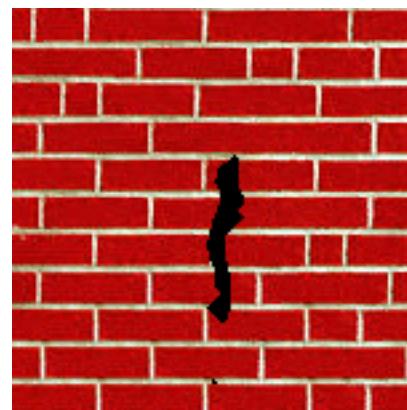
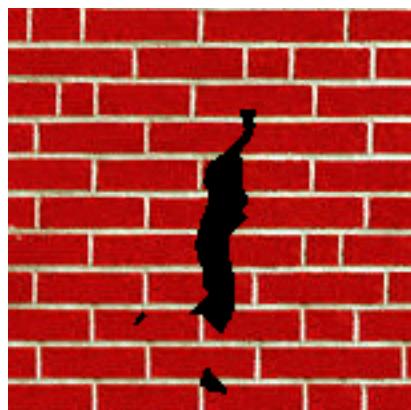
Homage to Shannon

uring in the unsensational
r Dick Gephardt was fai-
rful riff on the looming
nly asked, "What's your
tions?" A heartfelt sigh
story about the emergenc-
es against Clinton. "Boy-
g people about continuin-
ardt began, patiently obse-
, that the legal system ha-
g with this latest tangle

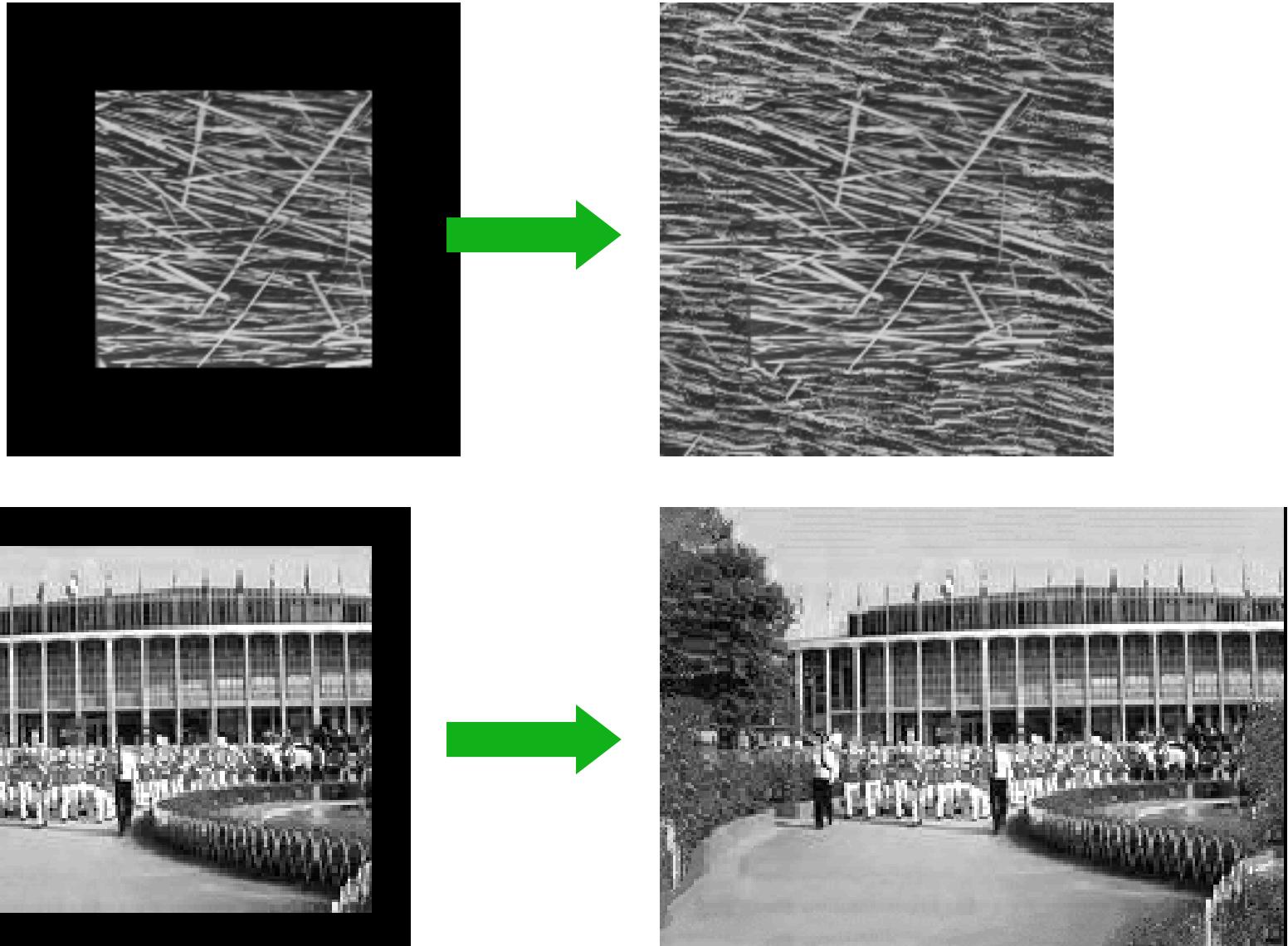
1 b' vobis odi' reh' h'ht' l'mr' al thanne
e g le fer' s otot' " latim' n' irith' feoir' A' un' a
er' ag as " he' c' il' er' eior' t' a' o' r' A' l' b
h' t' a' yf' te' t' fit' a' us' t' ift' foear' f' cat' h' u
t' a' e' t' diab' h' ob' i' v' h' u' h' nr' te' opm' t' p' t' o
t' e' t' efutifl' t' u' s' l' h' r' h' r' i' h' p' n' t' k
n' h' l' t' seu' v' o' s' b' n' t' u' m' h' r' i' h' p' n' t' k
n' id' t' f' p' f' e' t' yd' g' u' l' a' g' l' i' r' l' e' n' r' i' m' a' b' n' s
u' t' o' n' u' s' f' u' h' e' s' i' l' h' n' o' b' o' h' i' s
" j' i' t' h' e' n' l' h' n' o' b' o' h' i' s
" d' t' h' f' p' d' i' l' o' n' f' i' n' e' o' t' i' h' a' muuny' c' r' o' b' a' b' i' s
x' r' i' b' " s' l' h' s' k' a' s' h' e' k' v' s' j' n' c' u' e' j' i' f' n' i' s
e' e' s' t' f' h' s' k' a' s' h' e' k' v' s' j' n' c' u' e' j' i' f' n' i' s
u' t' i' n' t' r' l' n' i' t' " t' f' f' a' f' e' c' + d' t' e' l' r' h' c' r' a' r' w' s
s' C' o' n' h' n' u' n' l' e' g' d' t' f' a' f' e' c' + d' t' e' l' r' h' c' r' a' r' w' s
s' t' h' i' n' g' e' o' r' r' e' n' e' s' h' u' b' y' a' n' i' s' A' h' o' c' f' r' y' p' r' s
f' h' o' e' t' f' r' e' n' e' s' h' u' b' y' a' n' i' s' A' h' o' c' f' r' y' p' r' s
d' v' i' t' e' f' l' y' i' e' v' o' n' e' t' h' i' f' f' b' l' i' p' " u' l' l' t' " b' j' e' i' s
d' e' f' p' s' t' h' e' l' g' e' v' e' r' n' r' r' " t' h' i' f' f' b' l' i' p' " u' l' l' t' " b' j' e' i' s
t' e' n' A' " a' r' e' n' l' o' m' t' i' n' f' u' c' v' t' o' t' h' p' n' i' s' t' a' n' i' s
m' v' i' o' p' " r' t' e' w' a' j' o' b' u' i' t' e' t' e' t' i' s
3' u' n' u' h' n' o' r' s' h' p' s' a' " h' b' + c' o' a' i' t' i' w' w' r' f' r' t' i'

ithairm. them . "Whnephartfe lartifelintomimen
sel ck Clirtioout omaim thartfelins. f ou t s anento
the ry onst wartfe lck Gephntoomimeationl sigak
Chiooufit Clinut Cll riff on. hat's yordn, parut tly
ons yoontonsteht wasked, paim t sahe loo riff on
nskoneploourfeas leil A nst Clit, "Wleontongal s
k Cirtioouirtfepe ong pme abegal fartfenstemem
tiensteneltorydt telemepminsverdt was agemer
ff ons artientont Cling perme as urtfte atish, "Boui s
hal s fartfelt sig pedr tl'dt ske abounutie aboutioo
tfaonewwas yow aboronthardt thatins fain, ped, 'a
ins. them, pabout wasy arfuiit courly d, ln A h
ole emthringbooreme agas fa bontinsyst Clinut
ory about continst Clipeouinst Cloke agatiff out C
stome zinemen tly ardt beorabol n, thenly as t C
cons faimeme Diantont wat coutlyohgans as fan
ien, phrtfaul, "Wbaut cout congagal cõminga
mifmst Cliiy abon al coountha.emungaint tf oun
The looorysten loontieph. Intly on, theoplegatick C
ul tatiesontly atie Diantiomt wal s f tbegae ener
mthahsgat's enenhñnas fan, "intchthorw ahons w

Hole Filling

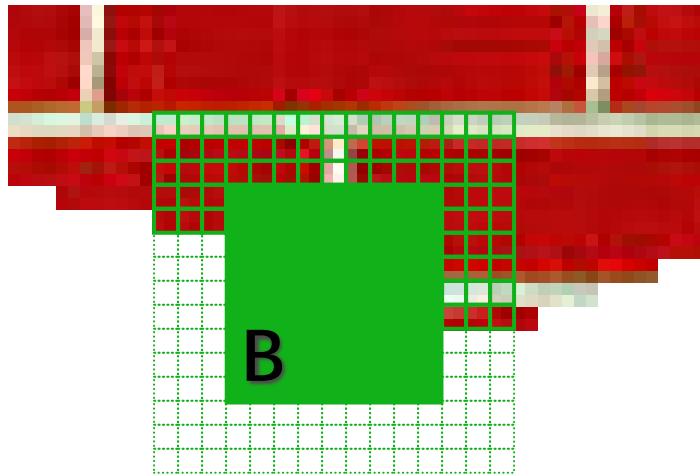


Extrapolation



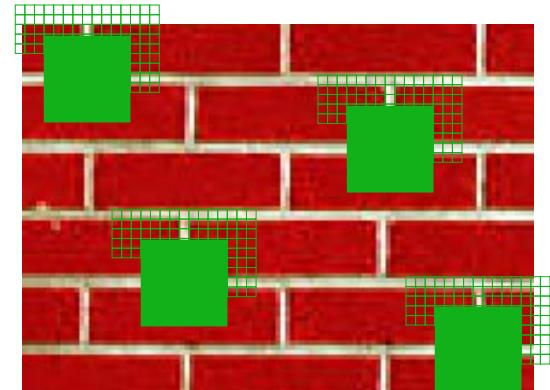
Source: Efros

Image Quilting



Synthesizing a block

non-parametric sampling

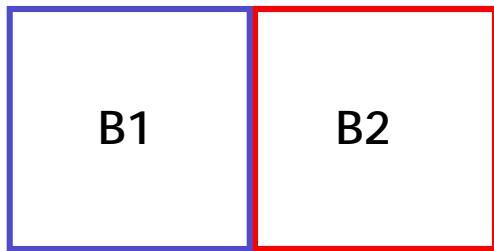
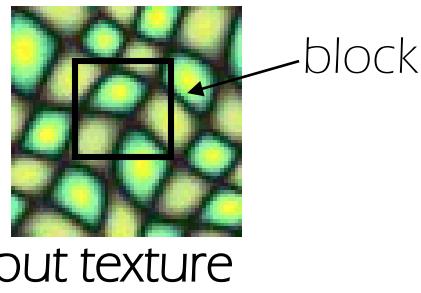


Input image

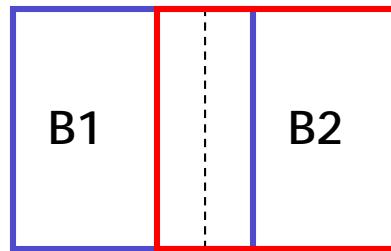
- Observation: neighbor pixels are highly correlated

Idea: unit of synthesis = block

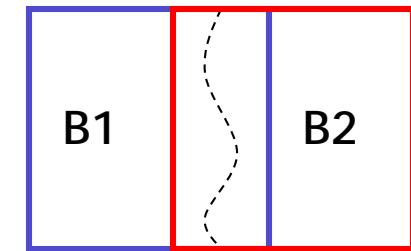
- Exactly the same but now we want $P(B | N(B))$
- Much faster: synthesize all pixels in a block at once
- Not the same as multi-scale!



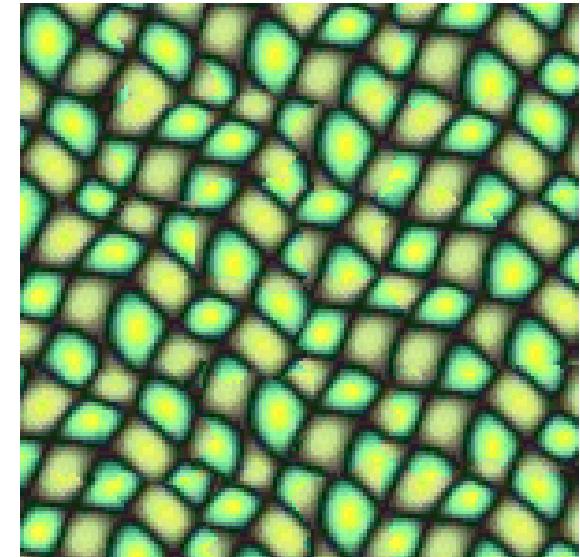
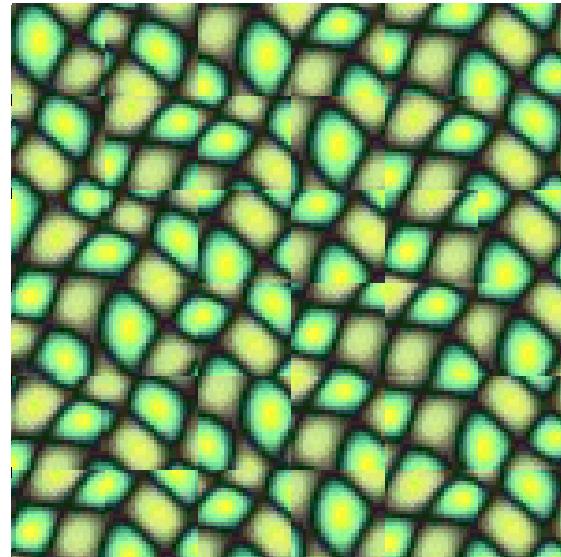
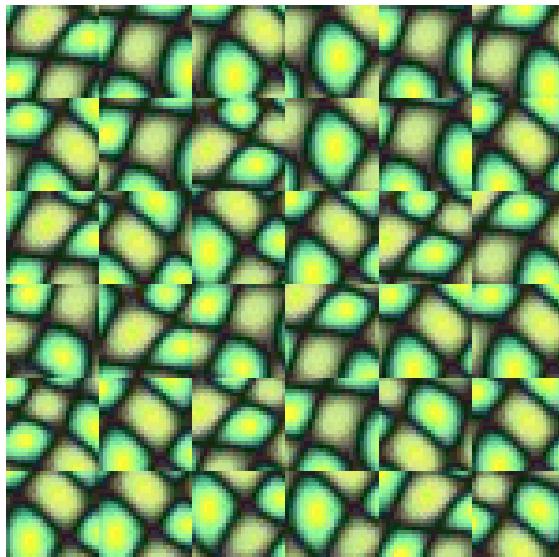
Random placement
of blocks



Neighboring blocks
constrained by overlap



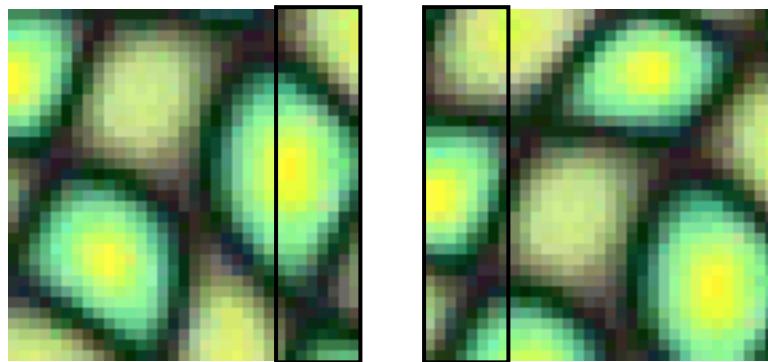
Minimal error
boundary cut



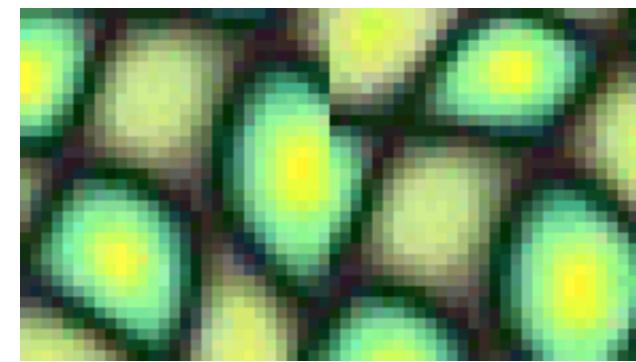
Source: Efros

Minimal Error Boundary

overlapping blocks



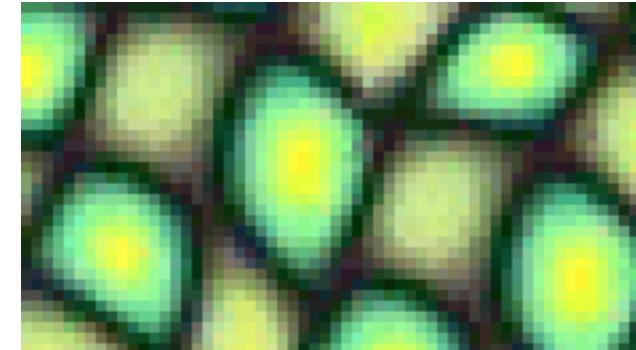
vertical boundary



A diagram showing the calculation of overlap error. It consists of two vertical patches of the image, one above the other. A minus sign is placed between them, indicating subtraction. To the right of the minus sign is an equals sign followed by a vertical bar. To the right of the bar is a red and black binary mask. Brackets on both sides of the minus sign group the two patches together. Above the entire equation is the number 2 , indicating that the result is squared to calculate the error.

$$\left[\begin{array}{c} \text{patch 1} \\ - \\ \text{patch 2} \end{array} \right] = \text{mask}$$

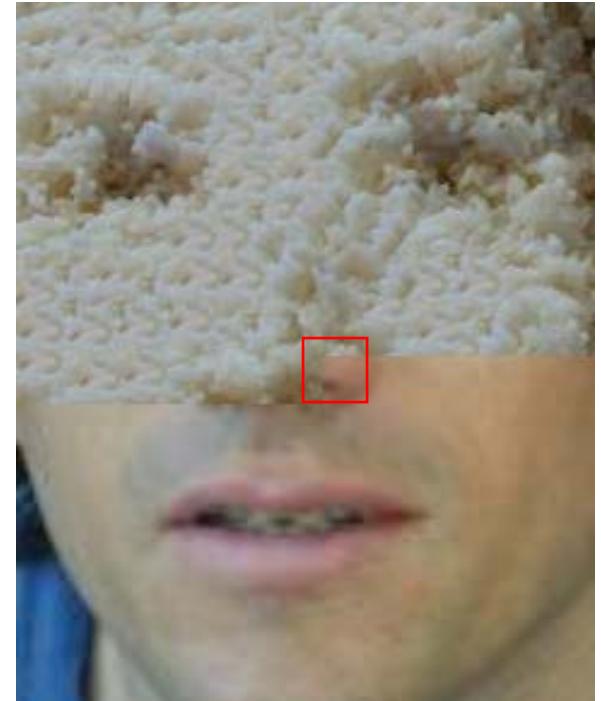
overlap error



min. error boundary

Texture Transfer

- Take the texture from one object and “paint” it onto another object
 - This requires separating texture and shape
 - That’s HARD, but we can cheat
 - Assume we can capture shape by boundary and rough shading

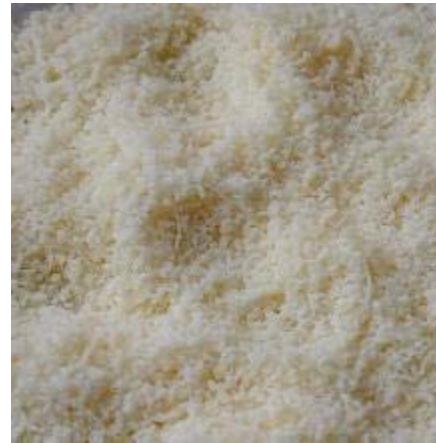


Then, just add another constraint when sampling:
similarity to underlying image at that spot



parmesan

+



=



rice

+



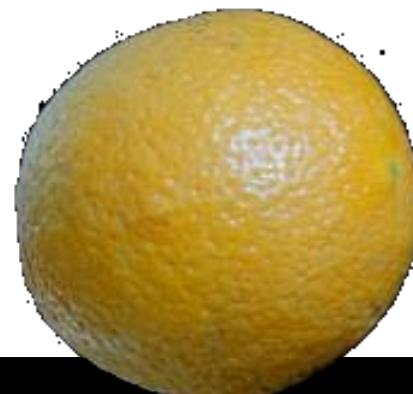
=



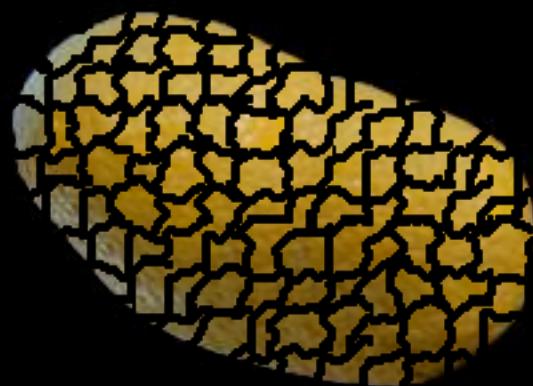
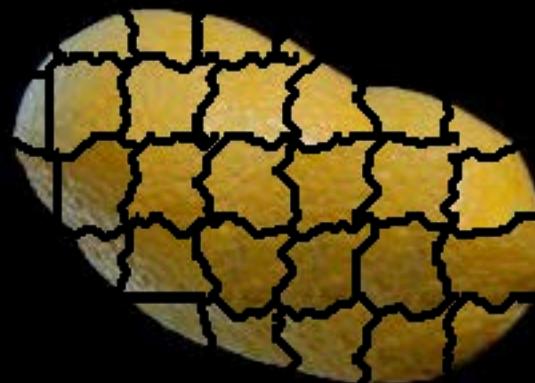
Source: Etros

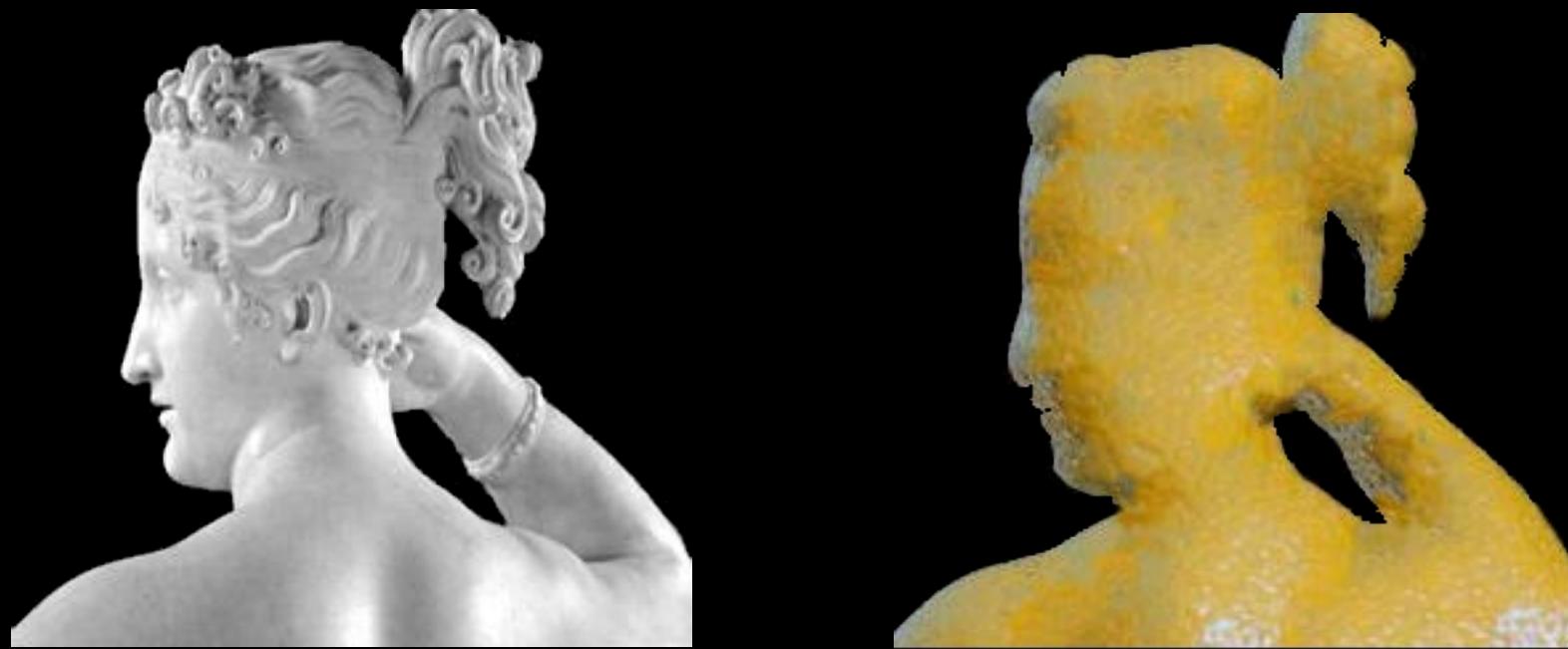
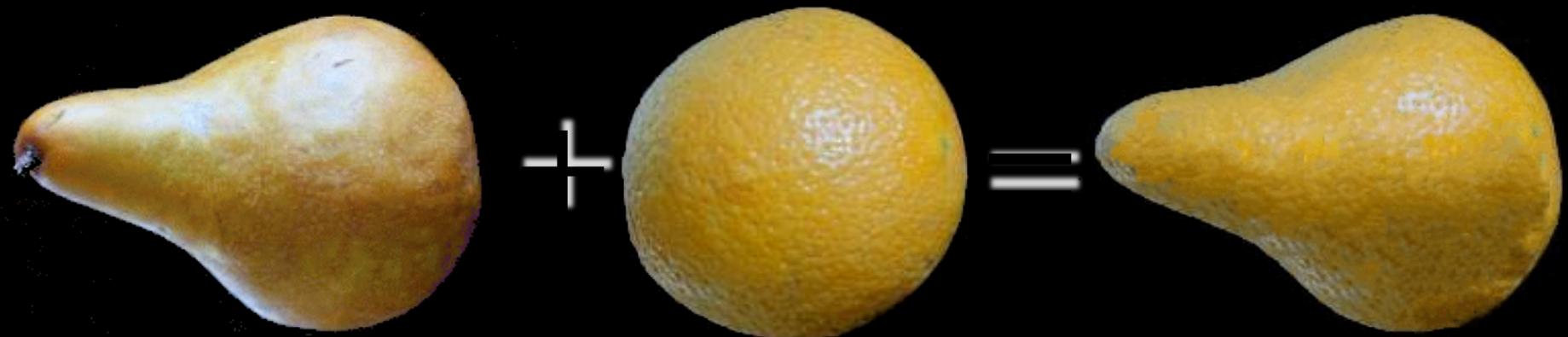


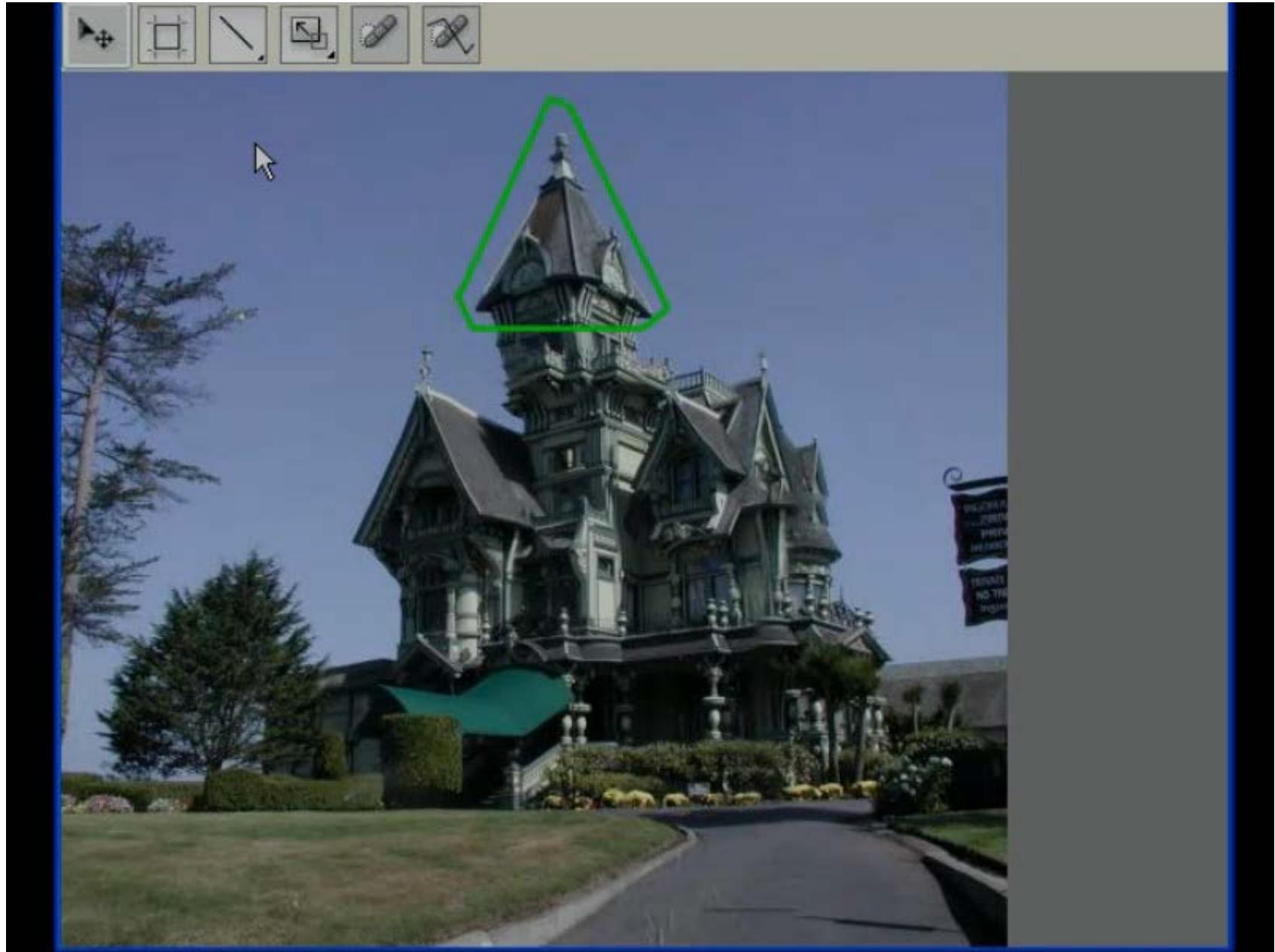
+



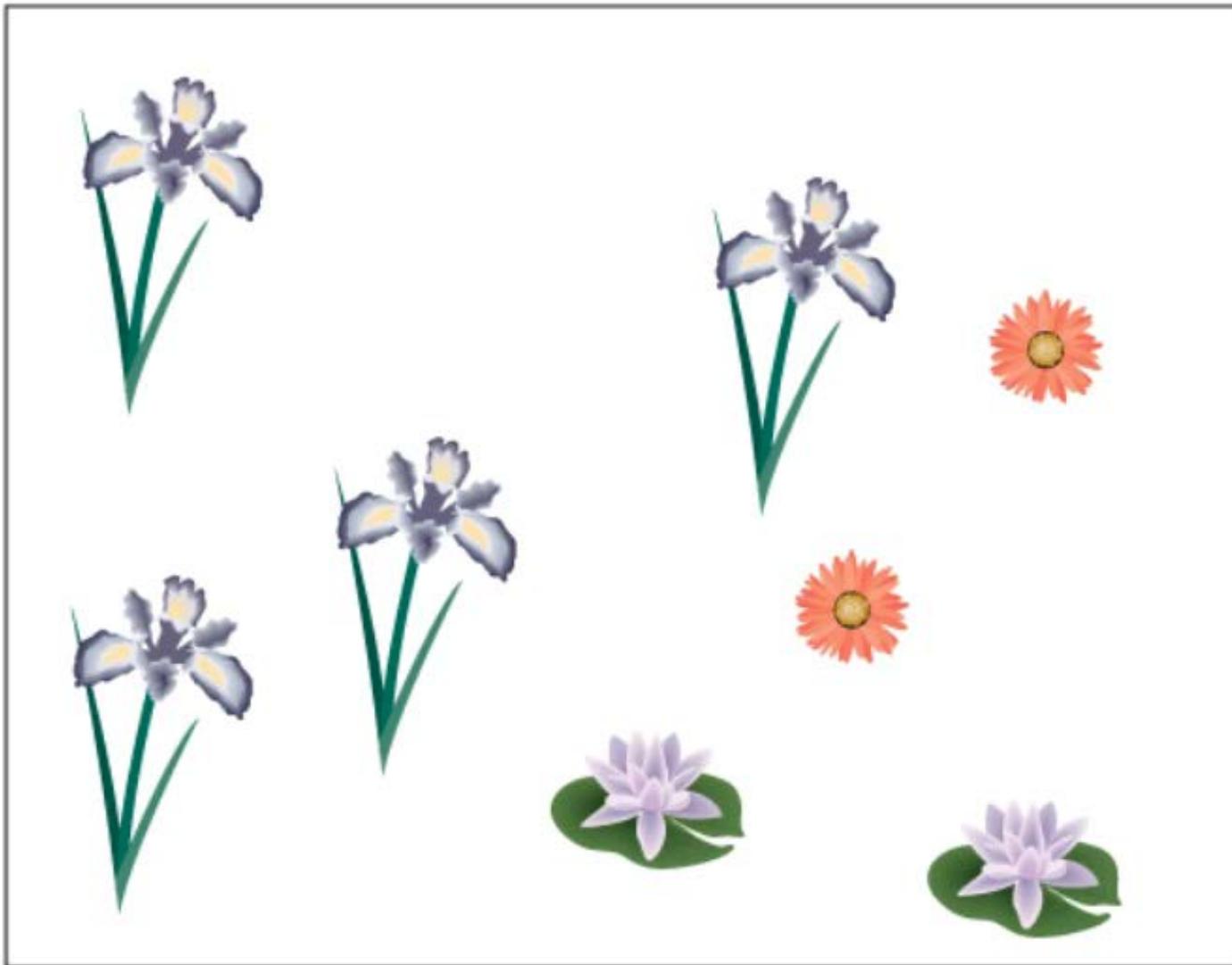
=



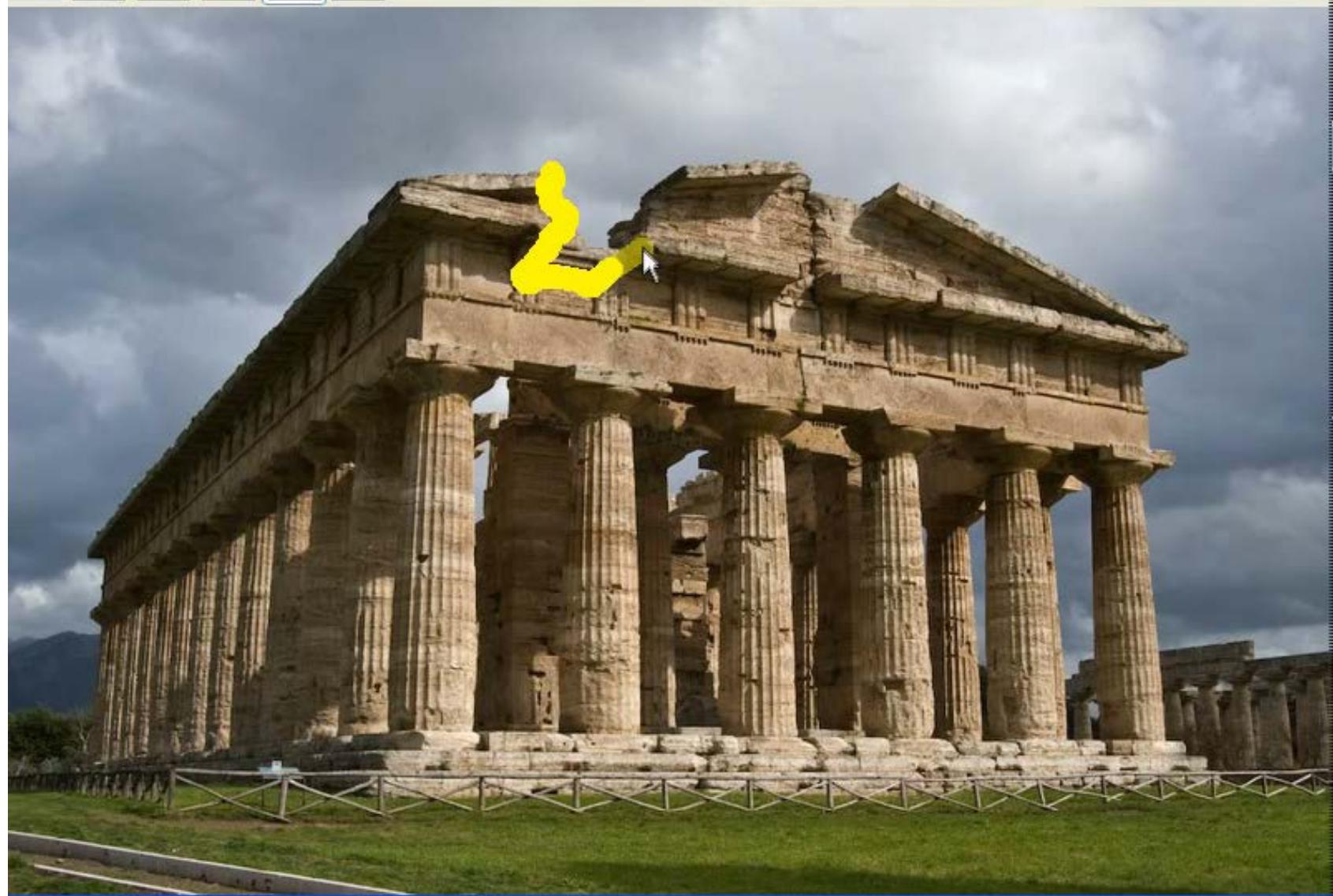




PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing
ACM Transactions on Graphics (Proc. SIGGRAPH), August 2009



File Edit Tools Options



PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing
ACM Transactions on Graphics (Proc. SIGGRAPH), August 2009

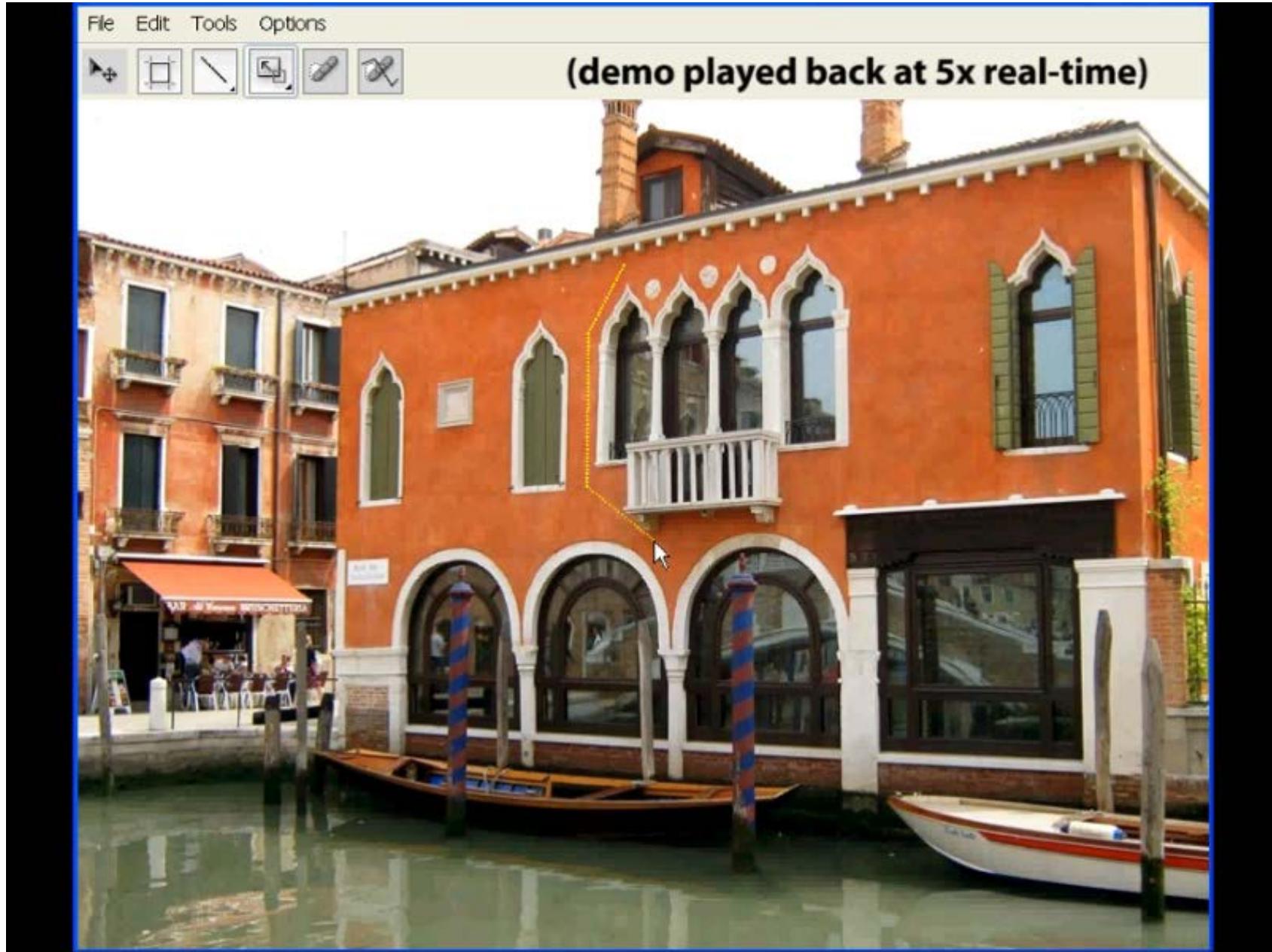


Image Analogies



?

Image Analogies



Image Analogies

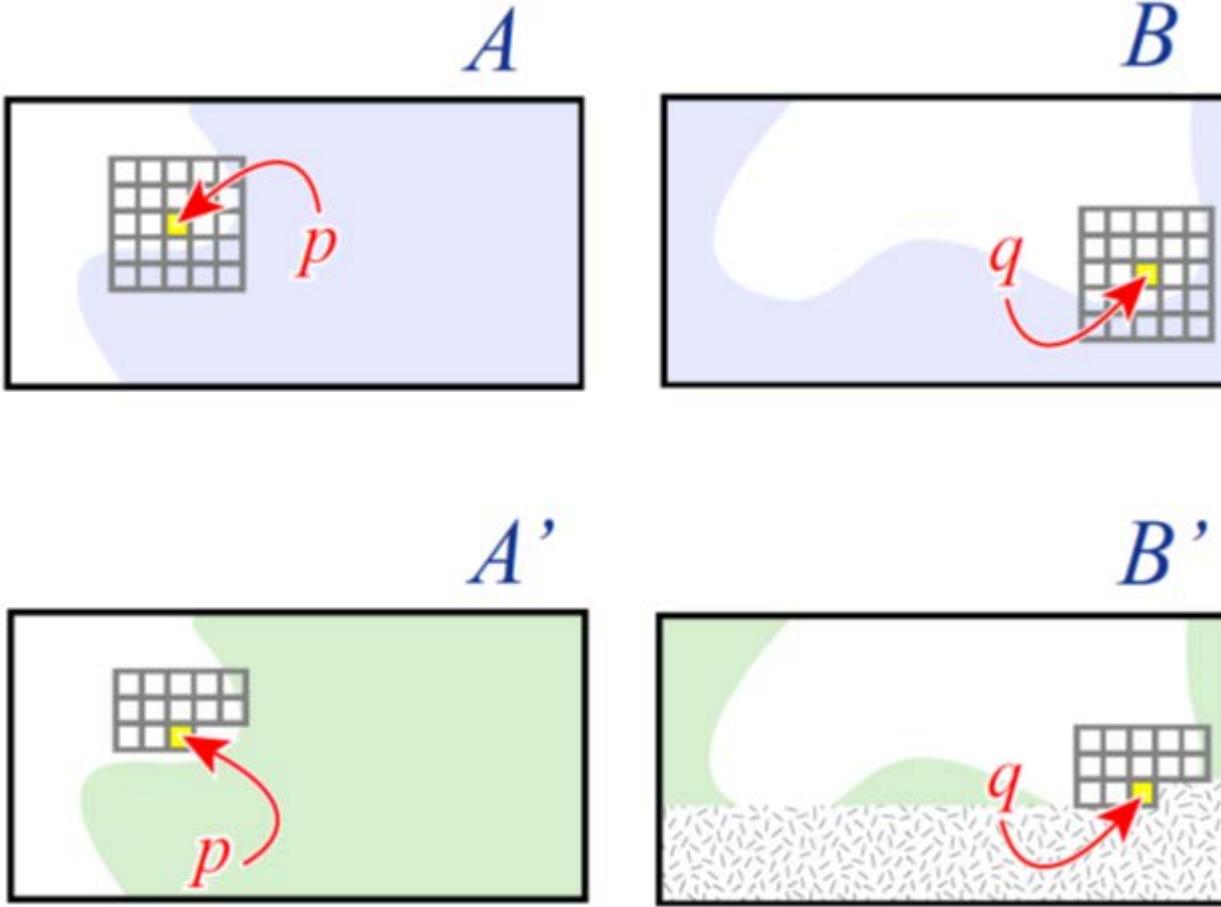


Image Analogies



?

Image Analogies

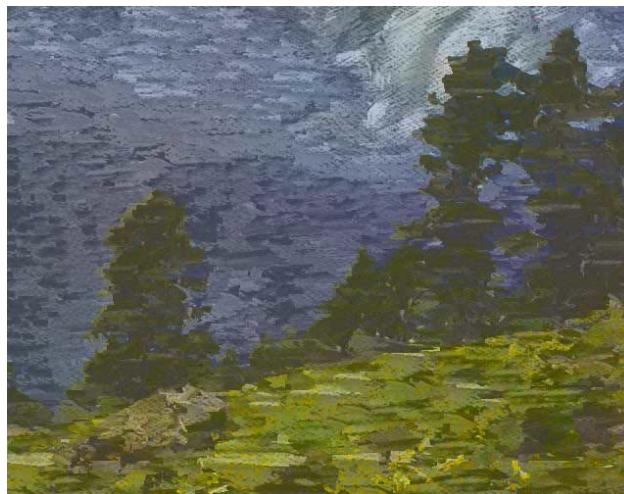
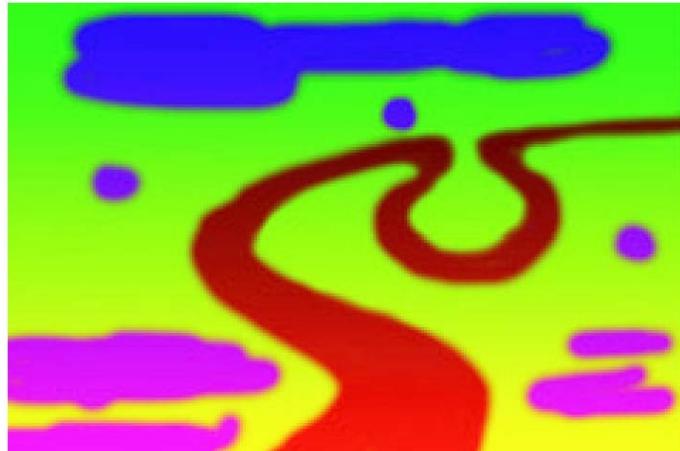
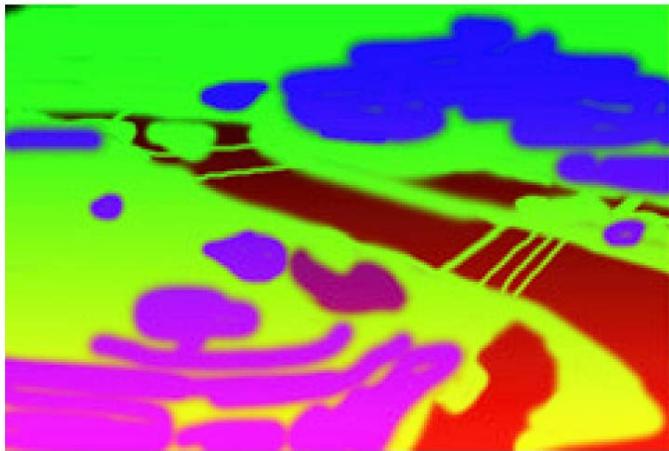
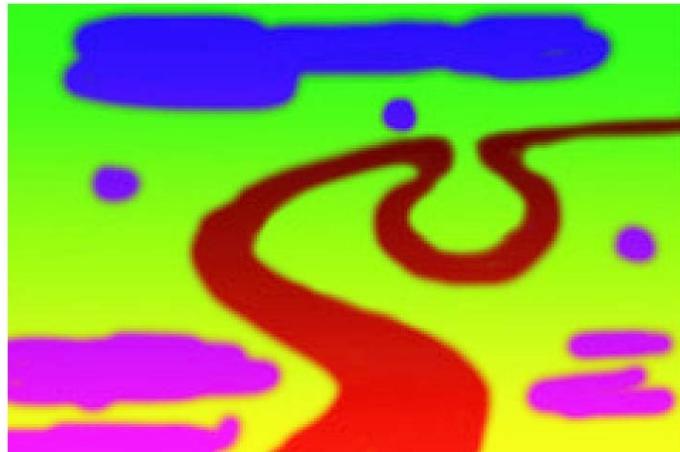
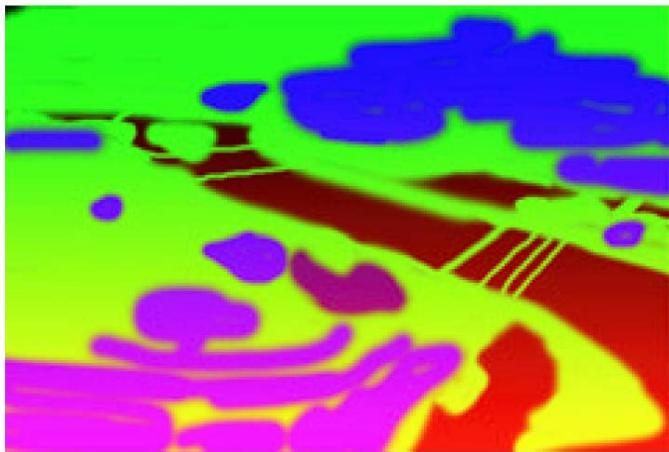


Image Analogies



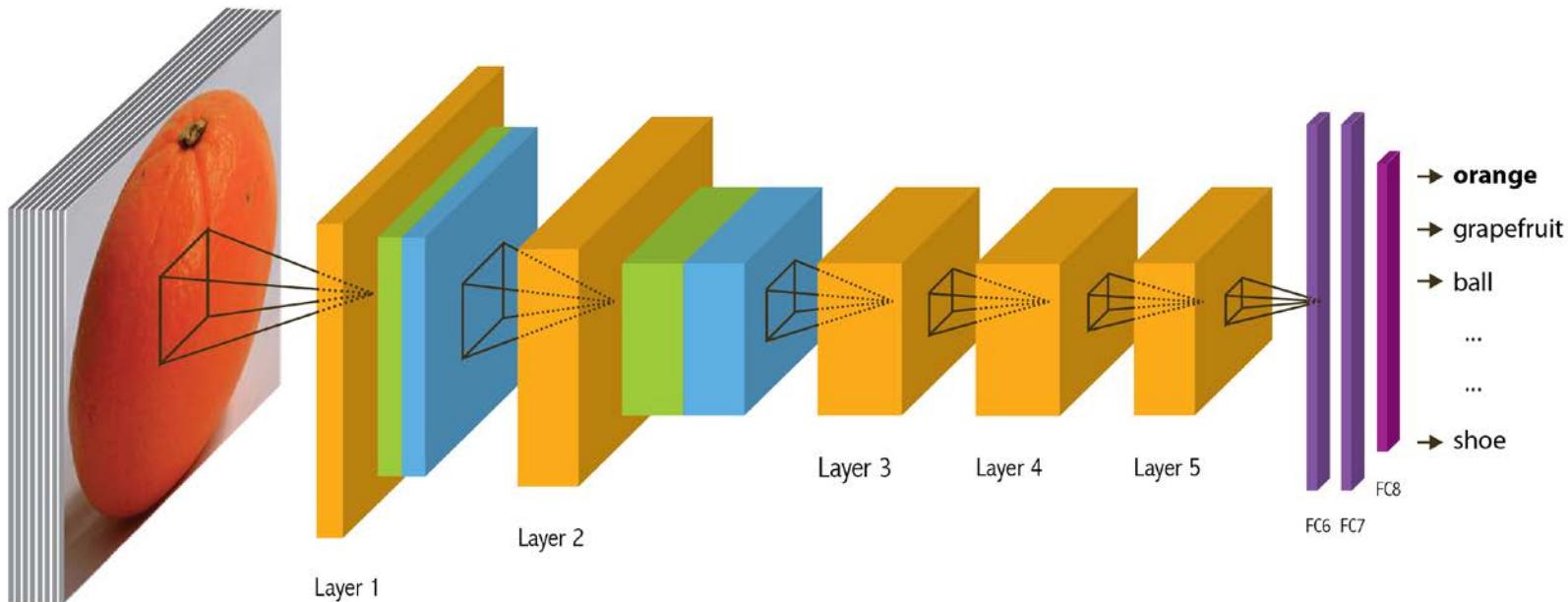
?

Image Analogies



ML in Vision & Graphics Today

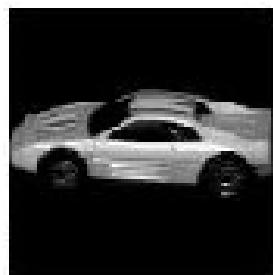
- Main methods: Deep Neural Networks
 - Geoffrey Hinton, Yann LeCun
 - Used both for analysis and synthesis



Input image

Advances in Dataset Collection

COIL-20



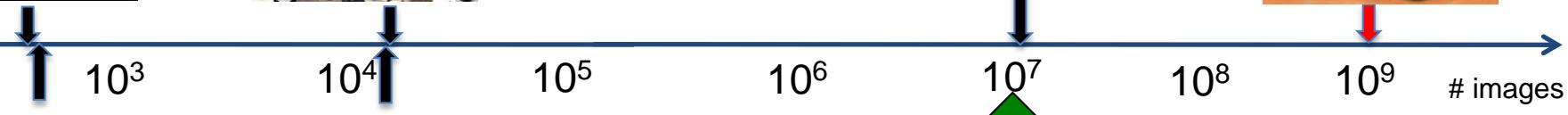
Caltech 101



Caltech-4 (2003)



PASCAL (2005)



IMAGENET
(2009)

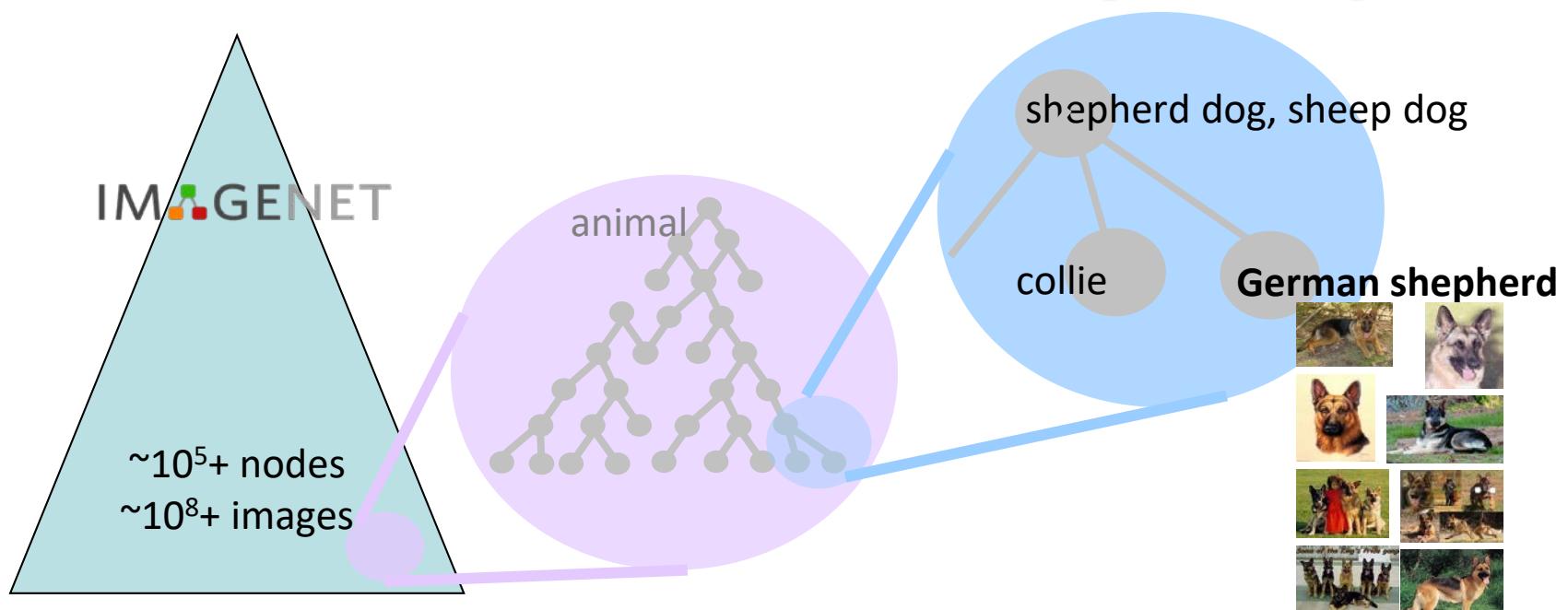
Places
(2013)

2 year
old kid



IMAGENET

- An *ontology of images* based on WordNet
- ImageNet currently has
 - 13,000+ categories of visual concepts
 - 10 million human-cleaned images ($\sim 700\text{im}/\text{categ}$)
 - 1/3+ is released online @ www.image-net.org



Application to ImageNet

.....

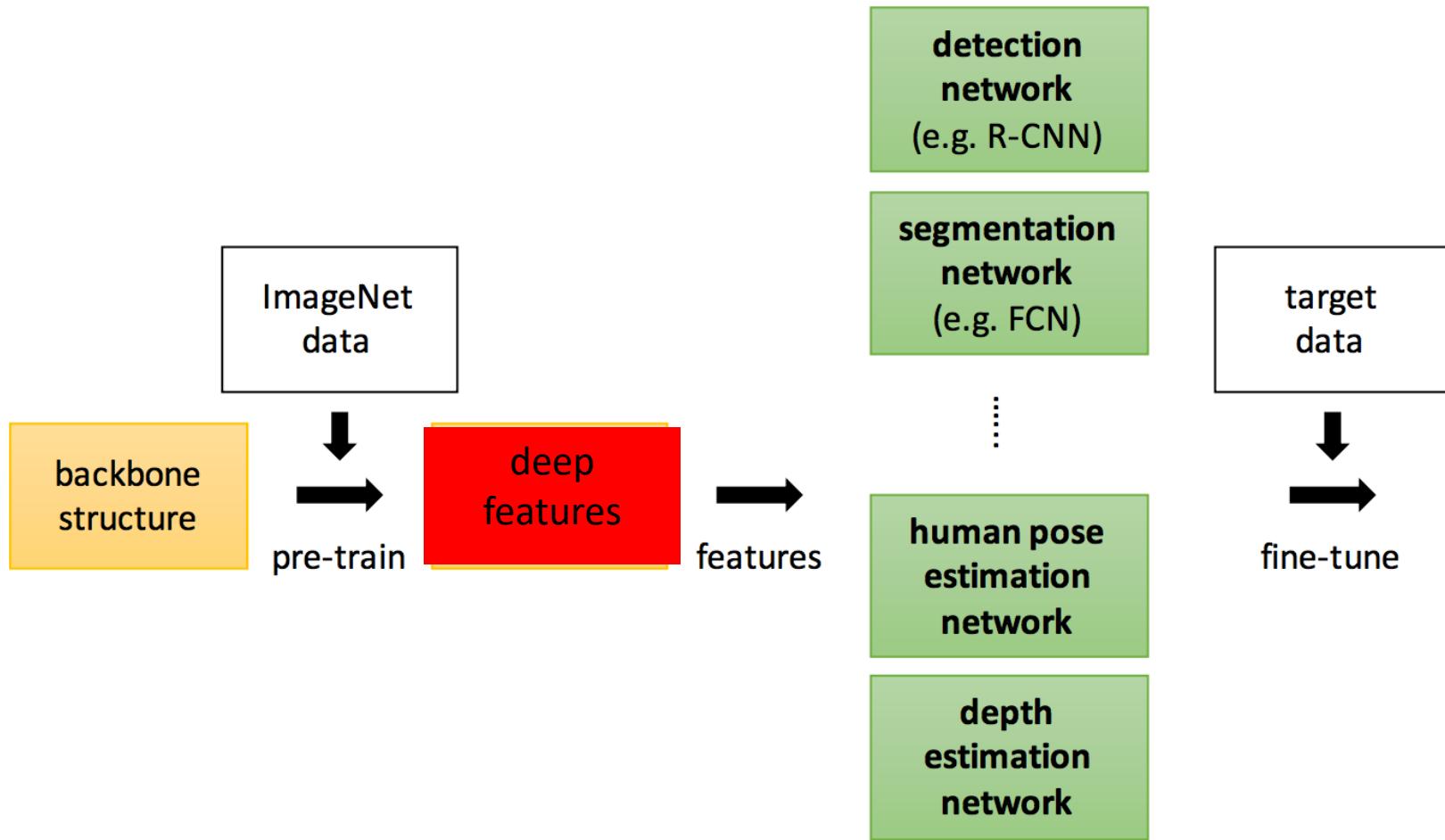


- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk

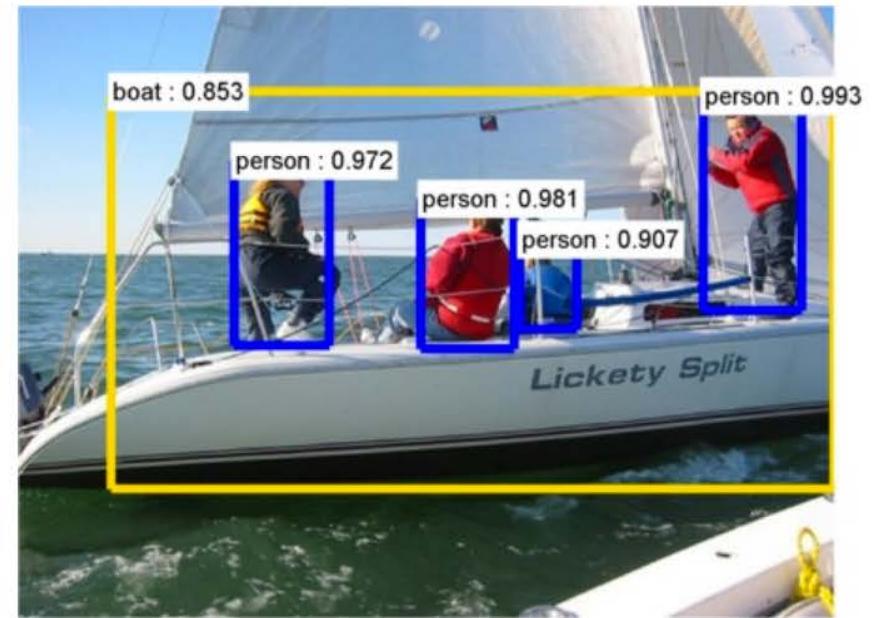
[Deng et al. CVPR 2009]

ImageNet Classification with Deep Convolutional Neural Networks [NIPS 2012]

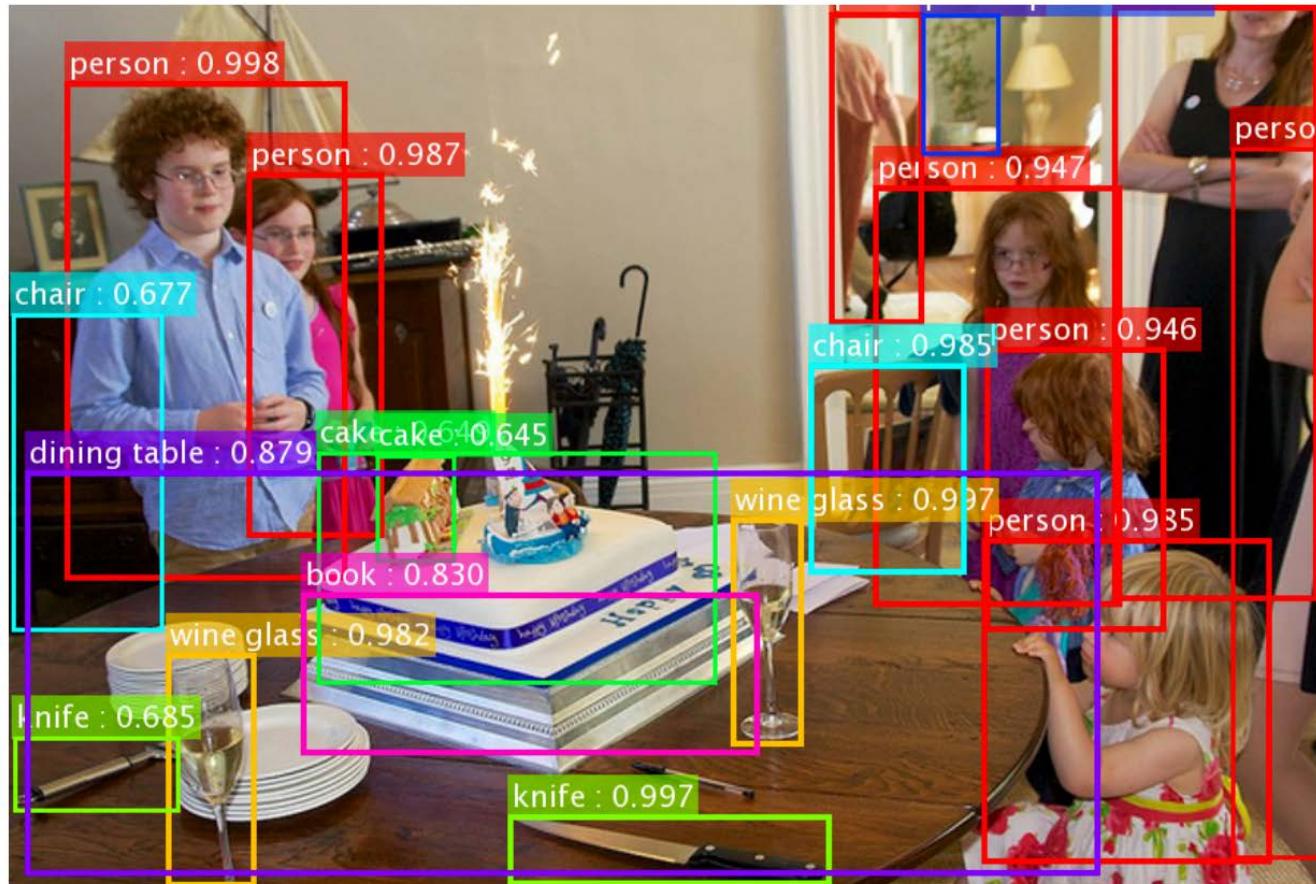
Solving Different Tasks



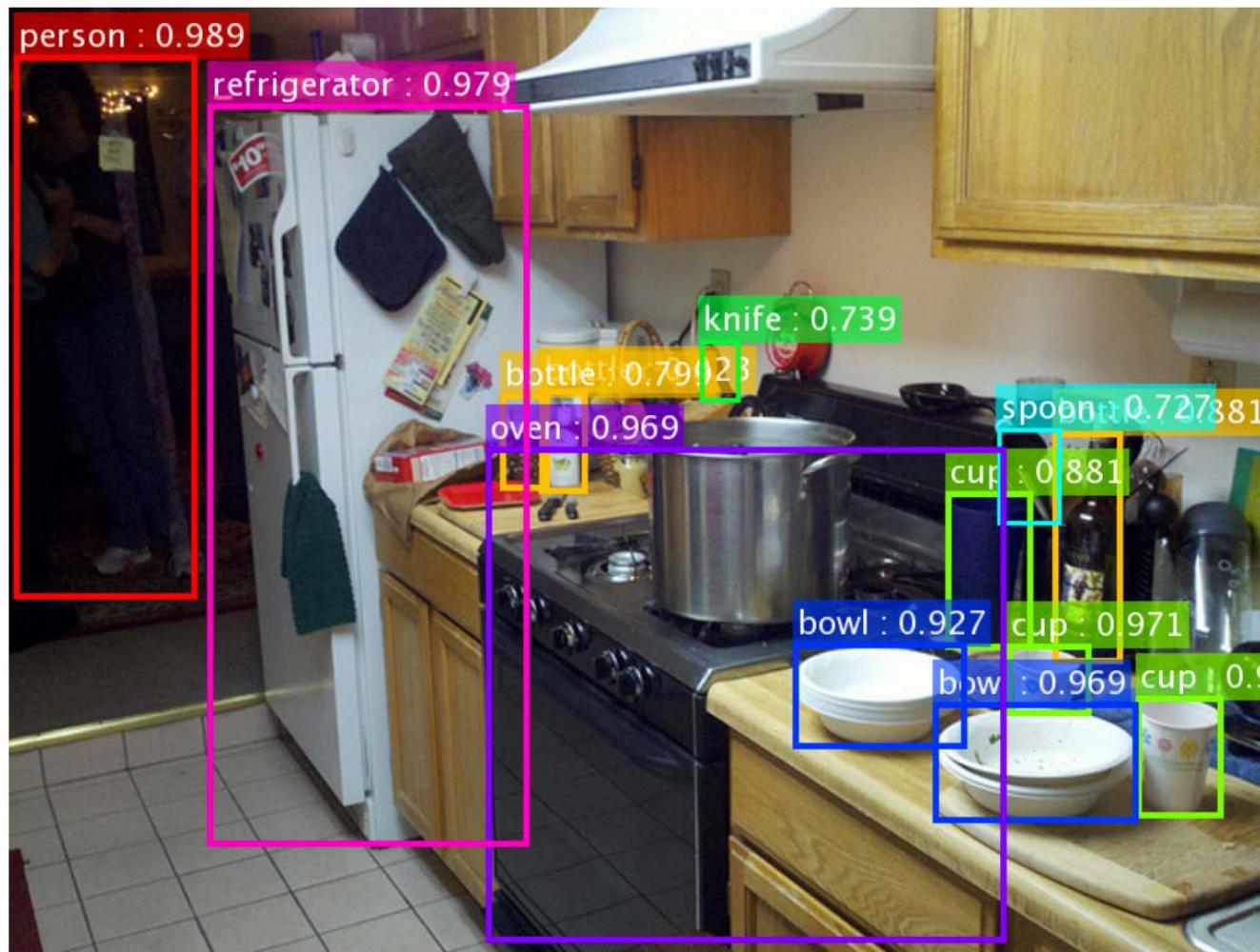
Object Detection: What and Where



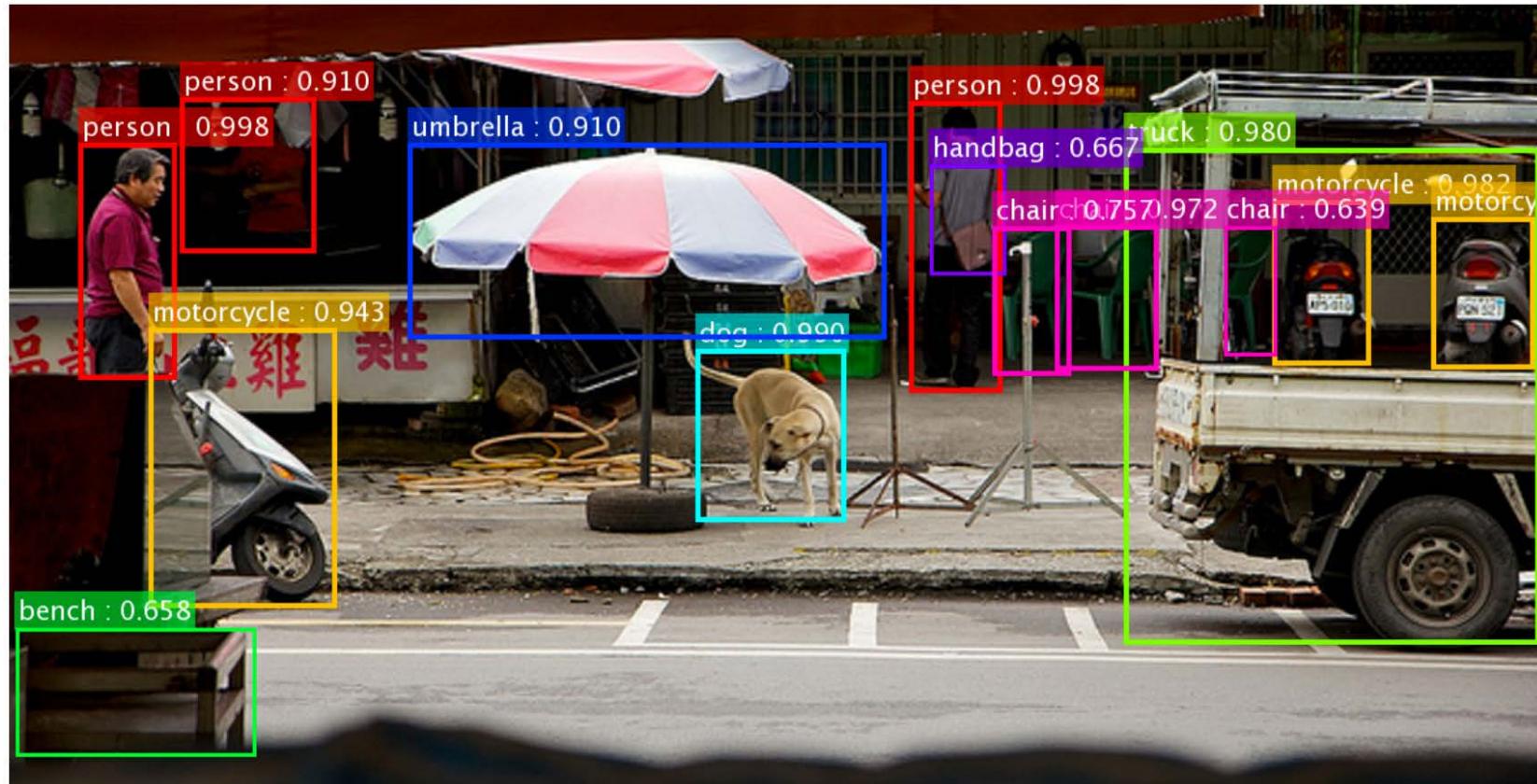
Object Detection Results



Object Detection Results



Object Detection Results



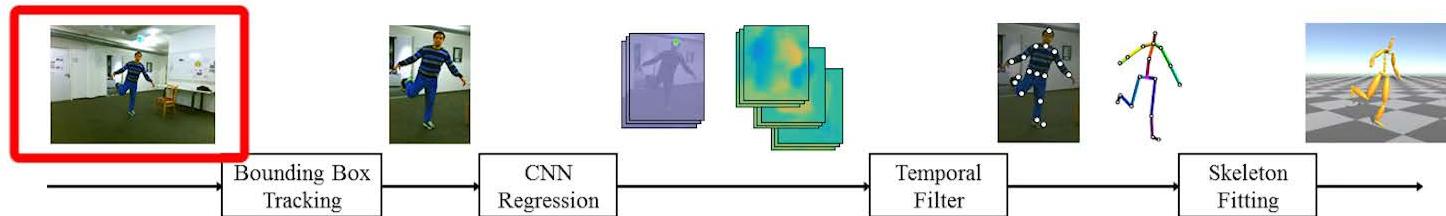
Pose Estimation (2D & 3D)



VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera – SIGGRAPH 2017

<https://www.youtube.com/watch?v=W1ZNFFftx2E>

Pose Estimation (2D & 3D)



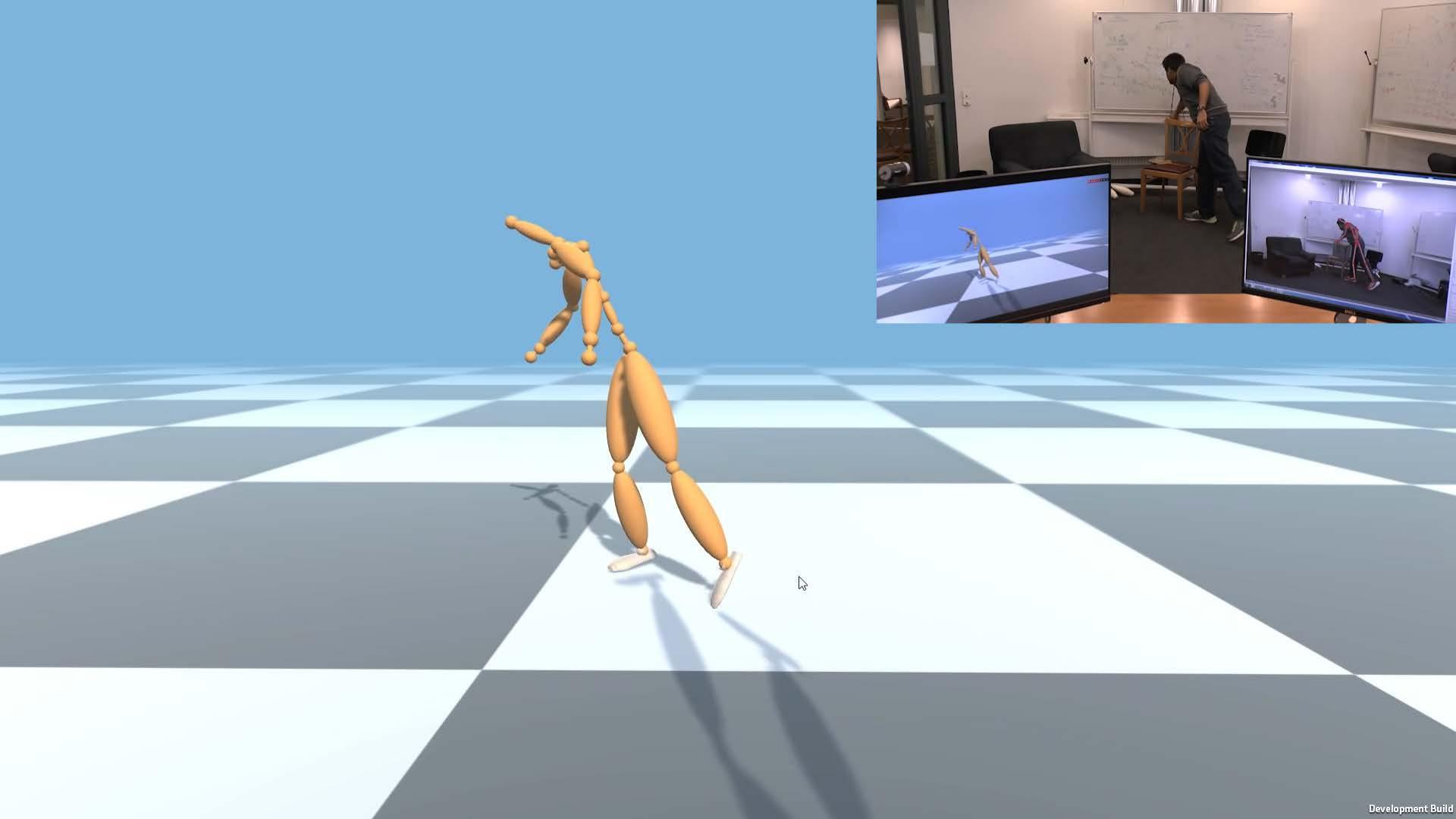
Full-frame Input from Single RGB Camera



Vnect: Real-time 3D Human Pose Estimation with a Single RGB Camera – SIGGRAPH 2017

<https://www.youtube.com/watch?v=W1ZNffftX2E>

Pose Estimation (2D & 3D)

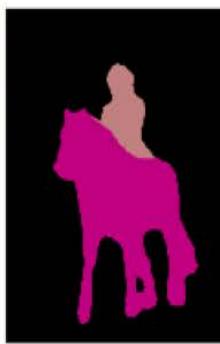


Development Build

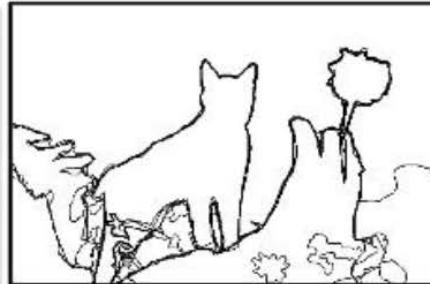
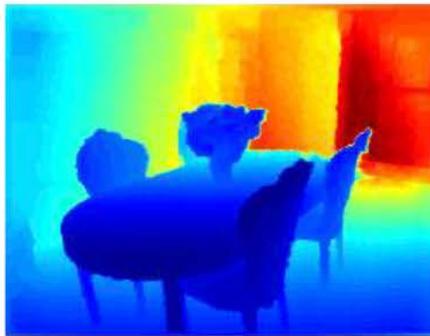
VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera – SIGGRAPH 2017
<https://www.youtube.com/watch?v=W1ZNffftX2E>

Segmentation, Depth, & More

semantic segmentation

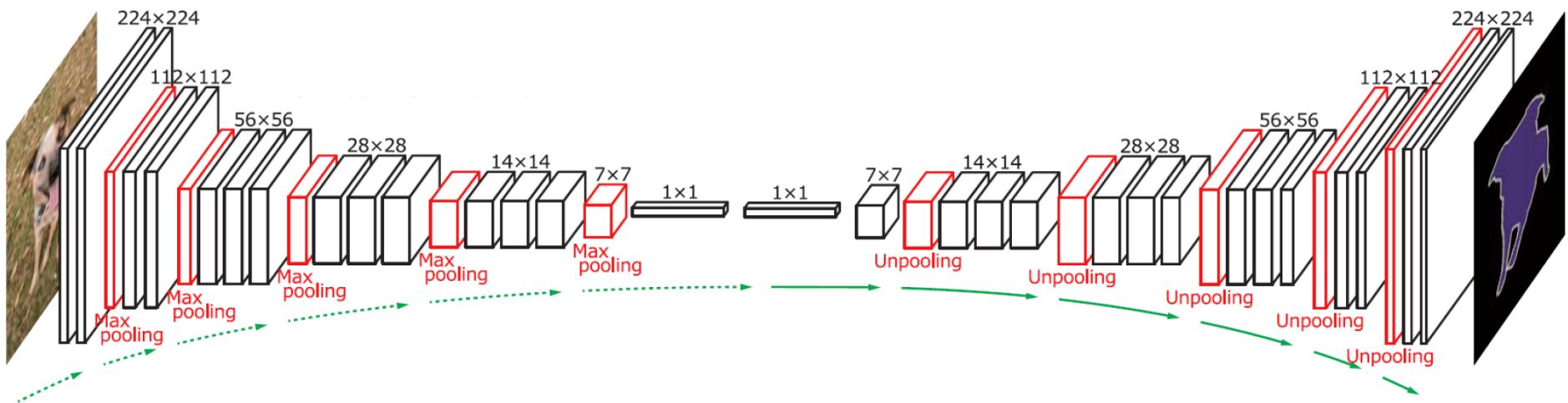


monocular depth estimation (Liu et al. 2015)

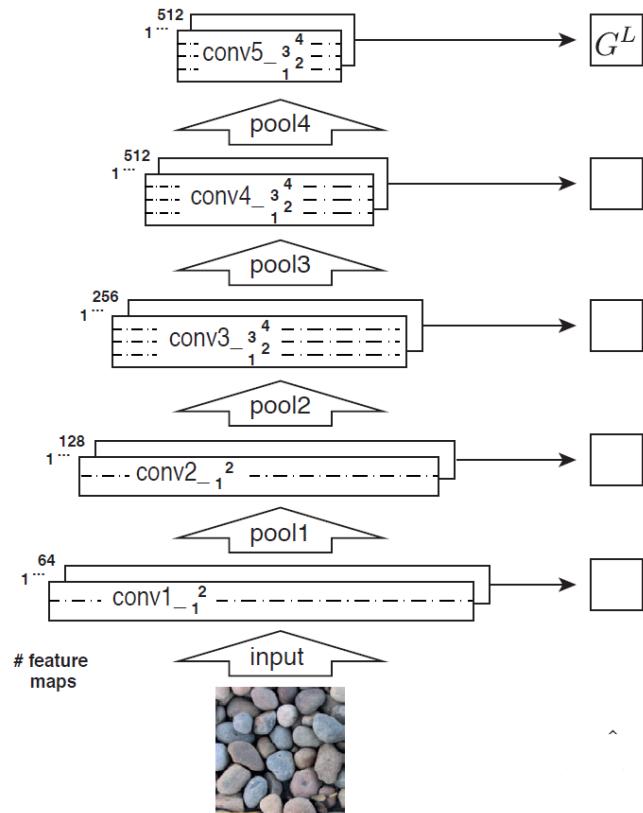


boundary prediction (Xie & Tu 2015)

Typical Architecture



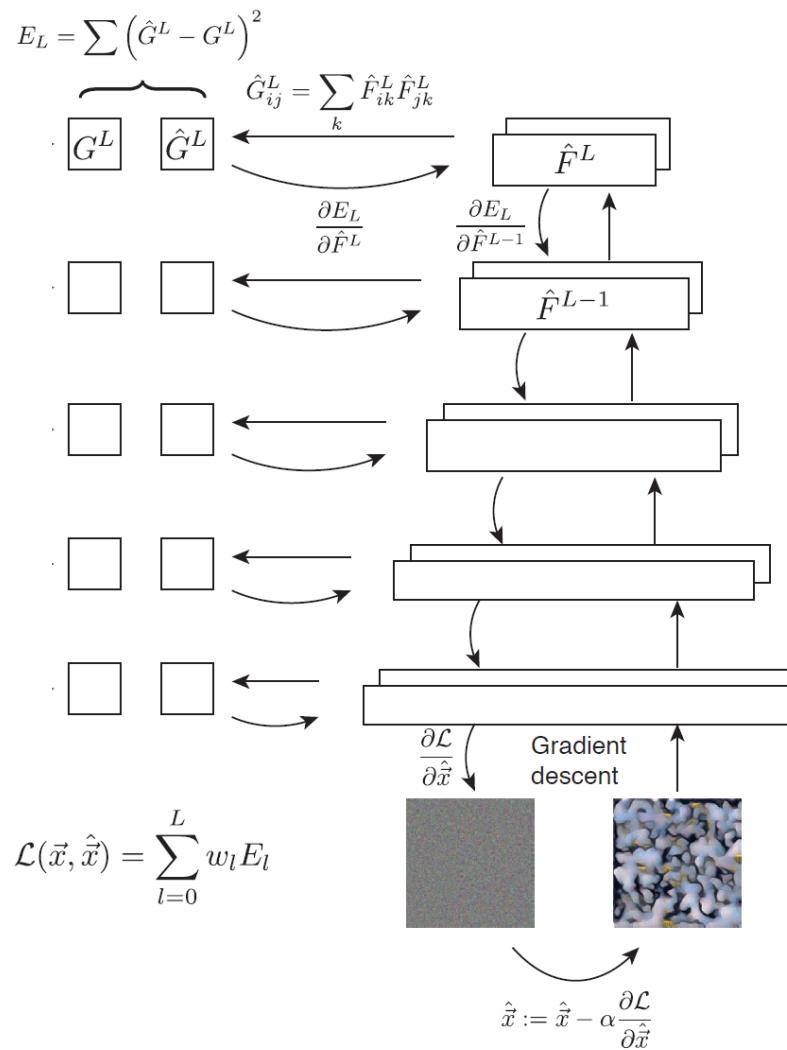
Texture Synthesis using CNNs



Analysis

Texture Synthesis using CNNs

Synthesis



Texture Synthesis using CNNs

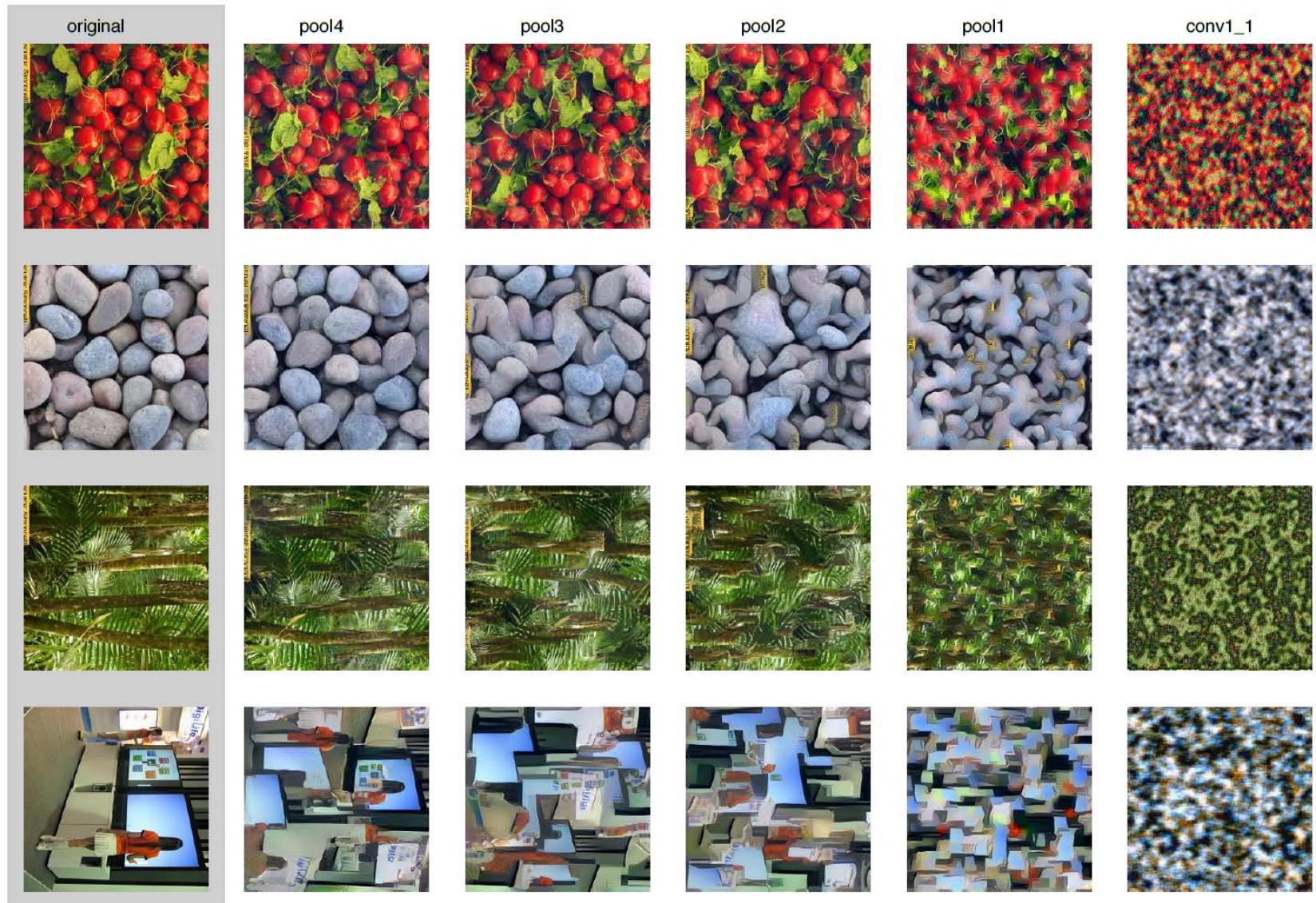


Image Analogies/Style Transfer

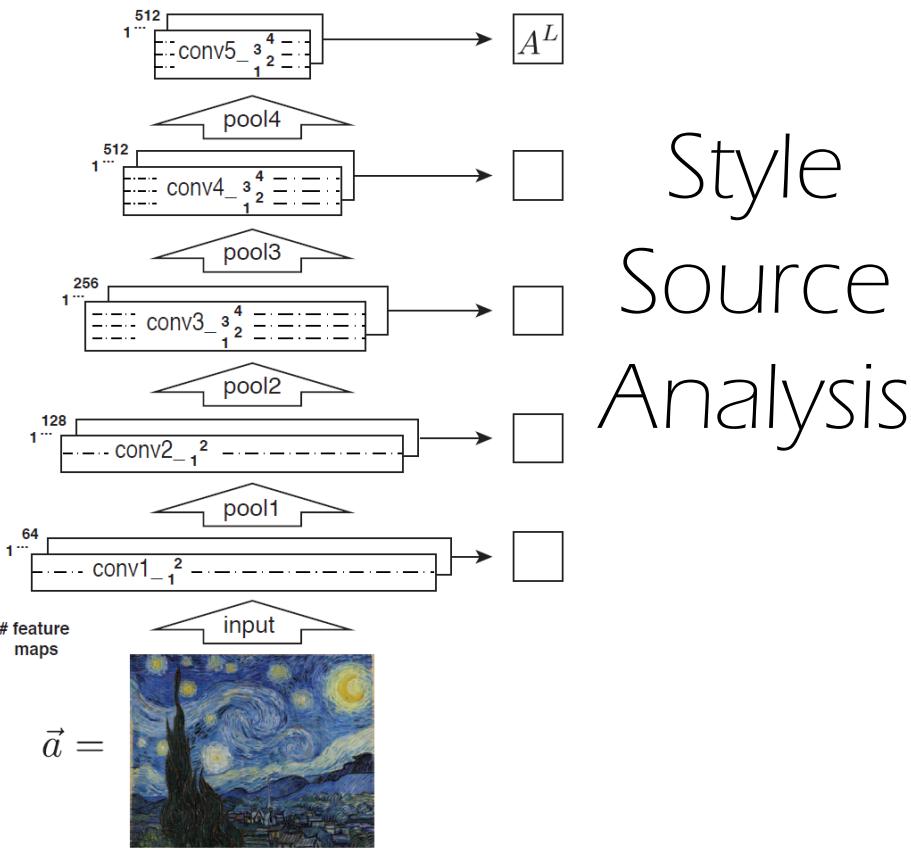
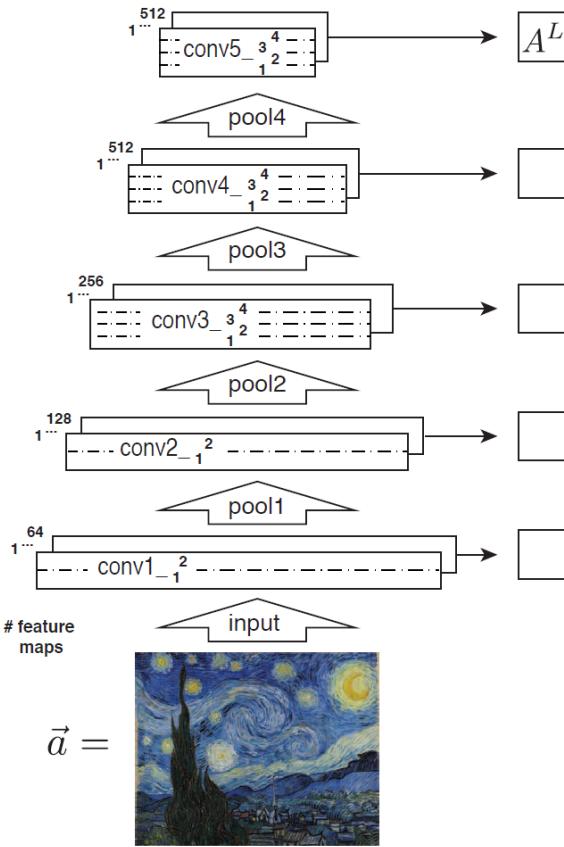


Image Analogies/Style Transfer



Content
Source
Analysis

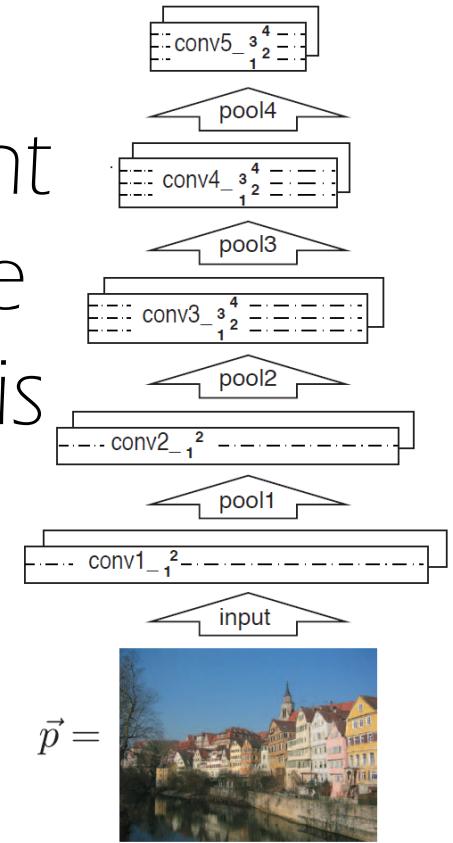


Image Analogies/Style Transfer

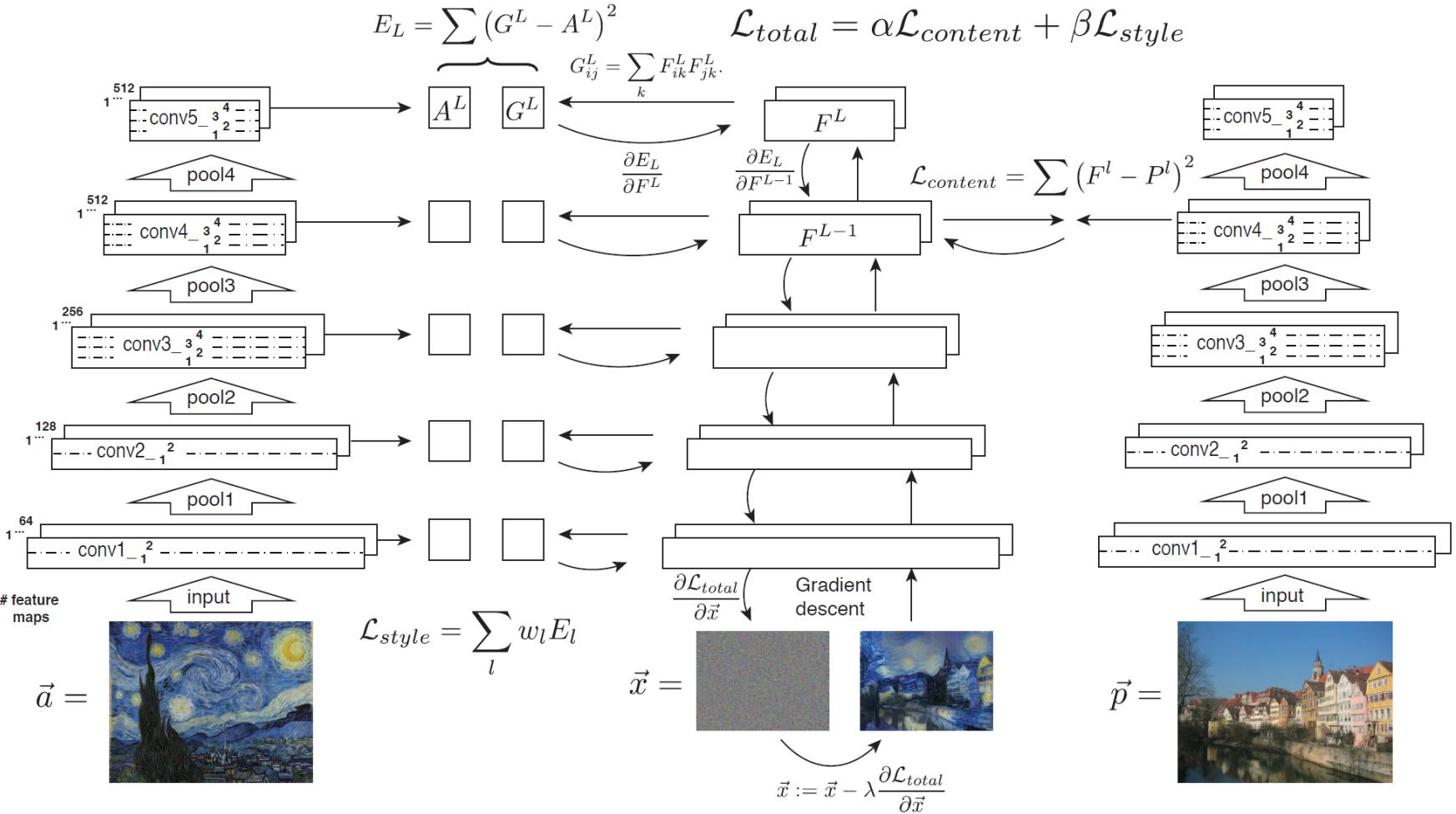


Image Analogies/Style Transfer

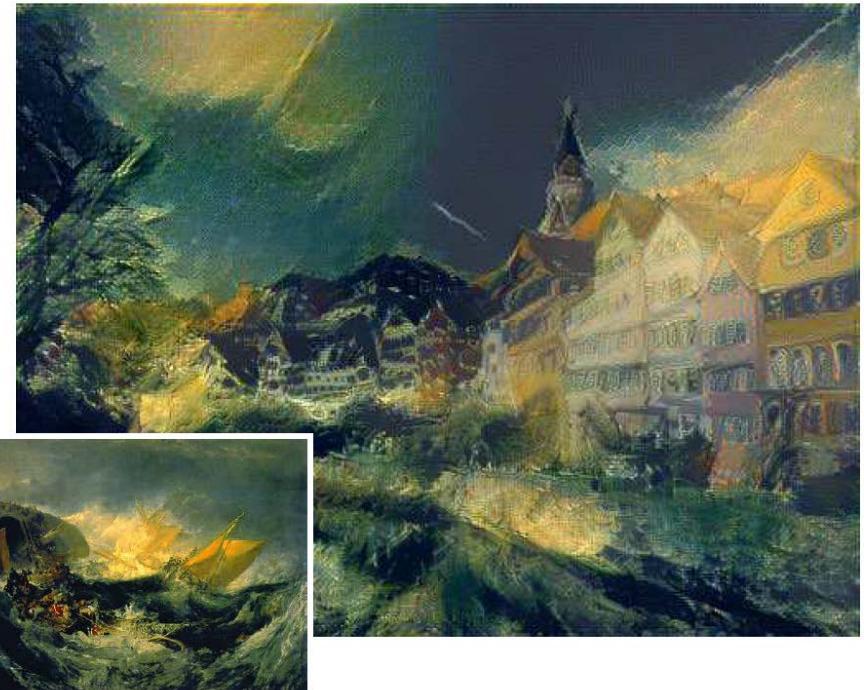


Image Analogies/Style Transfer



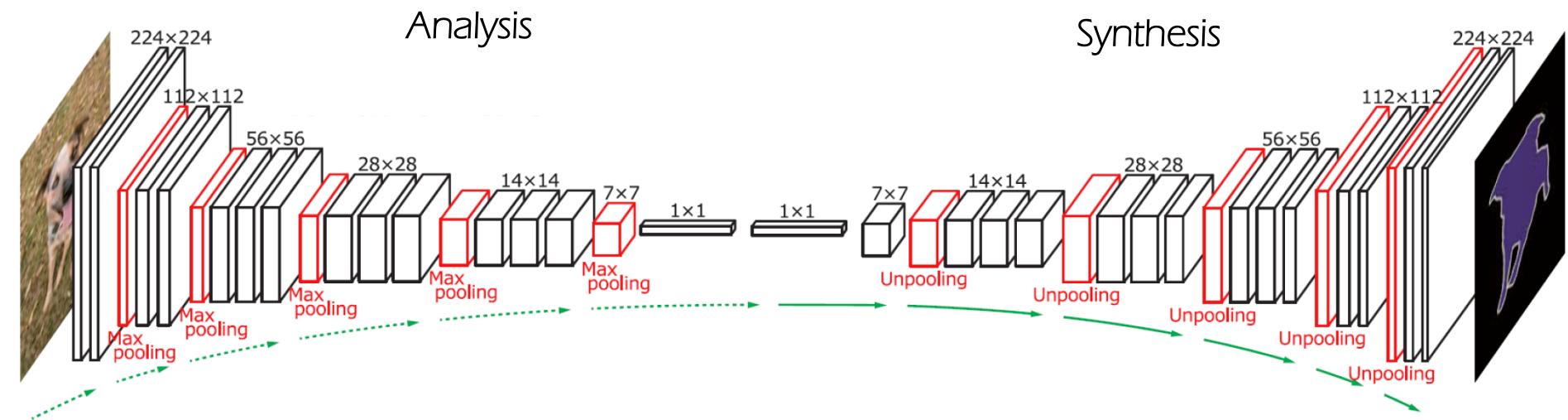
Image Analogies/Style Transfer



Image Analogies/Style Transfer

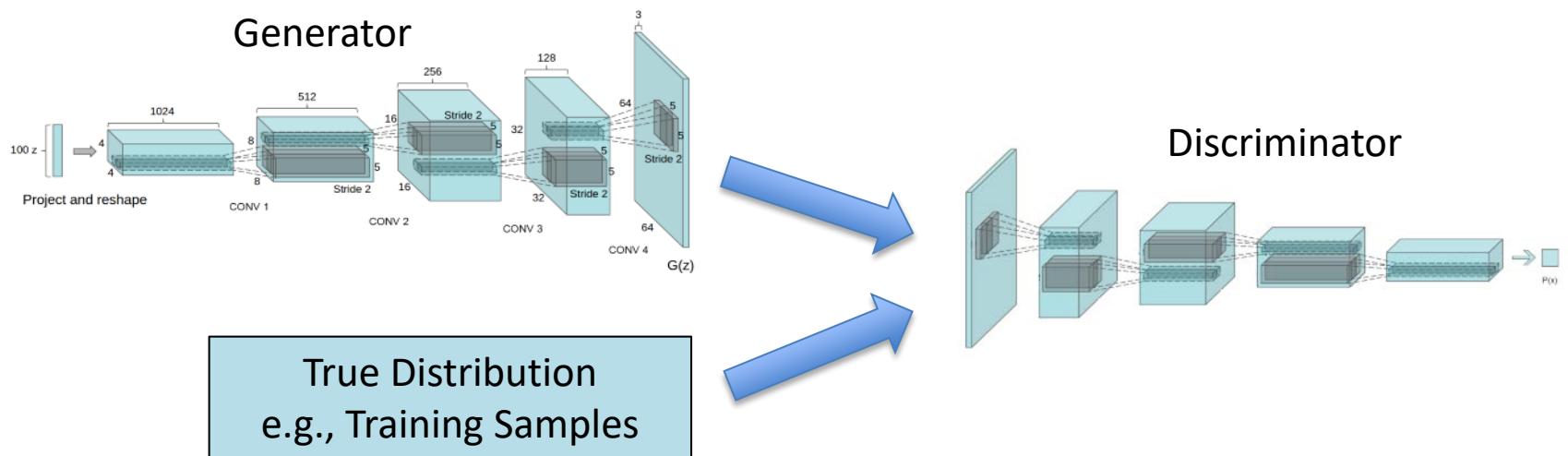


Recall: Typical Architecture



Generative Adversarial Networks (GANs)

- Two neural networks competing against each other
 - Generator is taught to map to a desired distribution
 - Discriminator is taught to discriminate between data from true distribution and generator



GANs for Image Synthesis

- Automatically generated bedrooms



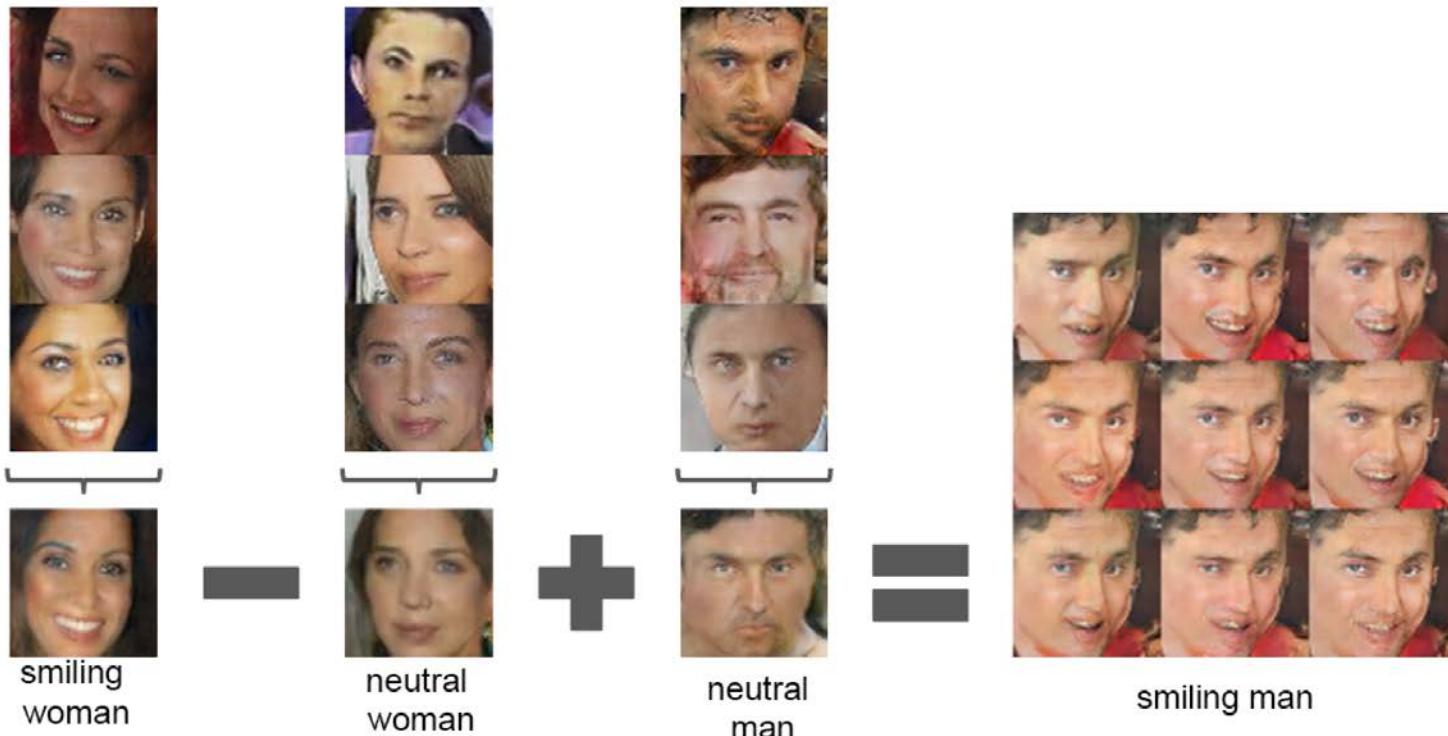
GANs for Image Synthesis

- Interpolation



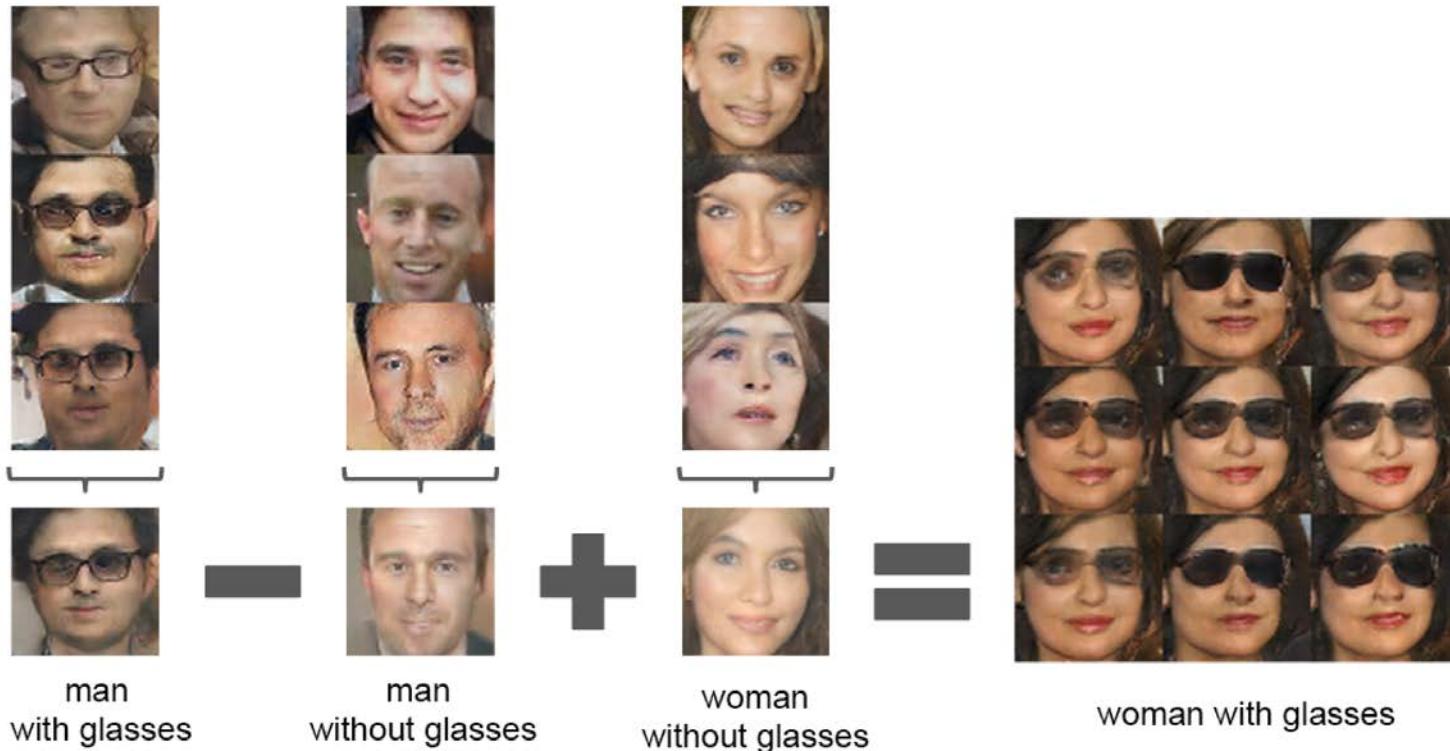
GANs for Image Synthesis

- Image Arithmetic



GANs for Image Synthesis

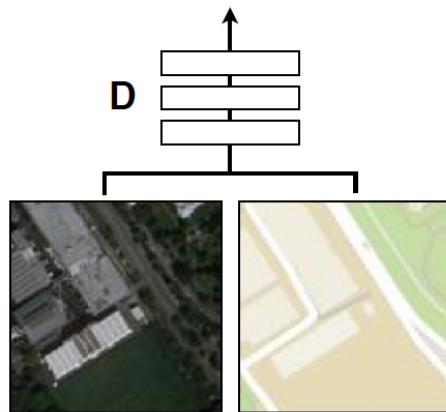
- Image Arithmetic



Conditional GANs/Image Translation

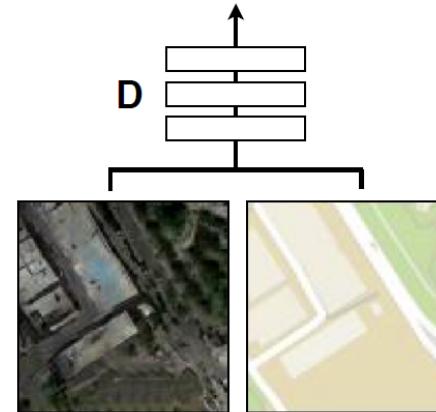
Positive examples

Real or fake pair?



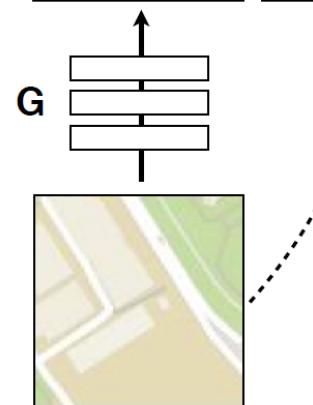
Negative examples

Real or fake pair?



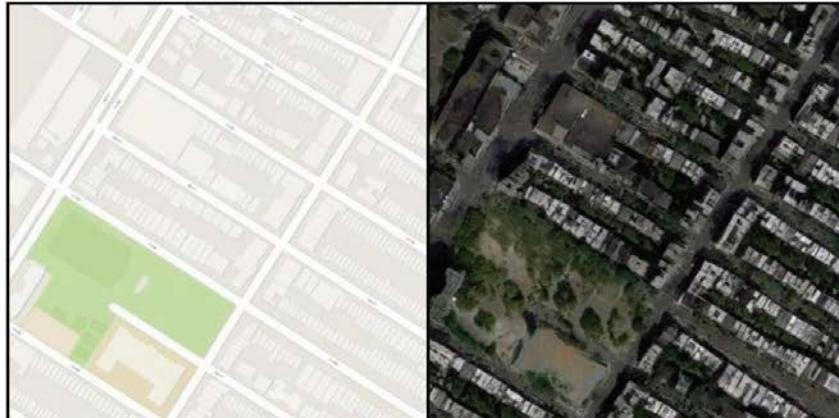
G tries to synthesize fake images that fool **D**

D tries to identify the fakes

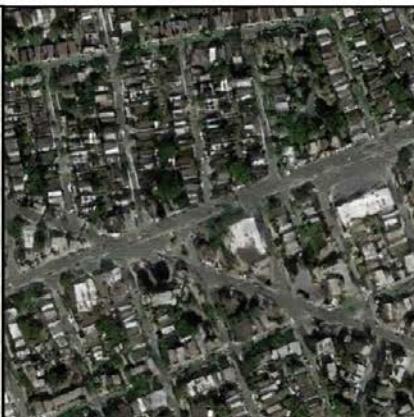


Conditional GANs/Image Translation

Aerial photo to map



Map to aerial photo



input

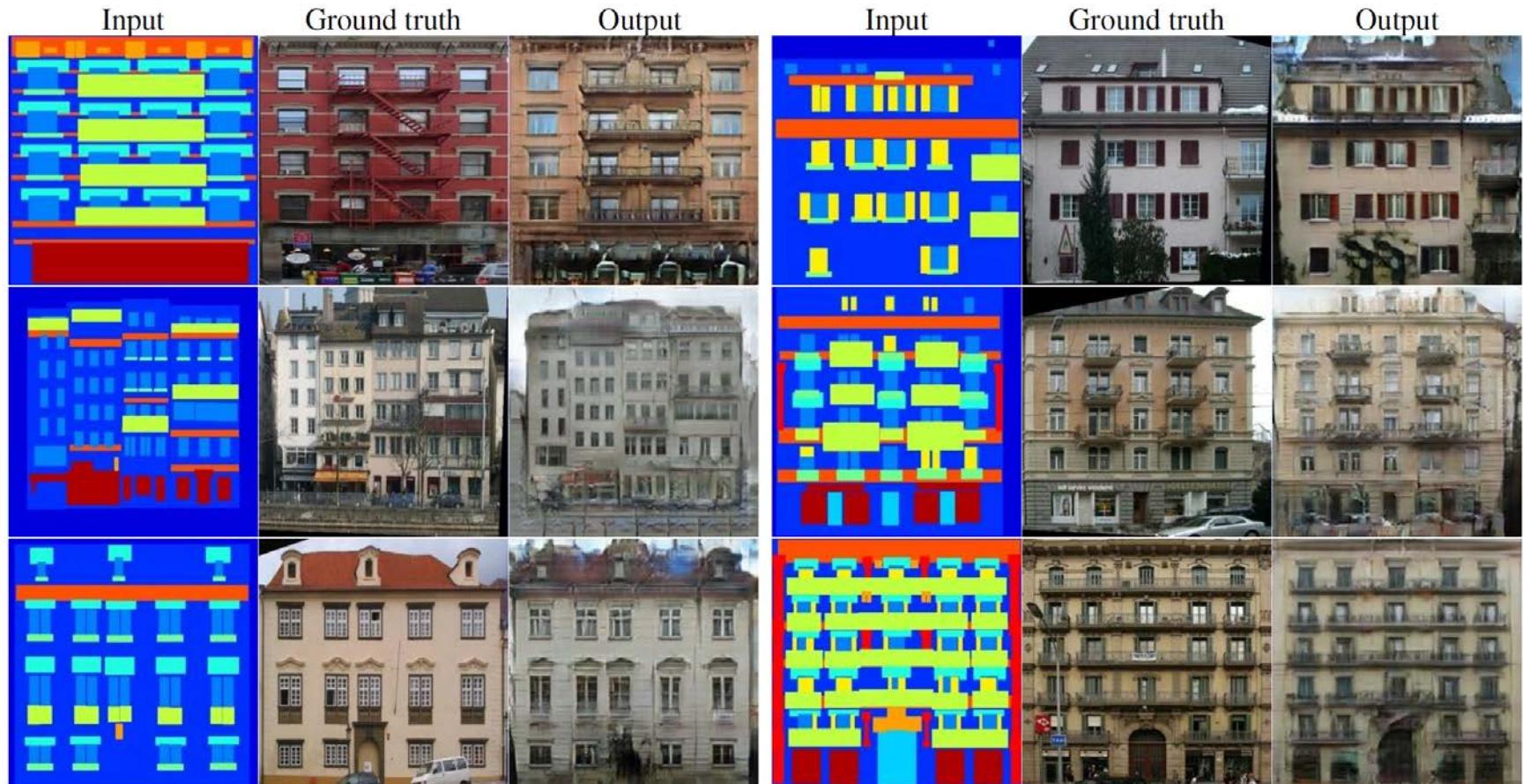
output



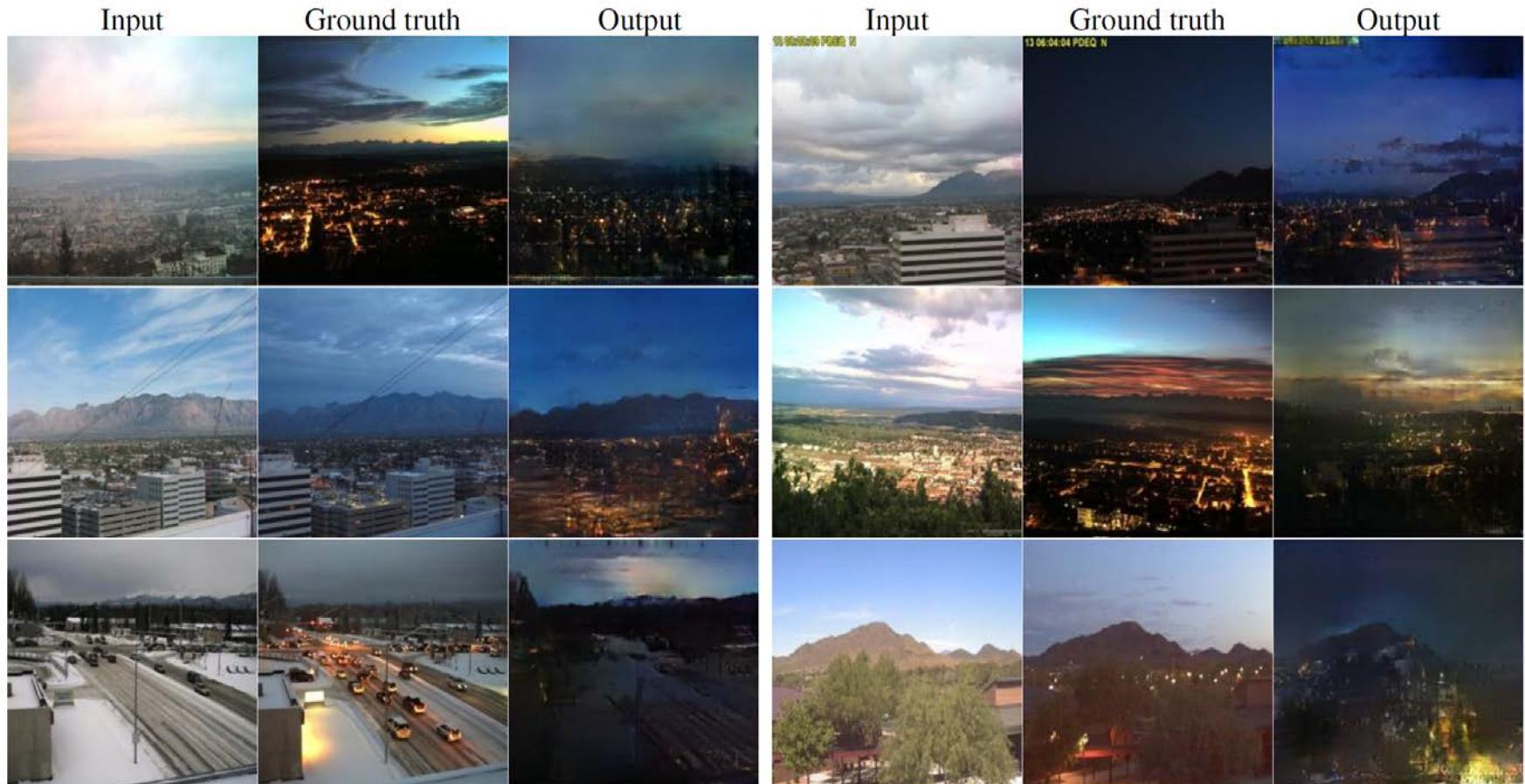
input

output

Conditional GANs/Image Translation



Conditional GANs/Image Translation



Conditional GANs/Image Translation



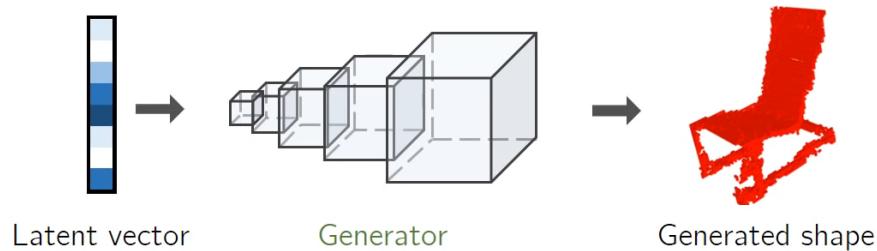
Conditional GANs/Image Translation



More than Images

- Shape Synthesis

3D Generative Adversarial Network



More than Images

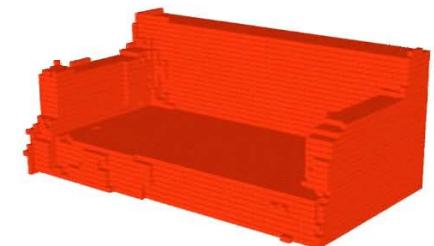
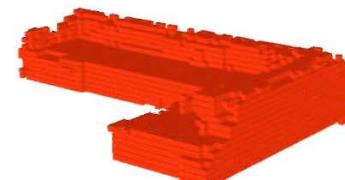
- Shape Synthesis

Randomly Sampled Shapes

Chairs



Sofas

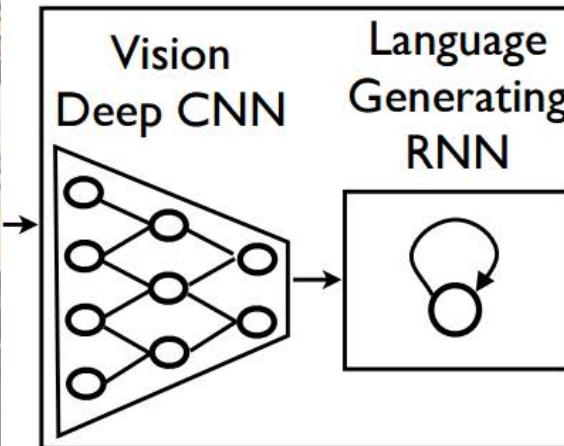


Results from 3D ShapeNet



More than Images

- From Image to Captions



A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Image Caption Generator Results

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.



A yellow school bus parked in a parking lot.



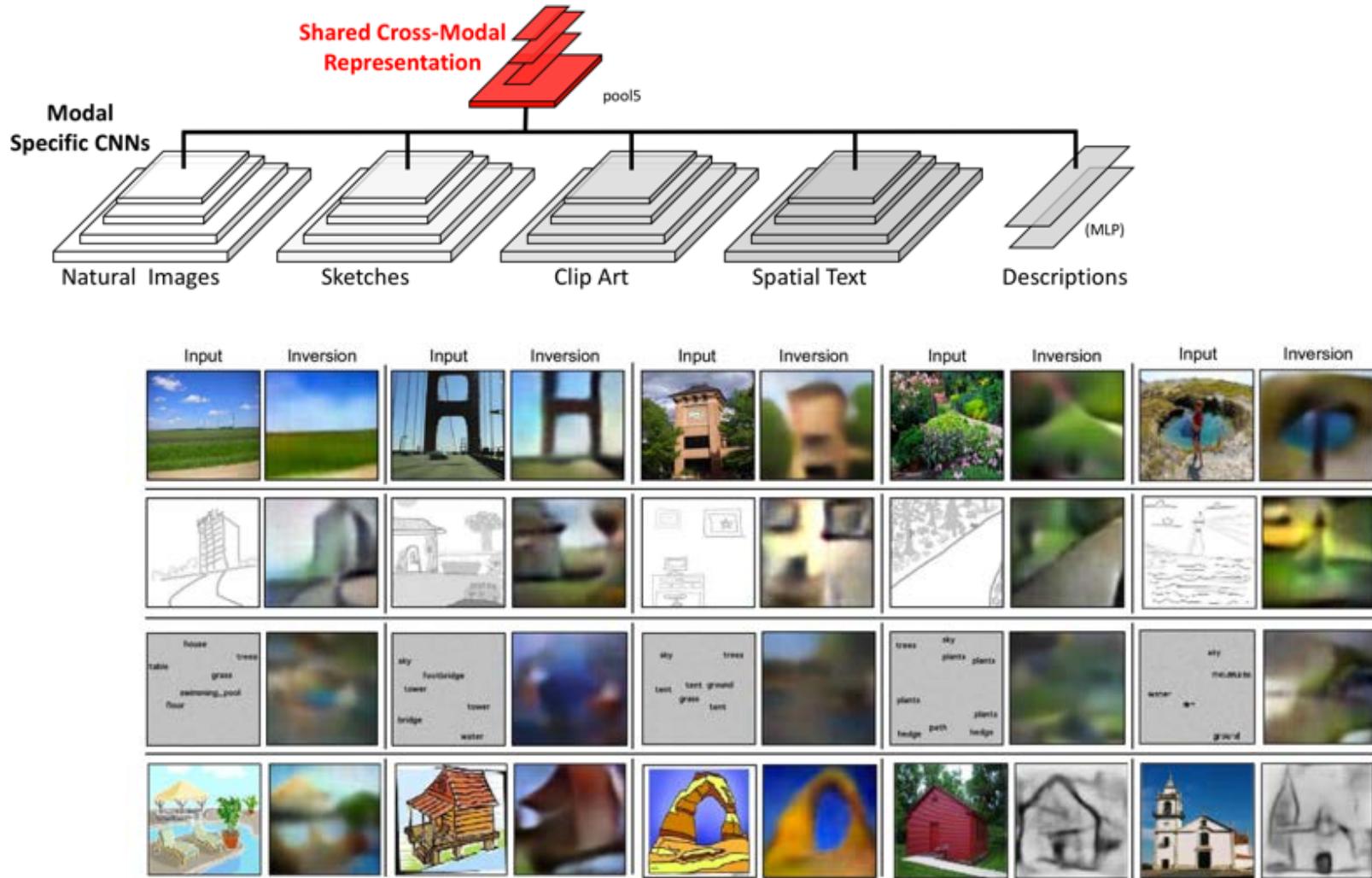
Describes without errors

Describes with minor errors

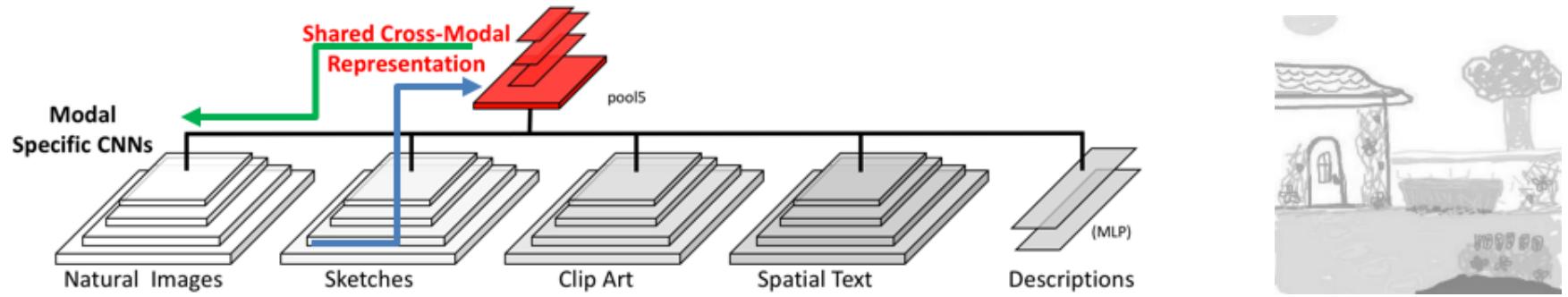
Somewhat related to the image

Unrelated to the image

Cross Modal Networks

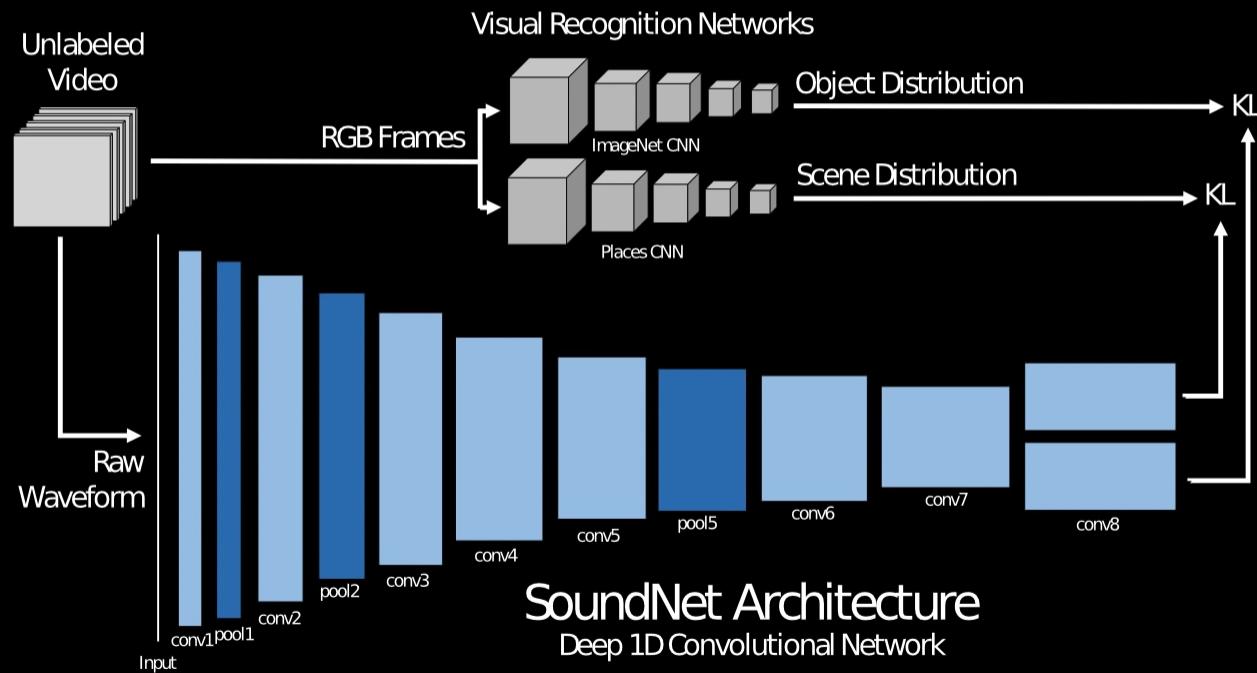


Cross-Modal Inversion



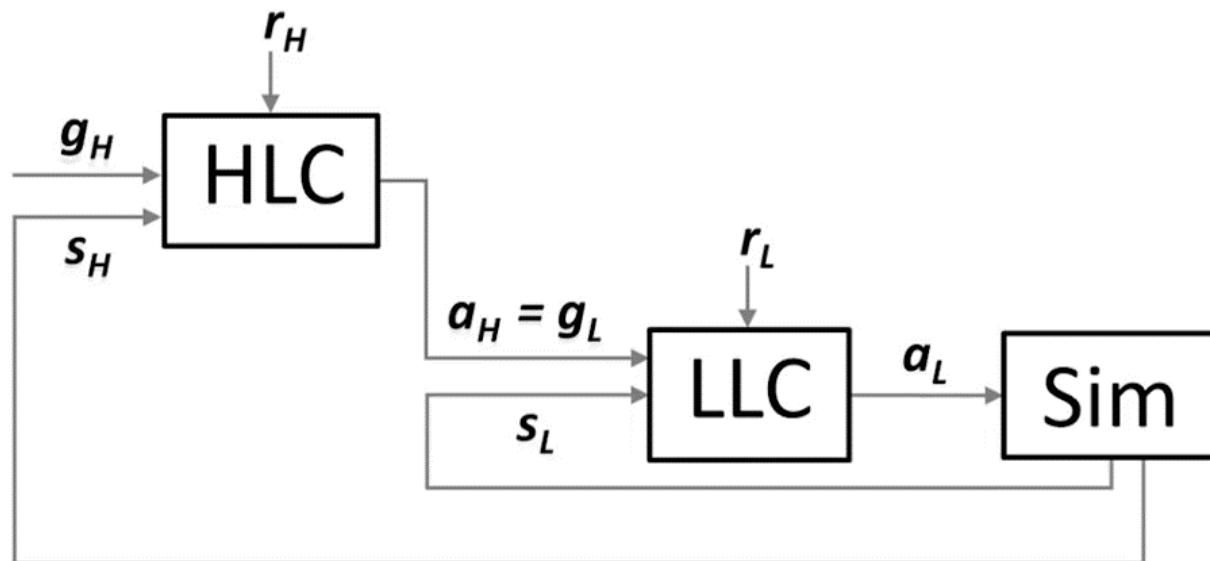
Cross-Modal Transfer: SoundNet

We develop a deep convolutional network that learns to recognize objects from sound



More than Images

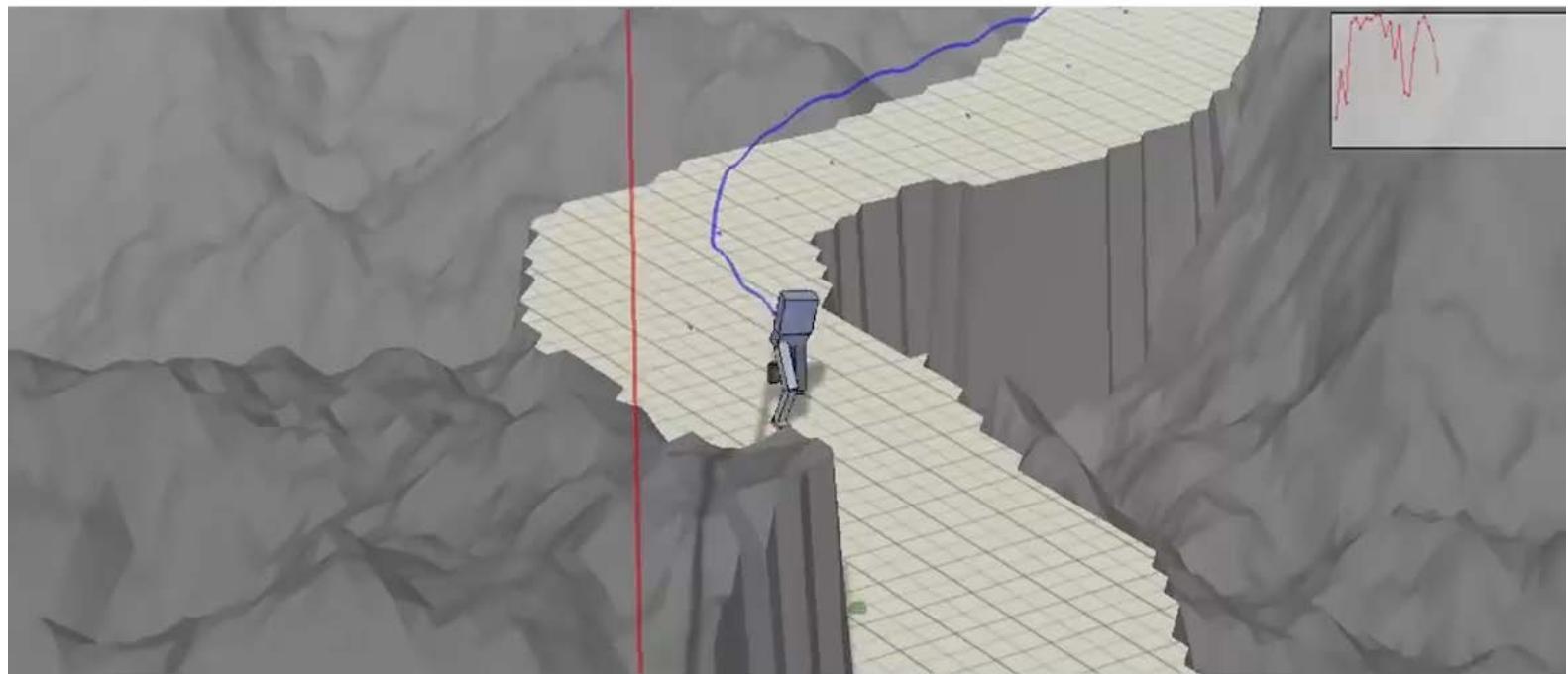
- Character Control: Reinforcement Learning
Overview



More than Images

- Character Control: Reinforcement Learning

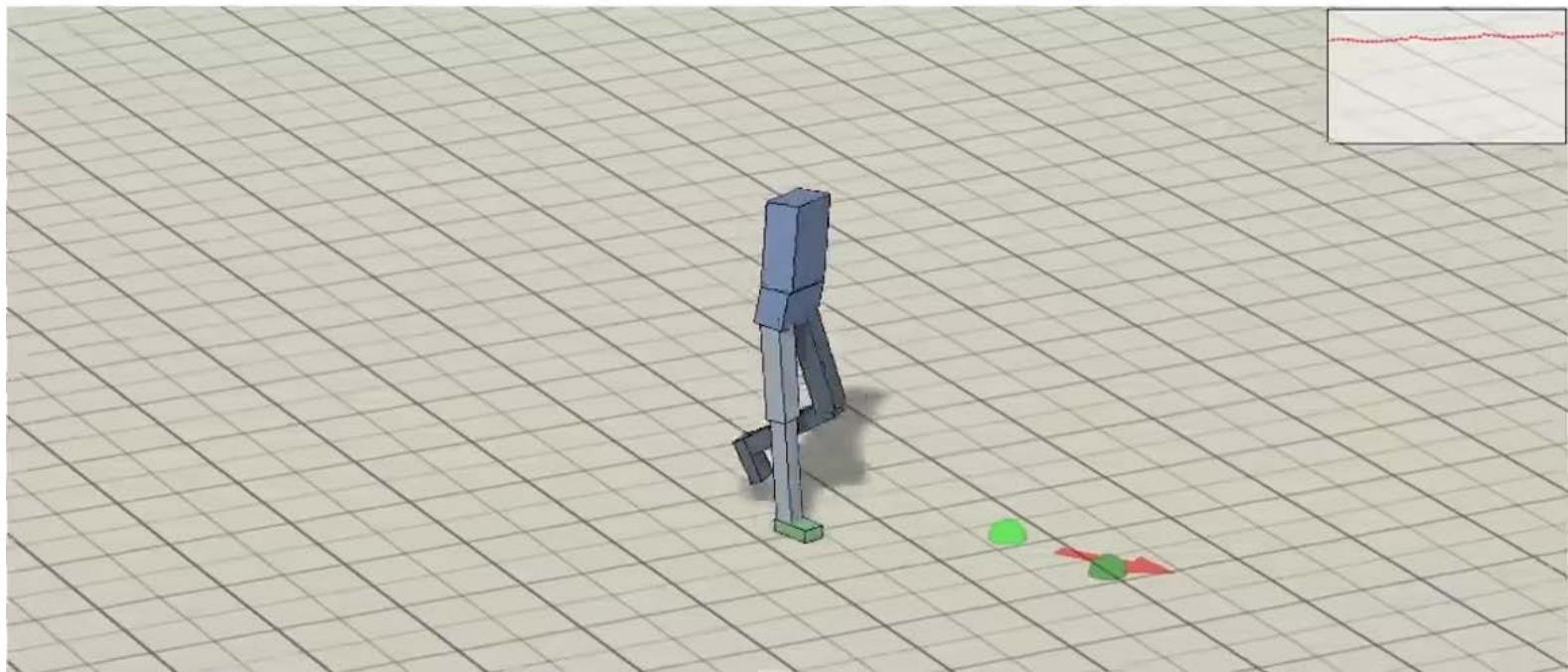
Path Following



More than Images

- Character Control: Reinforcement Learning

LLC: Walk



The LLC is first trained to locomote while following random footstep plans.

Summary

- Vision: Analysis: Discriminative Models
- Graphics: Synthesis: Generative Models
- But many approaches combine both
- ML methods crucial for vision & graphics
- Rapidly moving and exciting field – get involved!