Computer Vision

Lecture 6: Segmentation

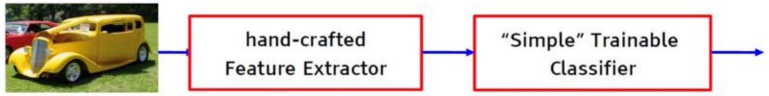
Oliver Bimber

JOHANNES KEPLER UNIVERSITY LINZ

Last Week: Feature Extraction

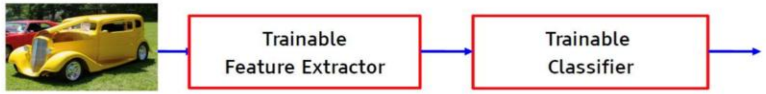
Model-Based vs. Learning-Based Feature Extraction

- Fixed engineered features (or kernels) + trainable classifier



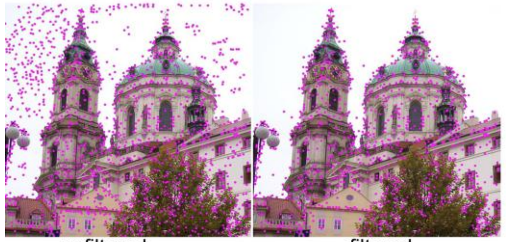
VS.

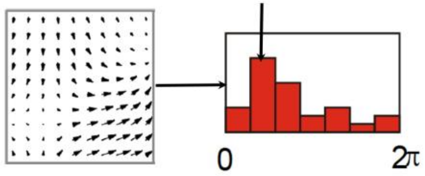
- End-to-end learning / feature learning / deep learning



JOHANNES KEPLER UNIVERSITY LINZ

Example: Scale-Invariant Feature Transform (SIFT)



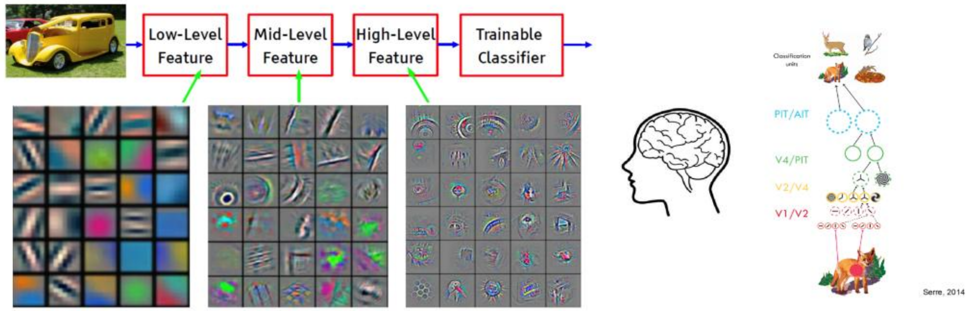


(3) Filter out Features in low-contrast Regions (Noise)

(4) Determine Feature (i.e., Gradient) Orientations and sort them into Histogram (largest Bin = main Orientation)

JOHANNES KEPLER UNIVERSITY LINZ

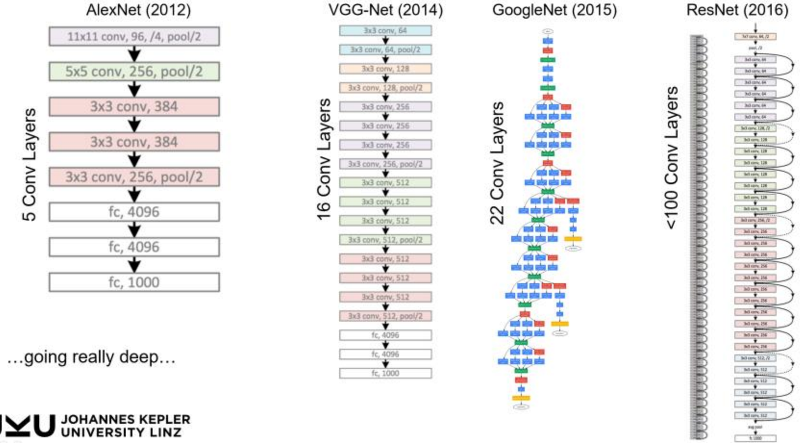
From Low-Level to High-Level Features



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

JOHANNES KEPLER UNIVERSITY LINZ

Development of Architectures



JOHANNES KEPLER UNIVERSITY LINZ

Course Overview

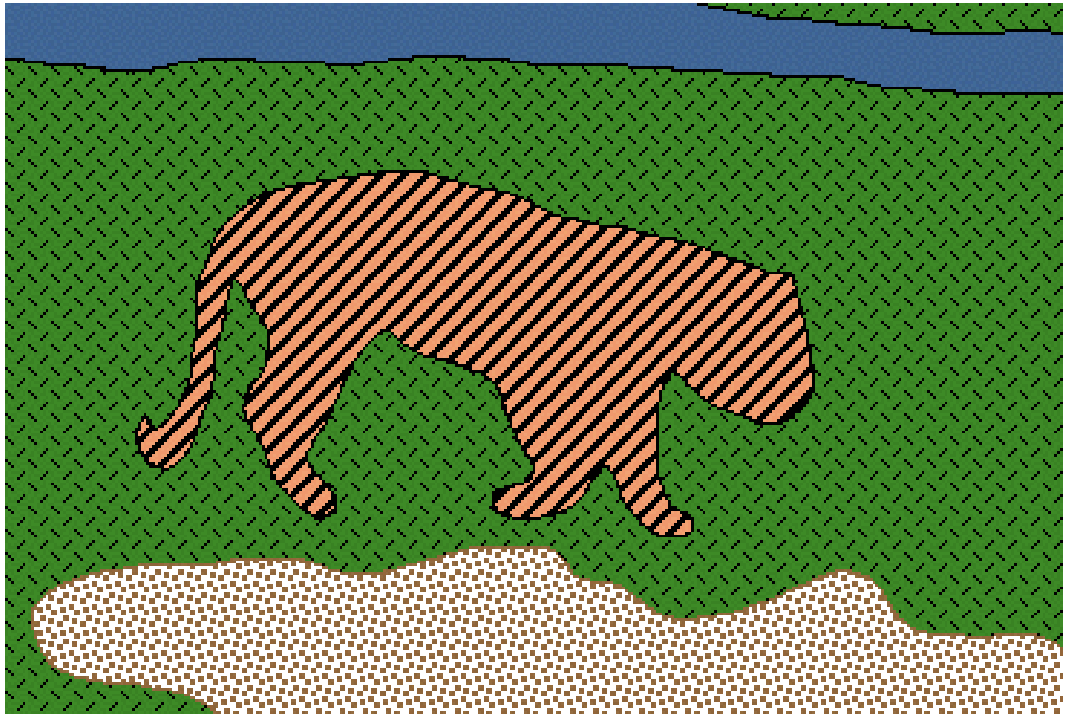
CW	Topic	Date	Place	Lab
41	Introduction and Course Overview	07.10.2025	Zoom	Lab 1
42	Capturing Digital Images	14.10.2025	Zoom	Lab 2
43	Digital Image Processing	21.10.2025	Zoom	Assignment 1
44	Machine Learning	28.10.2025	Zoom	
45	Feature Extraction	04.11.2025	Zoom	Open Lab 1
46	Segmentation	11.11.2025	Zoom	Assignment 2
47	Optical Flow	18.11.2025	Zoom	Open Lab 2
48	Object Detection	25.11.2025	Zoom	Assignment 3
49	Multi-View Geometry	02.12.2025	Zoom	Open Lab 3
50	3D Vision	09.12.2025	Zoom	Assignment 4
3	Trends in Computer Vision	13.01.2026	Zoom	
4	Q&A	20.01.2026	Zoom	Open Lab 4
5	Exam	27.01.2026	HS1 (Linz), S1/S3 (Vienna), S5 (Bregenz)	
9	Retry Exam	24.02.2026	tba	

Research Examples:

Today 1:45pm, HS1 JKU, or:

https://jku.zoom.us/j/97166662151?pwd=woIaNes9BcIitV6Vui32jBHTIyJ4VA.1

What is Segmentation?



Identify Groups of Pixels that go together

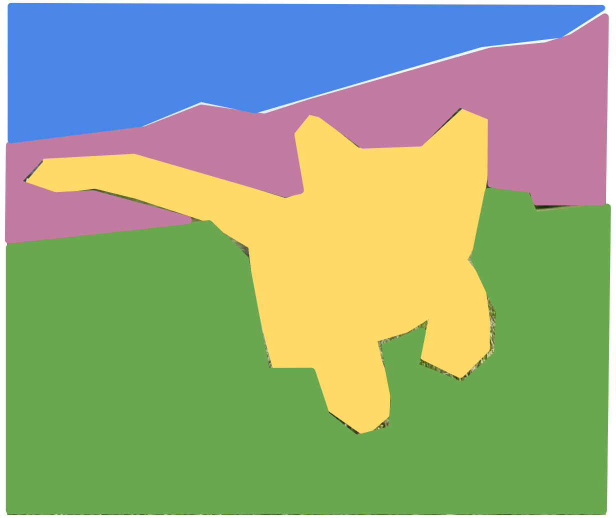
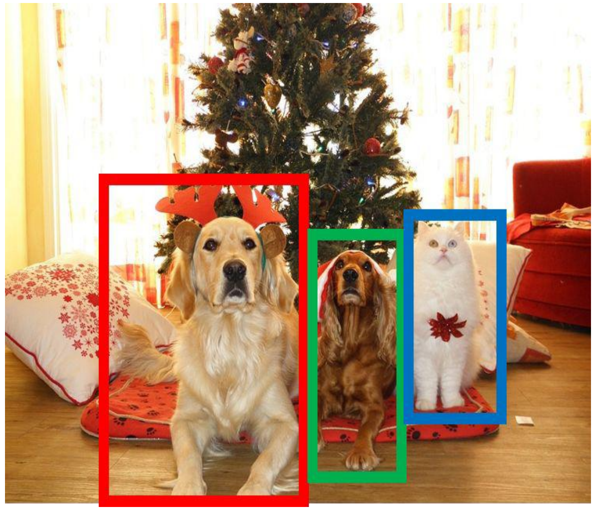
Types of Segmentation

Semantic (by Classes)

Instance (by Objects)

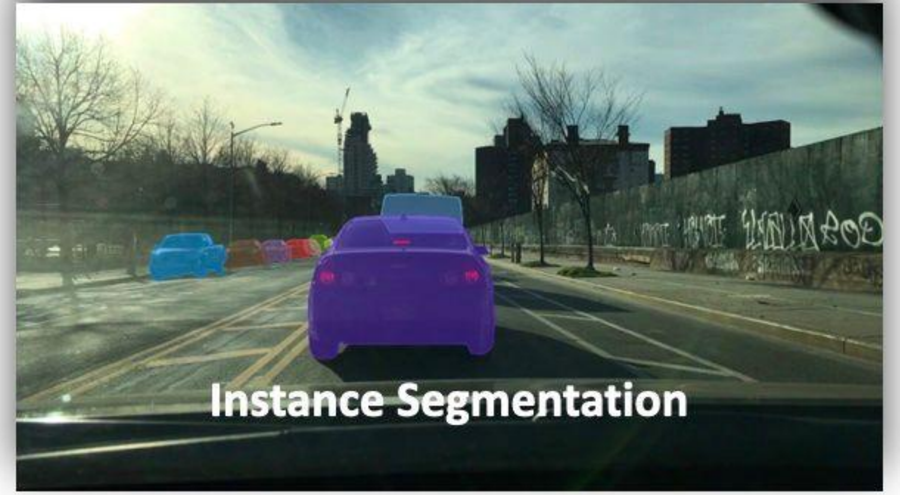
Panoptic (by Classes and Objects)

Classification vs. Segmentation

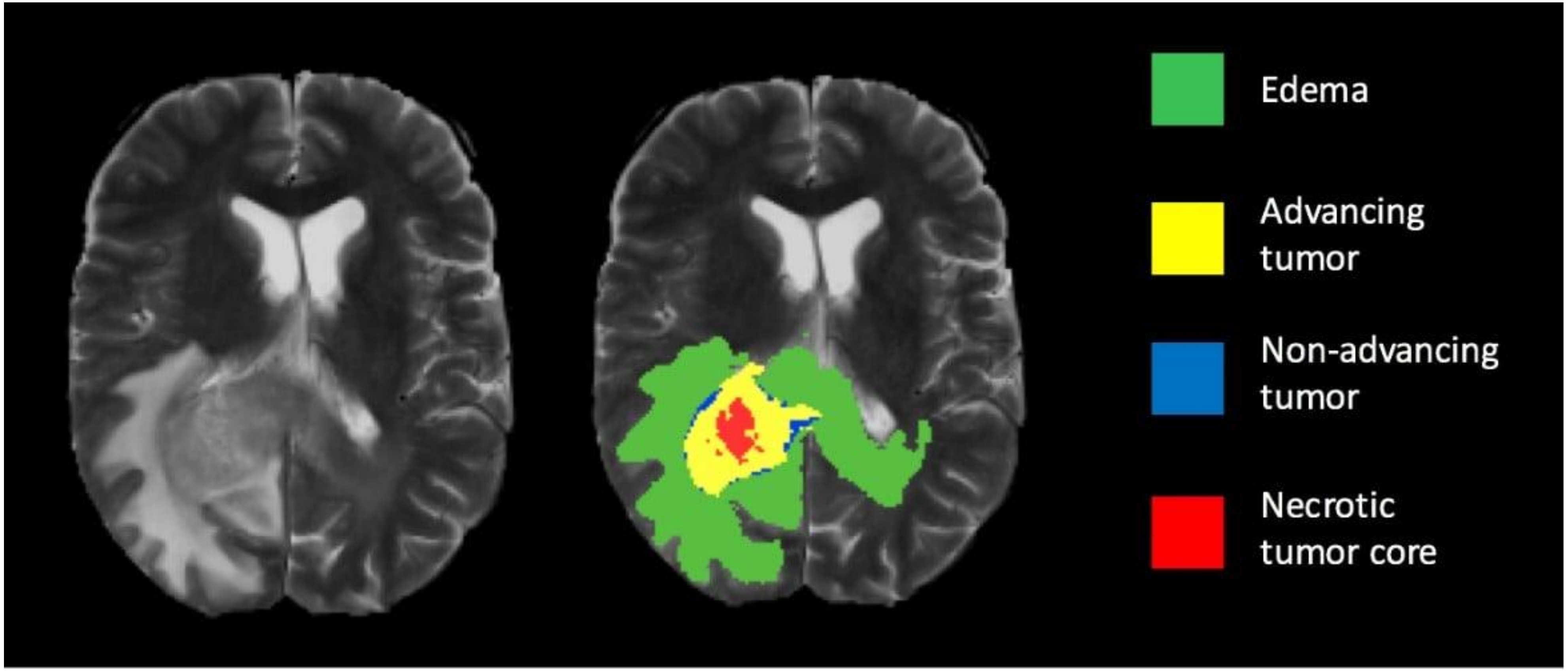
Classification	Semantic Segmentation	Object Detection	Instance Segmentation
			
CAT	GRASS, CAT, TREE, SKY	DOG, DOG, CAT	DOG, DOG, CAT
No spatial extent	No objects, just pixels	Multiple Object	

Example: Image Editing

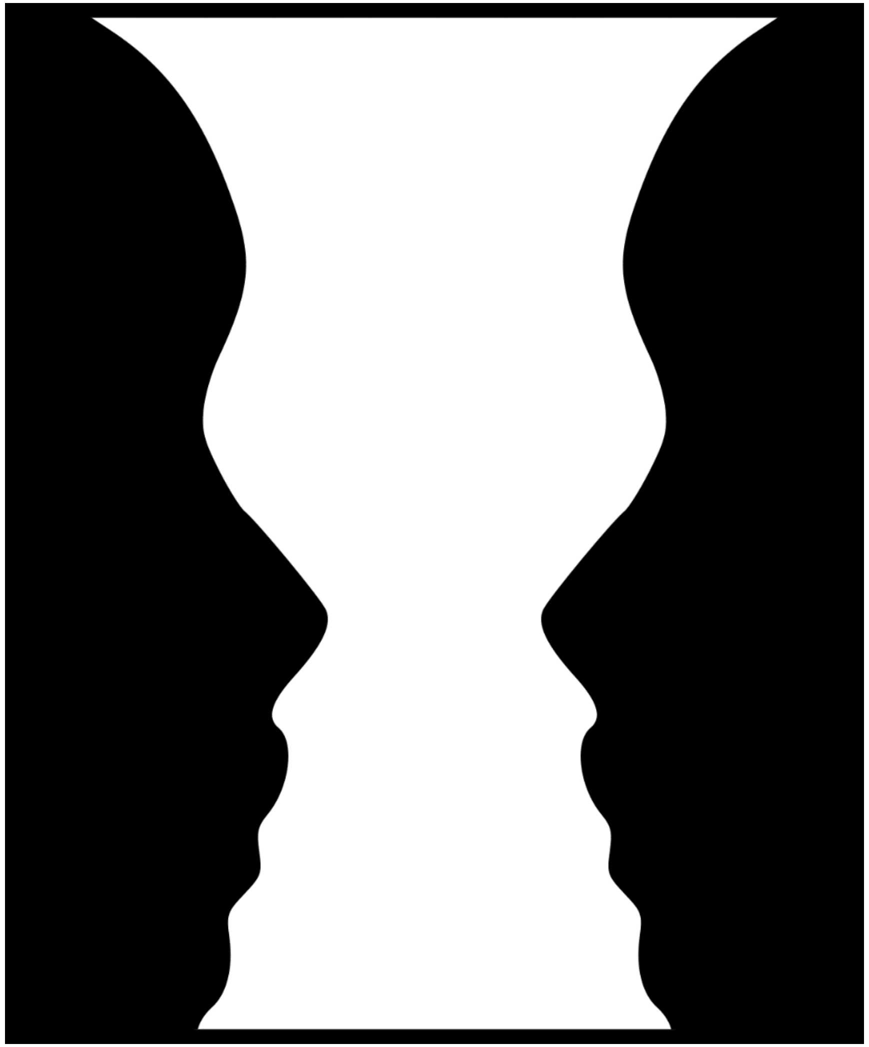
Example: Autonomous Driving

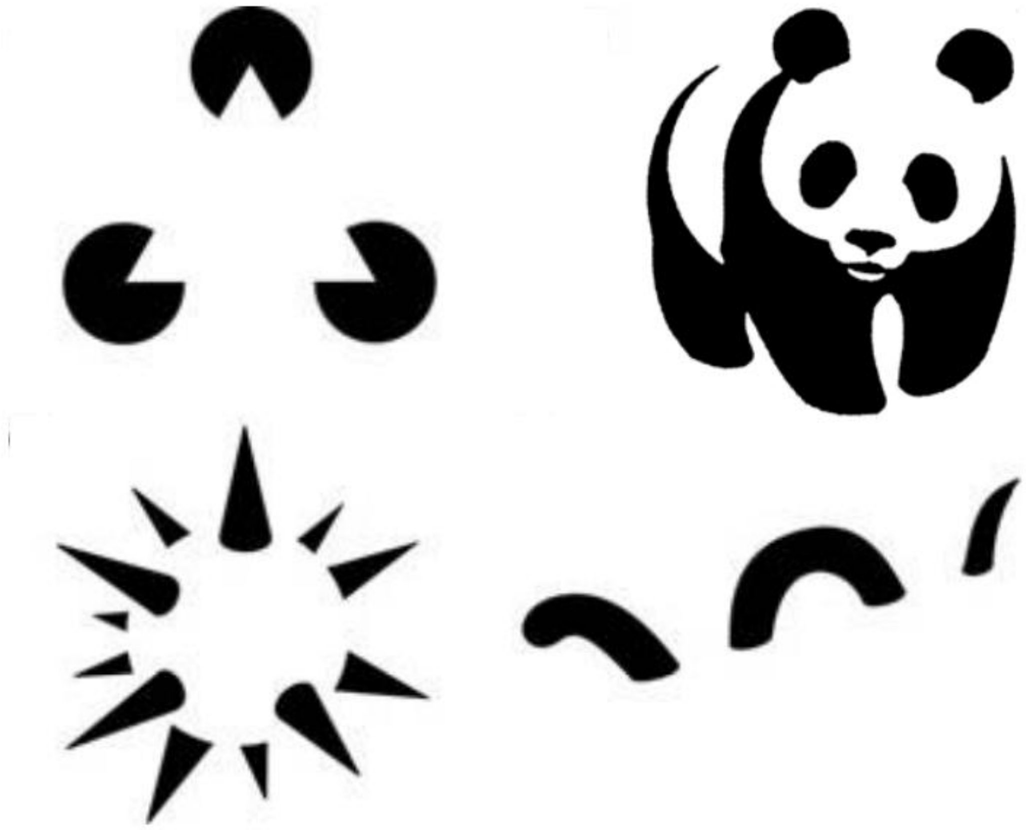


Example: Medical Imaging



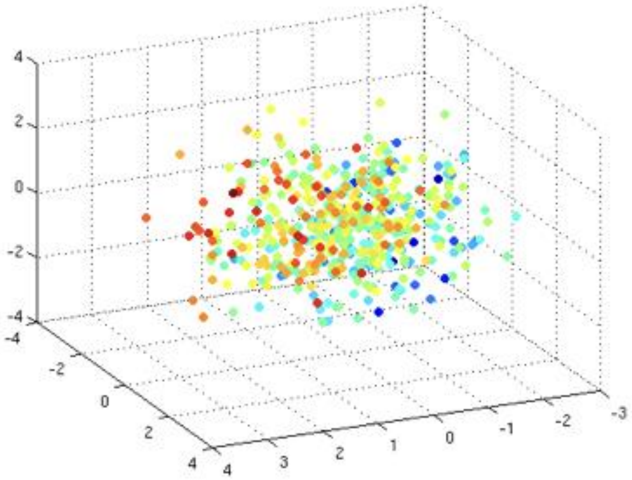
Segmentation done by Humans





Illusory or subjective contours are perceived

Segmentation by Clustering

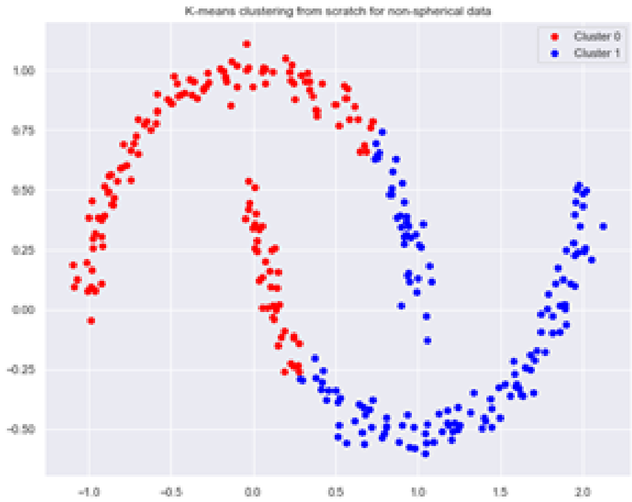


- Pixels are Points in a high-dimensional Space, eg.:
  - Color: 3d
  - Color + Location: 5d

- Cluster Pixels into Segments, eg.:
  - K-Means Clustering

Example: K-Means Clustering

$$\Phi_{(\text{cluster},\text{data})} = \sum_{i \in \text{cluster}} \left\{ \sum_{j \in \text{cluster}(i)} (x_j - c_i)^T (x_j - c_i) \right\}$$



K-means:
choose k points by random to act as cluster centers
until cluster centers are unchanged
    allocate each point to nearest cluster (ensure that each cluster has at least one point)
    replace cluster centers with mean of new cluster points
end

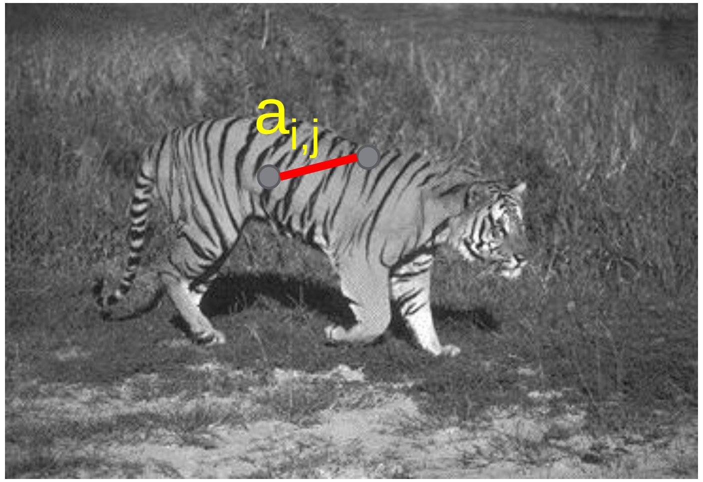
Example: K-Means Clustering

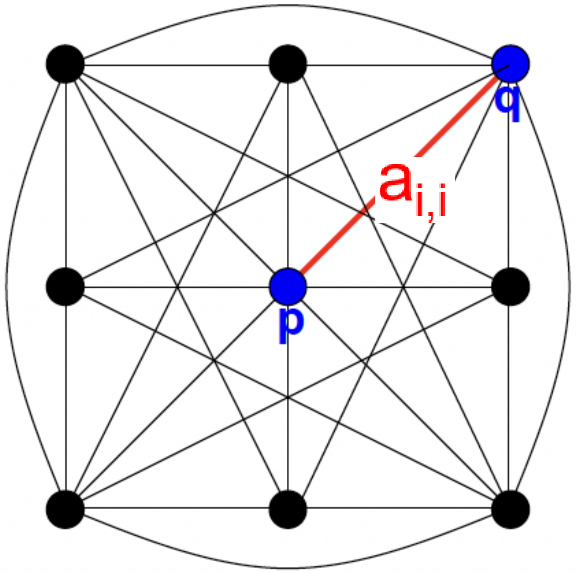
As K increases…

$$\Phi_{(\text{cluster},\text{data})} = \sum_{i \in \text{cluster}} \left\{ \sum_{j \in \text{cluster}(i)} (x_j - c_i)^T (x_j - c_i) \right\}$$

K-means:
choose k points by random to act as cluster centers
until cluster centers are unchanged
    allocate each point to nearest cluster (ensure that each cluster has at least one point)
    replace cluster centers with mean of new cluster points
end

Segmentation with Graphs

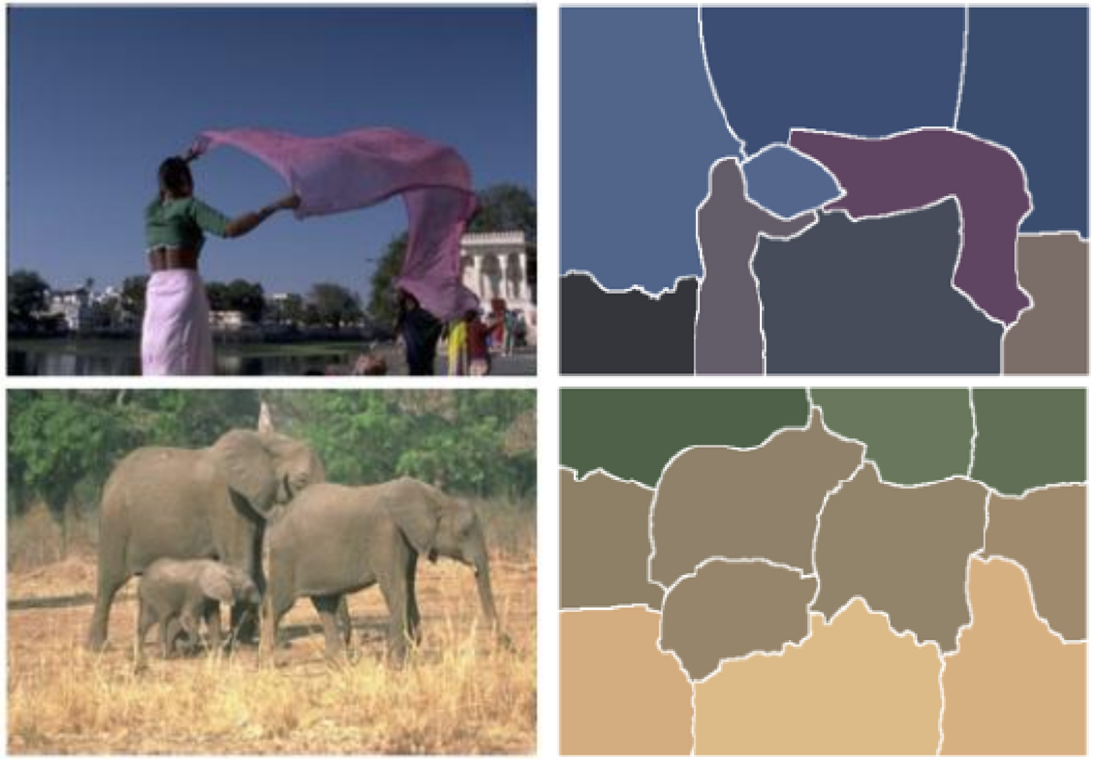




- Pixels are Nodes in a Graph
  - Edge between Pairs of Pixels (i,j)
  - Affinity Weight $a_{i,i}$ for each Edge measures “Similarity”

- Cluster Pixels into Segments, eg.:
  - Clustering by Graph Eigenvectors, Graph Cut, Grab Cut

Example: Clustering by Graph Eigenvectors



Let's assume your Graph is represented with an Affinity Matrix **A** (size: **NxN** for **N** Elements)

A good Cluster is one where Elements that are strongly connected to the Cluster also have large Values connecting one another in the Affinity Matrix

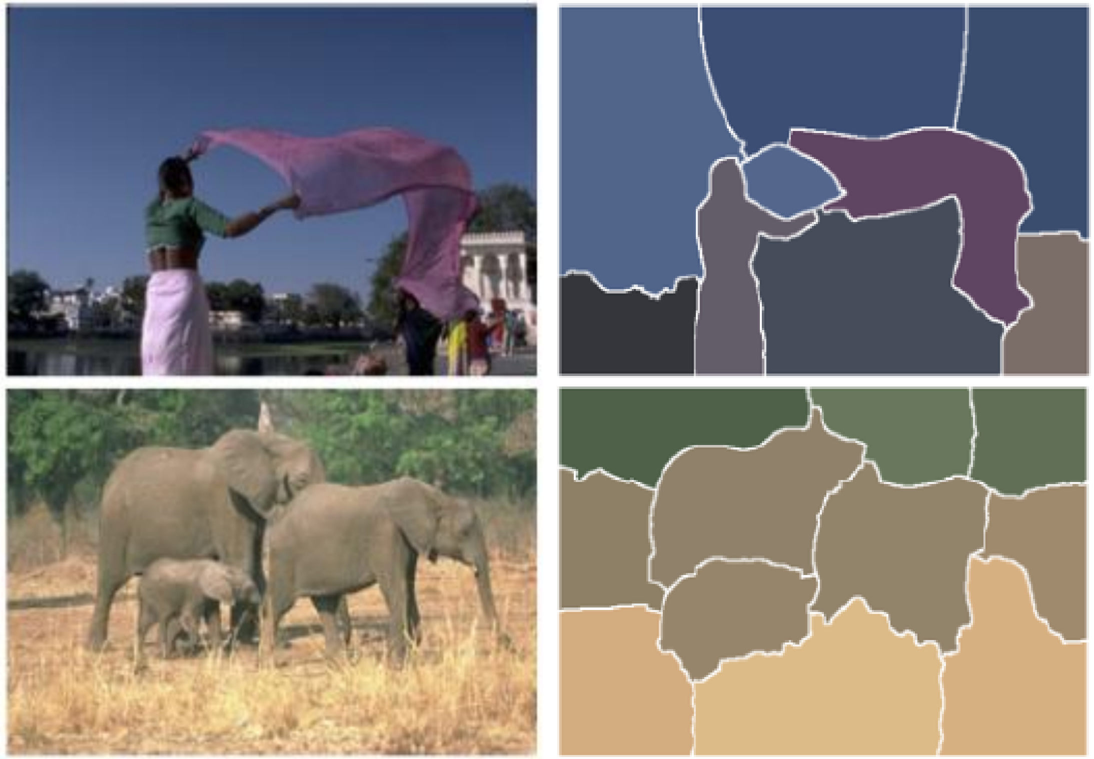
Let's assume you make an Assignment for Cluster **n** using a Weight Vector $\mathbf{w_n}$ (with **N** elements, high values indicate strong connectivity to the Cluster)

When is use Assignment a good one?

Our objective from above rephrased mathematically:

$$\text{Max}(\ w_n^T A w_n = \lambda\ )$$

Example: Clustering by Graph Eigenvectors



affinity between element i and j

X

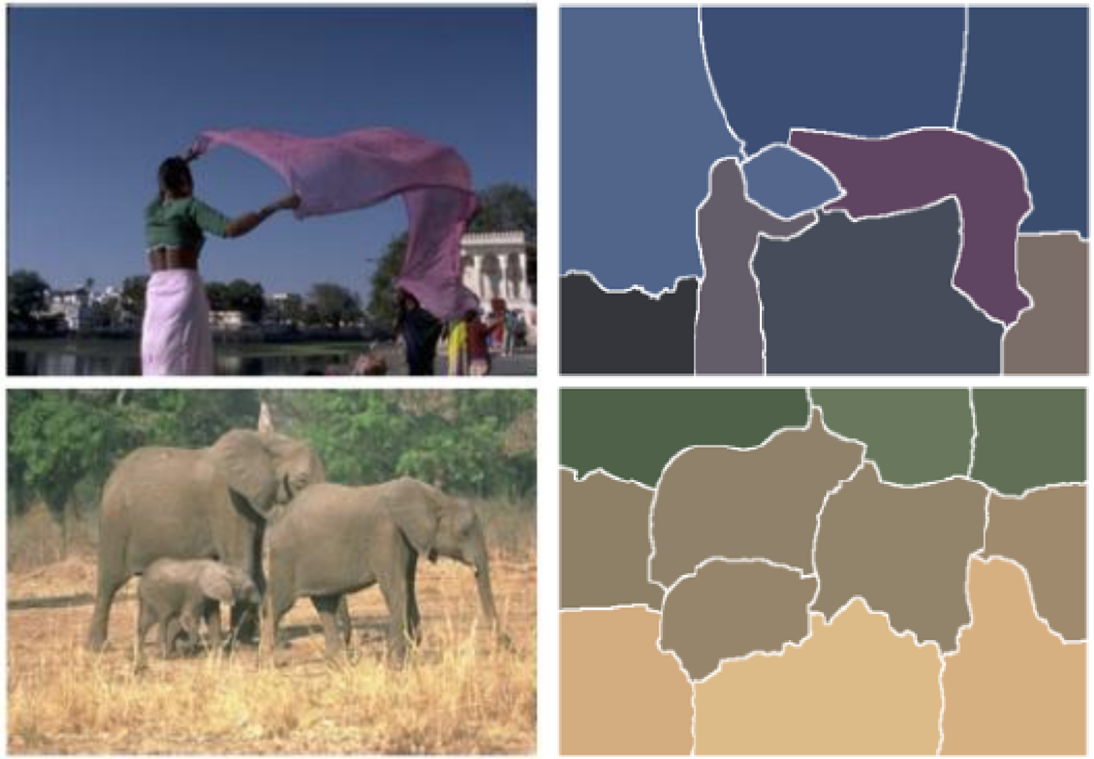
association of elements j with cluster n

$$\boxed{w_n^T}\boxed{Aw_n} = \lambda \qquad Aw_n = \lambda w_n$$

association of elements j with cluster n X

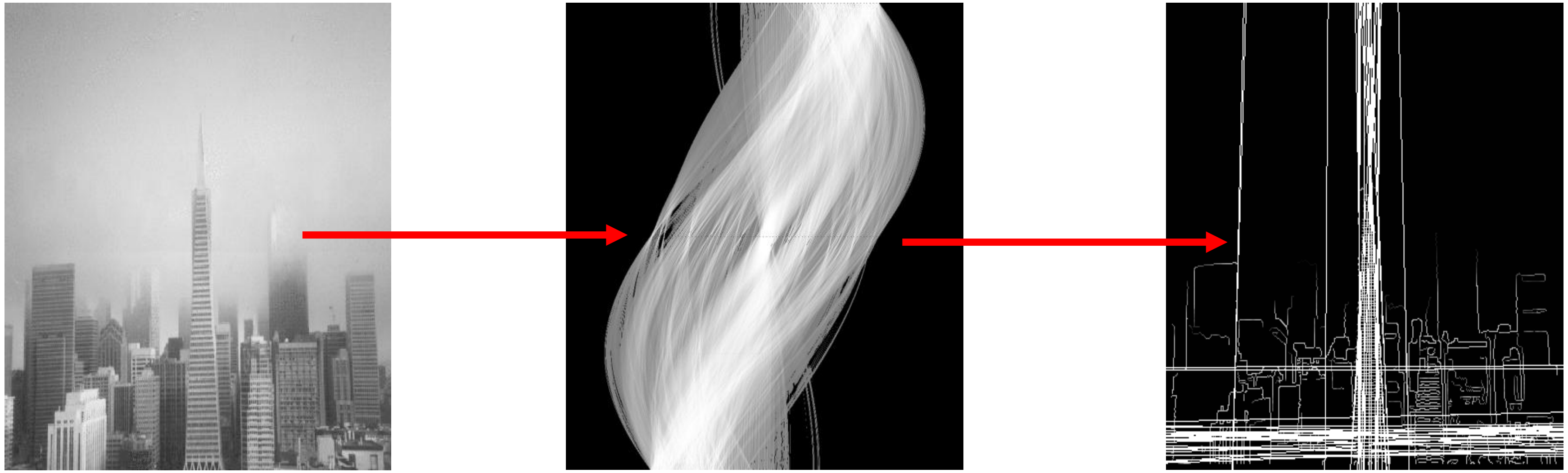
$$\begin{bmatrix} w_{n,0} \dots w_{n,j} \end{bmatrix} \begin{bmatrix} a_{0,0} & \dots & a_{i,0} \\ \dots & \dots & \dots \\ a_{0,j} & \dots & a_{i,j} \end{bmatrix} \begin{bmatrix} w_{n,0} \\ \dots \\ w_{n,j} \end{bmatrix}$$

Example: Clustering by Graph Eigenvectors



Clustering by Graph Eigenvectors:
construct affinity matrix A
compute eigenvalues and eigenvectors of A
until there are sufficient clusters
    take the eigenvector corresponding to next largest eigenvalue
    assign elements to cluster (multiply eigenvector with A and threshold)
    zero out all clustered elements in A
end

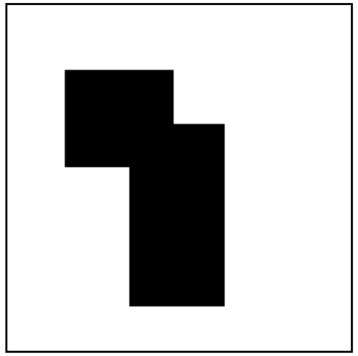
Segmentation by Fitting

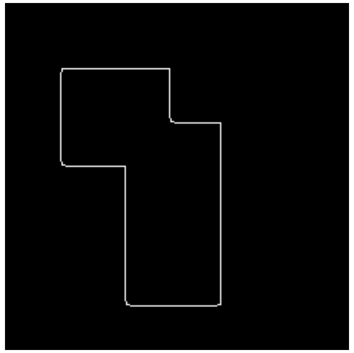


- Knowing the Shapes of the Segments
  - Segments can be parameterized

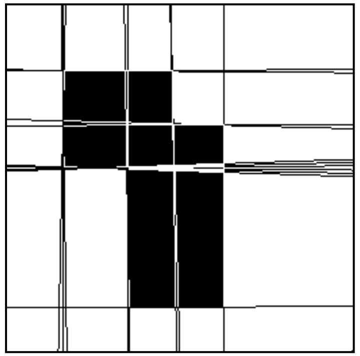
- Transform Pixels into Parameter Space, eg.:
  - Hough Transform

Example: Hough Transform

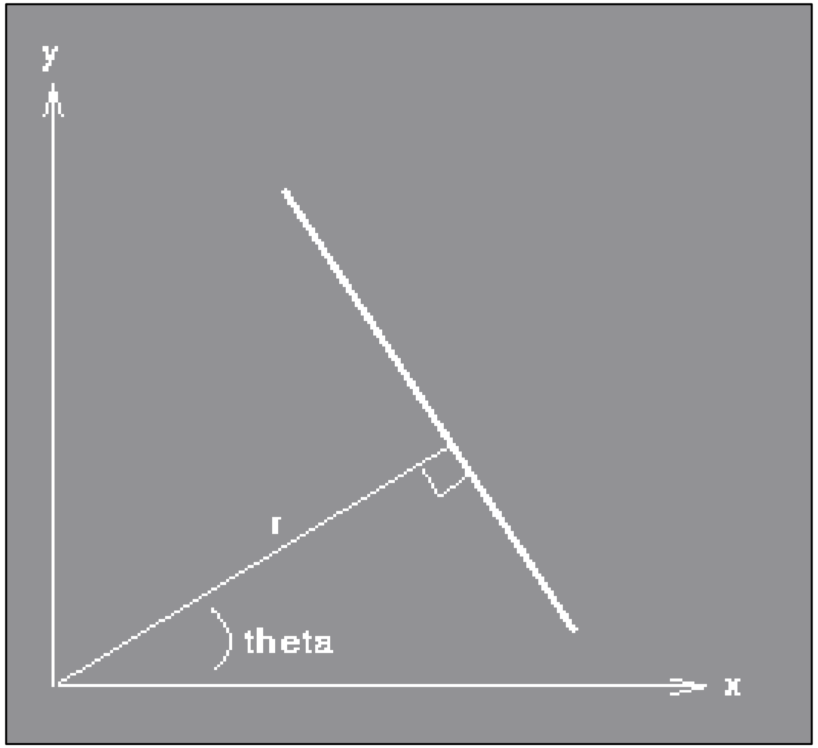

original image


edge detector


line space

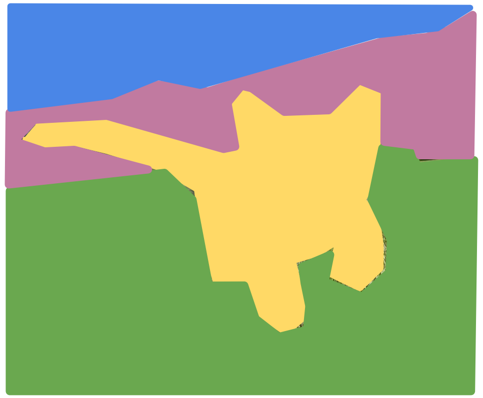

fitted lines

$$x \cos\theta + y \sin\theta + r = 0$$



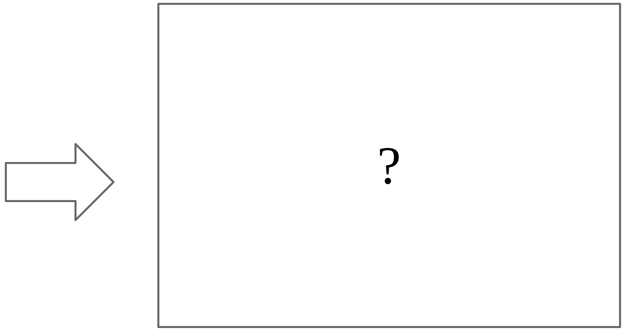
line space parameterization

Segmentation by Learning



GRASS, CAT,
TREE, SKY, ...

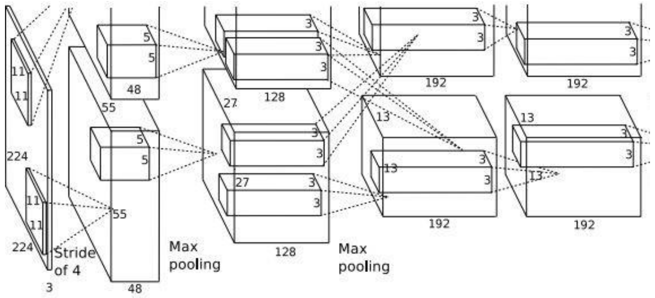
Paired Training Data: for each Training Image, each Pixel is labeled with a Semantic Category



At Test Time, Classify each Pixel of a new Image

Segmentation using CNNs

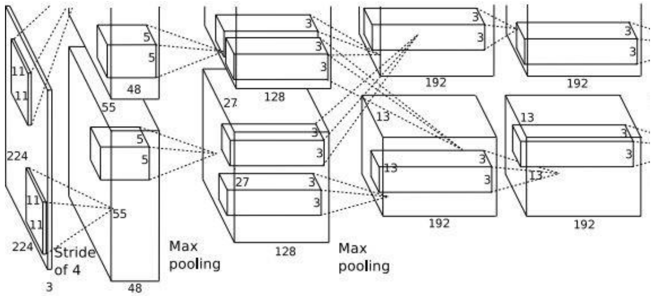
Full image



An intuitive Idea: encode the entire Image with a CNN, and do Semantic Segmentation on top

Segmentation using CNNs

Full image

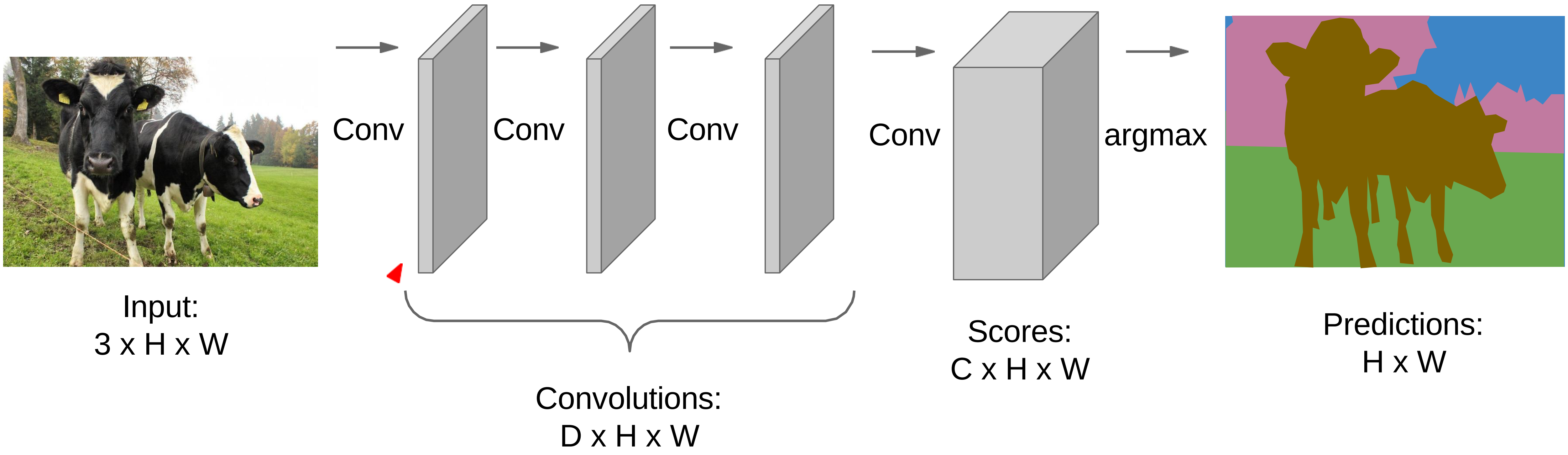


An intuitive Idea: encode the entire Image with a CNN, and do Semantic Segmentation on top

Problem: Classification Architectures often reduce Feature spatial Sizes to go deeper, but Semantic Segmentation requires the Output Size to be the same as Input Size

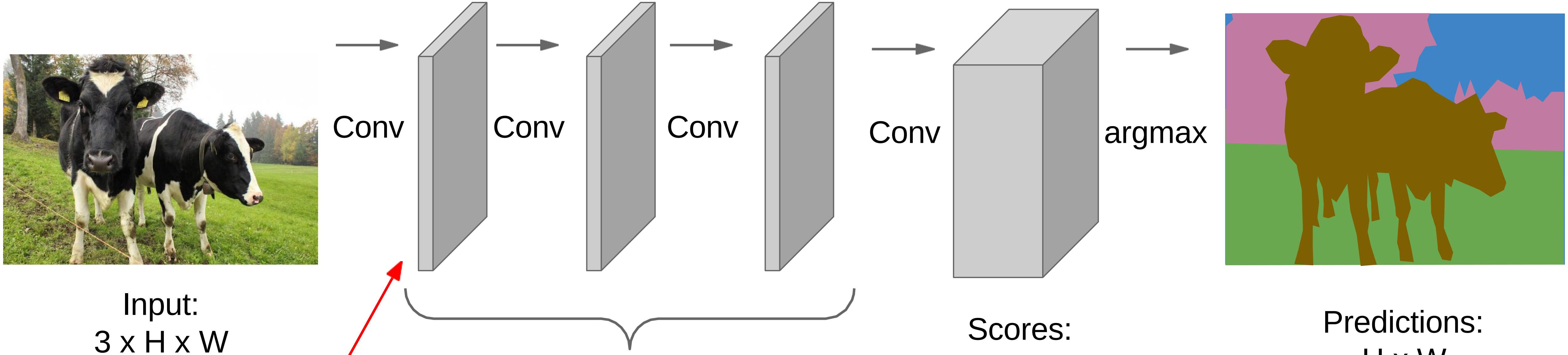
Segmentation using CNNs

Design a Network with only Convolutional Layers
without Downsampling Operators to make
Predictions for Pixels all at once!



Segmentation using CNNs

Design a Network with only Convolutional Layers without Downsampling Operators to make Predictions for Pixels all at once!



Input:
3 x H x W

Convolutions:
D x H x W

Scores:
C x H x W

Predictions:
H x W

Problem: Convolutions at original Image Resolution will be very expensive ...

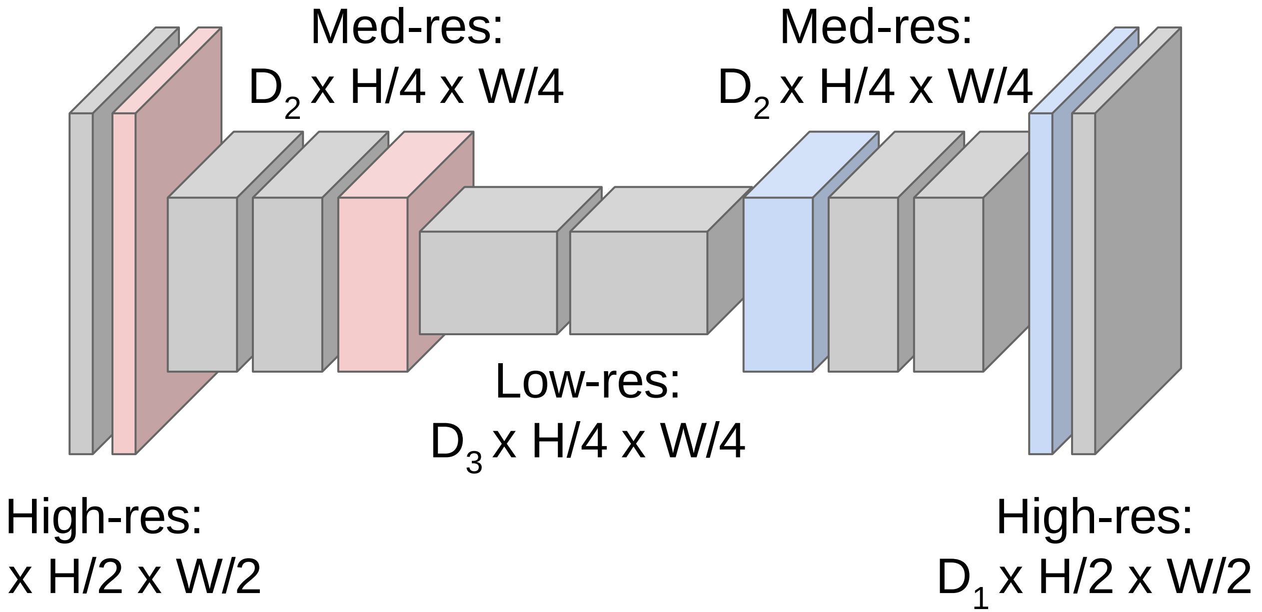
Segmentation using CNNs

Downsampling: Pooling, Strided Convolution

Design Network as a Bunch of Convolutional Layers, with **Downsampling** and **Upsampling** inside the Network!

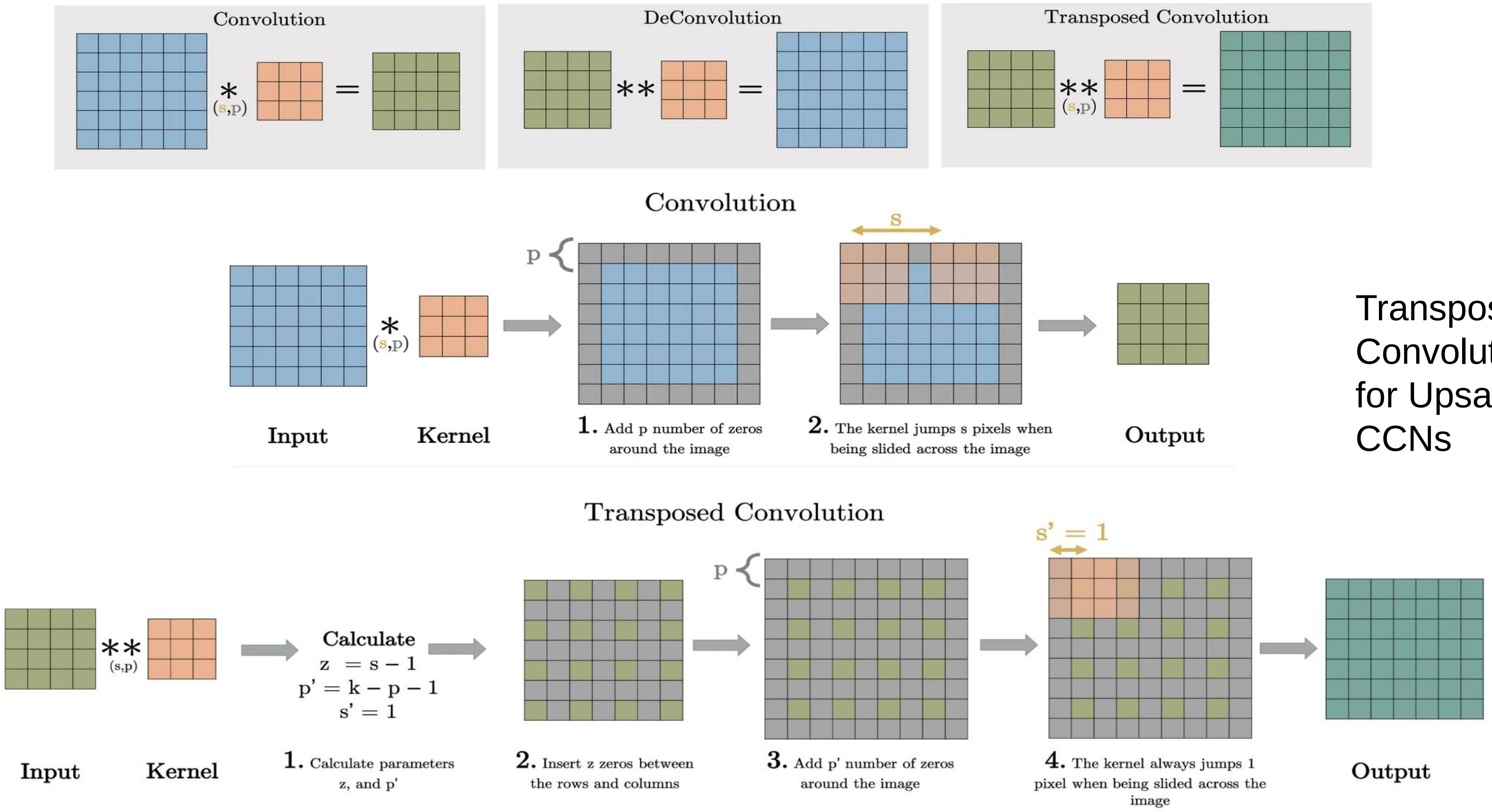
Upsampling: Strided Transposed Convolution (or Unpooling)

Input: 3 x H x W



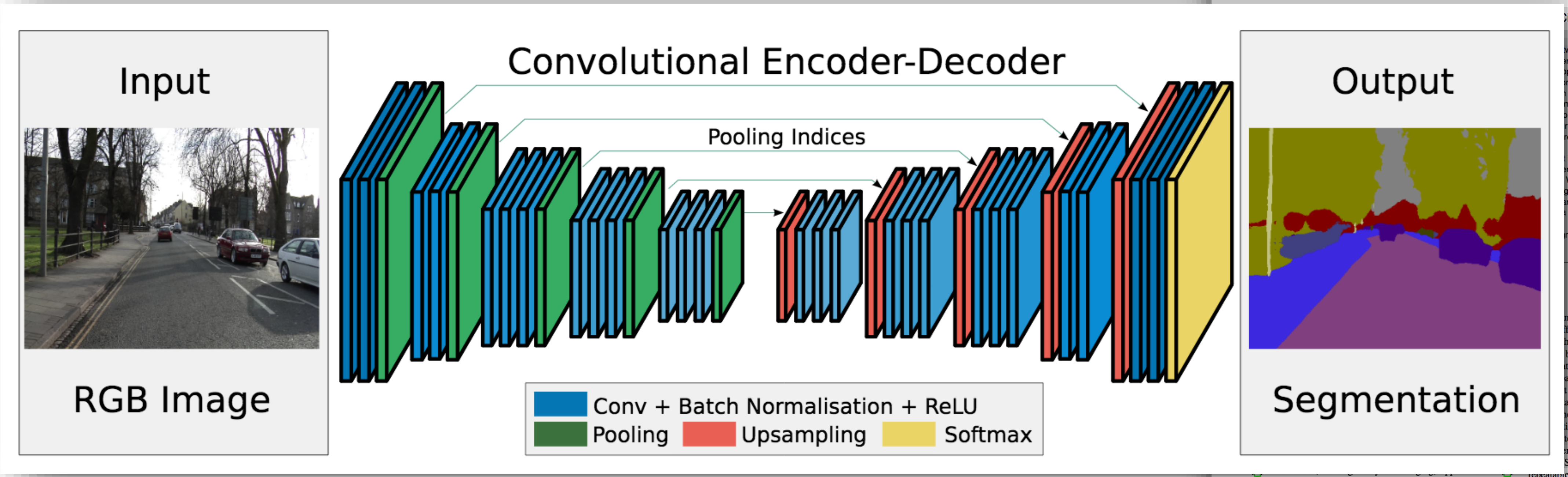
Predictions: H x W

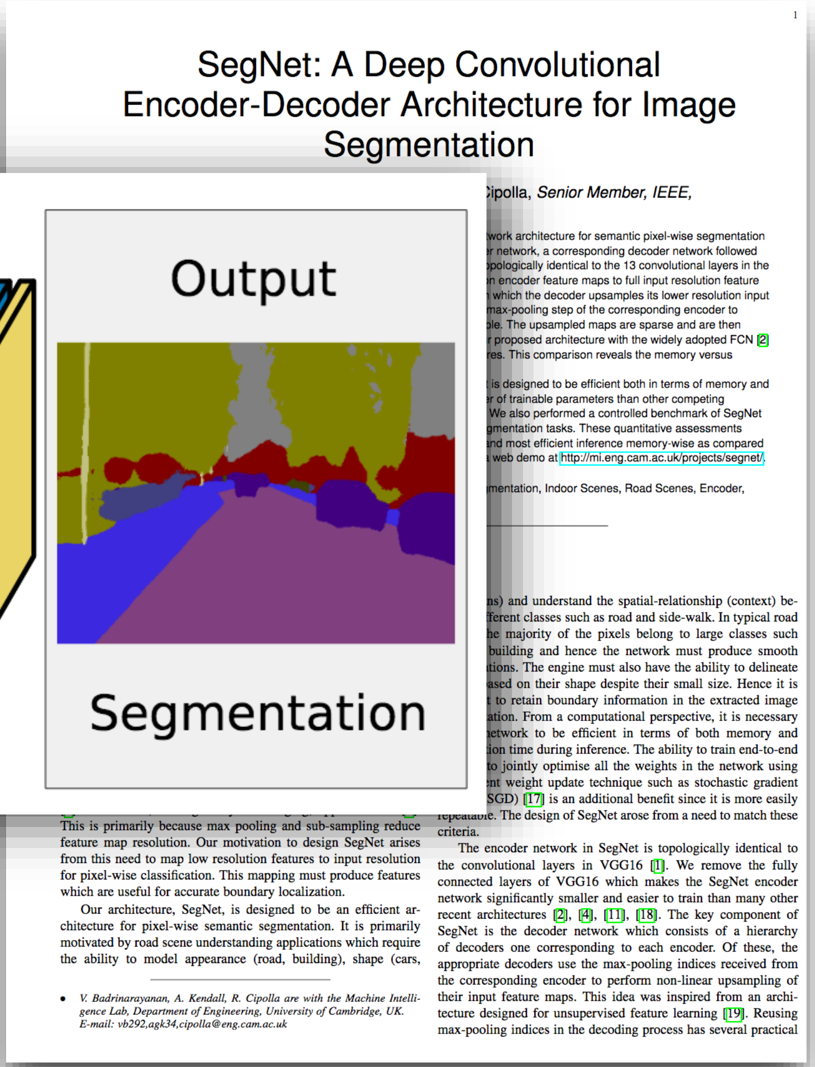
Recap: Transposed Convolution



Transposed Convolution used for Upsampling in CCNs

Autoencoder: Encoder-Decoder Architectures





SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

...ipolla, *Senior Member, IEEE,*

...work architecture for semantic pixel-wise segmentation
...r network, a corresponding decoder network followed
...pologically identical to the 13 convolutional layers in the
...n encoder feature maps to full input resolution feature
...n which the decoder upsamples its lower resolution input
...max-pooling step of the corresponding encoder to
...ble. The upsampled maps are sparse and are then
...r proposed architecture with the widely adopted FCN [2]
...res. This comparison reveals the memory versus

... is designed to be efficient both in terms of memory and
...r of trainable parameters than other competing
... We also performed a controlled benchmark of SegNet
...gmentation tasks. These quantitative assessments
...nd most efficient inference memory-wise as compared
... web demo at http://mi.eng.cam.ac.uk/projects/segnet/.

...mentation, Indoor Scenes, Road Scenes, Encoder,

This is primarily because max pooling and sub-sampling reduce feature map resolution. Our motivation to design SegNet arises from this need to map low resolution features to input resolution for pixel-wise classification. This mapping must produce features which are useful for accurate boundary localization.

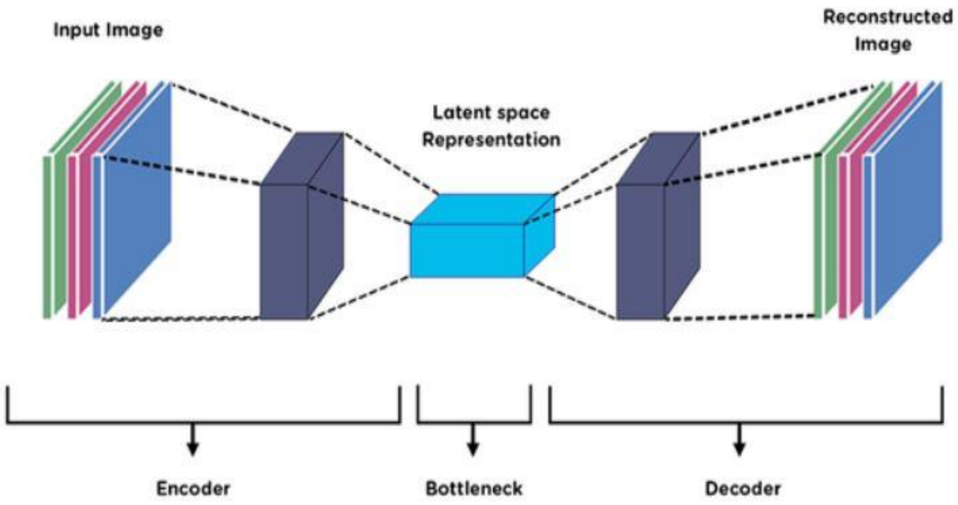
Our architecture, SegNet, is designed to be an efficient architecture for pixel-wise semantic segmentation. It is primarily motivated by road scene understanding applications which require the ability to model appearance (road, building), shape (cars,

...ns) and understand the spatial-relationship (context) be... ...fferent classes such as road and side-walk. In typical road ...he majority of the pixels belong to large classes such ...building and hence the network must produce smooth ...tions. The engine must also have the ability to delineate ...ased on their shape despite their small size. Hence it is ...t to retain boundary information in the extracted image ...ation. From a computational perspective, it is necessary ...etwork to be efficient in terms of both memory and ...ion time during inference. The ability to train end-to-end ...to jointly optimise all the weights in the network using ...nt weight update technique such as stochastic gradient ...SGD) [17] is an additional benefit since it is more easily repeatable. The design of SegNet arose from a need to match these criteria.

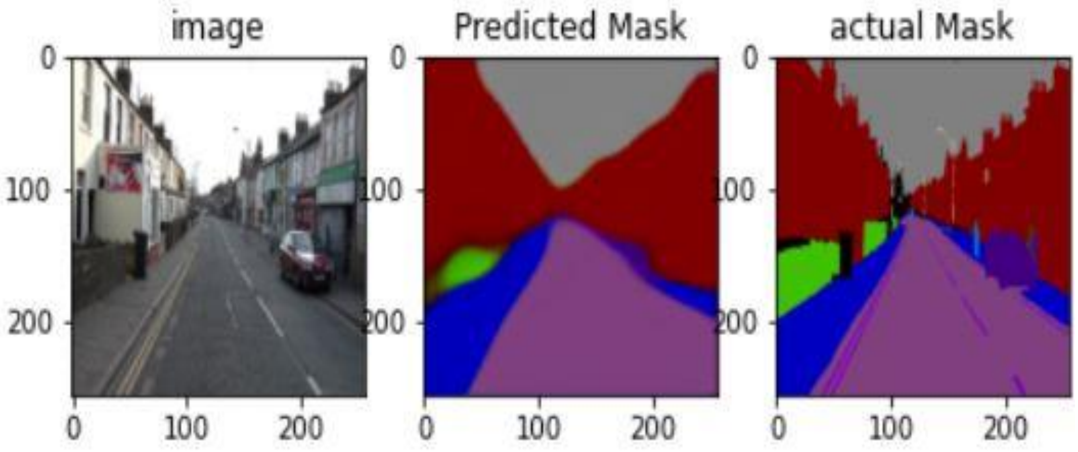
The encoder network in SegNet is topologically identical to the convolutional layers in VGG16 [1]. We remove the fully connected layers of VGG16 which makes the SegNet encoder network significantly smaller and easier to train than many other recent architectures [2], [4], [11], [18]. The key component of SegNet is the decoder network which consists of a hierarchy of decoders one corresponding to each encoder. Of these, the appropriate decoders use the max-pooling indices received from the corresponding encoder to perform non-linear upsampling of their input feature maps. This idea was inspired from an architecture designed for unsupervised feature learning [19]. Reusing max-pooling indices in the decoding process has several practical

- V. Badrinarayanan, A. Kendall, R. Cipolla are with the Machine Intelligence Lab, Department of Engineering, University of Cambridge, UK. E-mail: vb292,agk34,cipolla@eng.cam.ac.uk

Unconnected Autoencoders



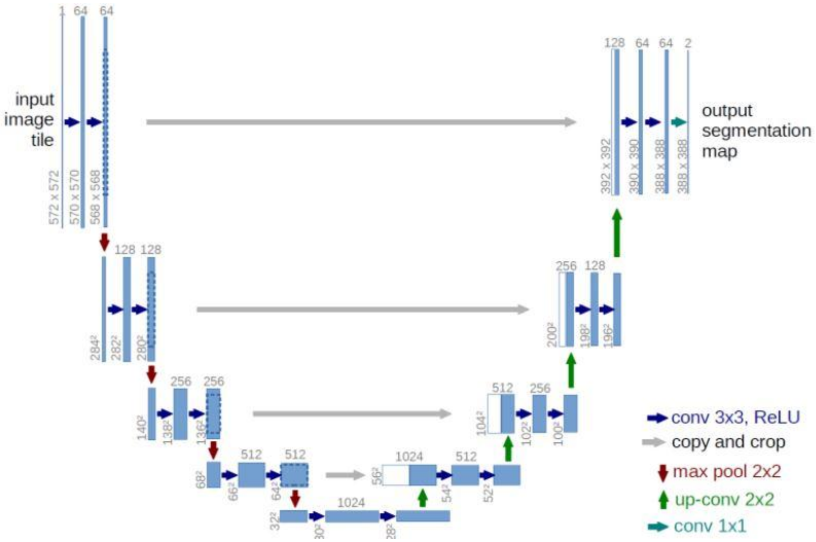
Unconnected Autoencoder



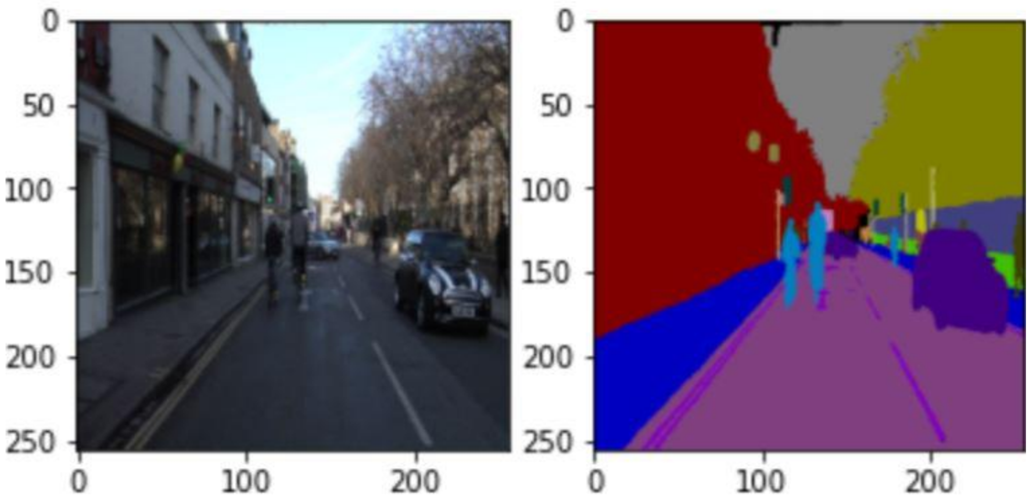
Imprecision in predicted Segmentation

However, convolution and pooling during encoding and transposed convolution during decoding leads to a imprecise feature prediction

Connected Autoencoders (U-Nets)



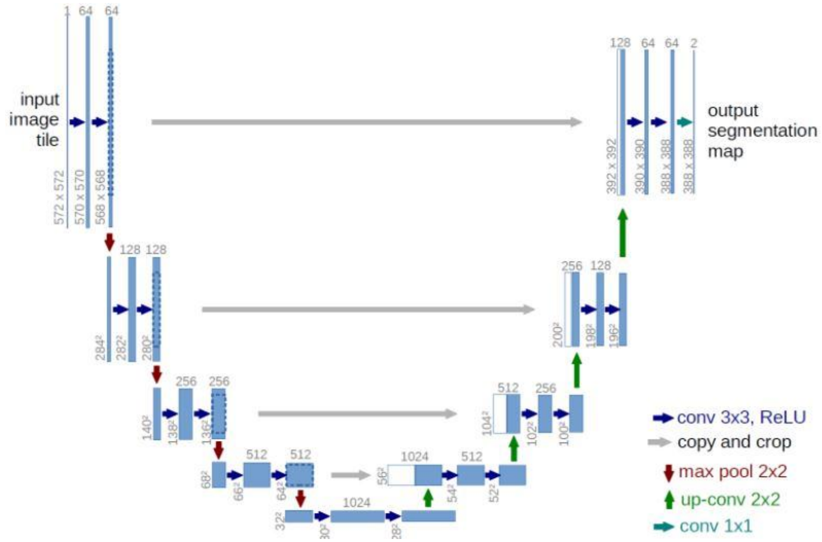
Connected Autoencoder (U-Net)



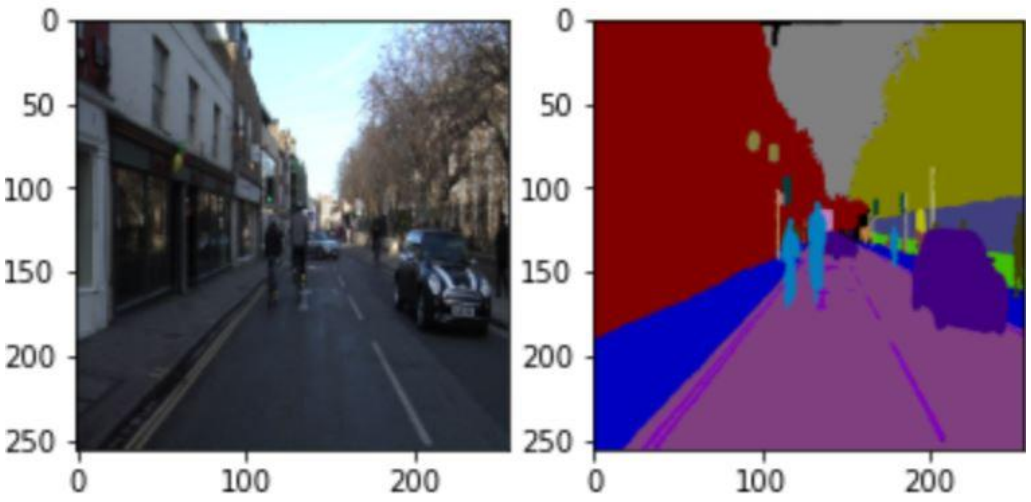
Predicted Segmentation

- U-Nets overcome this problem by connecting corresponding encoder-decoder layers with skip connections:
  - the output of an encoder level is skip-connected (concatenation) with the input of the corresponding decoder level

Connected Autoencoders (U-Nets)



Connected Autoencoder (U-Net)



Predicted Segmentation

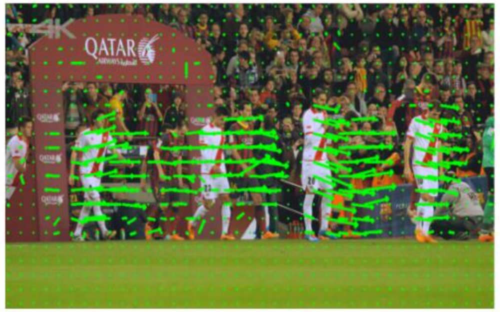
- Why conv 1x1?
  - linear projection of stack of features (dimensionality reduction)

Course Overview

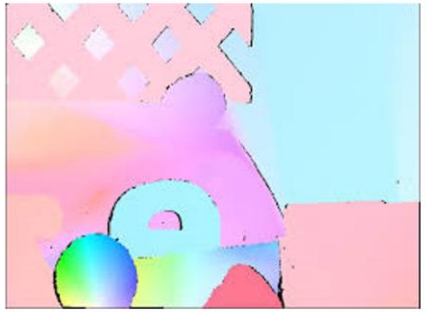
CW	Topic	Date	Place	Lab
41	Introduction and Course Overview	07.10.2025	Zoom	Lab 1
42	Capturing Digital Images	14.10.2025	Zoom	Lab 2
43	Digital Image Processing	21.10.2025	Zoom	Assignment 1
44	Machine Learning	28.10.2025	Zoom	
45	Feature Extraction	04.11.2025	Zoom	Open Lab 1
46	Segmentation	11.11.2025	Zoom	Assignment 2
47	Optical Flow	18.11.2025	Zoom	Open Lab 2
48	Object Detection	25.11.2025	Zoom	Assignment 3
49	Multi-View Geometry	02.12.2025	Zoom	Open Lab 3
50	3D Vision	09.12.2025	Zoom	Assignment 4
3	Trends in Computer Vision	13.01.2026	Zoom	
4	Q&A	20.01.2026	Zoom	Open Lab 4
5	Exam	27.01.2026	HS1 (Linz), S1/S3 (Vienna), S5 (Bregenz)	
9	Retry Exam	24.02.2026	tba	

Next Week: Optical Flow

What is Optical Flow?



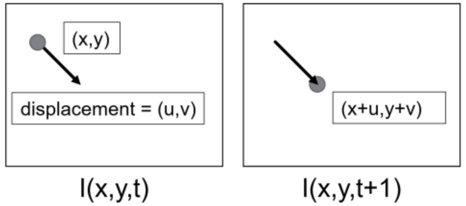
Sparse Optical Flow (Flow Vector per Region)



Dense Optical Flow (Flow Vector per Pixel)

Motion Vector of Pixel in Time Series (two consecutive Video Frames at Times t and t+1)

The Optical Flow Equation



Brightness Constancy:

$$I(x+u, y+v, t+1) = I(x,y,t)$$

$$0 \approx I(x+u, y+v, t+1) - I(x,y,t)$$

Taylor Expansion:

$$\approx I(x,y,t+1) + I_x u + I_y v - I(x,y,t)$$

$$= I(x,y,t+1) - I(x,y,t) + I_x u + I_y v$$

Optical Flow Equation:

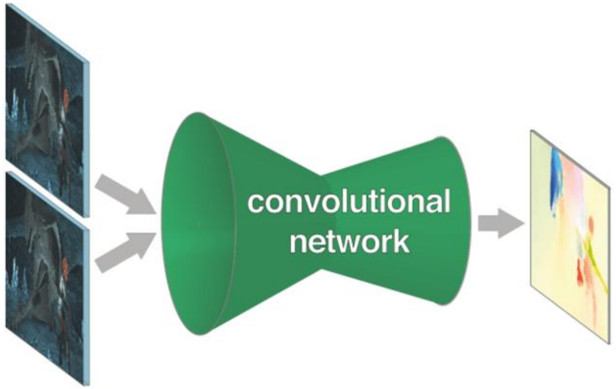
$$0 = I_t + I_x u + I_y v$$

$$= I_t + \nabla I \cdot [u,v]$$

I_t, I_x, I_y are partial derivatives of image intensity (gradients) in t,x,y

JOHANNES KEPLER UNIVERSITY LINZ

Optical Flow and Machine Learning

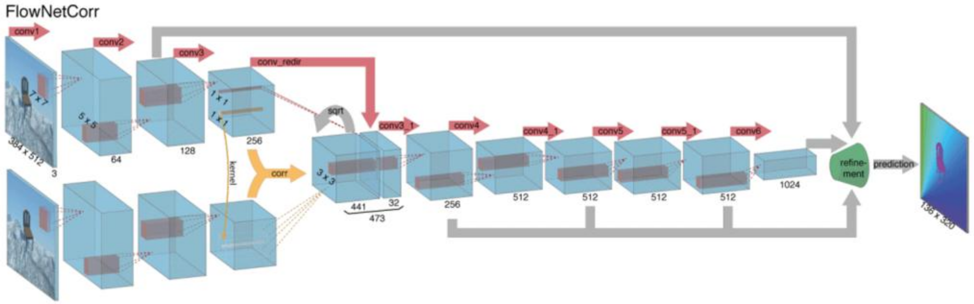


Encoder+Decoder Architectures (e.g. U-Nets)

JOHANNES KEPLER UNIVERSITY LINZ

Example: FlowNetCorr (Correlation)

https://lmb.informatik.uni-freiburg.de/Publications/2015/DFIB15/flownet.pdf



Input: 2 Tensors of individual RGB Images (Feature Maps are computed later → Correlation Layer)

JOHANNES KEPLER UNIVERSITY LINZ

JOHANNES KEPLER UNIVERSITY LINZ

Thank You

JOHANNES KEPLER
UNIVERSITY LINZ