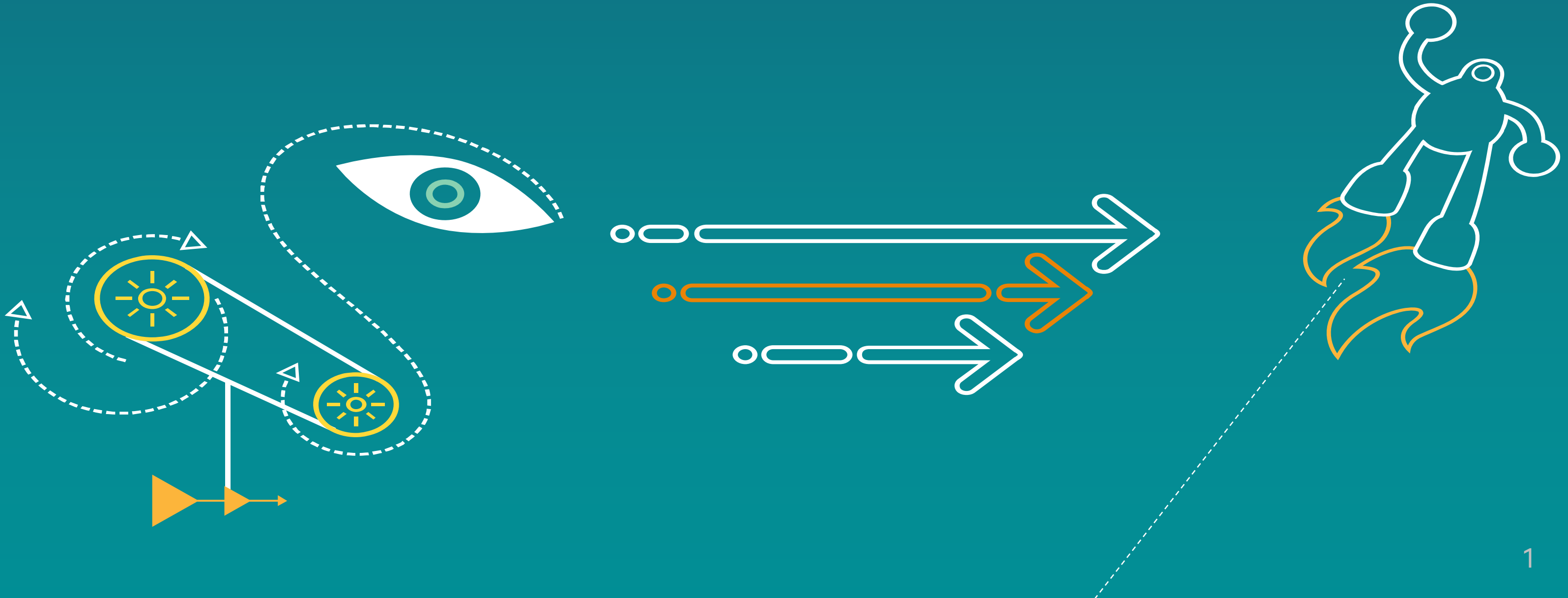
Koen E. A. van de Sande (Engineer, Staff, QTI), **Daniel Fontijne** (Engineer, Senior Staff, QTI), **Anthony Sarah** (Engineer, Sr. Staff/Manager, QTI), **Harro Stokman** (Director, Product Management, QTI), **R. Blythe Towal** (Engineer, Staff, QTI), **Cees G. M. Snoek** (Engineer, Principal, QTI)

Qualcomm Research Deep Net for Video Concept Localization

QUALCOMM

QUALCOMM TECHNOLOGIES, INC.

November 16, 2015



Overall summary

Run	Iframe F-score	Iframe Precision	Iframe Recall	Pixel F-score	Pixel Precision	Pixel Recall	Total
Gamora	1 concept	5 concepts		5 concepts	6 concepts	2 concepts	**19 concepts**
Rocket	1 concept						**1 concept**
Starlord							
Groot	2 concepts		1 concept				**3 concepts**

‘Gamora’ is best approach in 19 out of 60 possible comparisons

Inspiration from ImageNet

Box proposals with deep convolutional network features

FLAIR

Selective search

PCA-reduced Color SIFT

Fisher vectors

Spatial pyramid

Linear SVM

Hard negative mining

vd Sande et al. CVPR 2014

R-CNN

Selective search

Features from AlexNet

Features from VGGNet

Pre-train on 1,000 ImageNet categories

Linear SVM

Hard negative mining

Girschik et al. PAMI 2015

Multibox

Inception network for box proposals

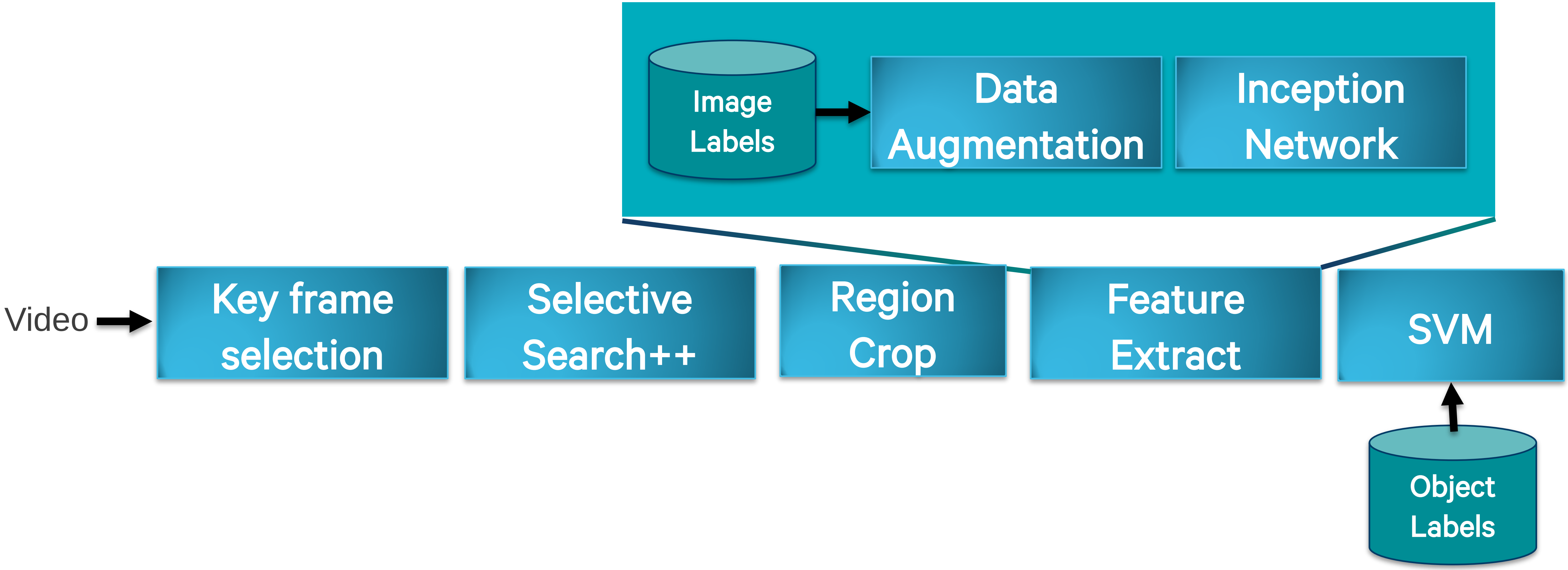
Features from Inception network

Pre-train on 1,000 ImageNet categories

Szegedy et al. CVPR 2015

Approach

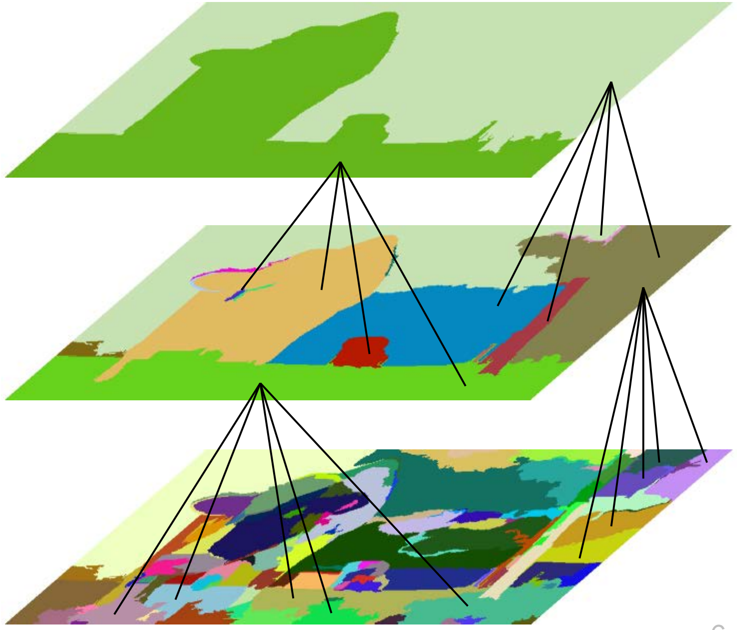
High-level overview of training system



vd Sande et al. ICCV 2011

Selective Search provides the box proposals

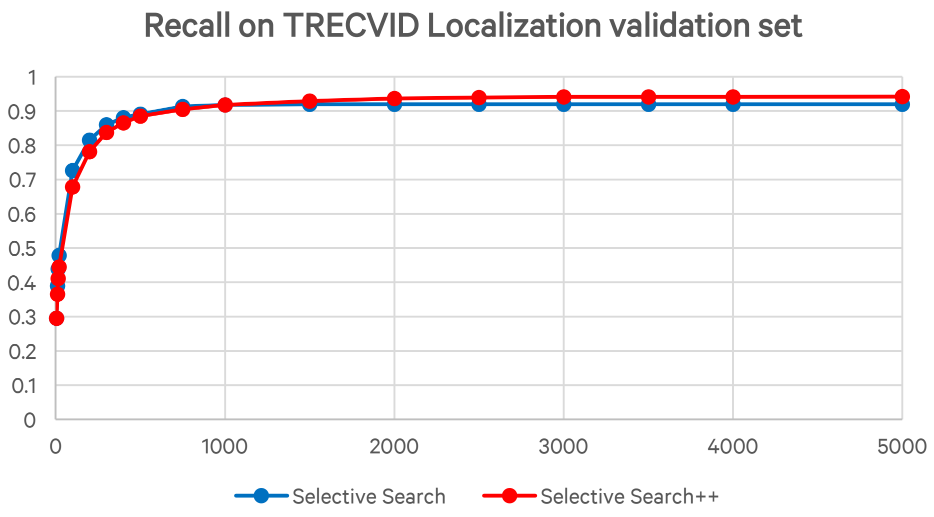
- Hierarchical segmentation of video frames based on low level features
- Merge adjacent superpixels based on a set of region similarity criteria
- Known to provide high-recall with a limited number of boxes
- Used by many groups on detection challenges

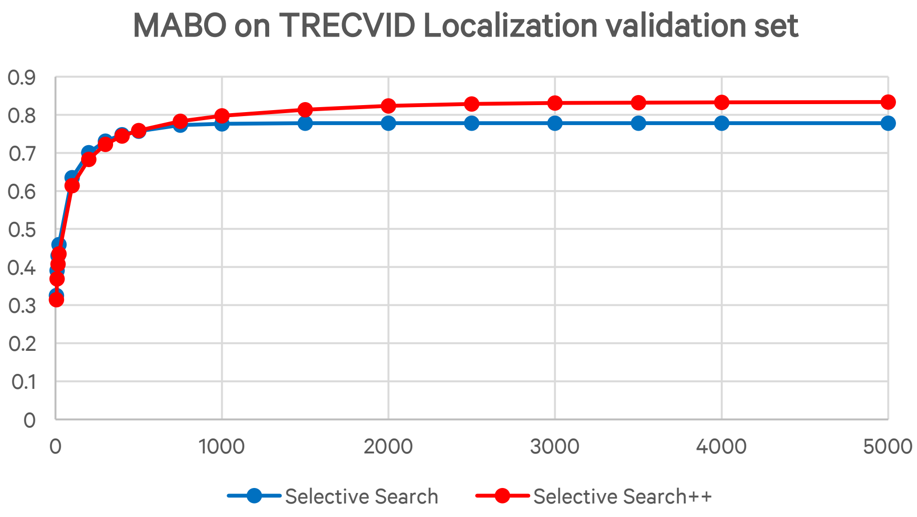


Code available at: http://koen.me/research/selectivesearch/

Selective Search++

- More region similarity criteria used and different thresholds for merging
- Higher recall for 1000+ boxes per image
- Higher Mean Average Best Overlap (MABO) for 500+ boxes





Szegedy et al. CVPR 2015

Feature extraction by Inception-style network

- Small 1x1 convolutions
- Convolution stride of two or one
- ReLU non-linearity
- Four max-pool layers
- Alex-style fully connected head
- Dropout
- Nine inception modules
- Batch normalization
- We rely on two best models from SIN task

Image labels

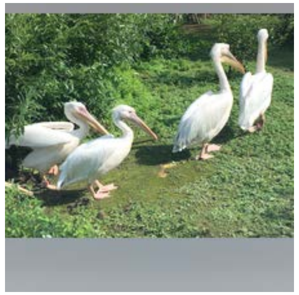
- All models are pre-trained on ImageNet
  - 1,000 standard ImagNet categories
  - 2,048: 1,024 categories better matching video concepts, plus 1,024 random categories
  - 4,096 same as above, plus more random categories

Wu et al., arXiv:1501.02876v4

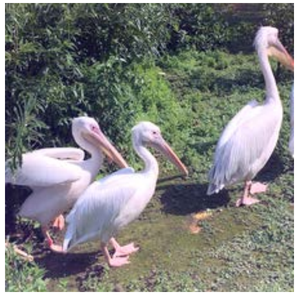
Data augmentation

Adding color casting and vignetting to default translation and mirroring

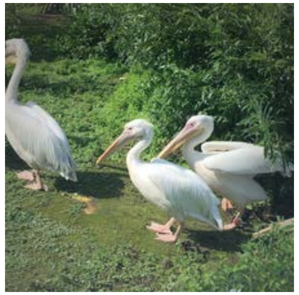
Original



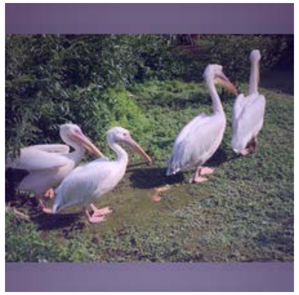
Translate/Mirroring



Color casting



Vignetting



All augmentations

- Random set of augmentations chosen for each image each time it is presented to the network for training

Object labels

Internal train and validation set used for our experiments

Object	Internal train set		Internal validation set	
	#positive images	#positive boxes	#positive images	#positive boxes
Airplane	1034	1545	183	248
Anchorperson	328	402	209	270
Boat/Ship	1132	1943	94	167
Bridges	993	1051	133	146
Bus	349	435	49	56
Computers	205	281	61	103
Flags	752	1061	370	88
Motorcycle	693	1097	66	95
Quadruped	1094	1483	384	55
Telephones	524	654	59	62
TOTAL	7104	9952	1608	1290

Predicting object labels

- Following the convention in the literature we train linear SVMs on the features from the classification models to classify boxes
  - Positive examples from object labels
  - Negative examples from random sampling of background regions

- We perform two rounds of hard negative mining

Fusion

- Our models exploit diversity in image labels
- We have two models available for non-weighted late fusion

Experiments

Value of deep learning features

Feature	mAP
Color Fisher with FLAIR	26.5
AlexNet trained on 1,000 ImageNet categories	29.9
Qualcomm network trained on 1,000 ImageNet categories	37.3

Qualcomm deep learning features much better than AlexNet

Value of image labels

Feature	mAP
Qualcomm network trained on 1,000 ImageNet categories	37.3
Qualcomm network trained on 2,048 ImageNet categories	39.8
Qualcomm network trained on 4,096 ImageNet categories	40.3

Learning on more object categories results in stronger features

Value of selective search++

Feature	Selective Search (mAP)	Selective Search++ (mAP)
Qualcomm Network - 2,048	39.8	40.2
Qualcomm Network - 4,096	40.3	42.4

Selective search++ further improves concept localization accuracy

Value of fusion

Feature	Selective Search (mAP)	Selective Search++ (mAP)
Qualcomm Network - 2,048	39.8	40.2
Qualcomm Network - 4,096	40.3	42.4
Qualcomm Network - 2,048 & 4,096	43.7	45.3

Fusion of our best two individual models provides another gain

Submissions

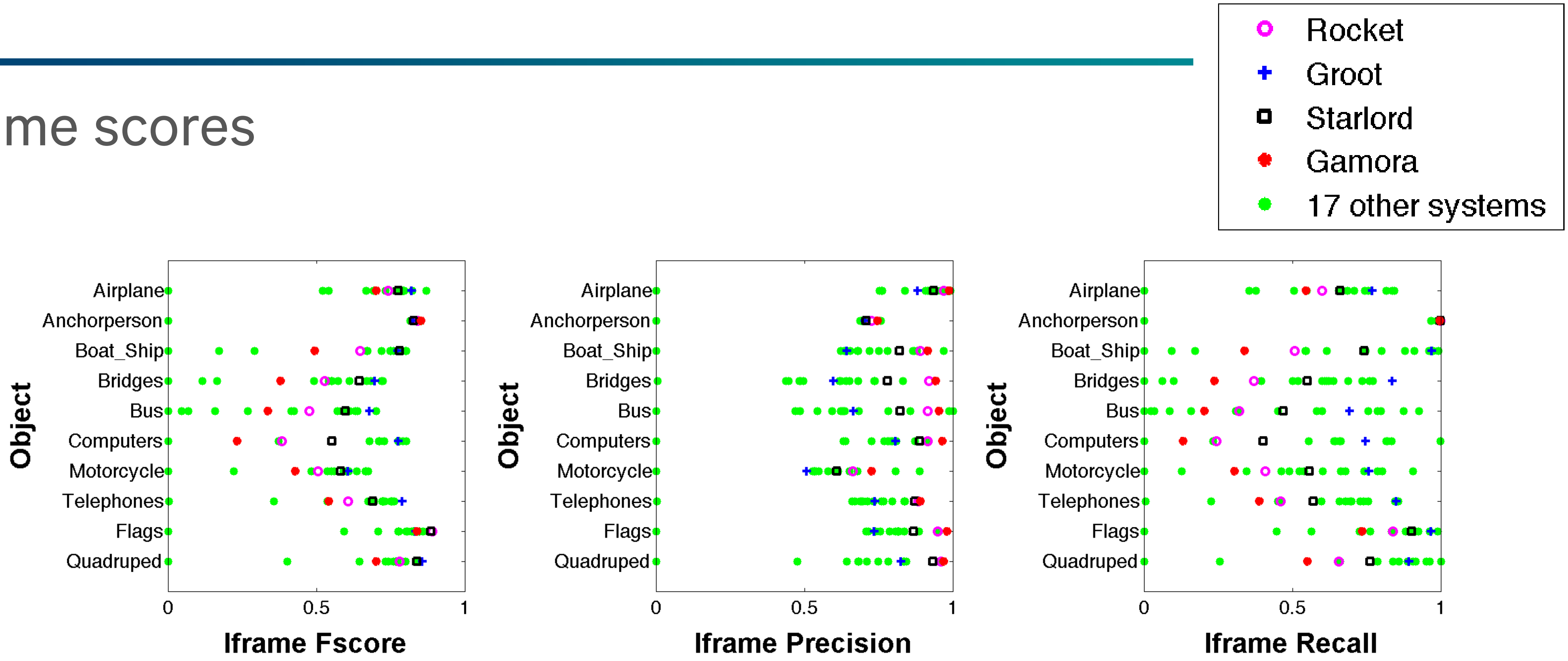
Overview of runs on internal validation set

Run	Threshold	Max # of boxes	Recall	Precision	F-score	mAP
Gamora	0.5	1	34%	55%	0.42	30.9
Rocket	0.0	1	41%	42%	0.41	35.0
Starlord	-0.5	1	47%	24%	0.32	38.1
Groot	-1.1	3	64%	7%	0.12	43.5

All our runs based on the same set of boxes and confidences

Different choices aim to optimize either precision, recall or balance both

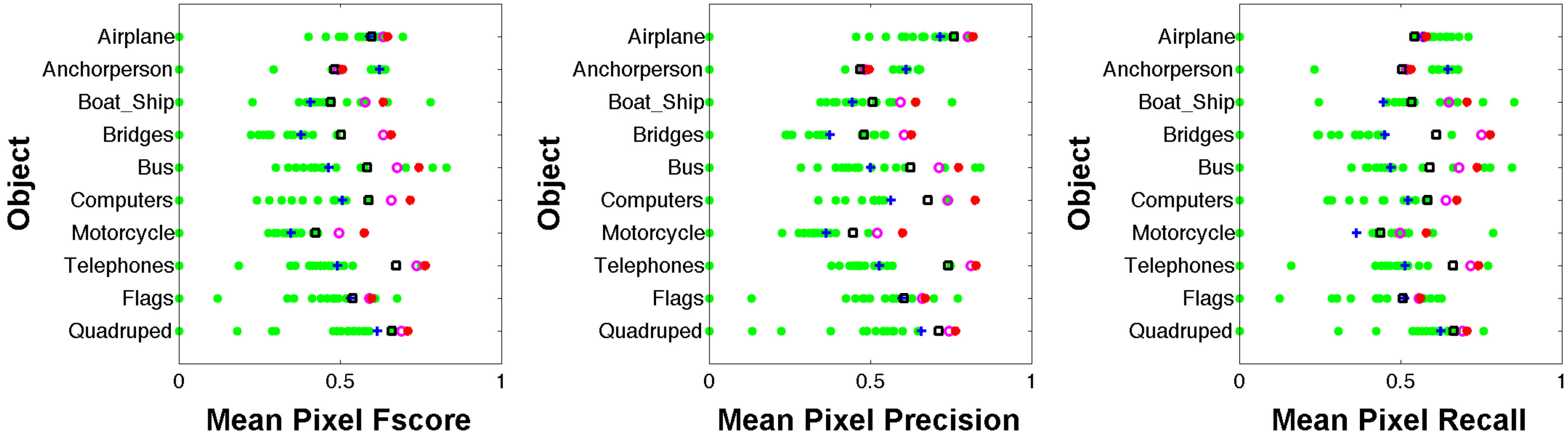
I-frame scores



High-recall run 'Groot' is penalized for predicting more than one box

High-precision run 'Gamora' is more likely to localize the object

Pixel scores



'Rocket' is meant to balance precision and recall, but is almost always outperformed by 'Gamora'

Overall summary

Run	Iframe F-score	Iframe Precision	Iframe Recall	Pixel F-score	Pixel Precision	Pixel Recall	Total
Gamora	1 concept	5 concepts		5 concepts	6 concepts	2 concepts	**19 concepts**
Rocket	1 concept						**1 concept**
Starlord							
Groot	2 concepts		1 concept				**3 concepts**

'Gamora' is best approach in 19 out of 60 possible comparisons

Conclusions

- High-recall box proposals and deep learned features powerful combination
- Advantageous to pre-train representation on more object categories

Thank you

Follow us on: f 🐦 in

For more information on Qualcomm, visit us at:
www.qualcomm.com & www.qualcomm.com/blog

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable.

Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.