Convolutional Neural Networks – Part 2 Fundamentals and Applications

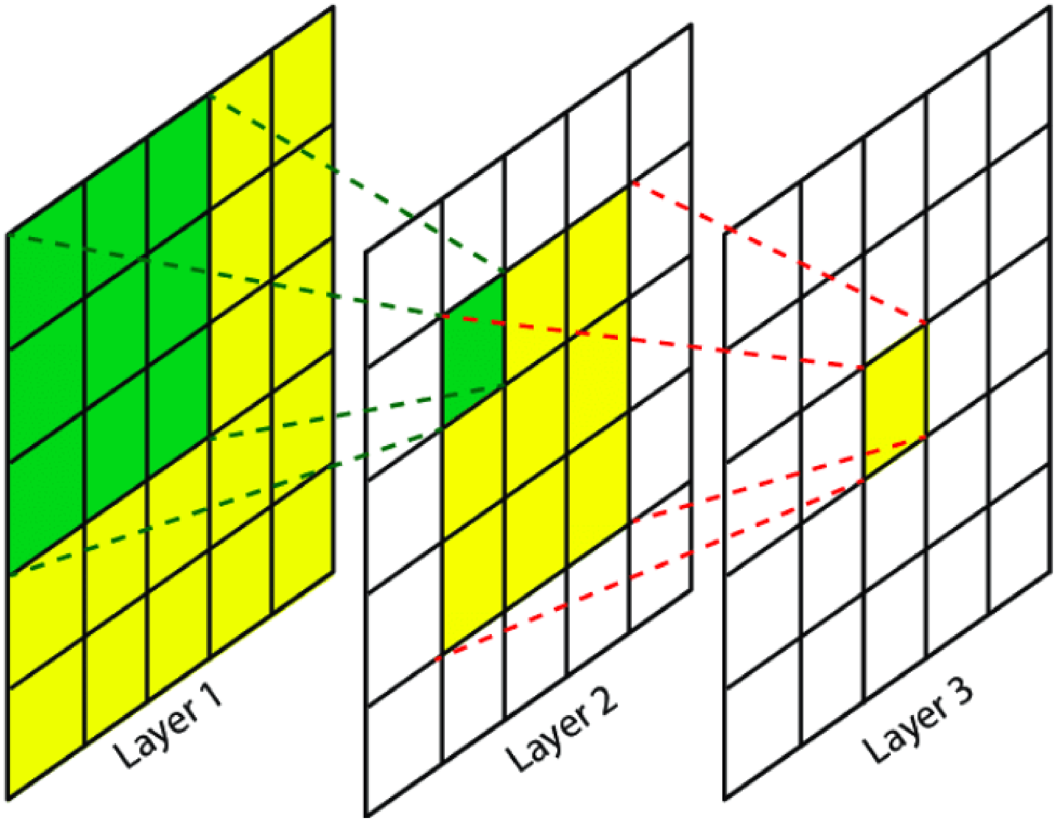
ECE 449

Outline

- Fundamentals
  - Receptive field
  - Convolution
    - Dilated conv
    - 1×1 conv
    - Depthwise conv
    - Group conv
- CV related Applications
  - Image Classification
  - Detection and Tracking
  - Pose Estimation
  - Segmentation
  - 3D and Localization
  - Image Reconstruction
  - …

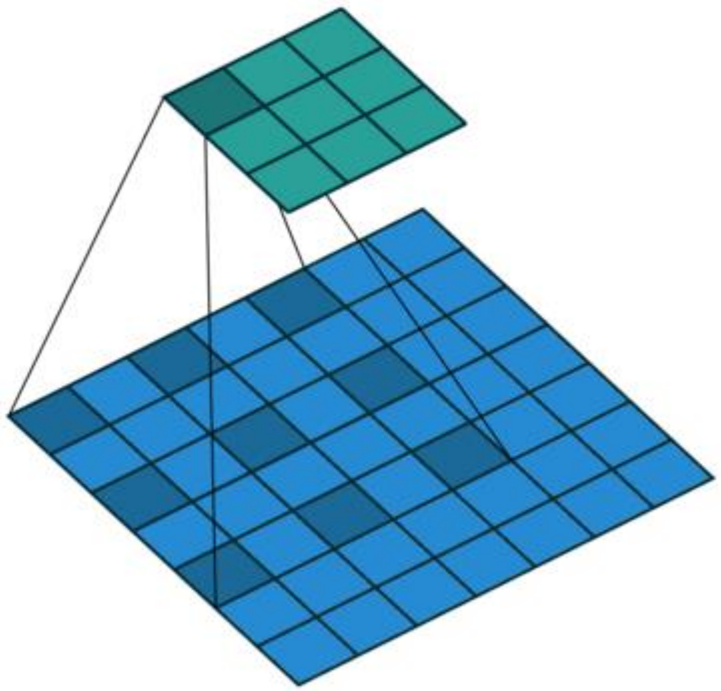
Receptive Field

- The receptive field in CNN is the region of the input space that affects a particular unit of the network.



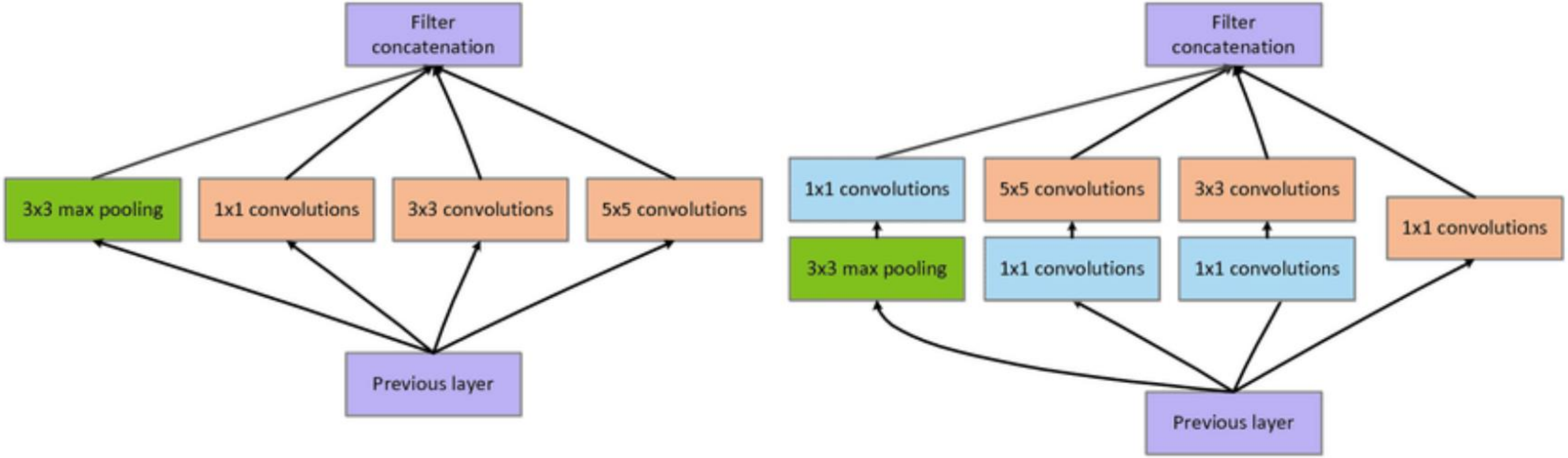
Dilated Conv

- Enlarge the receptive field
- Example, dilation=2



1×1 Conv

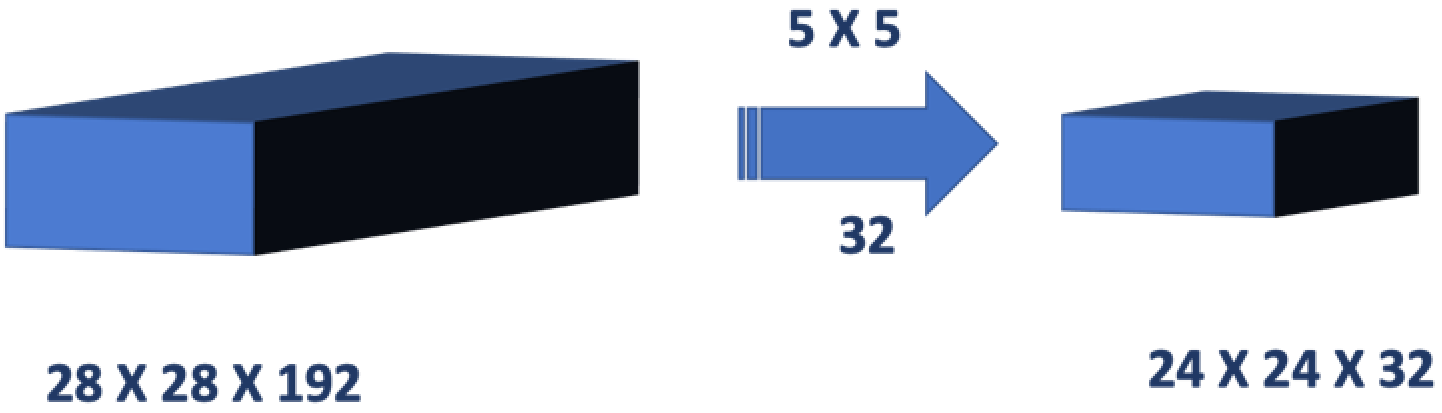
- Example, inception block



(a) Inception module, initial form (b) Inception module with dimension reductions

1×1 Conv

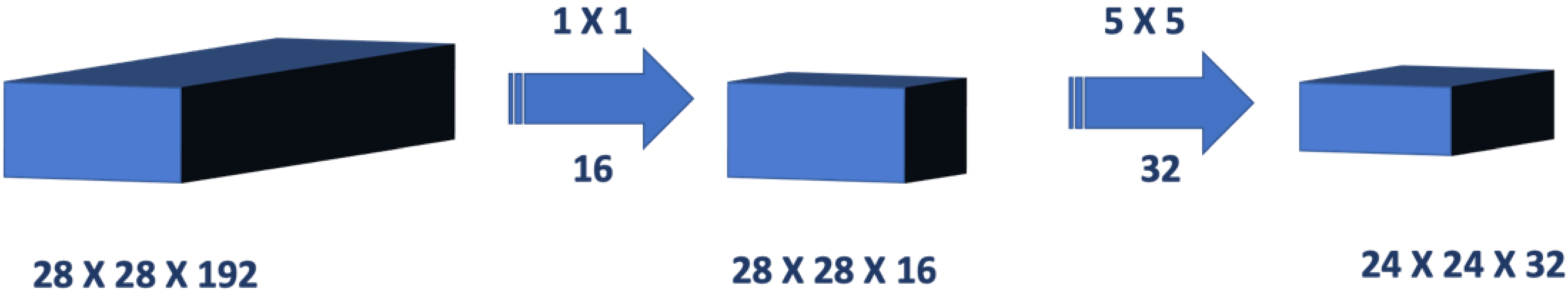
- Cross channel information aggregation
- Used for feature projection or dimension reduction



Operations: (5×5×192)×(24×24)×32=88.5M
Parameters: 5×5×192×32=153.6K

1×1 Conv

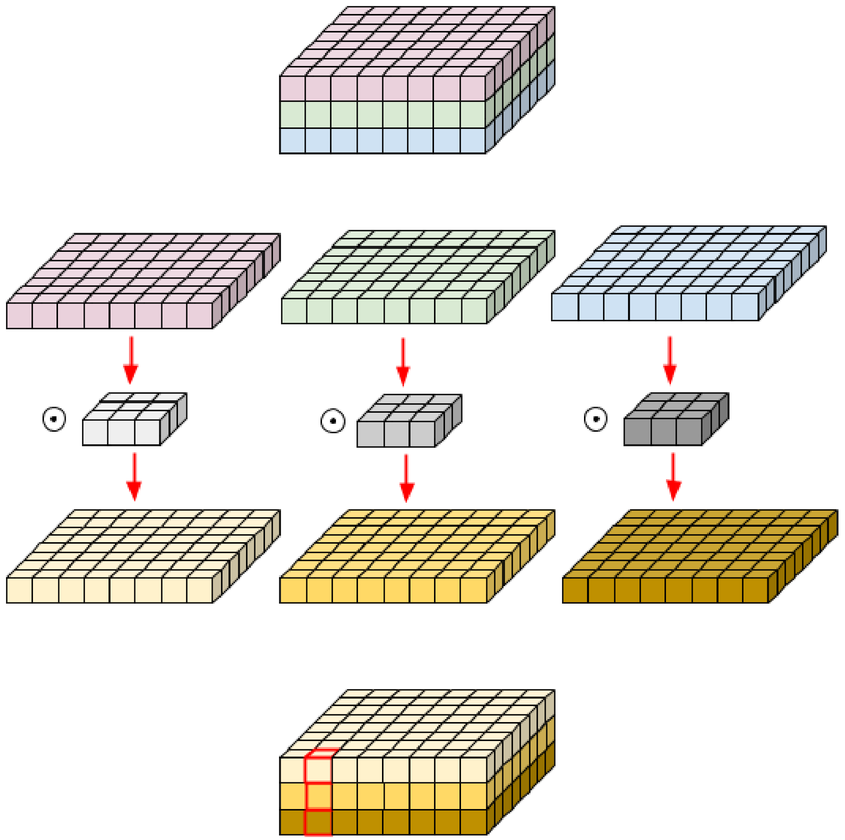
- Cross channel information aggregation
- Used for feature projection or dimension reduction



Operations: $(1\times1\times192)\times(28\times28)\times16+(5\times5\times16)\times(24\times24)\times32=2.4M+7.4M=9.8M$
Parameters: $1\times1\times192\times16+5\times5\times16\times32=3K+12.8K=15.8K$

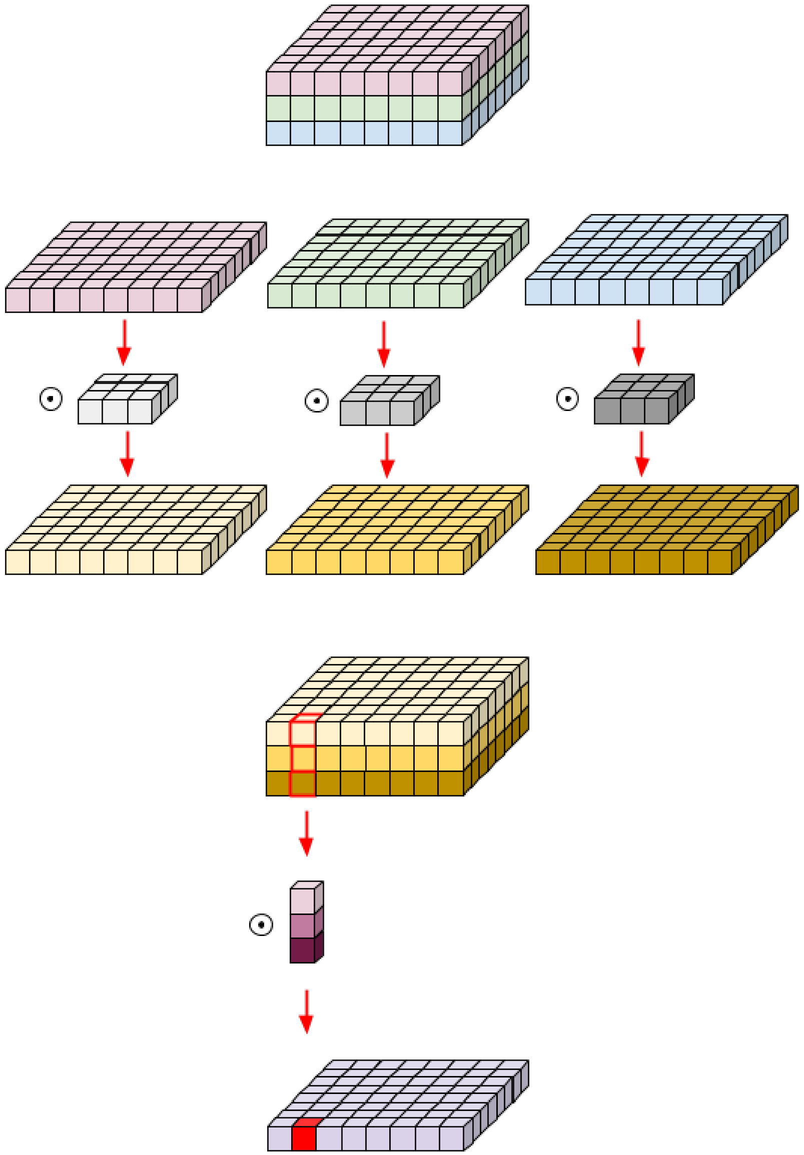
Depthwise Conv

- Each channel has its own kernel
  - Number of parameters reduced
  - Number of operations reduced
- What could be the limitation?
  - No information exchange across different channels



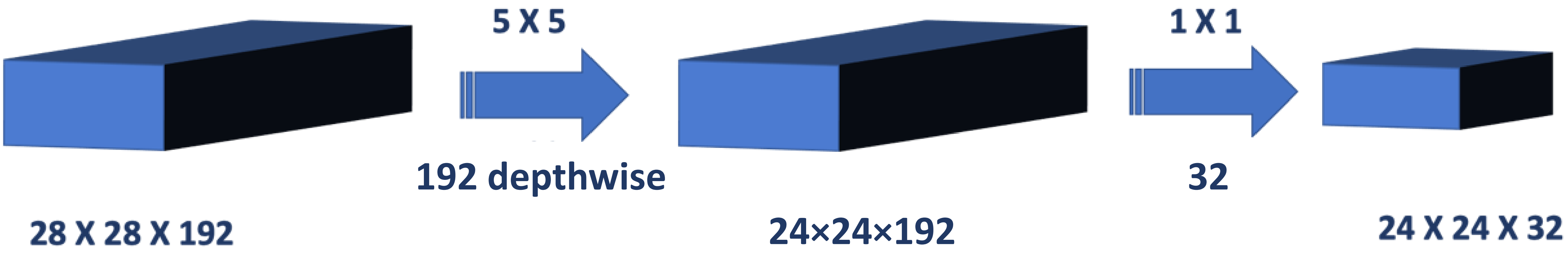
Depthwise Conv

- Combine depthwise conv with 1×1 conv



Depthwise Conv

- Example



Operations: $(5\times5)\times(24\times24)\times192+(1\times1\times192)\times(24\times24)\times32=2.8M+3.5M=6.3M$
Parameters: $5\times5\times192+1\times1\times192\times32=4.8K+6.1K=10.9K$

Depthwise Conv

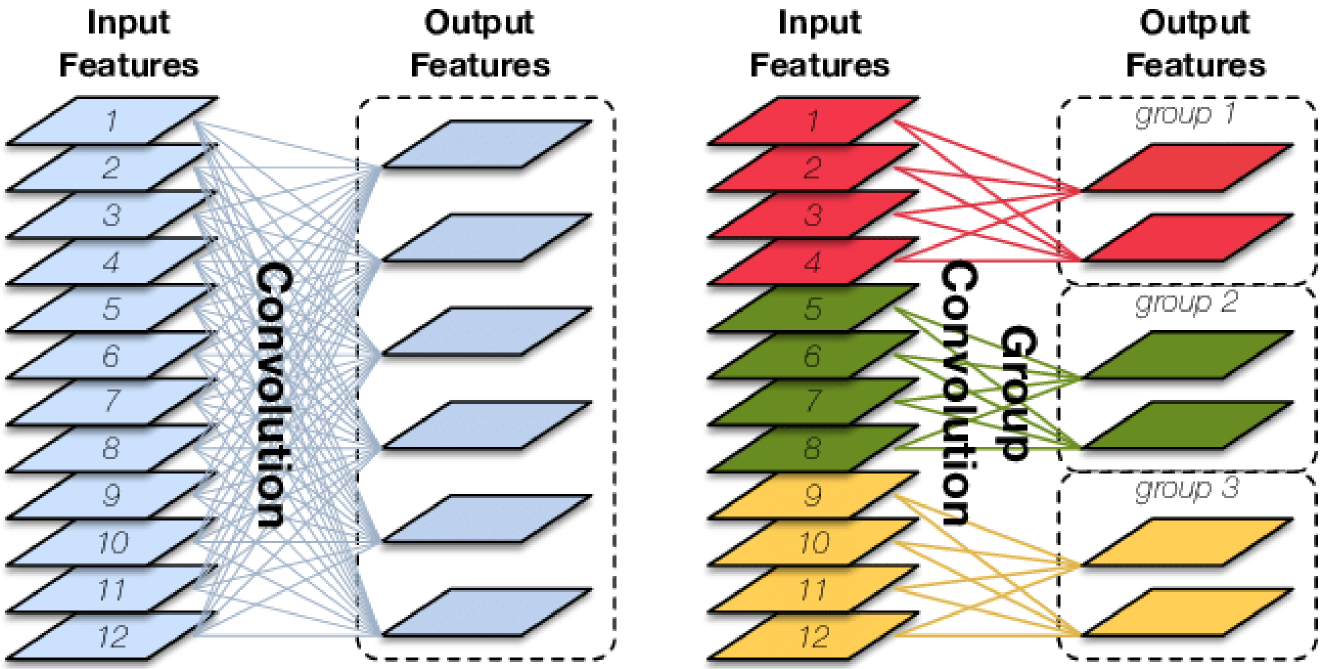
- MobileNet on ImageNet

Table 8. MobileNet Comparison to Popular Models

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

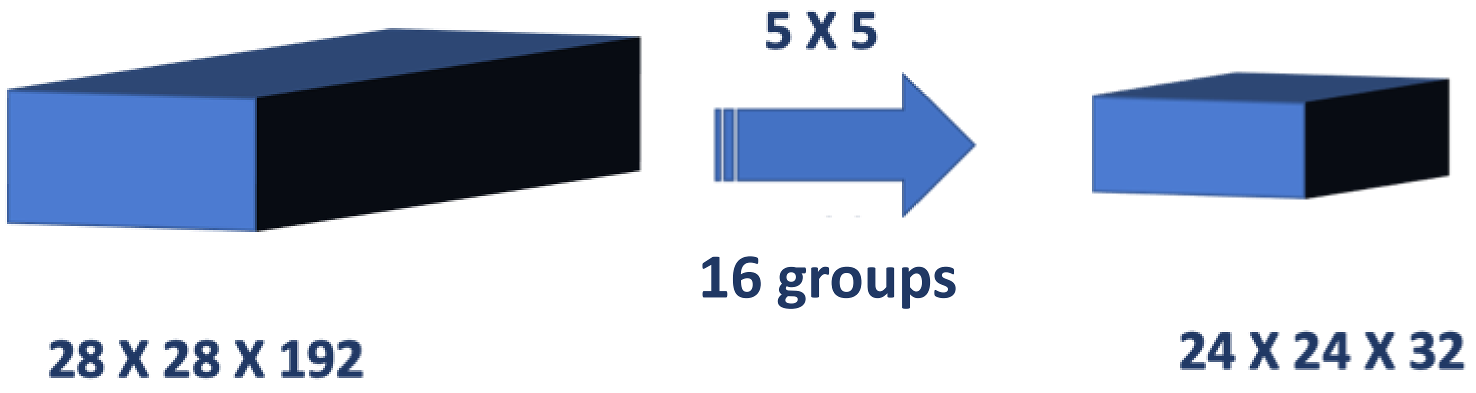
Group Conv

- Split channels into groups
  - Reduce the number of parameters
  - Usually still need 1×1 conv afterwards



Group Conv

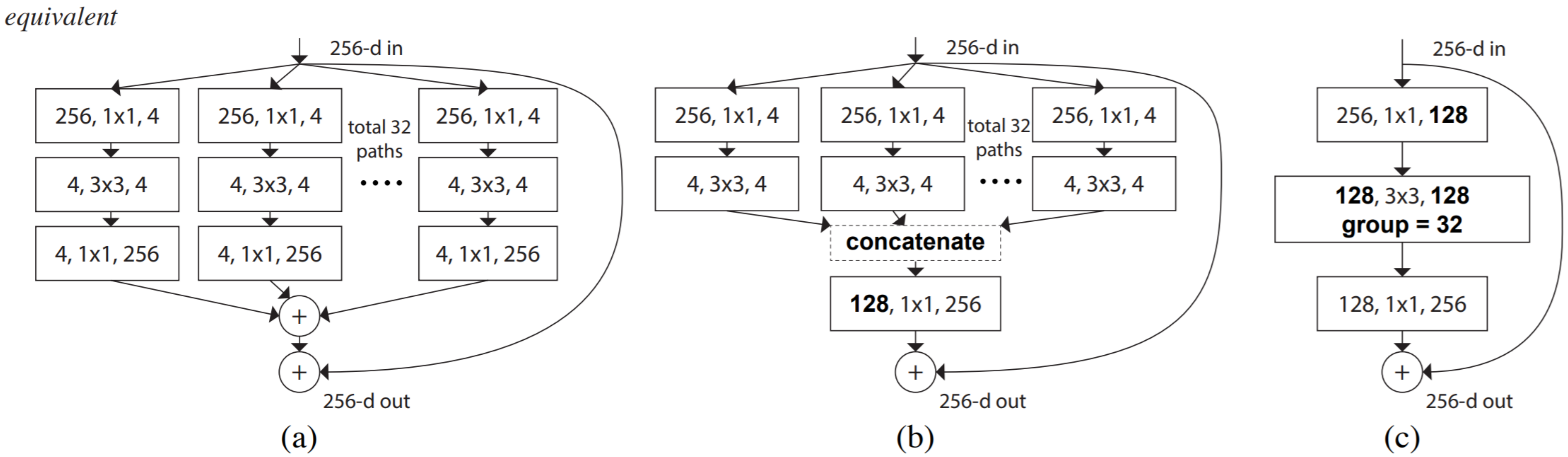
- Example



Each group has 12 input channels and 2 output channels
Operations: (5×5×12)×(24×24)×32=5.5M
Parameters: 5×5×12×2×16=9.6K

Group Conv

- ResNeXt



CV Related Topics

- Image Classification
- Detection and Tracking
- Pose Estimation
- Segmentation
- 3D and Localization
- Image Reconstruction
- …

Image Classification

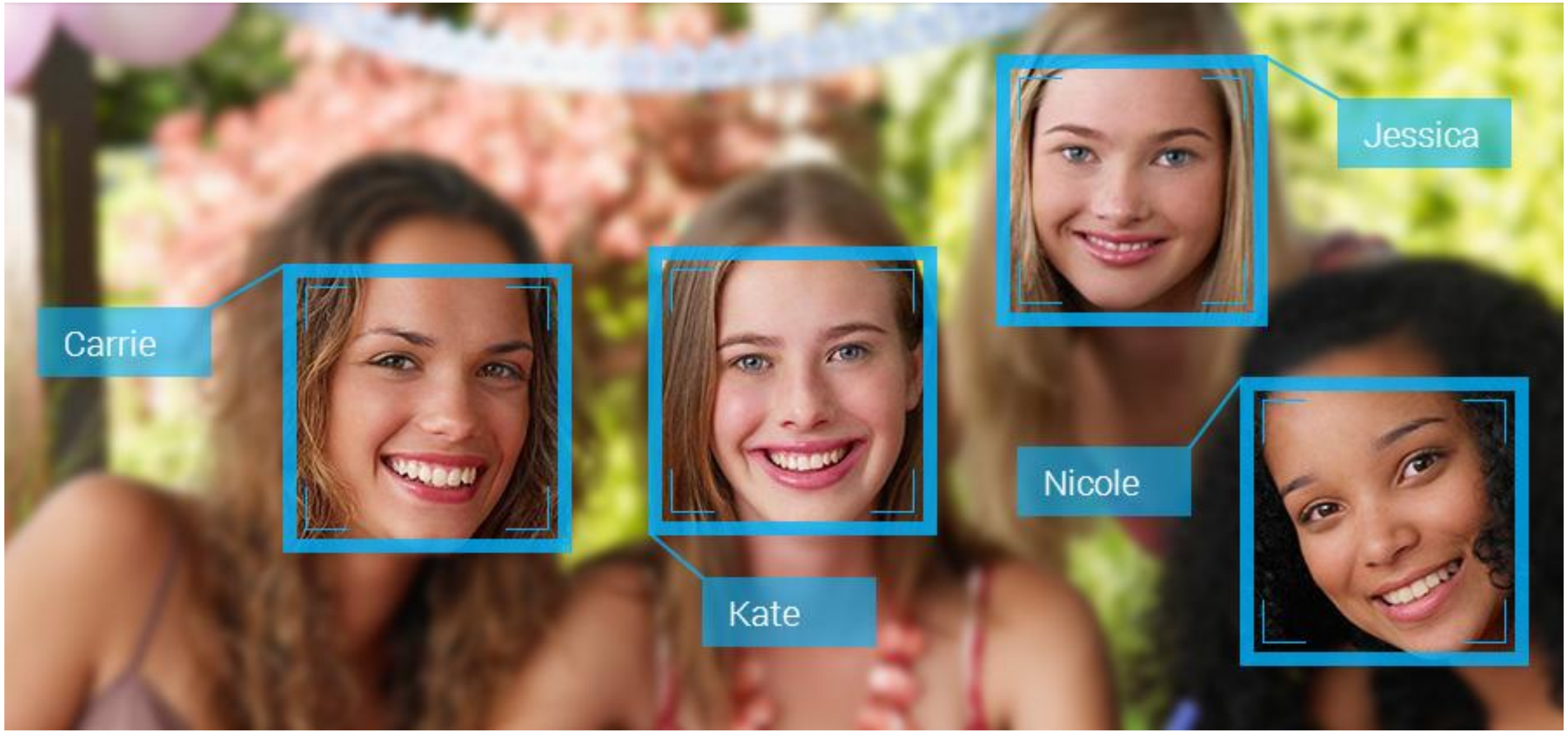
- Fine-grained image classification
- Face recognition
- Face verification

Fine-Grained Image Classification

- Classify sub-categories

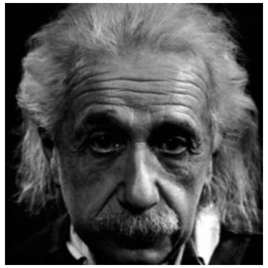
Face Recognition

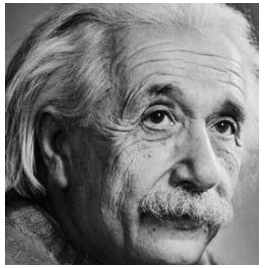
- Identify different faces



Face Verification

- Verify whether two faces from the same person





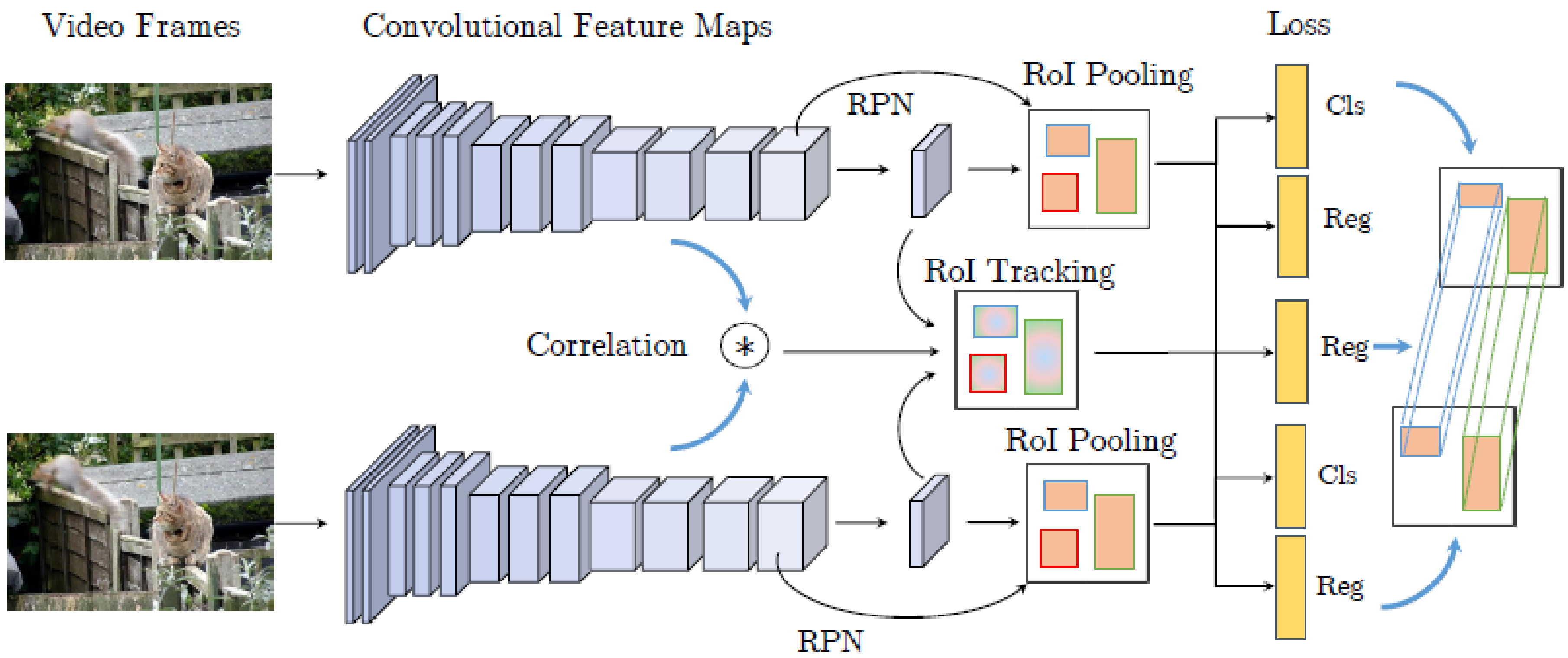
Are they from the same person?

Detection and Tracking

- Video object detection
- 3D object detection
- Visual tracking
- Multi-object tracking

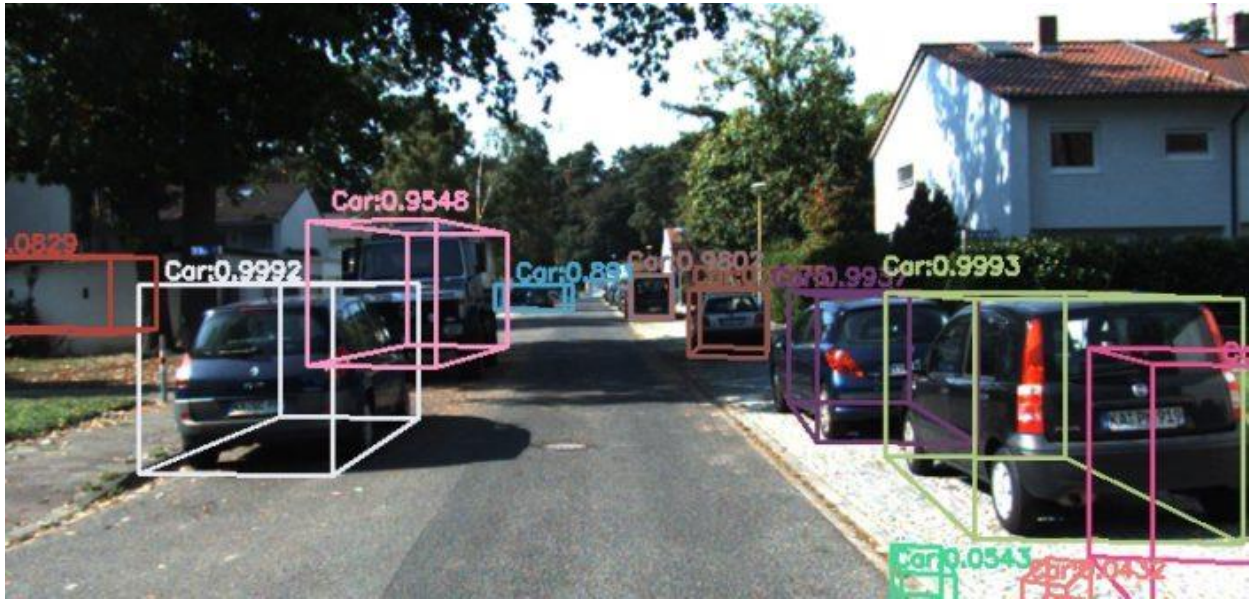
Video Object Detection

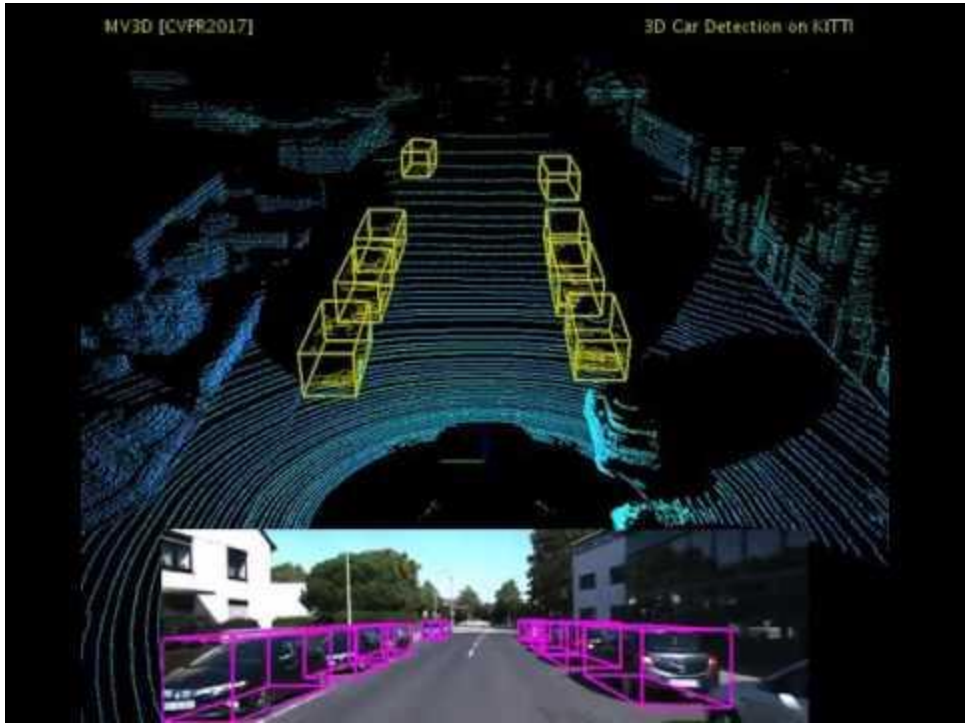
- Detect objects in sequential frames



3D Object Detection

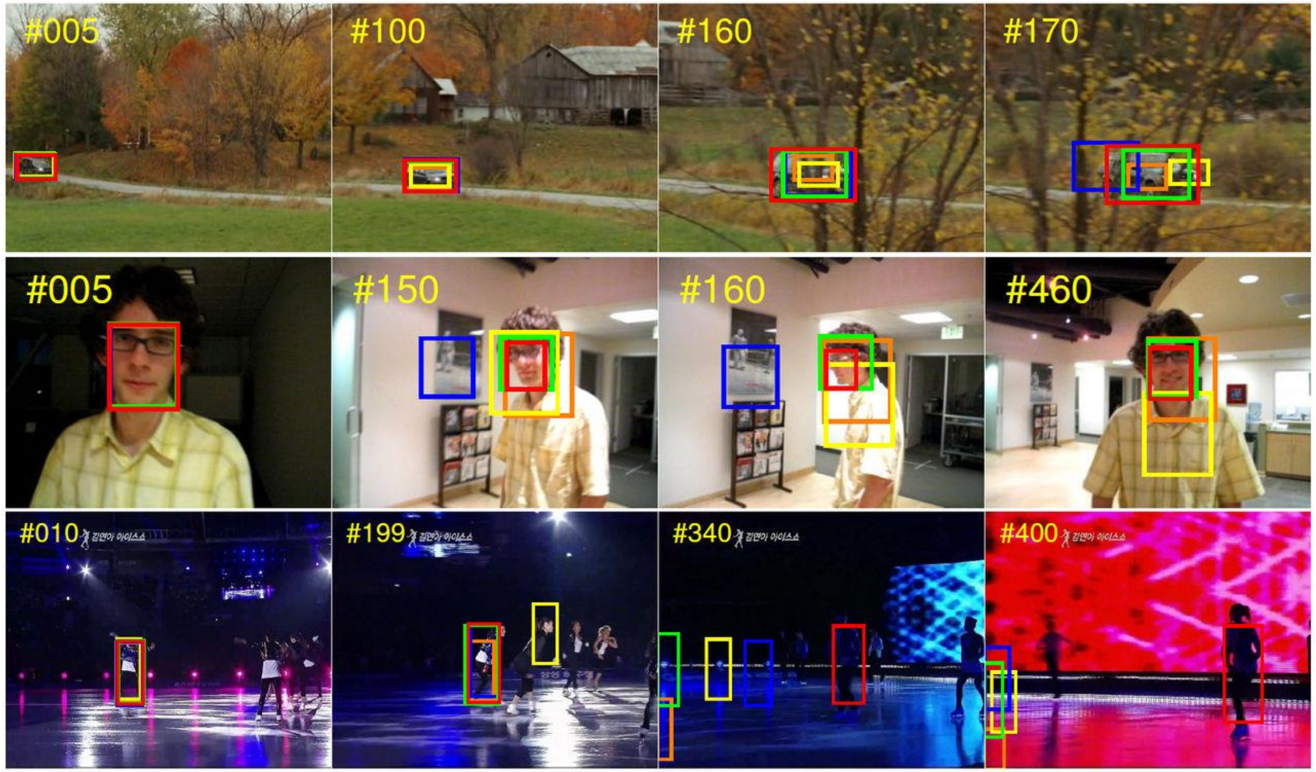
- Localize the 3D shape or position of the targets





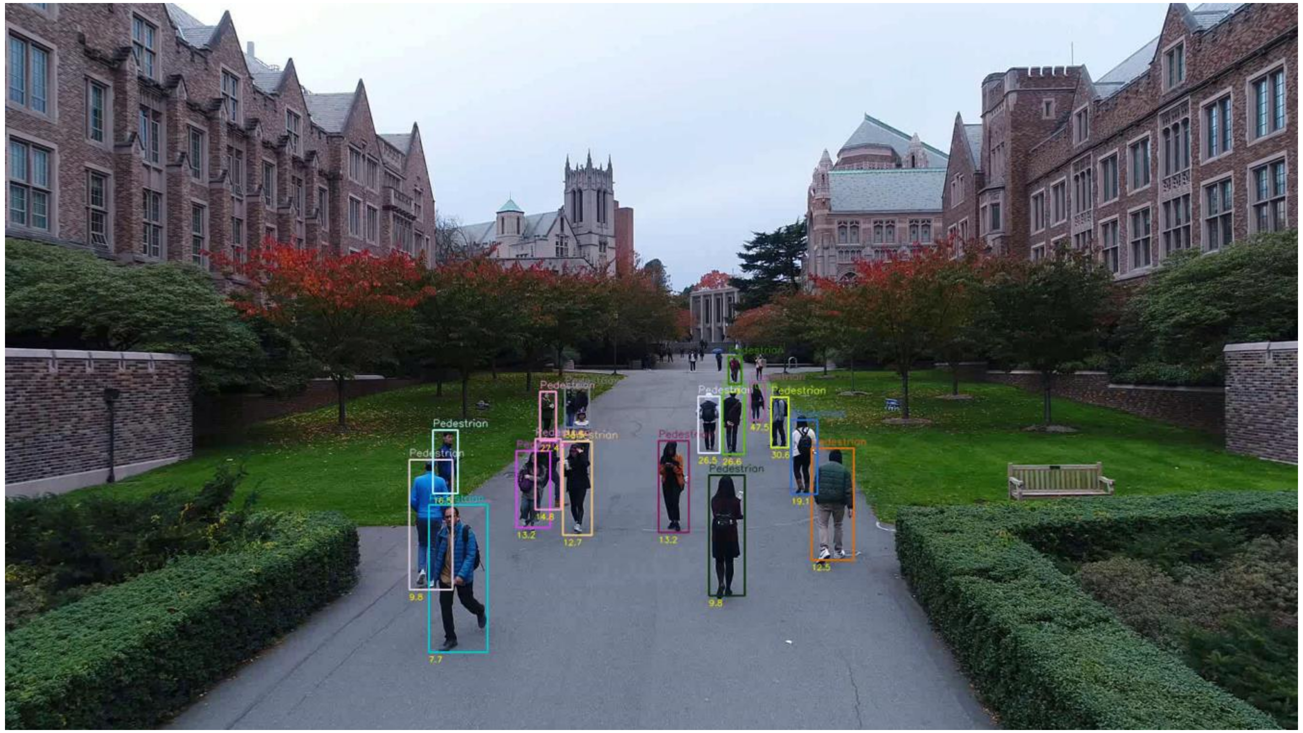
Visual Tracking

- Given the annotation of the first frame, detect the same object in following frames



Multi-Object Tracking

- Associate detected objects in the input video

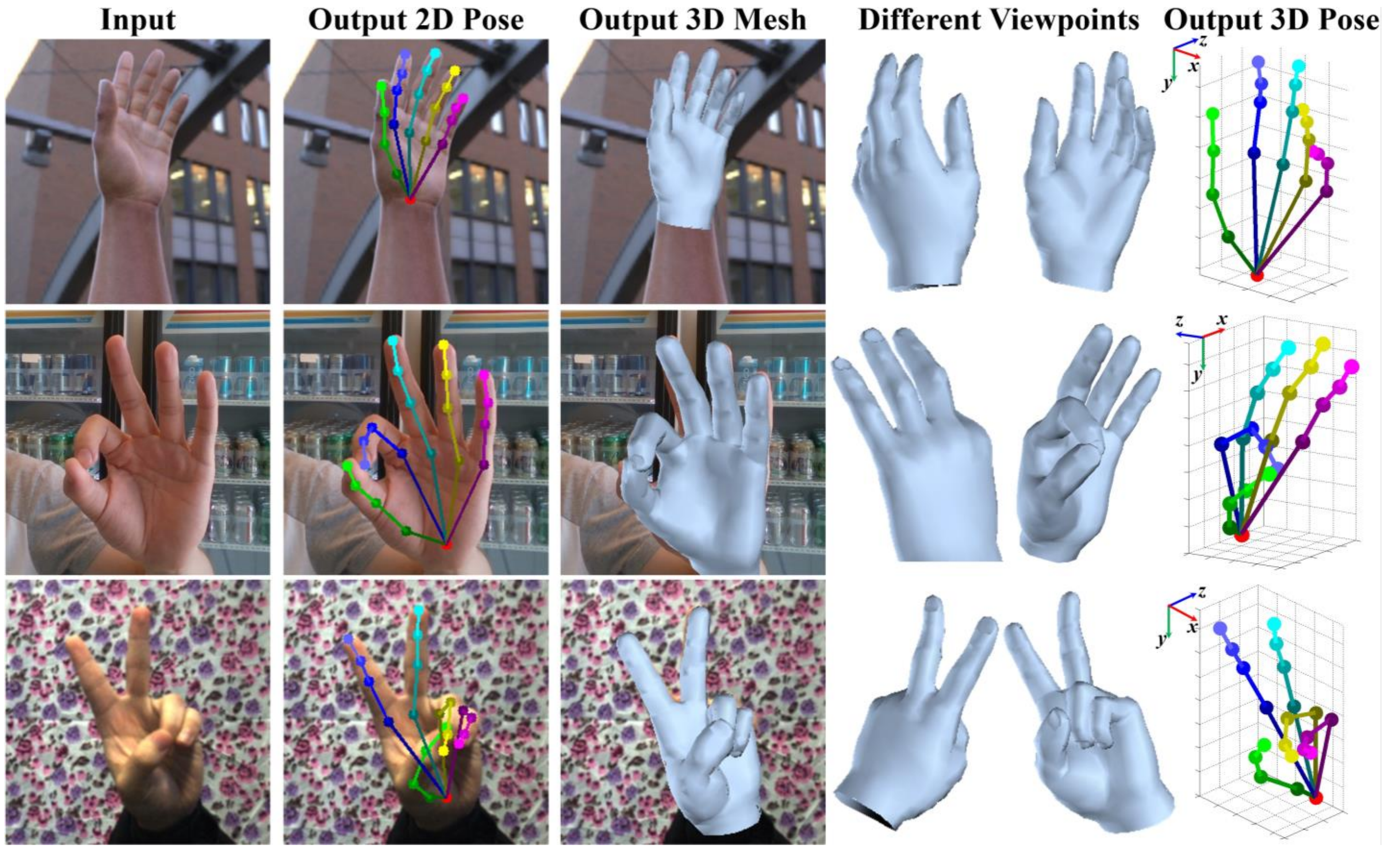


Pose Estimation

- Hand pose estimation
- Human pose estimation
- Car keypoint detection

Hand Pose Estimation

- Estimate 2D/3D hand pose from RGB image or depth image



Human Pose Estimation

- Estimate 2D/3D human pose

Car Keypoint Estimation

- Estimate 2D/3D car keypoint

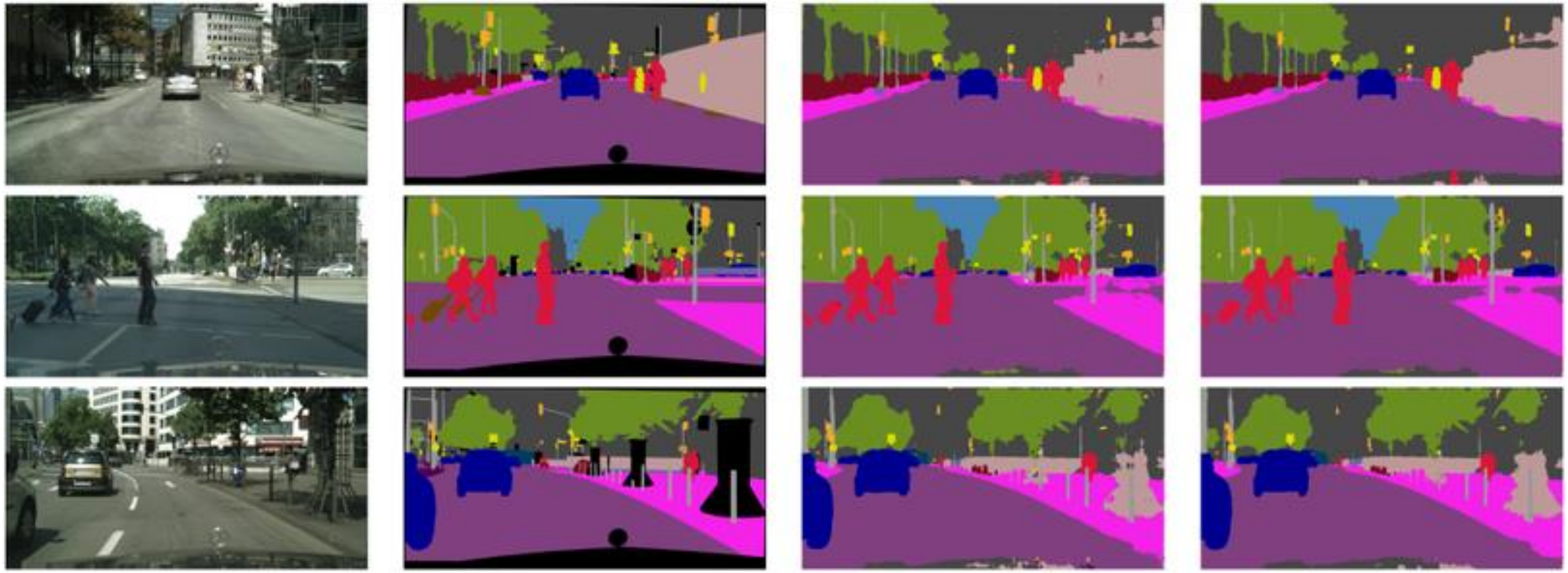
(a) Heatmaps **(b) Keypoints** **(c) Projections** **(d) 3D space**

Segmentation

- Semantic segmentation
- Instance segmentation
- Video object segmentation

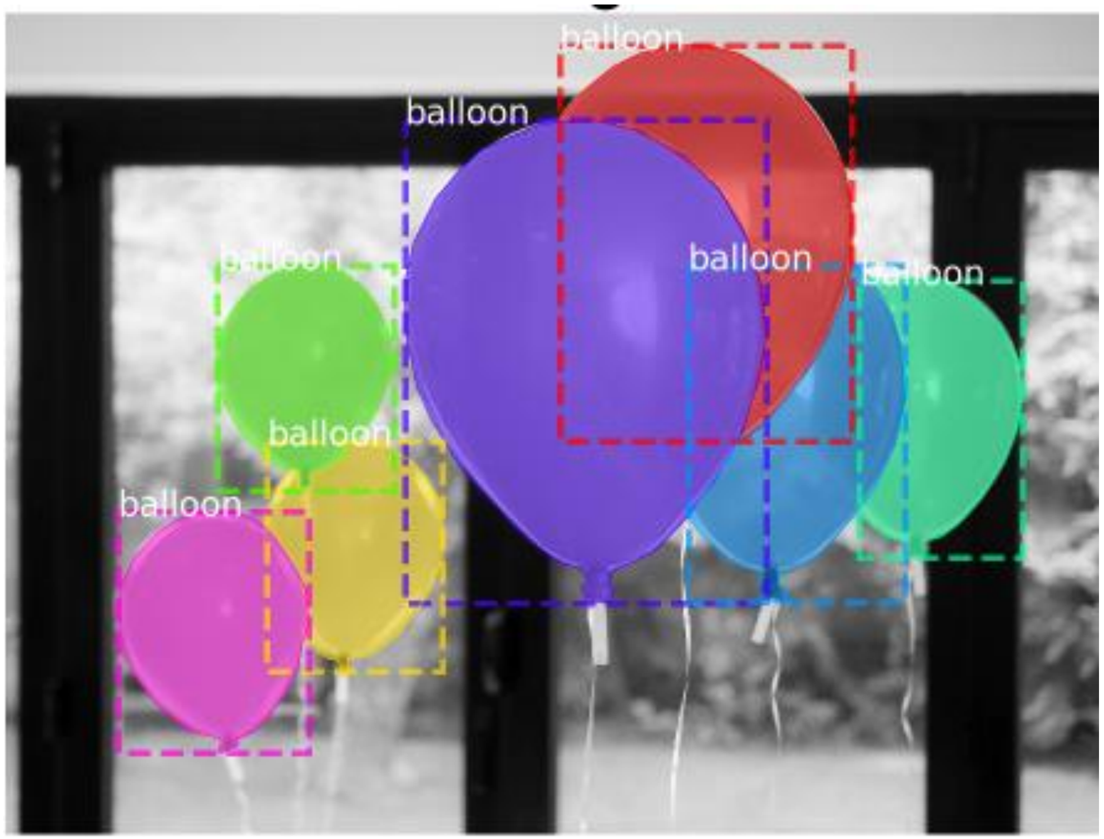
Semantic Segmentation

- Segment objects with class labels



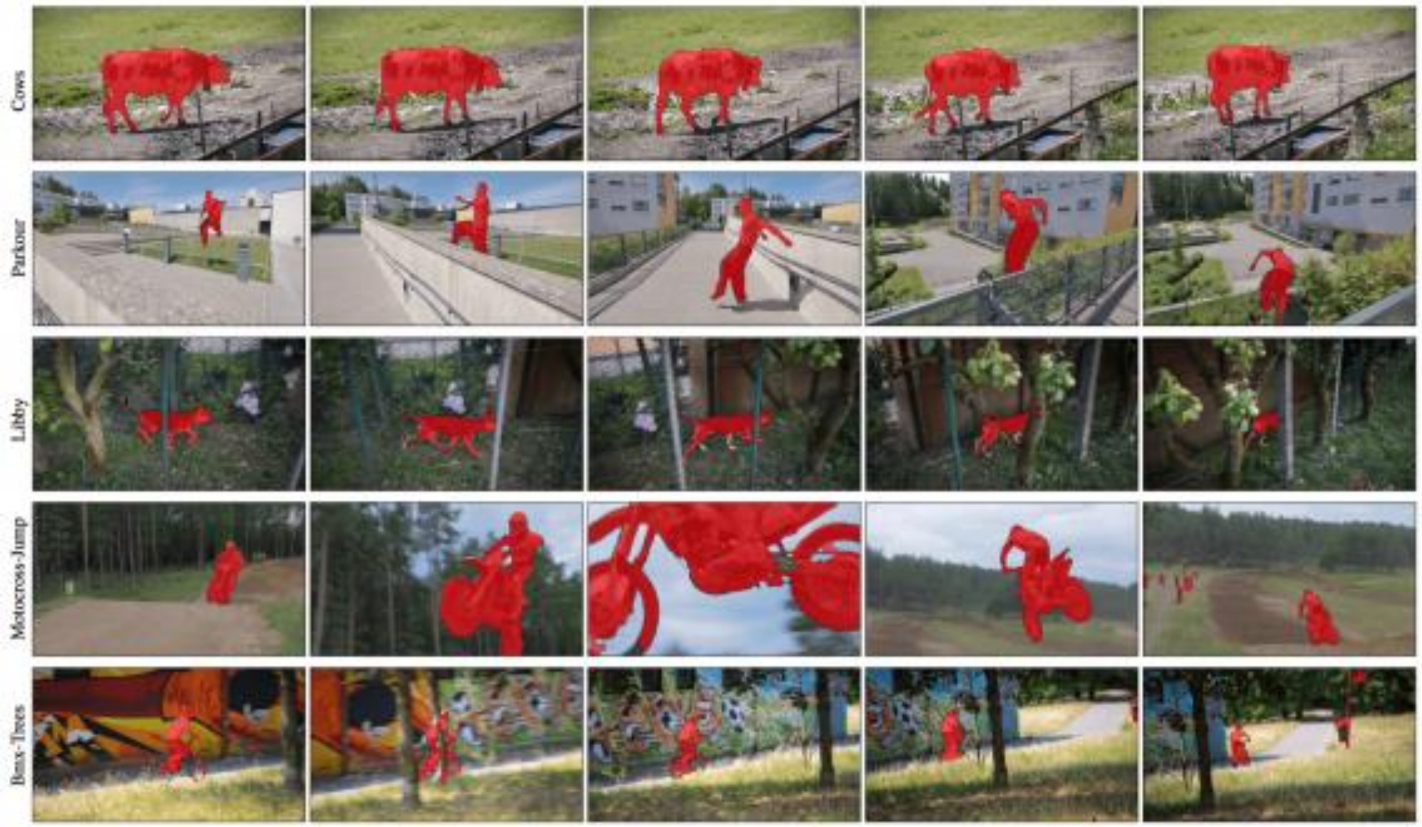
Instance Segmentation

- Segment objects with instance labels



Video Object Segmentation

- Segment objects in the video sequence

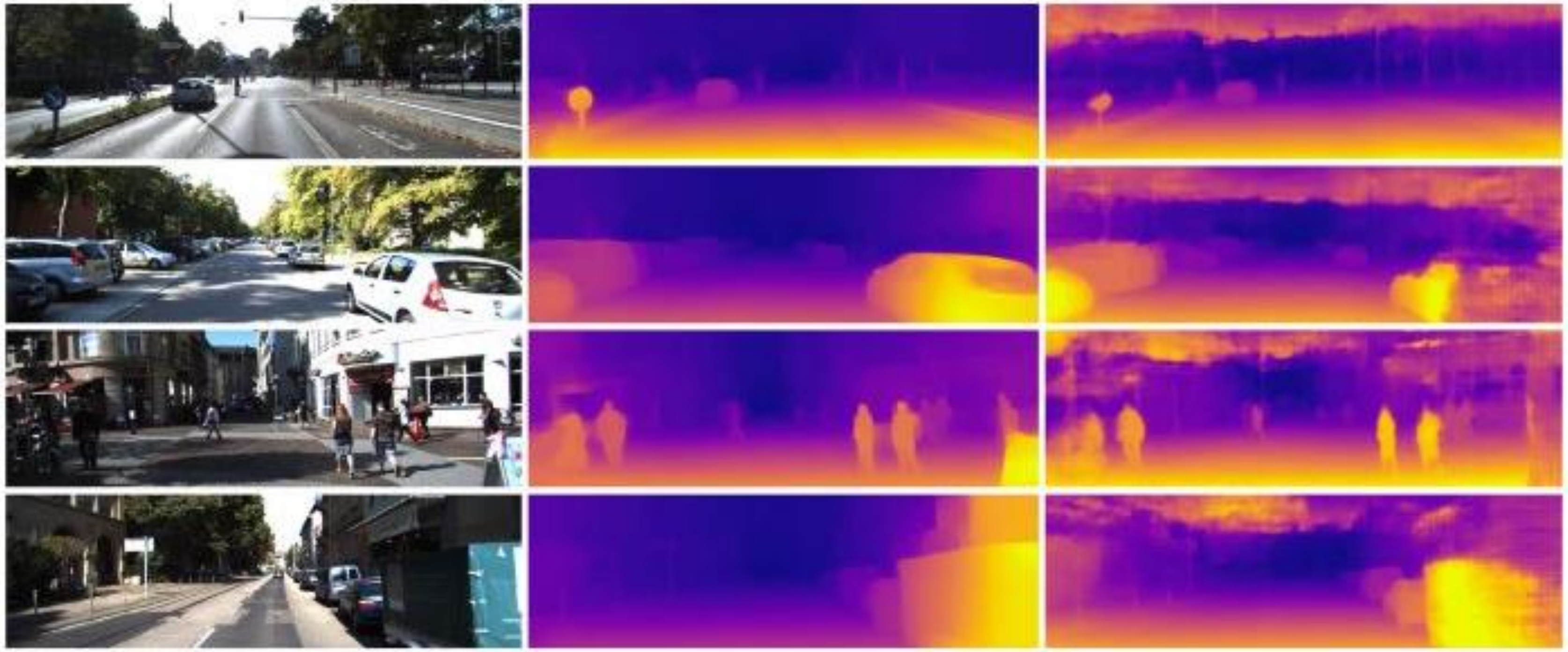


3D and Localization

- Depth map estimation
- Optical / scene flow estimation
- Camera pose estimation

Depth Map Estimation

- Estimate depth map from RGB images



Optical / Scene Flow Estimation

- Estimate the 2D/3D offsets between two images

(a) Color image 1

(b) Color image 2

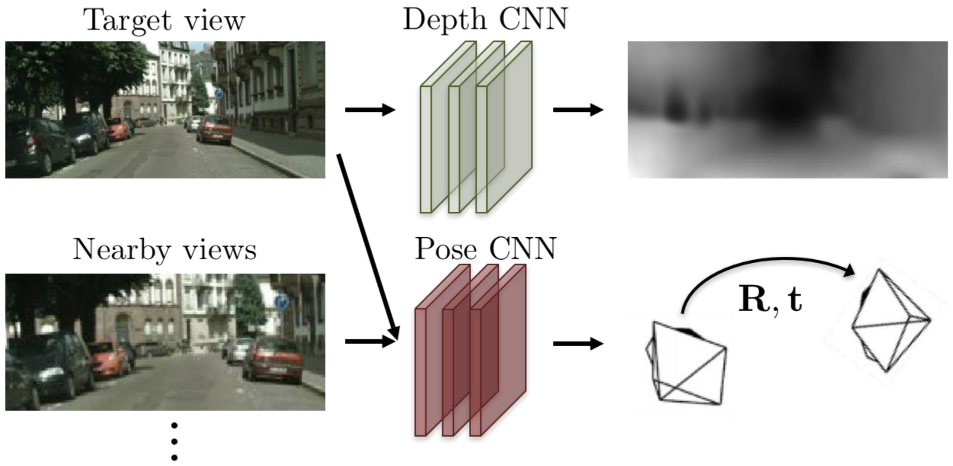
(c) Ground truth flow map

(d) Flow map (Ours-S-600k)

Camera Pose Estimation

- Estimate camera location and orientation based on sequential frames

(a) Training: unlabeled video clips.



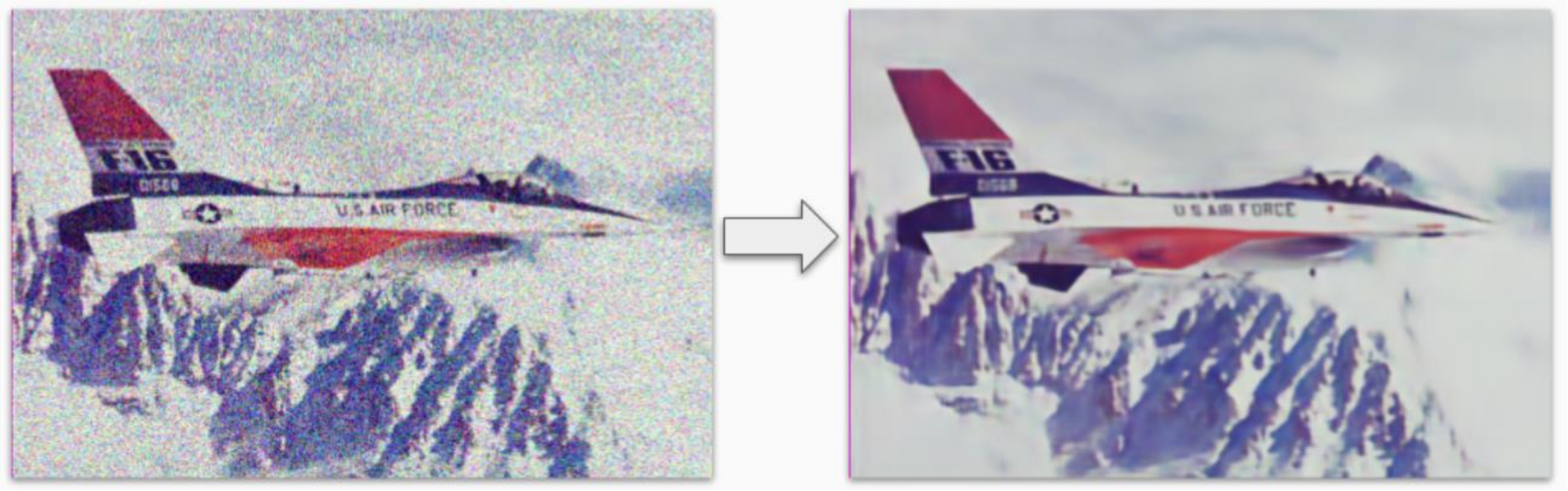
(b) Testing: single-view depth and multi-view pose estimation.

Image Reconstruction

- Image denoising
- Super-resolution
- Image inpainting

Image Denoising

- Reconstruct images with noise



Super-Resolution

- Reconstruct high resolution images from low resolution images

Image Inpainting

- Recover missing regions in the image