

Deep filter banks for texture recognition and segmentation

Mircea Cimpoi, University of Oxford

Subhransu Maji, UMASS Amherst

Andrea Vedaldi, University of Oxford



UNIVERSITY OF
OXFORD

Texture understanding

Indicator of materials properties, e.g. brick vs wooden



Complementary to shape



Correlated with identity but not the same



Kickstarted **orderless image representations** (e. g. Bag of words)

[Bajcsy et al. 73, Julesz 81, Ojala et al. 96, 02, Dana et al. 99, Leung and Malik 99, Varma and Zisserman 03, 05, Caputo et al. 05, Lazebnik et al. 05, 06, Timofte and Van Gool 12 Sharma et al. 12, Sifre and Mallat 13, Sharan et. al 09, 13]

Is there a relation between
texture representations
and deep convolutional neural networks?

Texture representations

Filters + histogramming

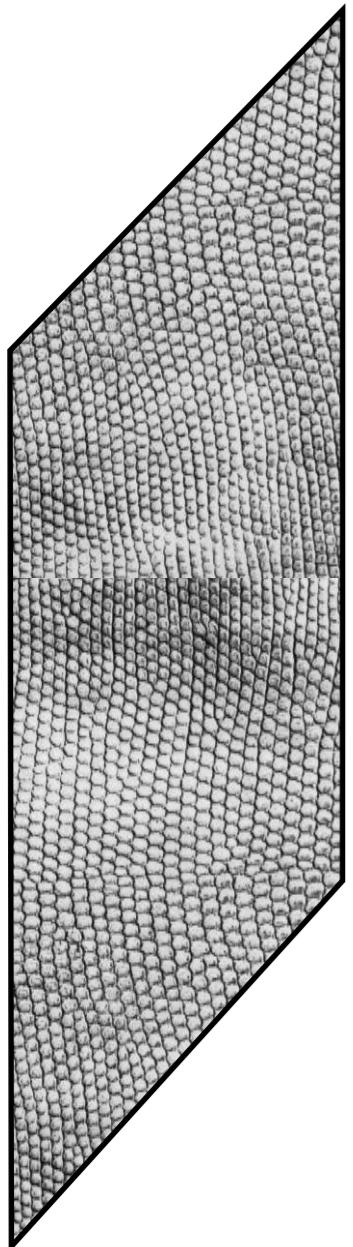


image x

[Leung and Malik 99, 01, Schmid 01, Varma and Zisserman 02, 05]

Texture representations

Filters + histogramming

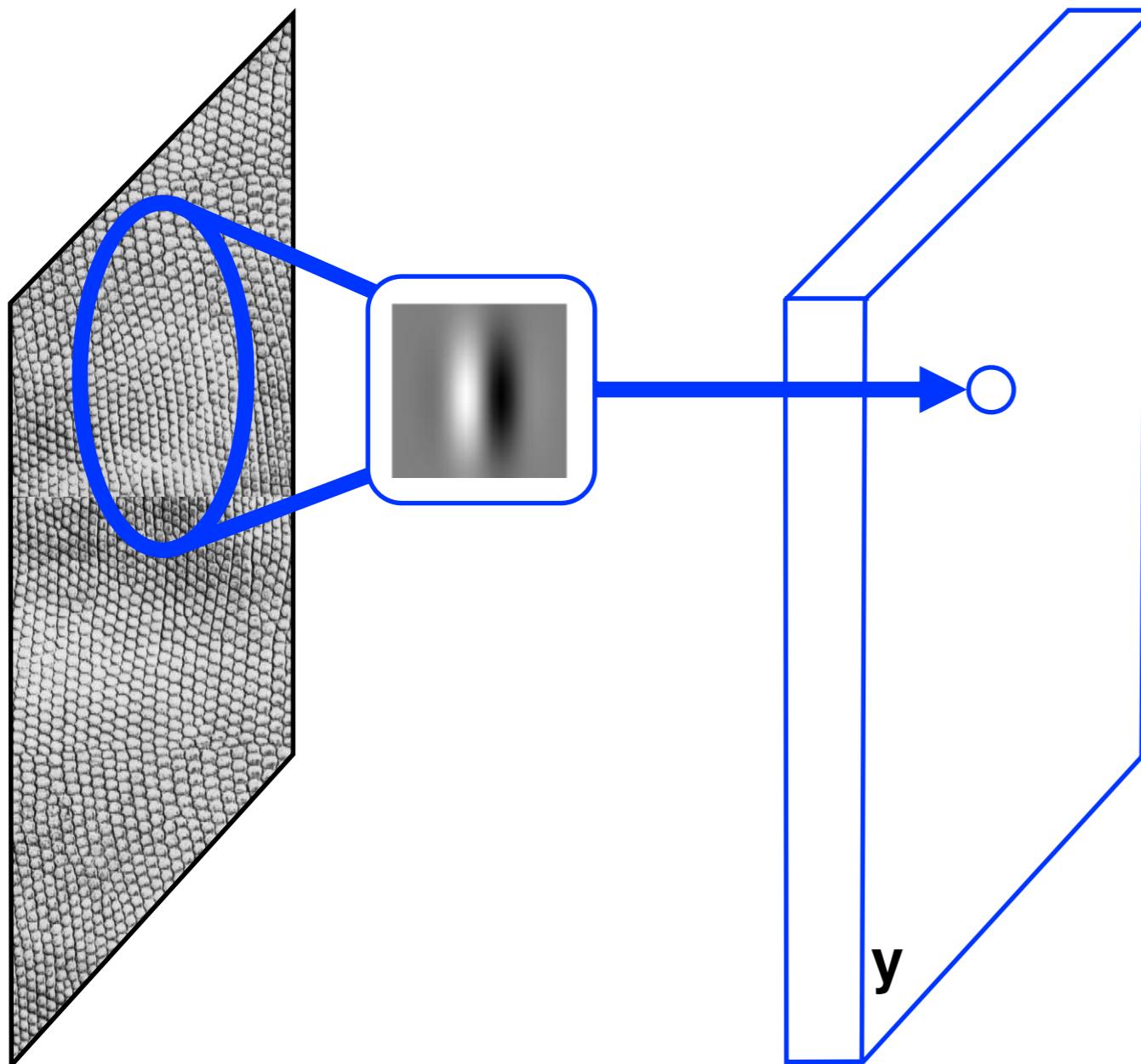


image x

Texture representations

Filters + histogramming

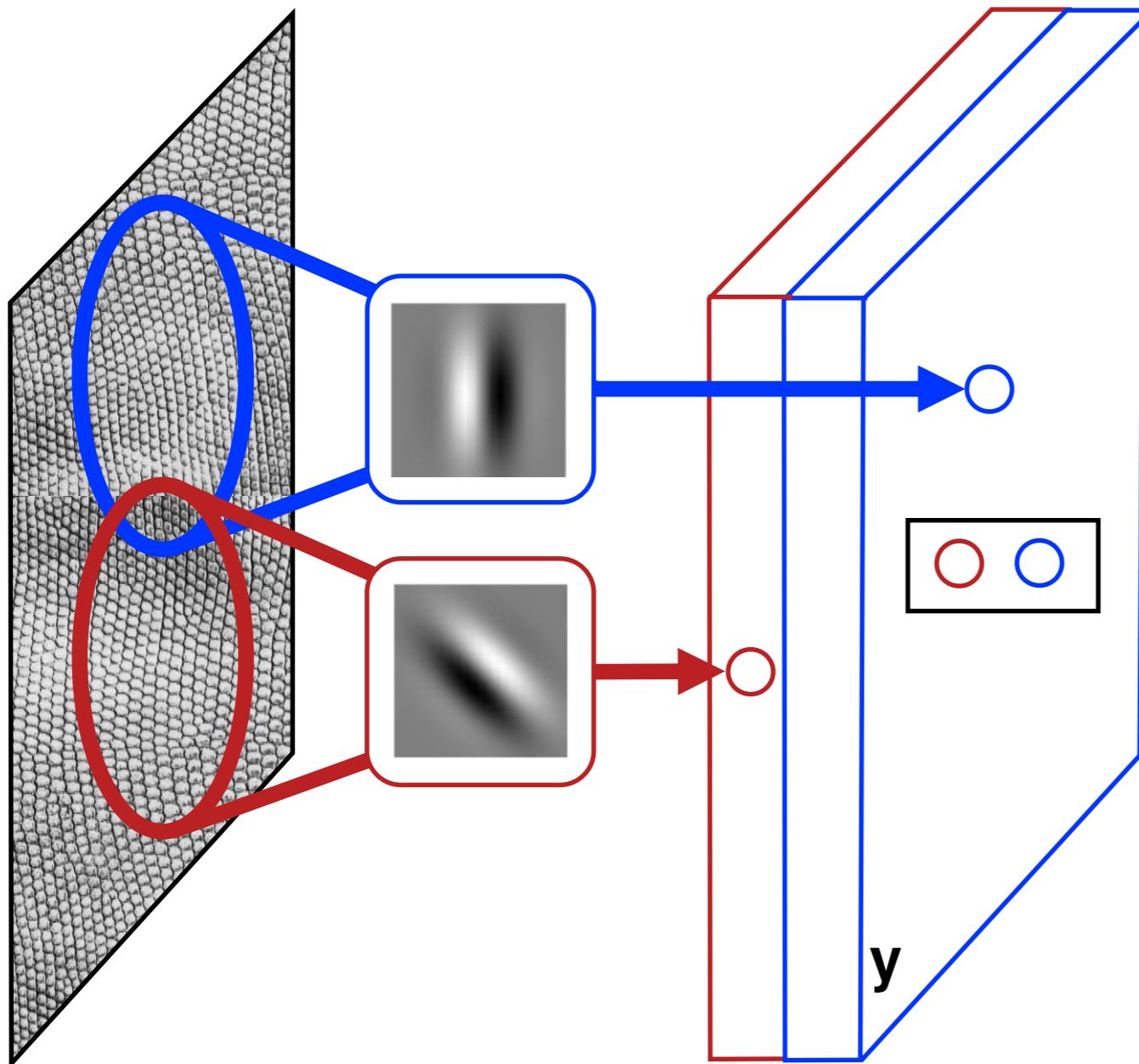


image x

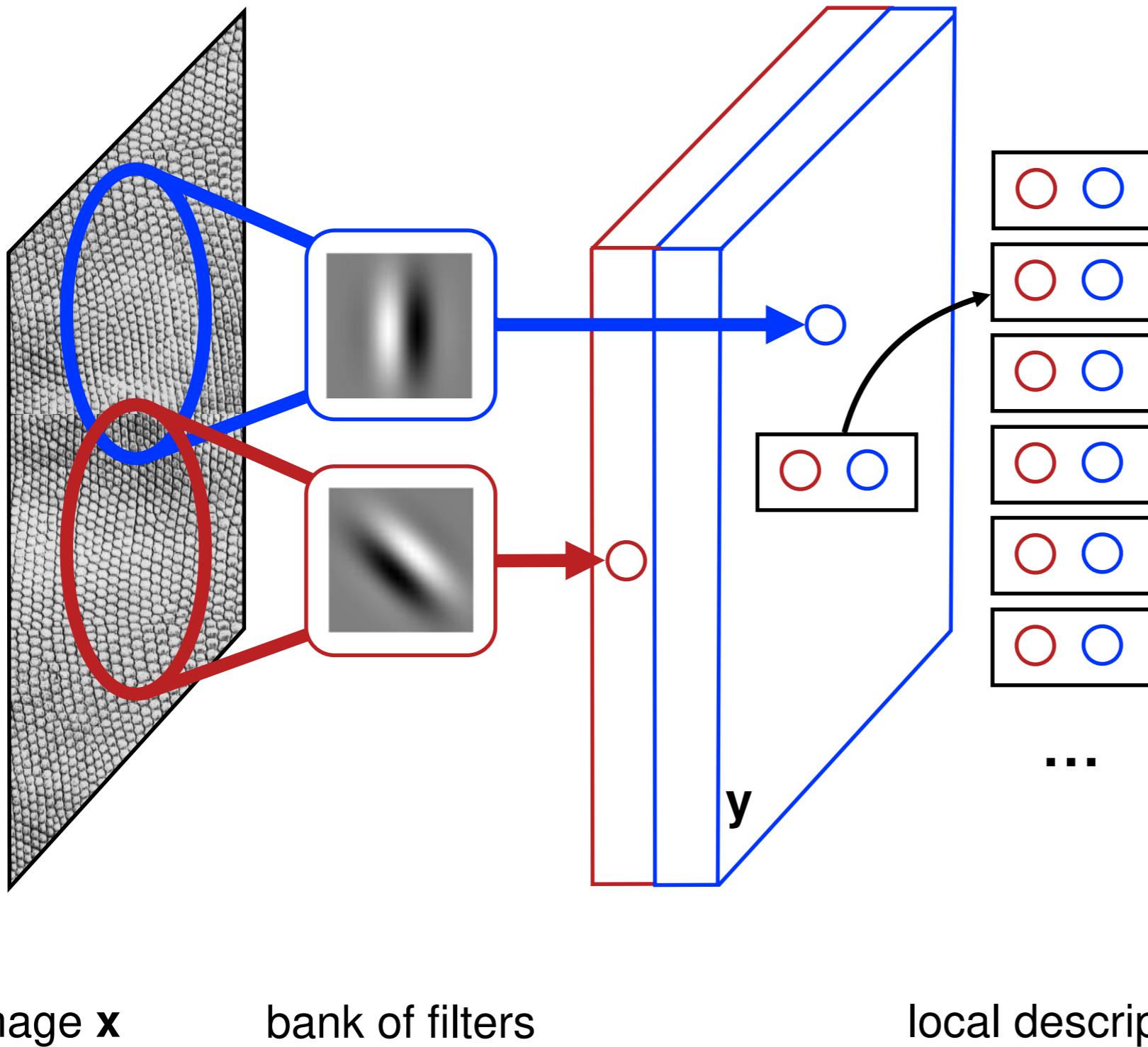
bank of filters

local descriptors

VQ + histogram

Texture representations

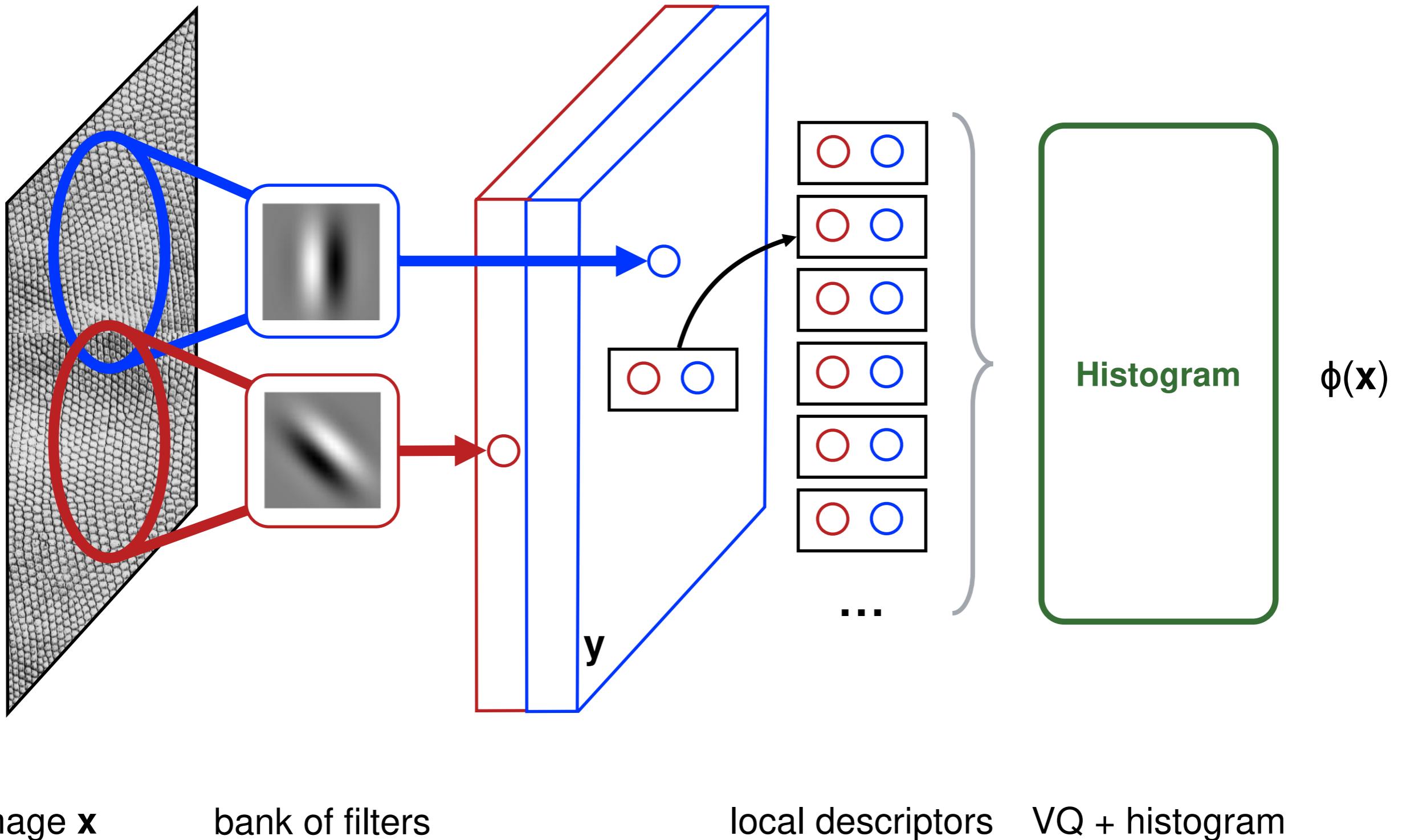
Filters + histogramming



[Leung and Malik 99, 01, Schmid 01, Varma and Zisserman 02, 05]

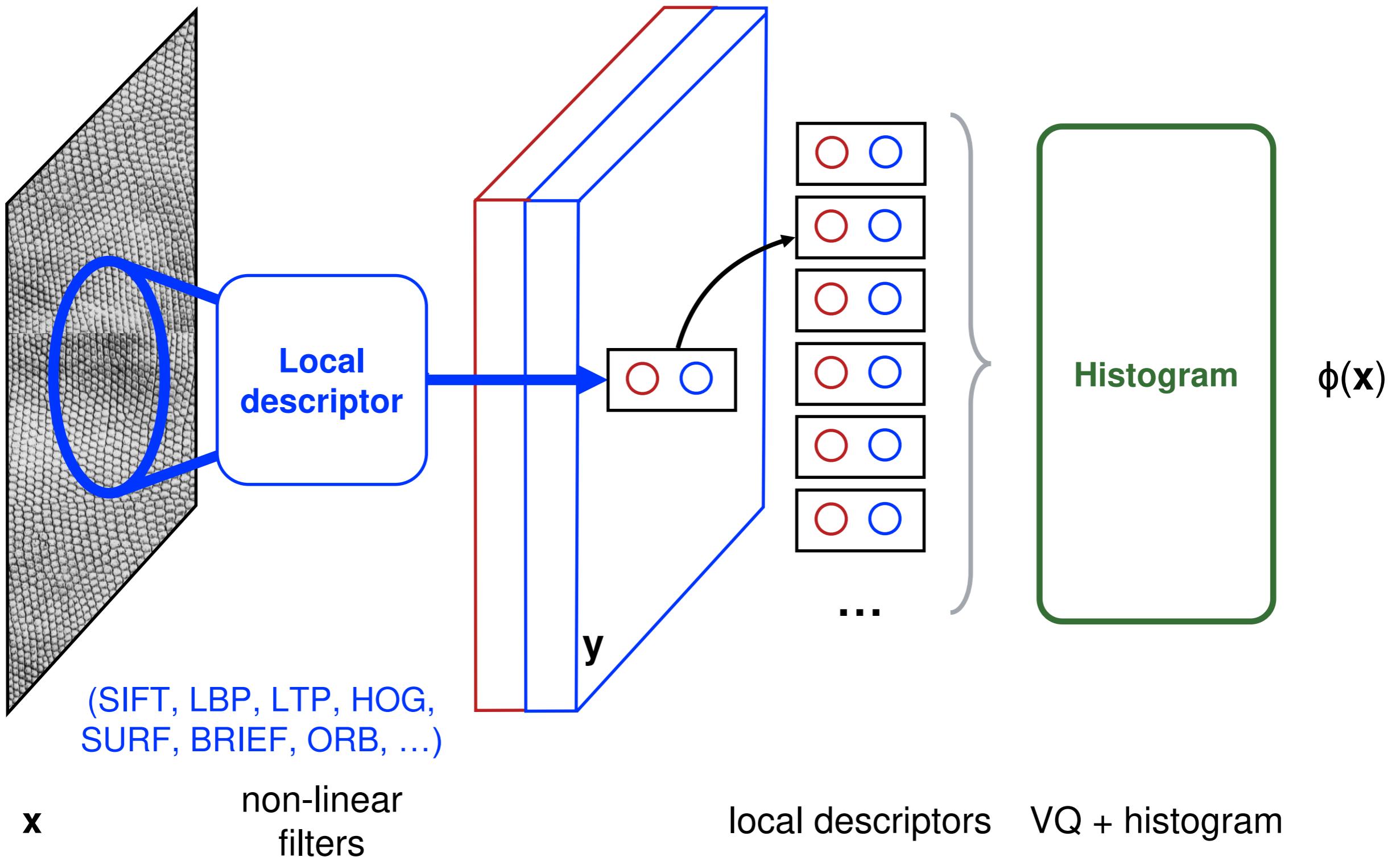
Texture representations

Filters + histogramming



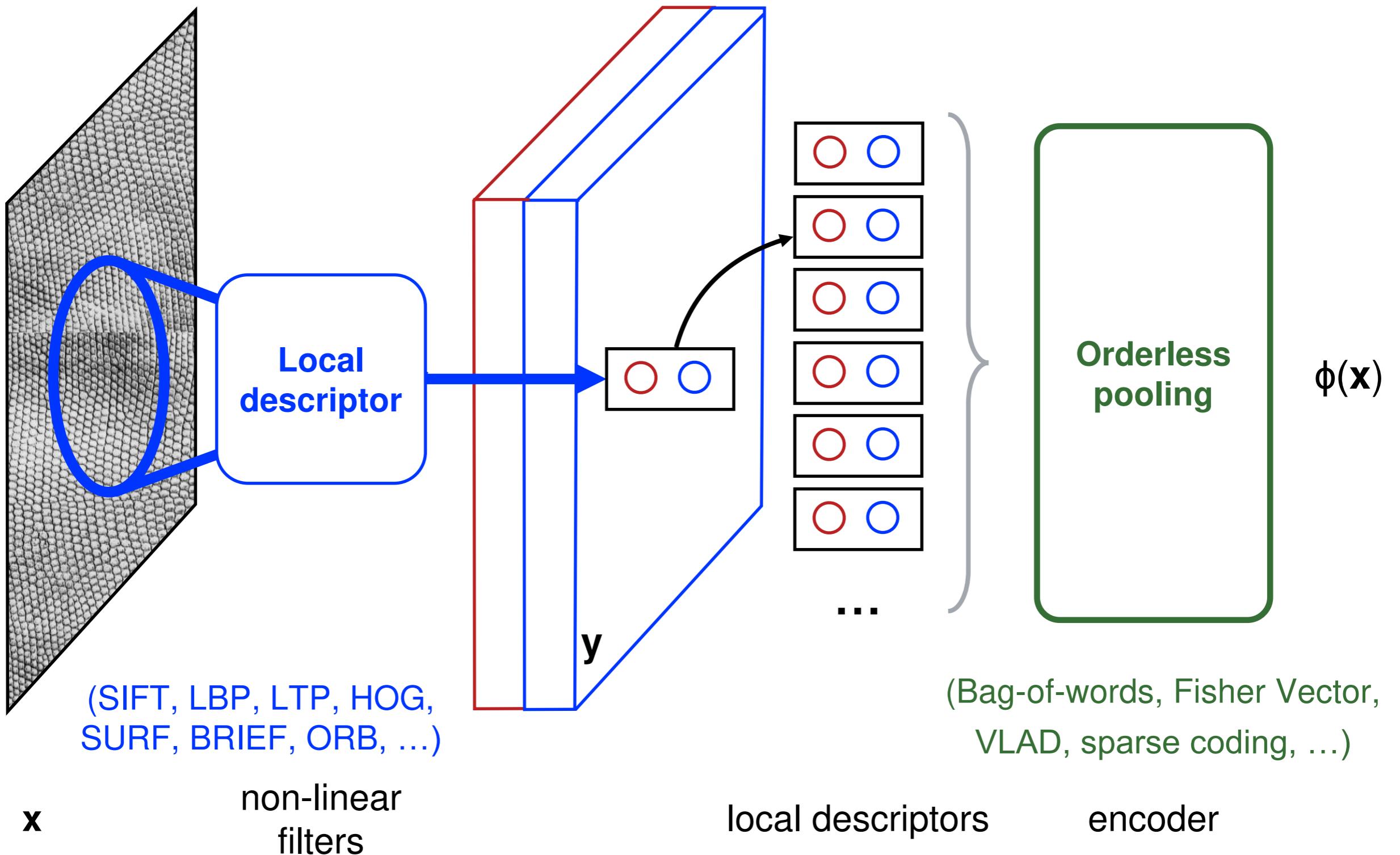
Texture representations

Filters may be non-linear

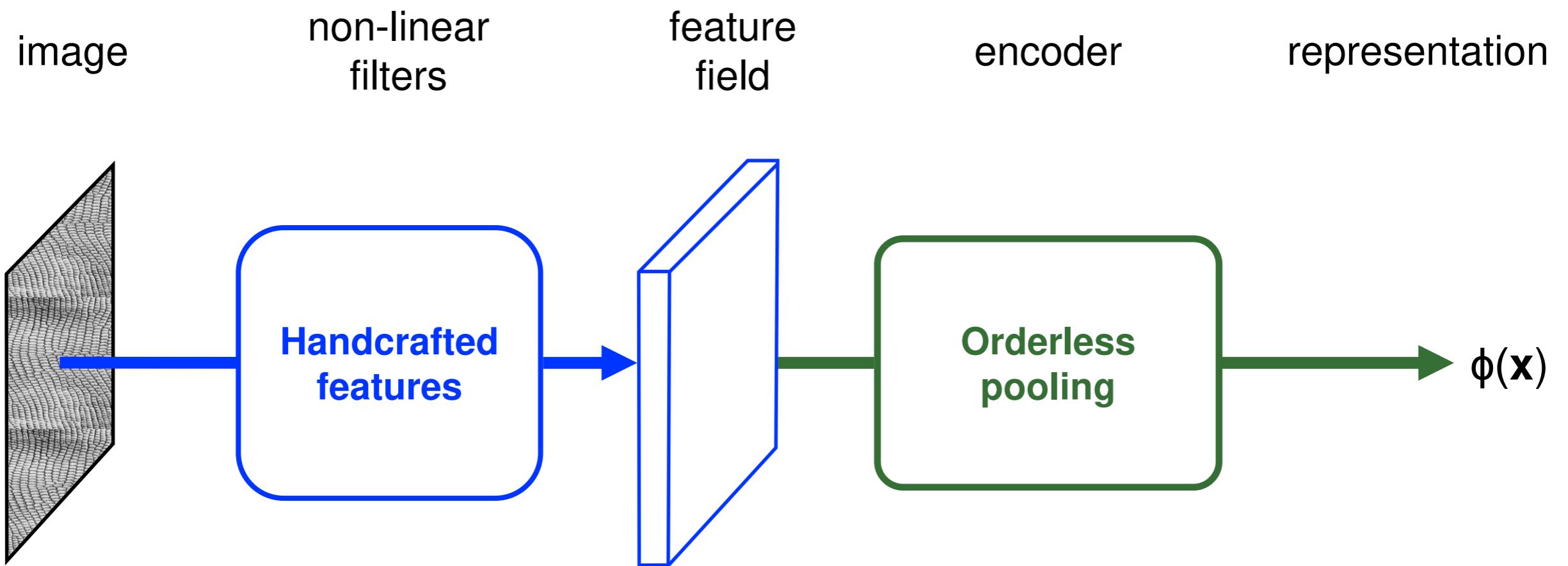


Texture representations

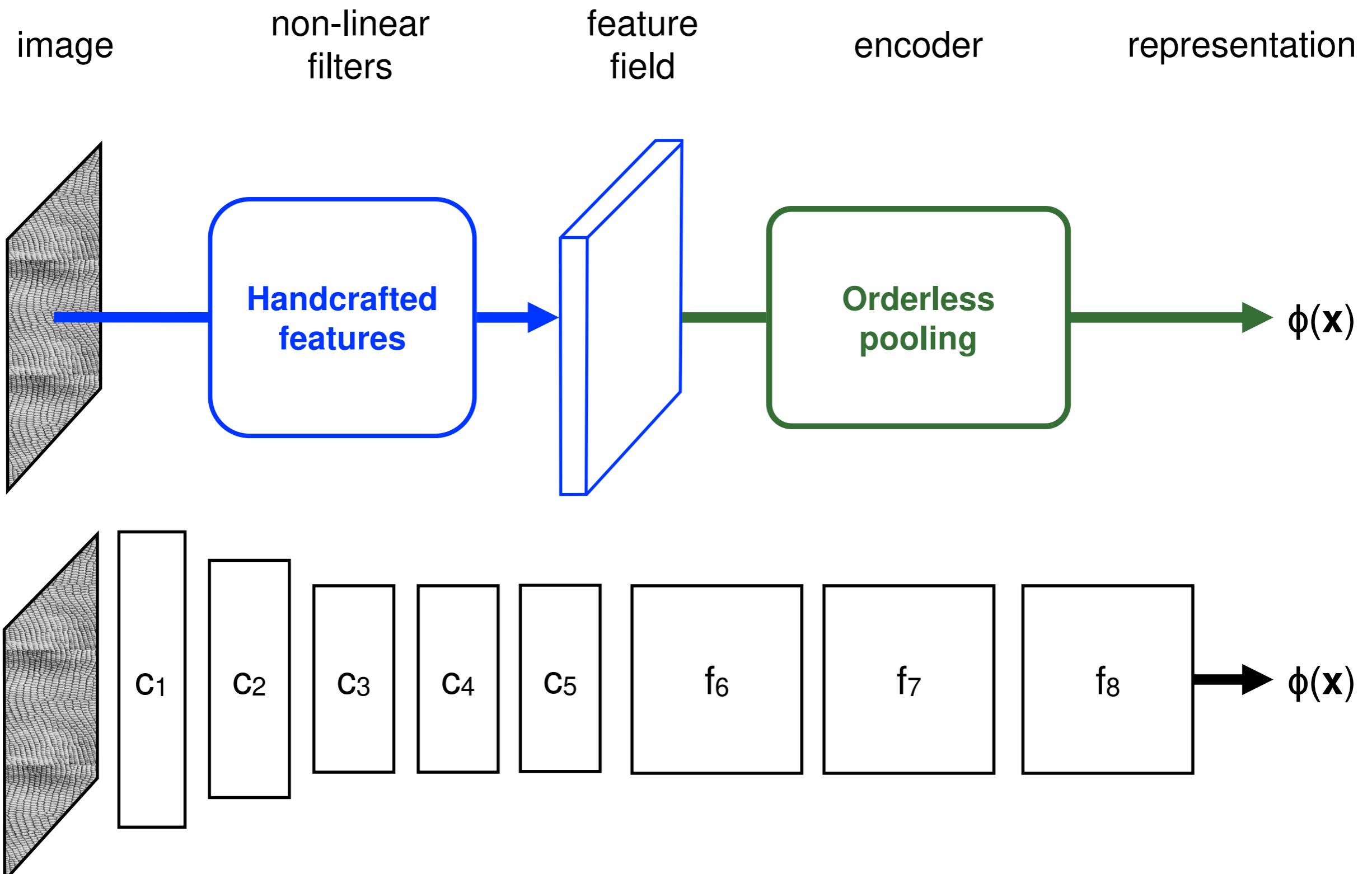
Replace histograms with an order-less pooling encoder



Texture representations vs CNNs

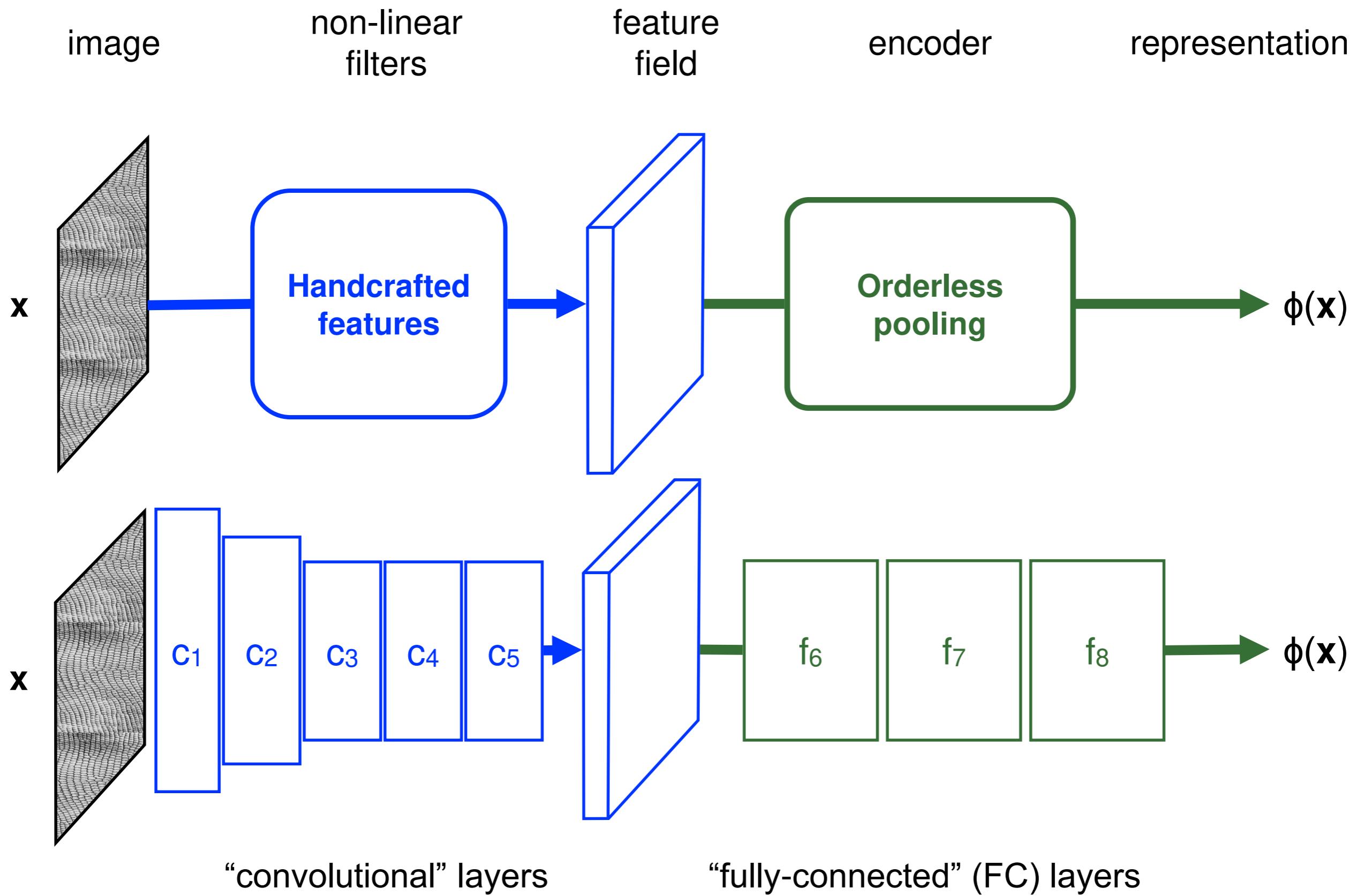


Texture representations vs CNNs

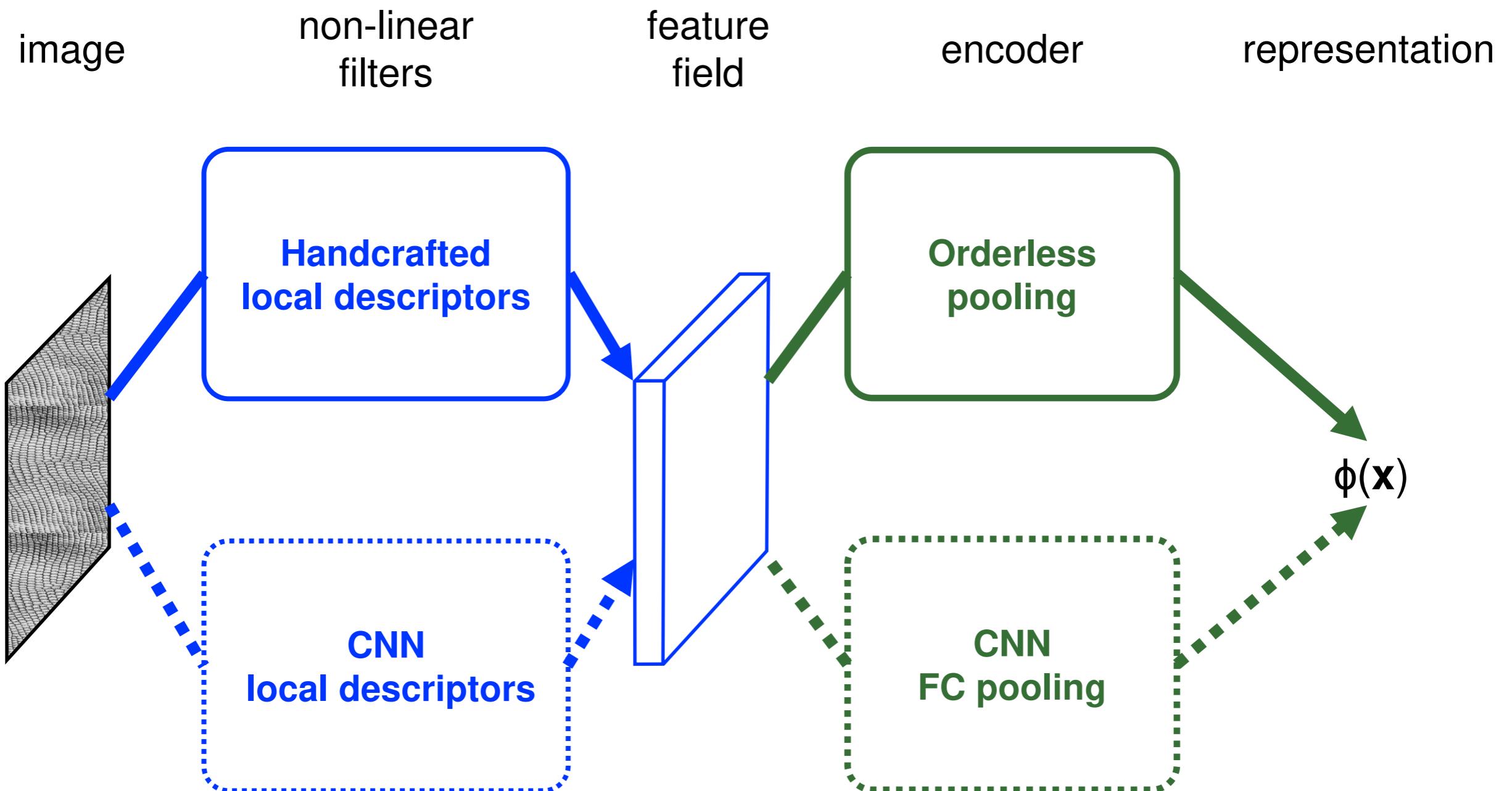


[Krizhevsky et al. 12]

Texture representations vs CNNs

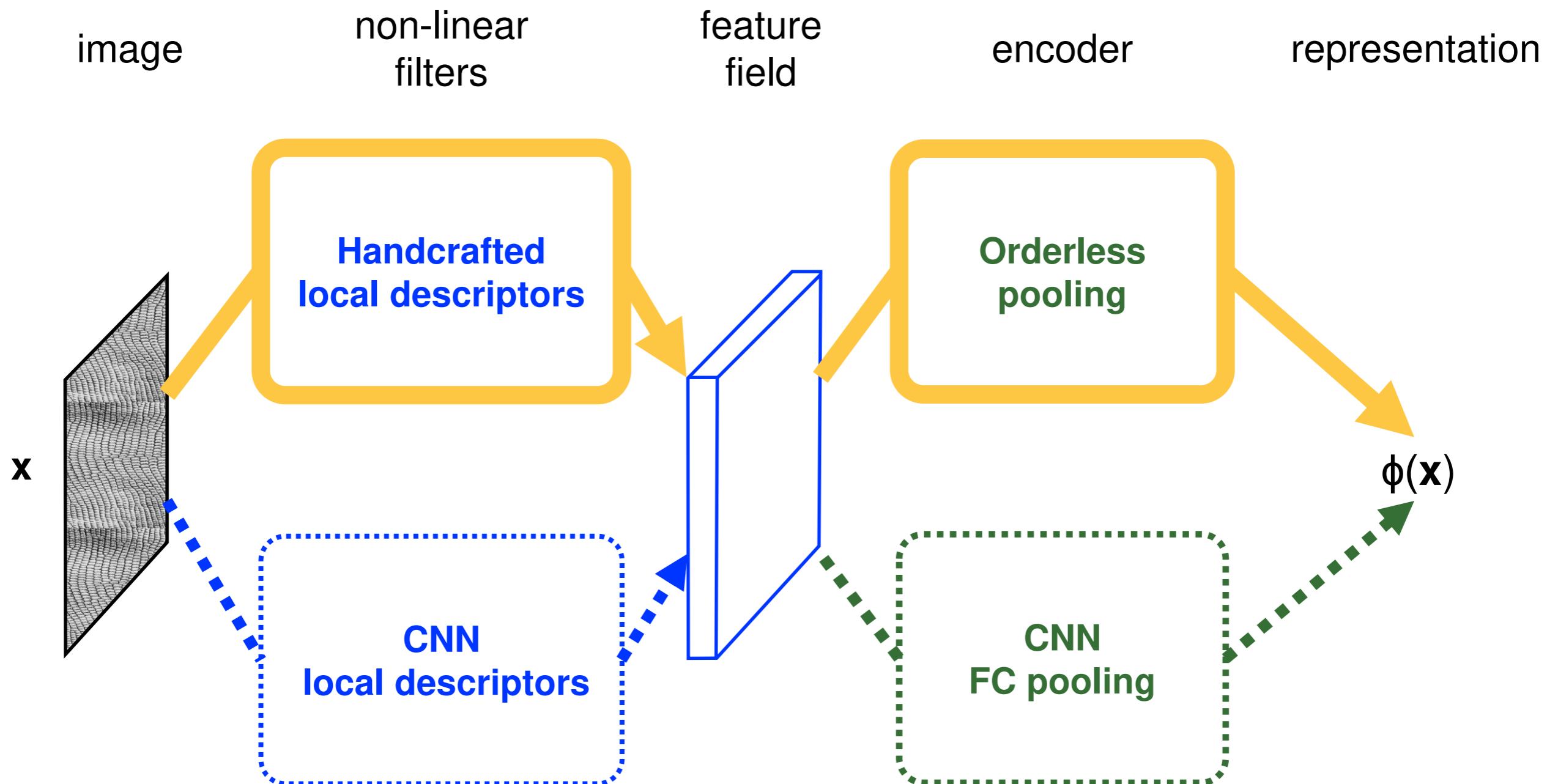


Mix and match



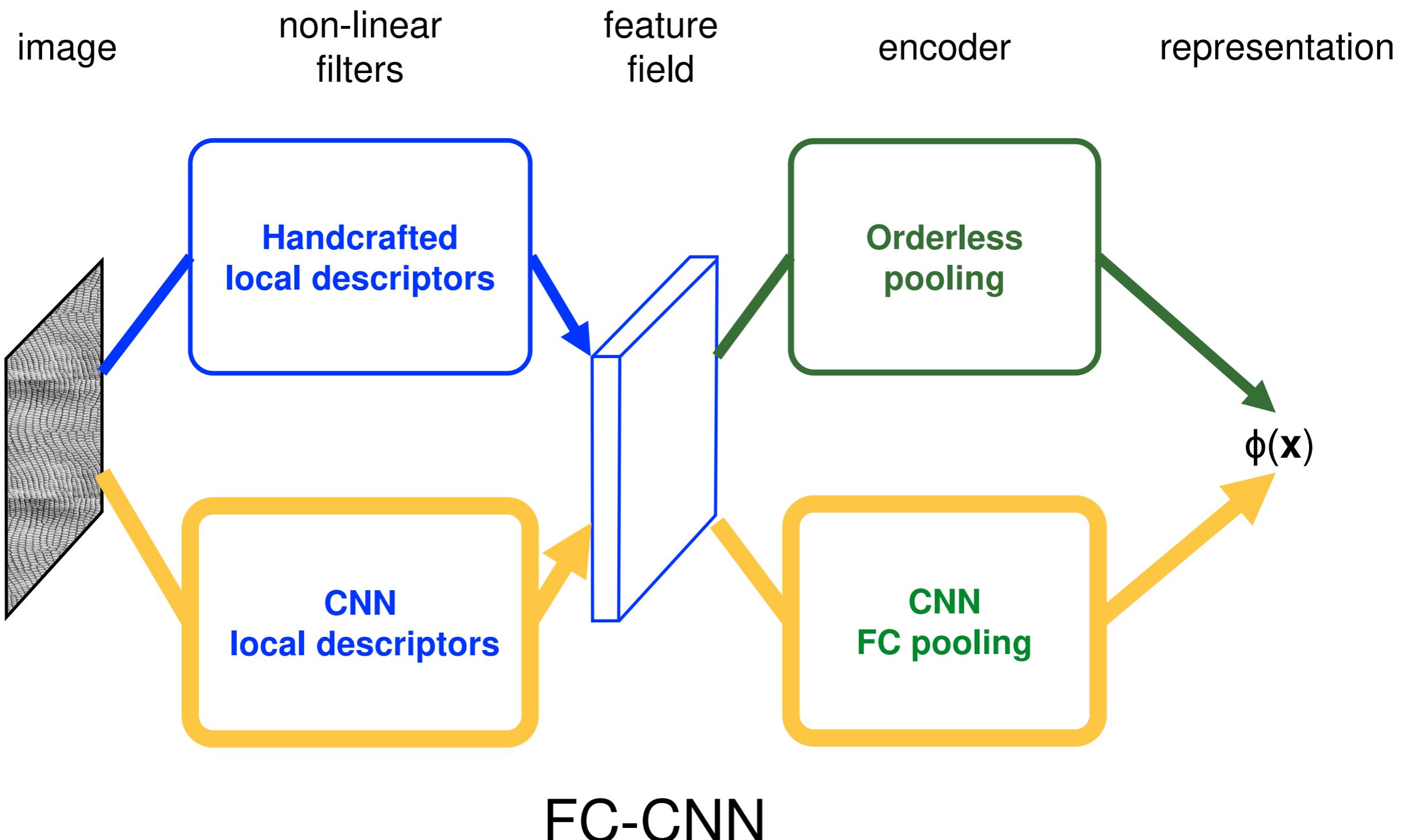
Mix and match

Standard texture representation



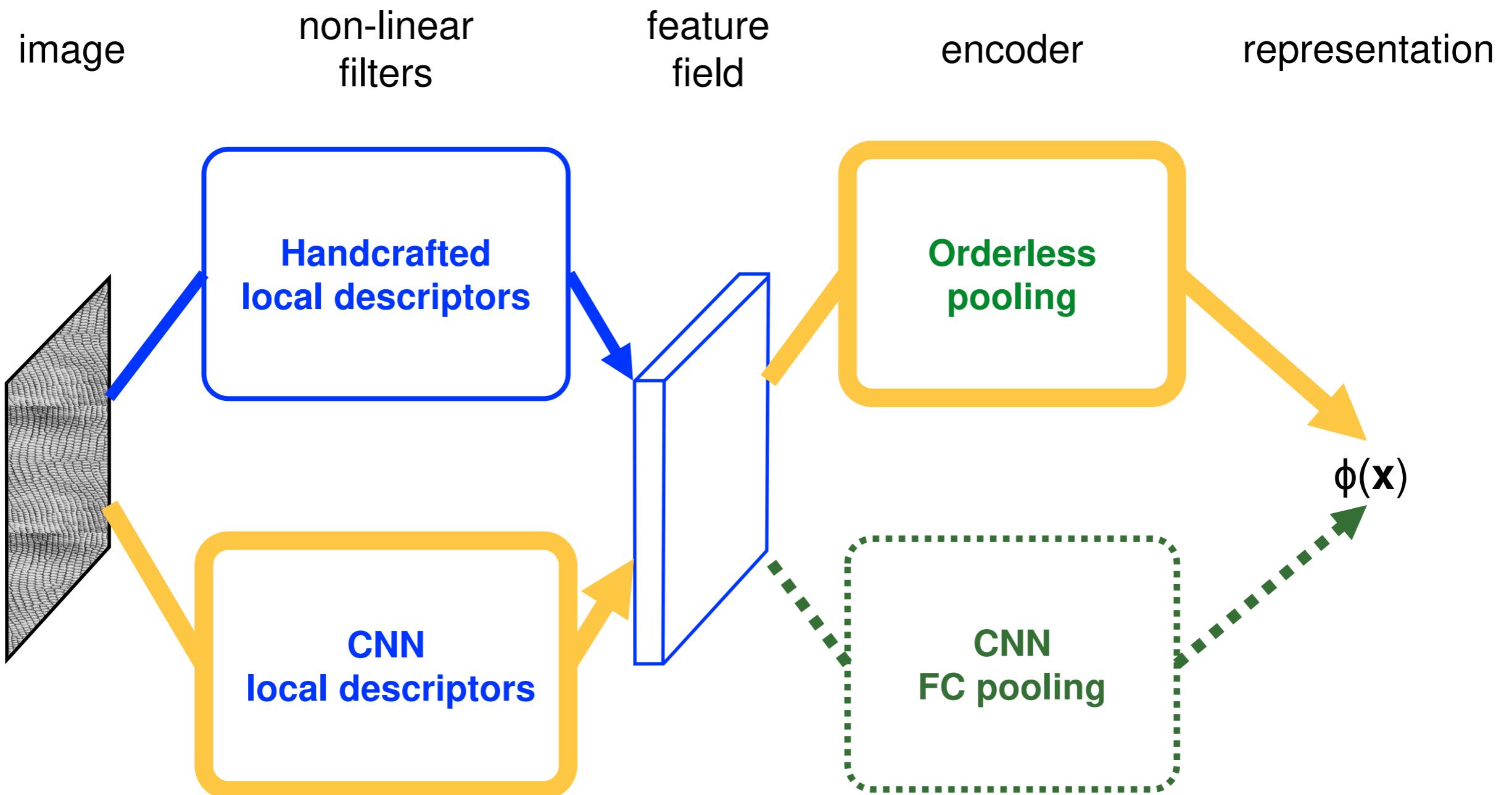
Mix and match

Standard application of CNN



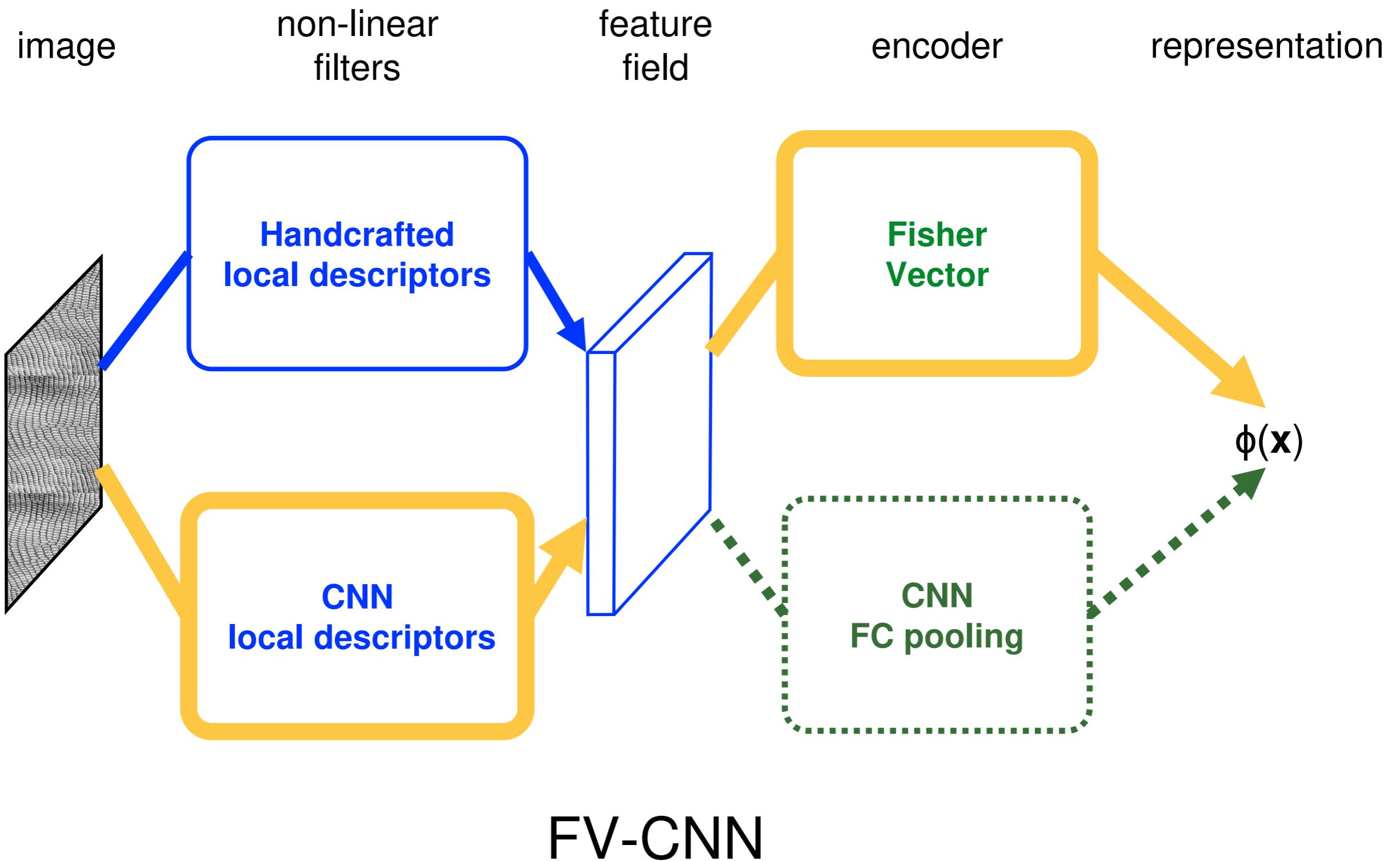
Mix and match

Order-less pooling of CNN local descriptors

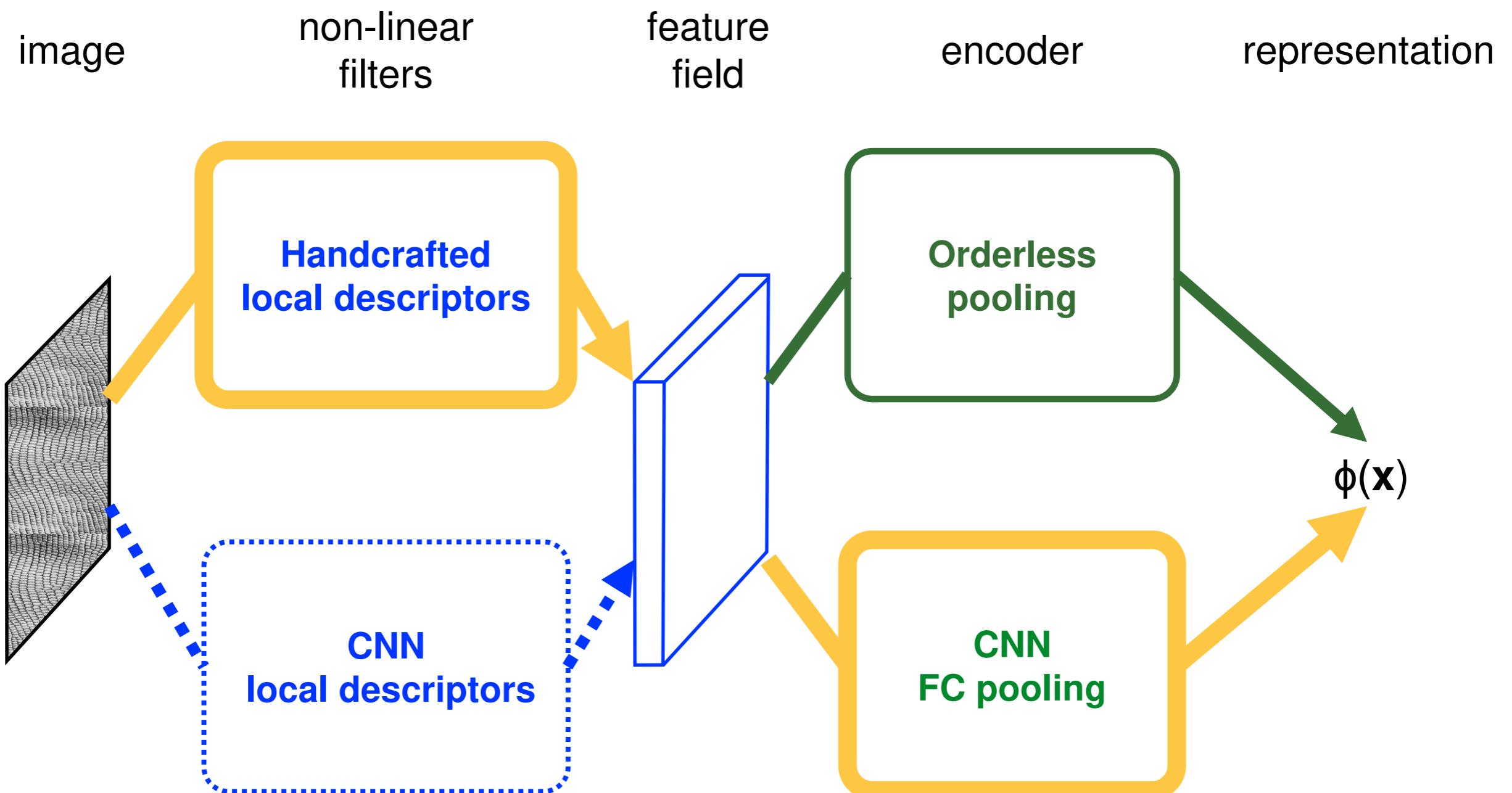


Mix and match

CNN descriptors pooled by Fisher Vector



Mix and match

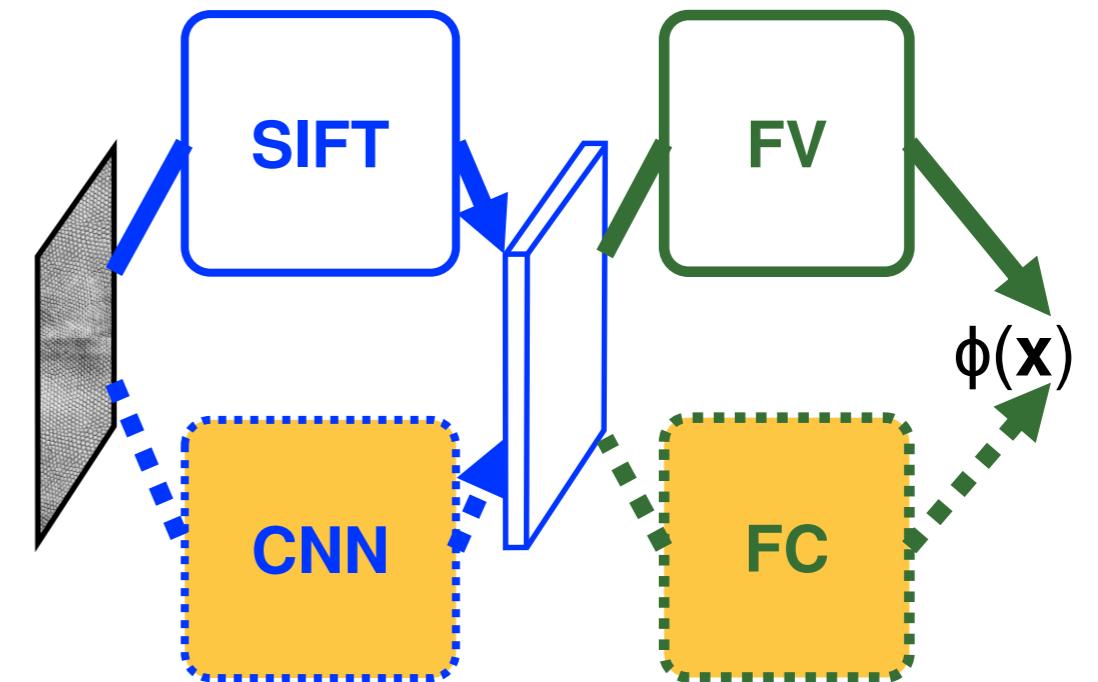


See [Perronnin and Larlus 15] Poster 2B-44

Tested modules

Baseline CNN models

- ▶ **Typical**
AlexNet [Krizhevsky et al.12]
VGG-M [Chatfield et al.14]
- ▶ **Deep**
VGG-VD [Simonyan Zisserman 14]



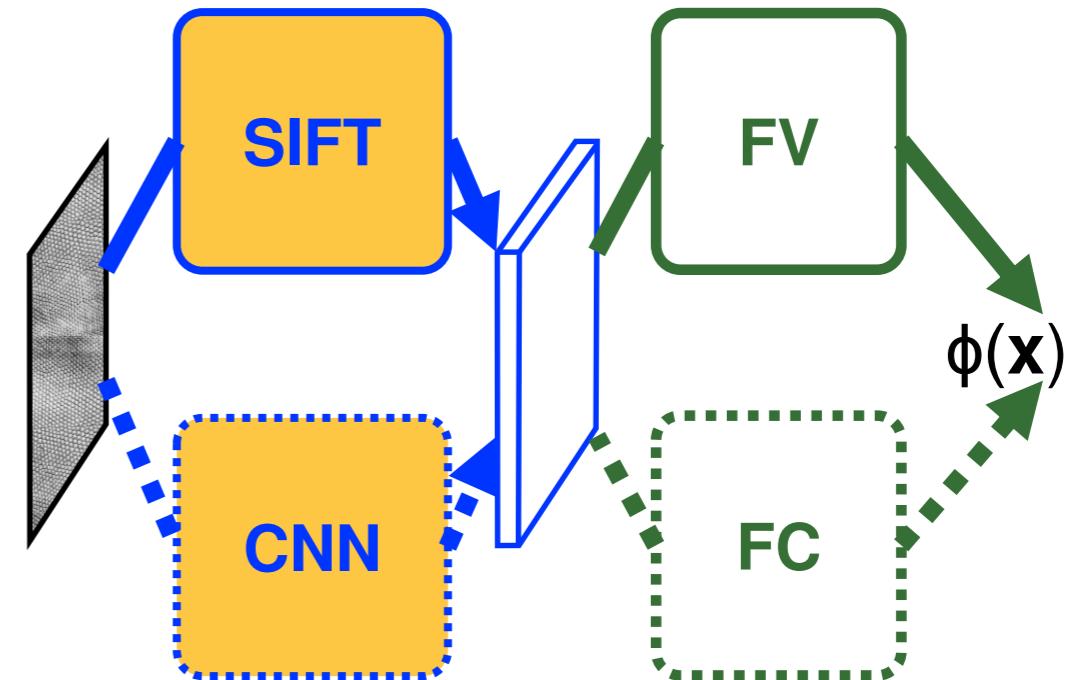
Tested modules

Baseline CNN models

- ▶ **Typical**
AlexNet [Krizhevsky et al.12]
VGG-M [Chatfield et al.14]
- ▶ **Deep**
VGG-VD [Simonyan Zisserman 14]

Local image descriptors

- ▶ **Handcrafted:** SIFT [Lowe 99]
- ▶ **Learned:** Convolutional layers of CNNs



Tested modules

Baseline CNN models

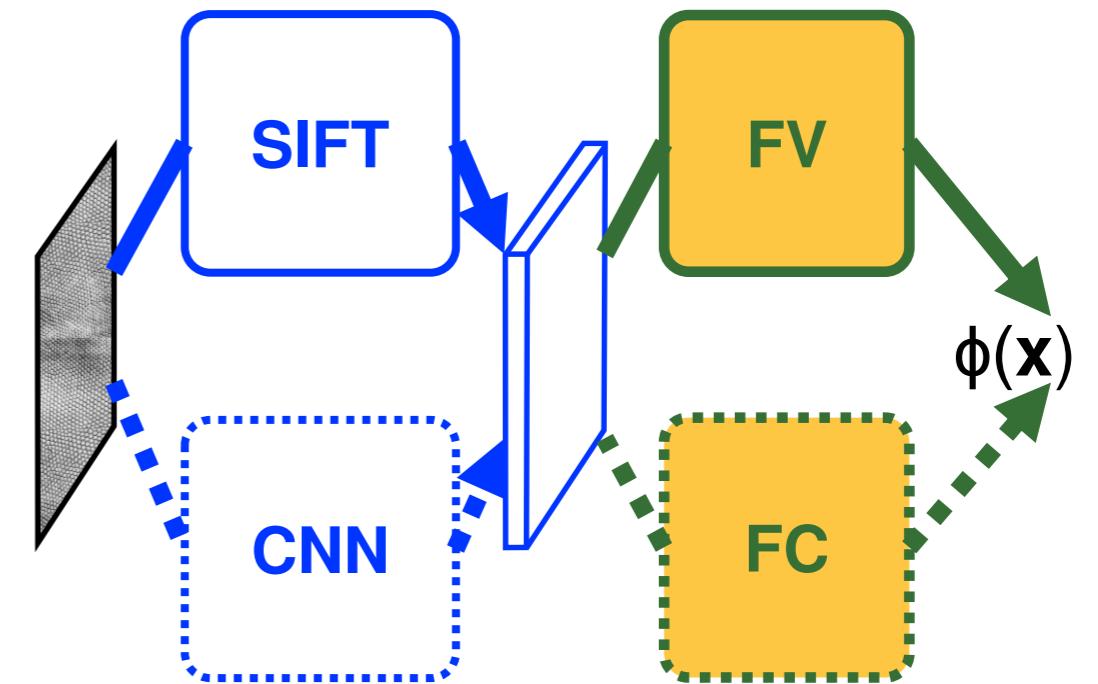
- ▶ **Typical**
AlexNet [Krizhevsky et al.12]
VGG-M [Chatfield et al.14]
- ▶ **Deep**
VGG-VD [Simonyan Zisserman 14]

Local image descriptors

- ▶ **Handcrafted:** SIFT [Lowe 99]
- ▶ **Learned:** Convolutional layers of CNNs

Pooling encoders

- ▶ **Classical**
Bag of Visual Words [Sivic and Zisserman 03, Csurka et al. 04]
Fisher Vector [Perronnin and Dance 07, Perronnin et al. 10]
- ▶ **CNN**
FC layers [Chatfield et al. 14, Girshick et al. 2014, Gong et al. 14, Razavin et al. 14]



Findings: what pooling CNNs is good for

25

How does FV-CNN perform compared to other descriptors?

How does FV-CNN handle region recognition?

What is the benefit of FV-CNN in domain-transfer?

Datasets and benchmarks

Material recognition (FMD)

[Liu et al.10, Sharan et al. 13]



Fine-grained recognition (CUB)

[Wah et al. 11]



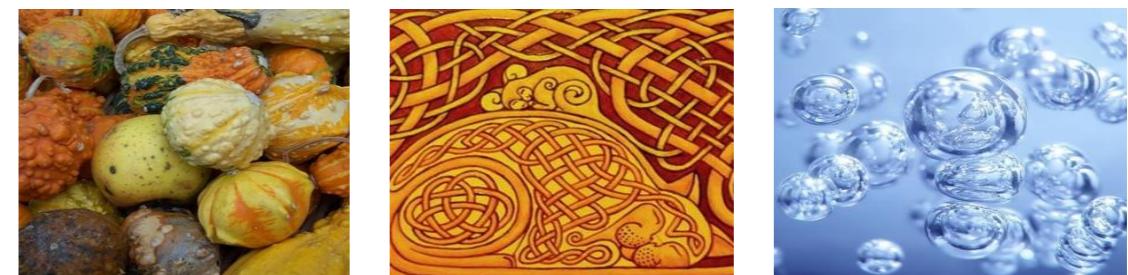
Object recognition (VOC07)

[Everingham et al. 07]



Texture attribute recognition (DTD)

[Cimpoi et al. 14]



Scene recognition (MIT Indoors)

[Quattoni and Torralba 09]

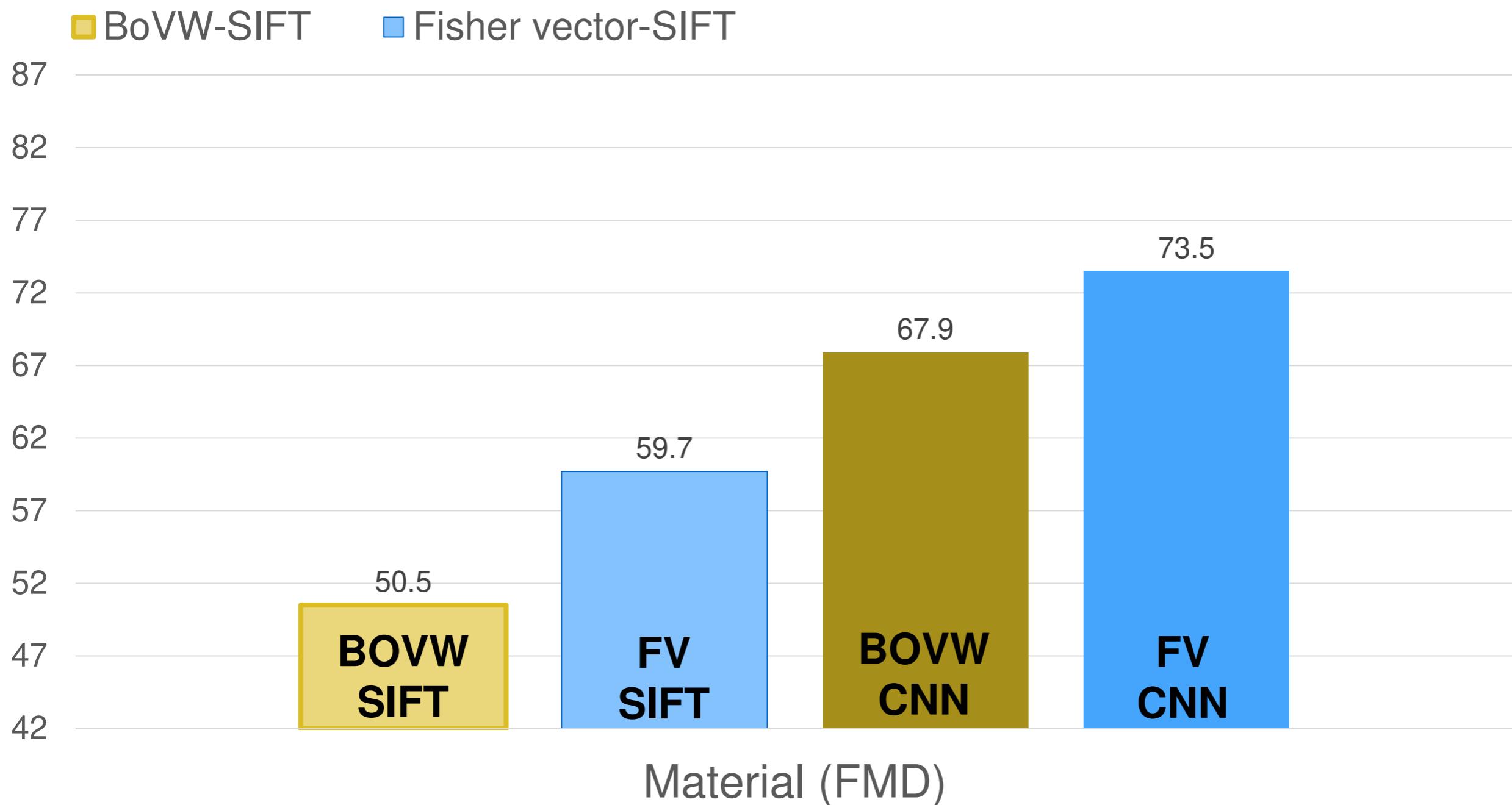


Things and stuff (MSRC)

[Criminisi 04, Shotton et al. 06]



Which feature and encoder?

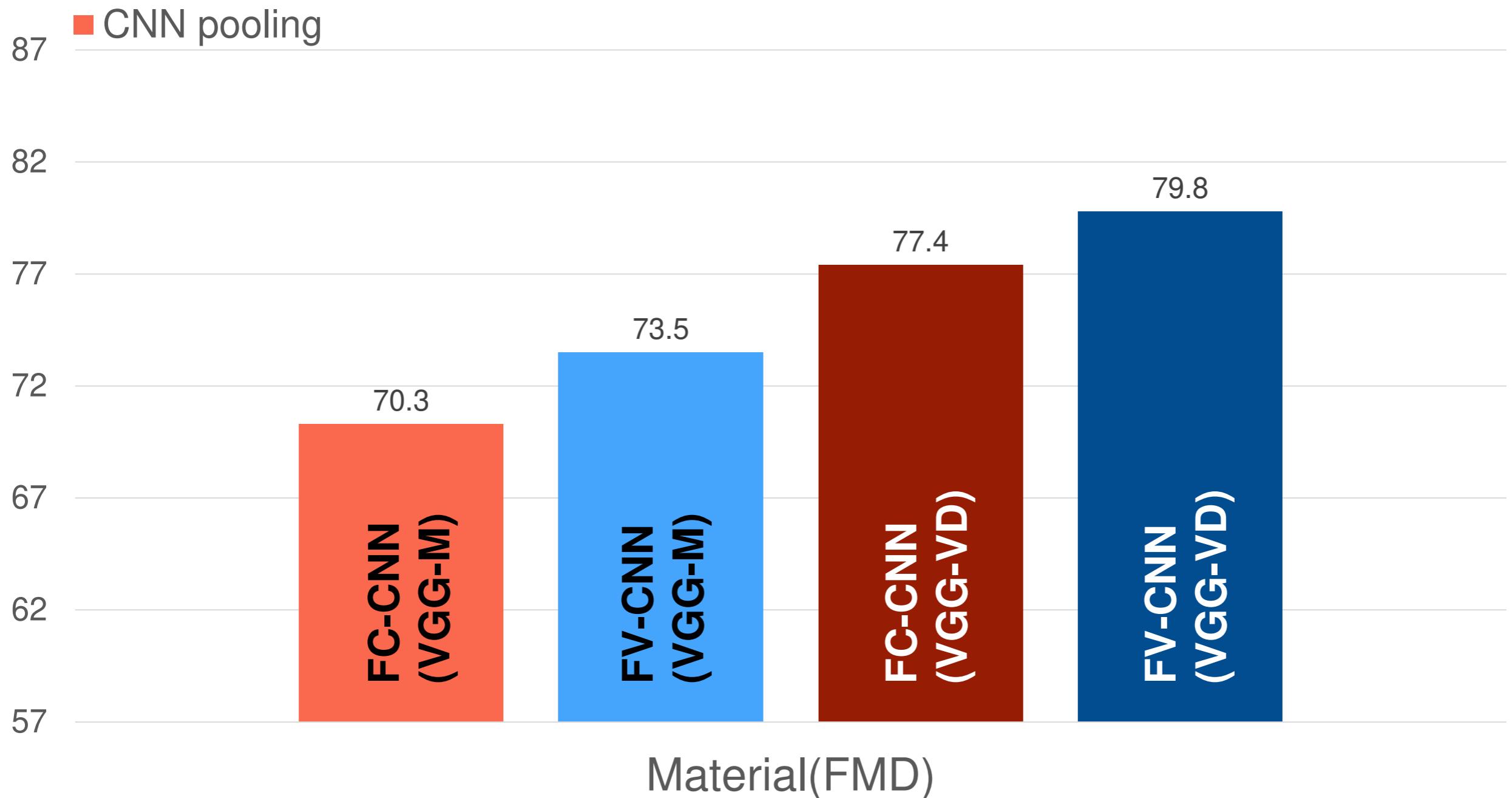


Finding 1) BoVW < FV

Material (FMD)

Finding 2) SIFT < CNN

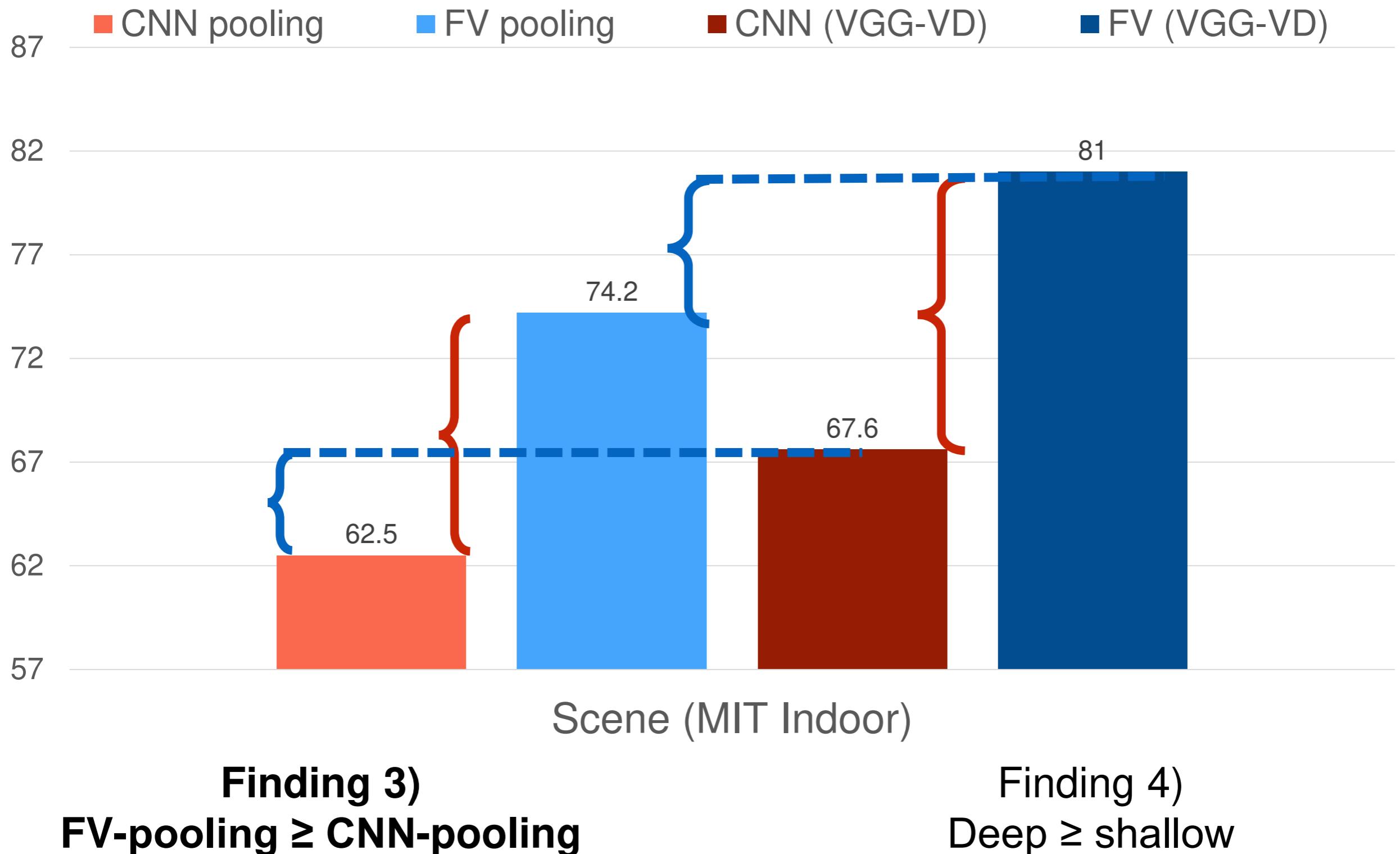
CNN vs Fisher Vector pooling



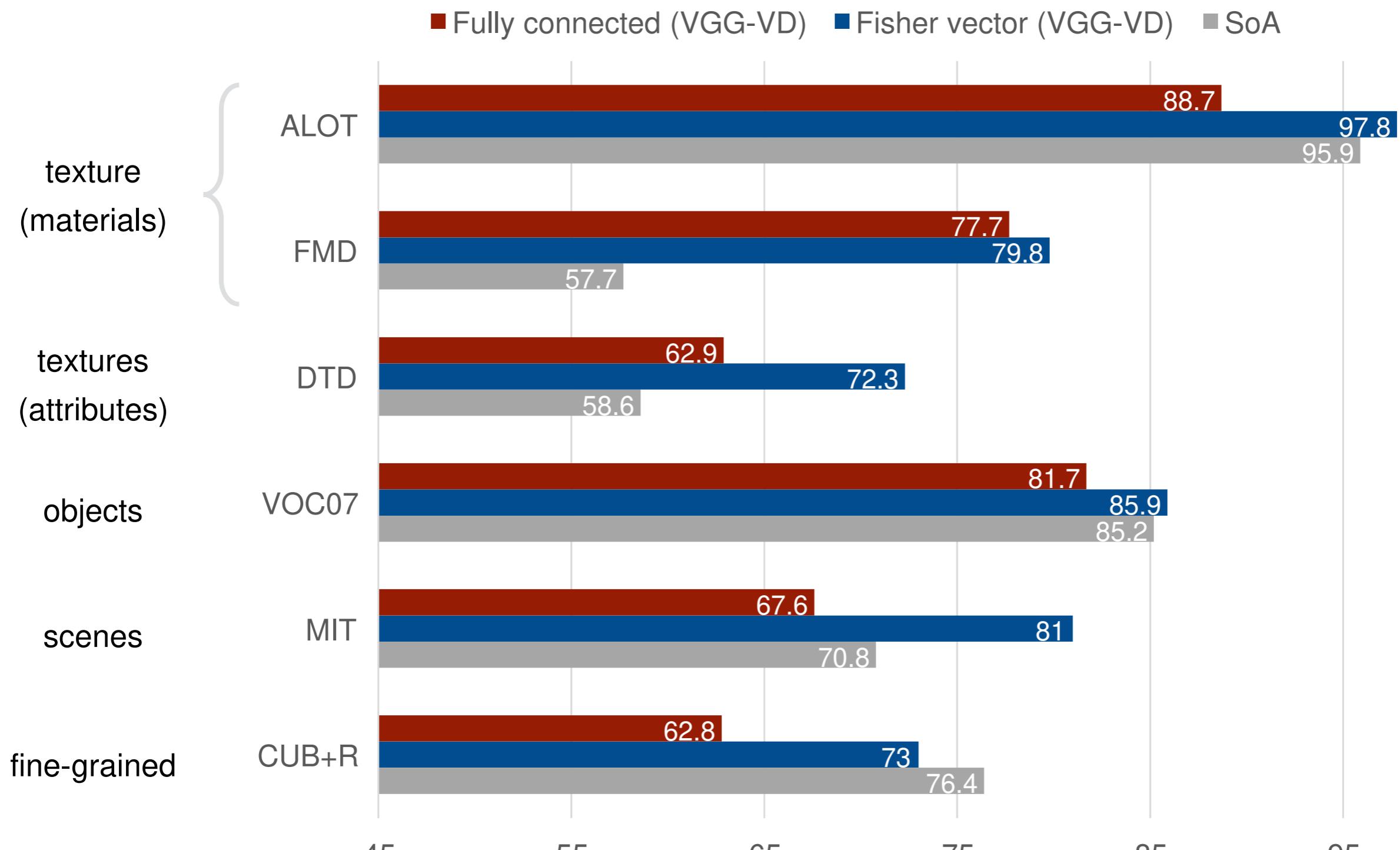
Finding 3)
FV-pooling \geq CNN-pooling

Finding 4)
Deep \geq shallow

CNN vs Fisher Vector pooling



Breadth of applicability



Finding 5) FV + CNN applies to many diverse domains

[Cimpoi et al. 14, Sulc and Matas 14, Sharan et al. 13, Wei and Levoy 14, Zhou et al. 14, Zhang et al. 14
Burghouts and Geusebroek 09, Sharan et al. 09, Everingham et al. 08, Quattoni and Torralba 09, Wah et al. 11]

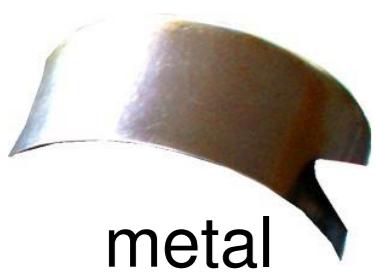
Findings: what pooling CNNs is good for

How does FV-CNN perform compared to other descriptors?

How does FV-CNN handle region recognition?

What is the benefit of FV-CNN in domain-transfer?

Texture recognition in the “wild” and “clutter” (OS)



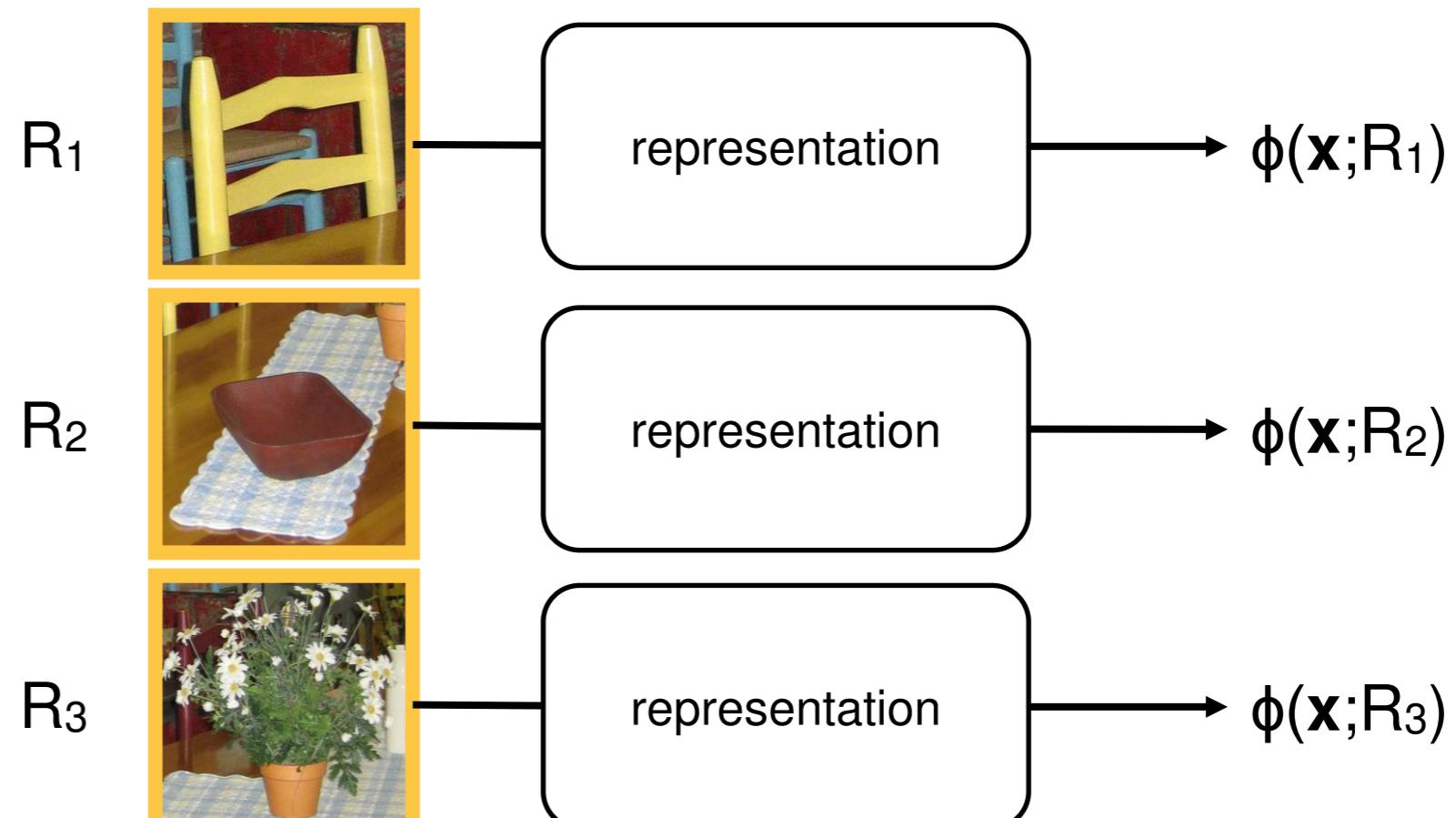
A new texture benchmark

- ▶ Based **OpenSurfaces dataset** [Bell et al. 13, 15]
- ▶ Textures in the wild (uncontrolled conditions)
- ▶ Textures in **clutter** (do not fill the image)

First extensive evaluation of texture material/attribute recognition of this kind

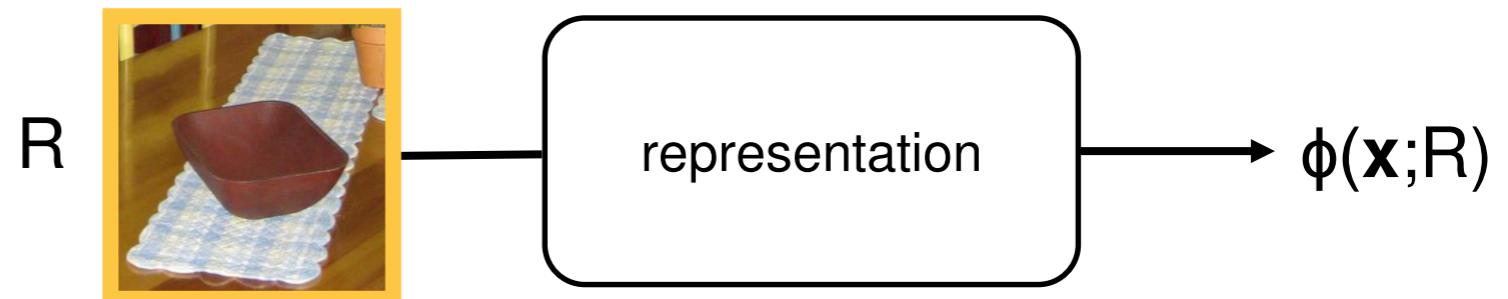
Regions: the crop & describe approach

E.g. R-CNN

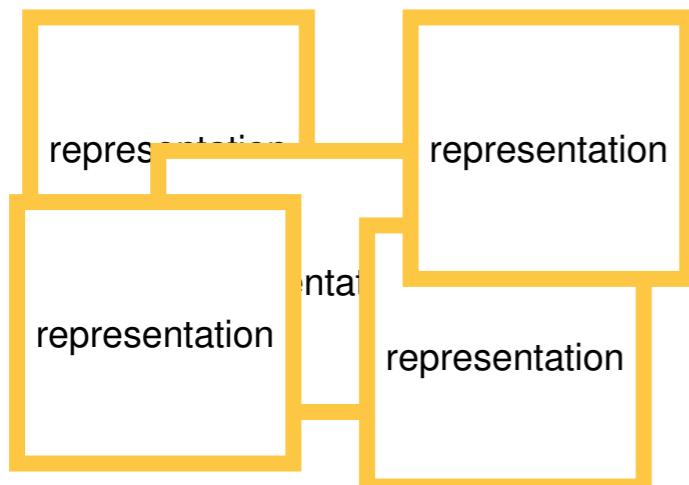


Pros: straightforward & universal construction

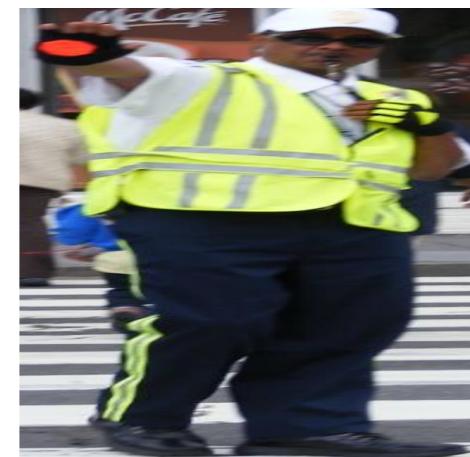
Crop & describe limitations



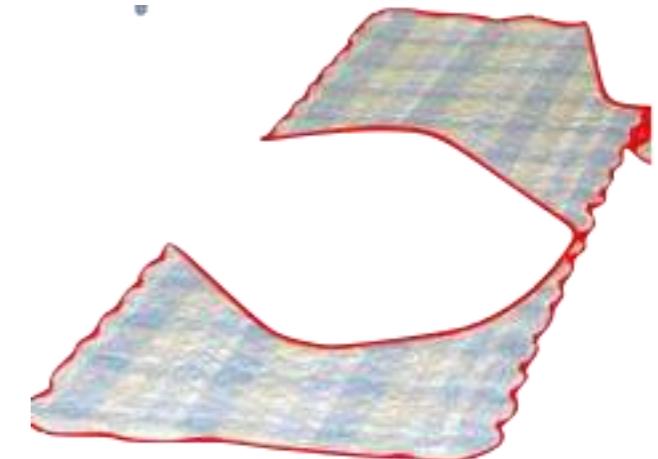
Expensive



May distort images

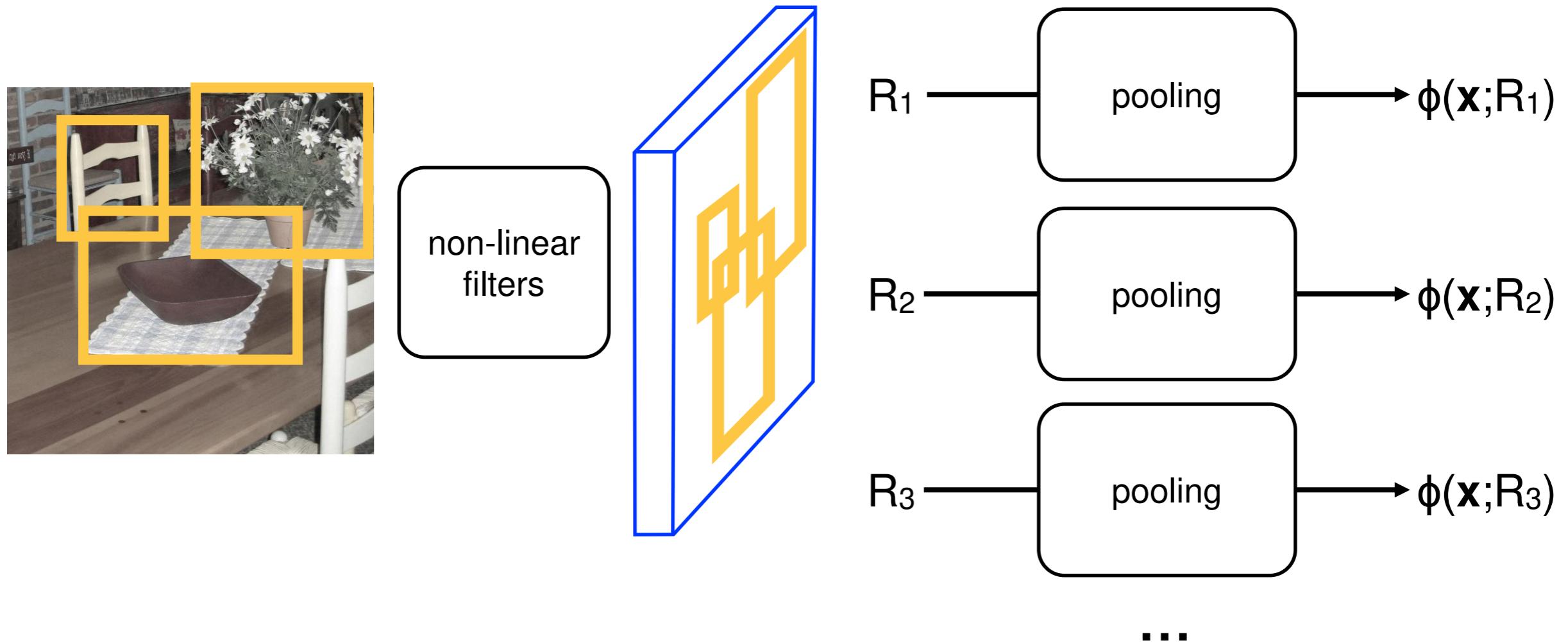


Can only do rectangles



Regions: the pooling encoder approach

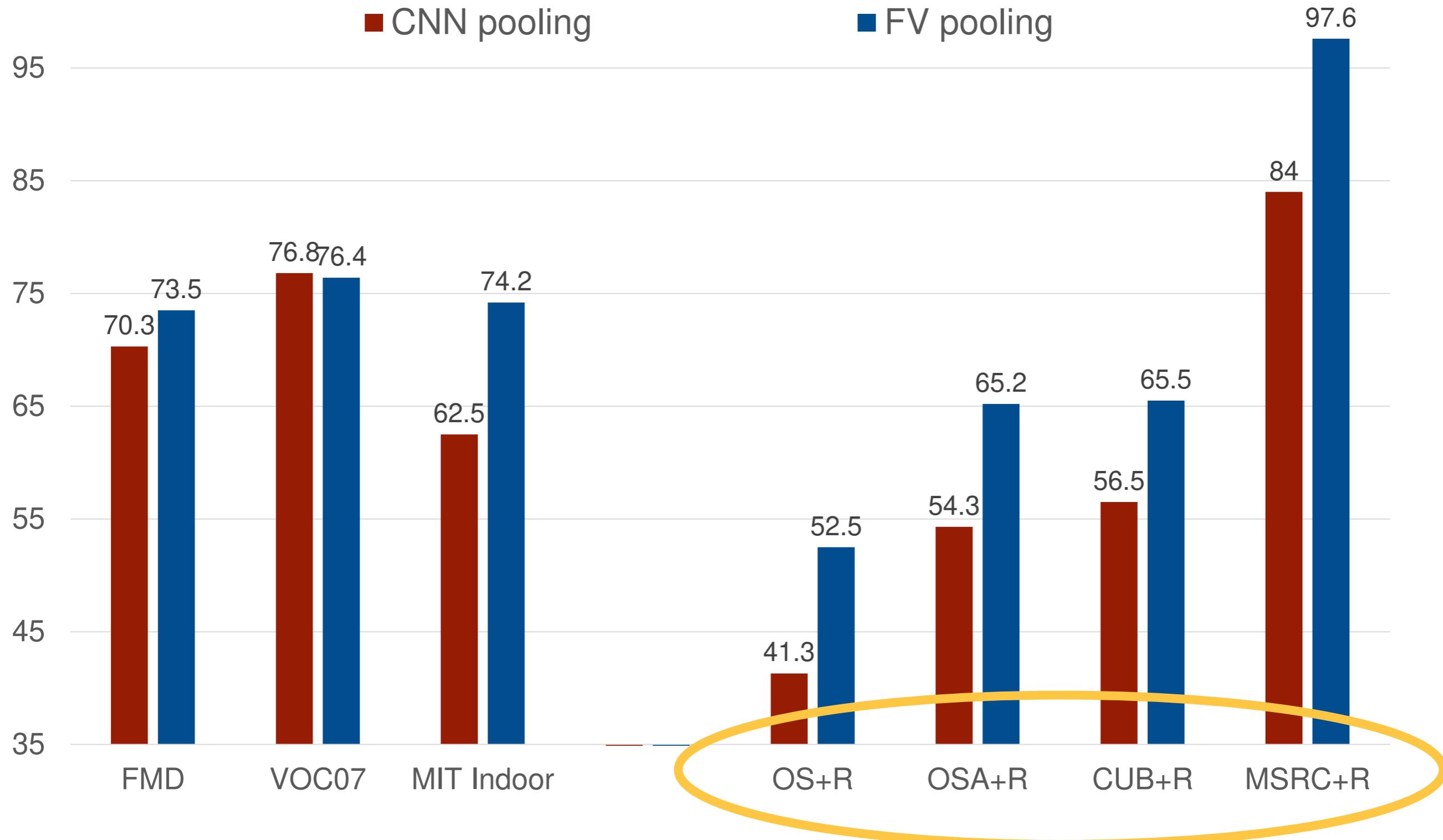
Share the local descriptors



Cons: restricted to a convolutional representation

Pros: fast, flexible, multiscale, and often more accurate

FV vs FC pooling *for regions*



**Finding 6) FV pooling >> CNN pooling for small, variable regions
(and faster too!)**

Findings: what pooling CNNs is good for

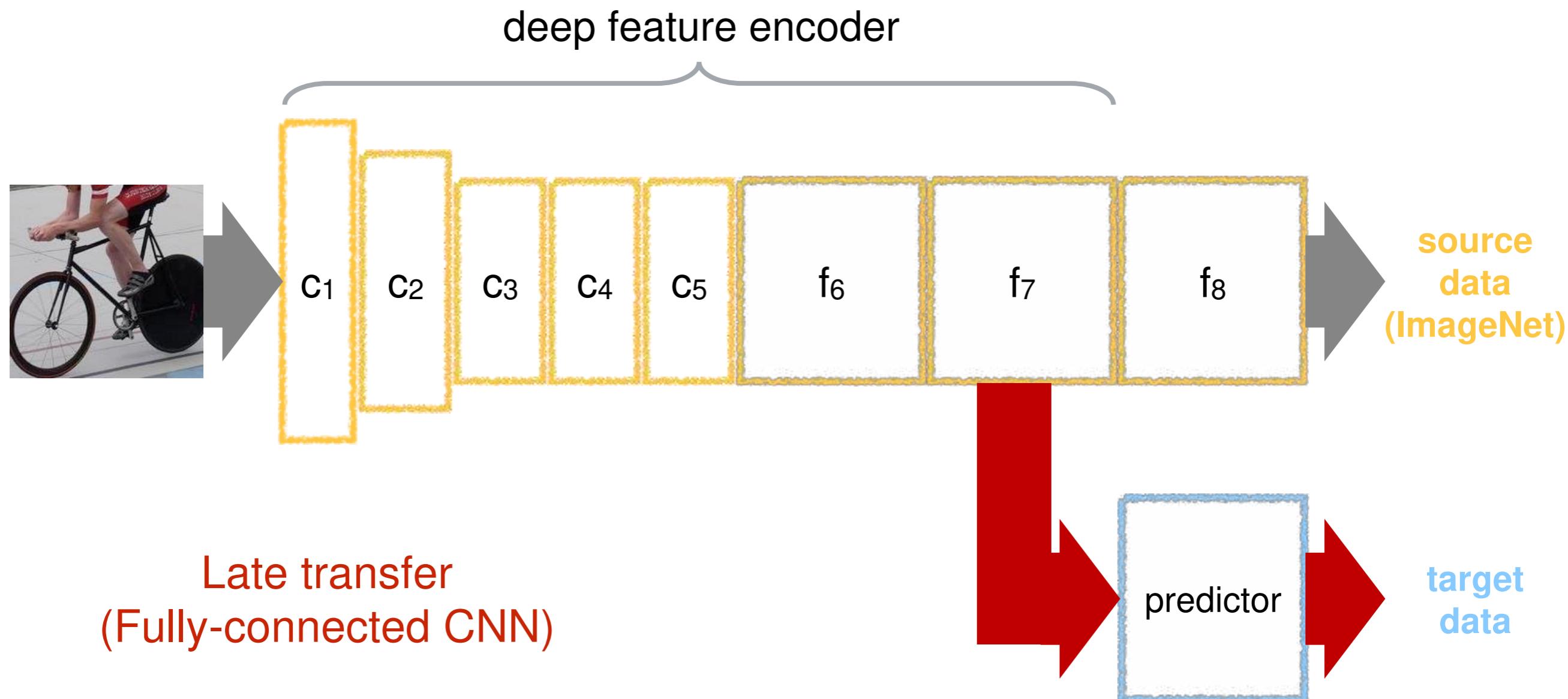
How does FV-CNN perform compared to other descriptors?

How does FV-CNN handle region recognition?

What is the benefit of FV-CNN in domain-transfer?

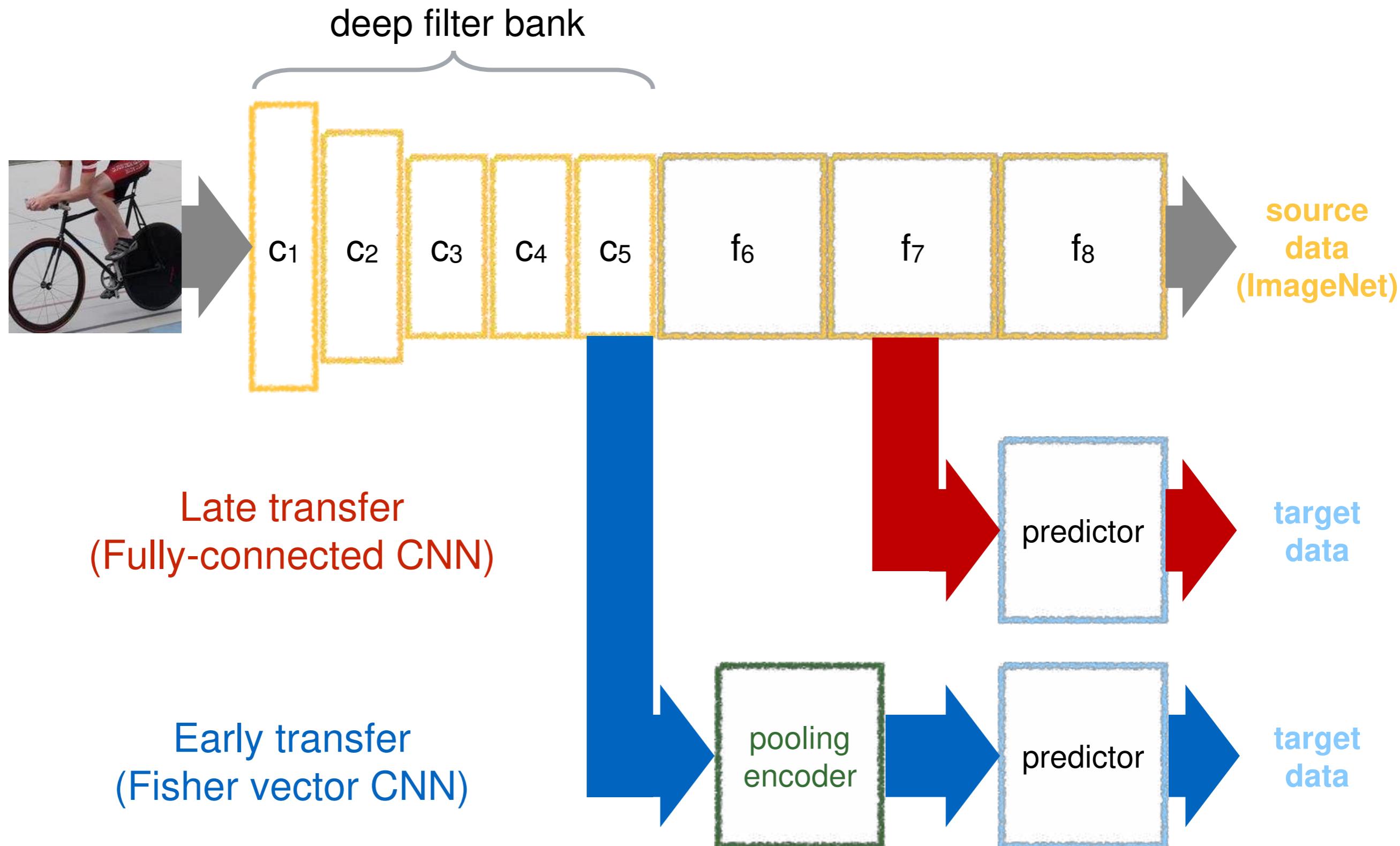
Late vs early transfer

Transfer either the fully connected or the convolutional layers

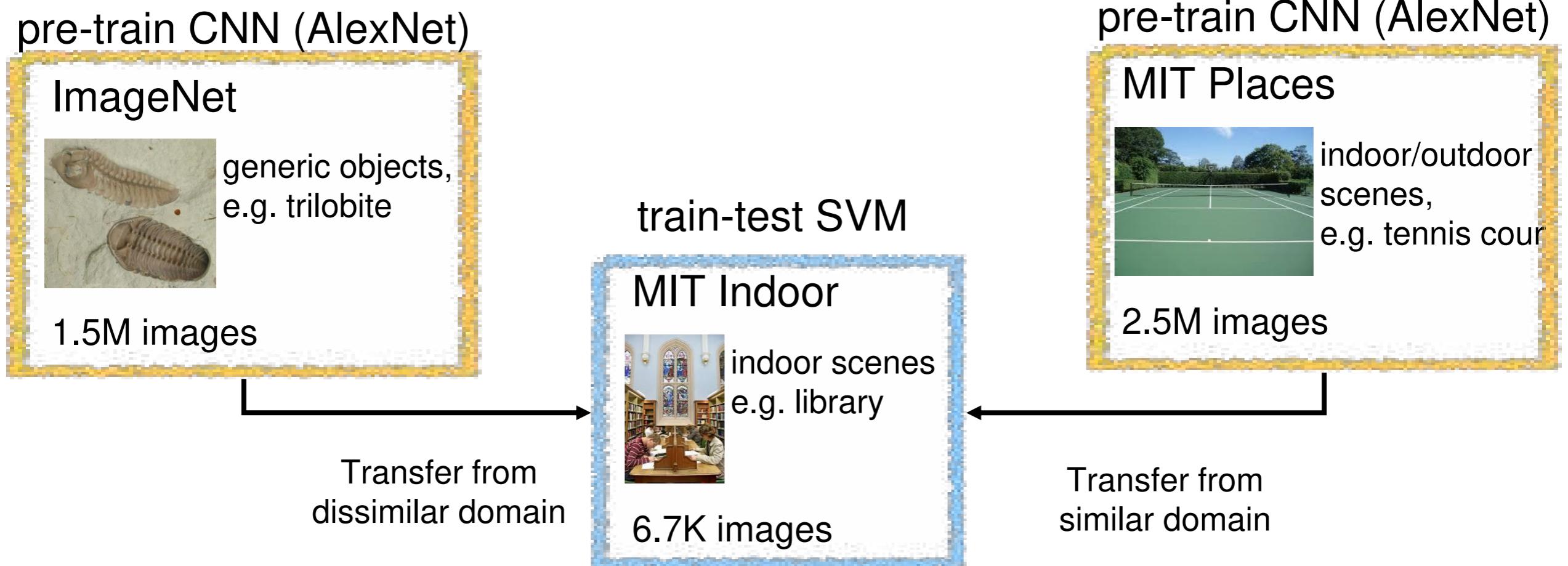


Late vs early transfer

Transfer either the fully connected or the convolutional layers



Early vs late transfer (FV-CNN)



VGG-VD

67.6%



81.0%

58.6%



69.7%

Late transfer
(Fully-connected CNN)

Early transfer
(Fisher vector CNN)

65.0%



67.6%

Summary

Hybrid architectures: Classical feature encoders can be used effectively as CNN building blocks, or inspire new ones

FV-CNN has several benefits

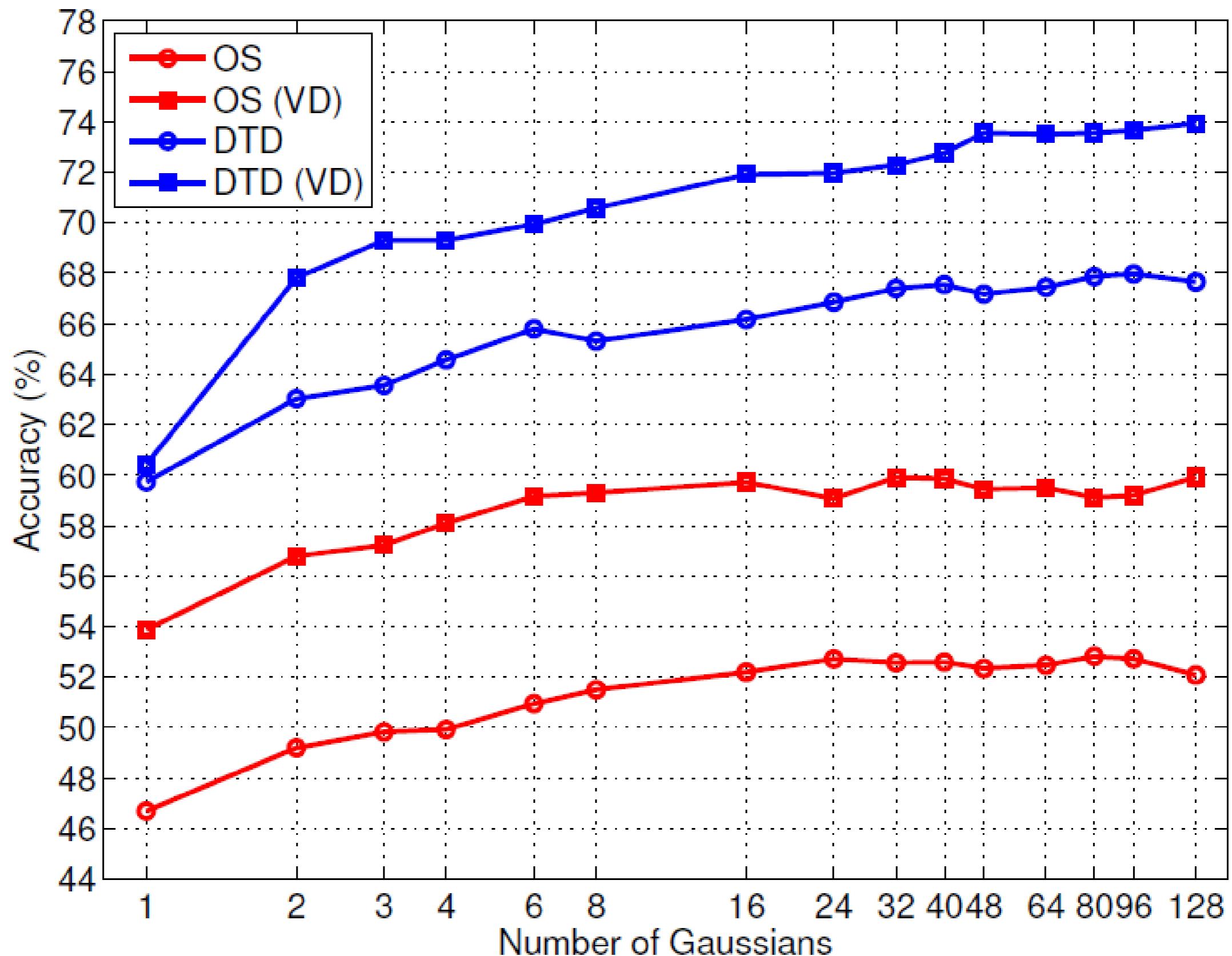
- ▶ Simple
- ▶ Excellent performance in diverse domain
- ▶ Works particularly well and efficiently with image regions
- ▶ Reduces the domain gap in transfer learning

A new **benchmark** for material and texture attribute recognition in clutter

Many more experiments in the paper, IJCV version, and DPhil thesis

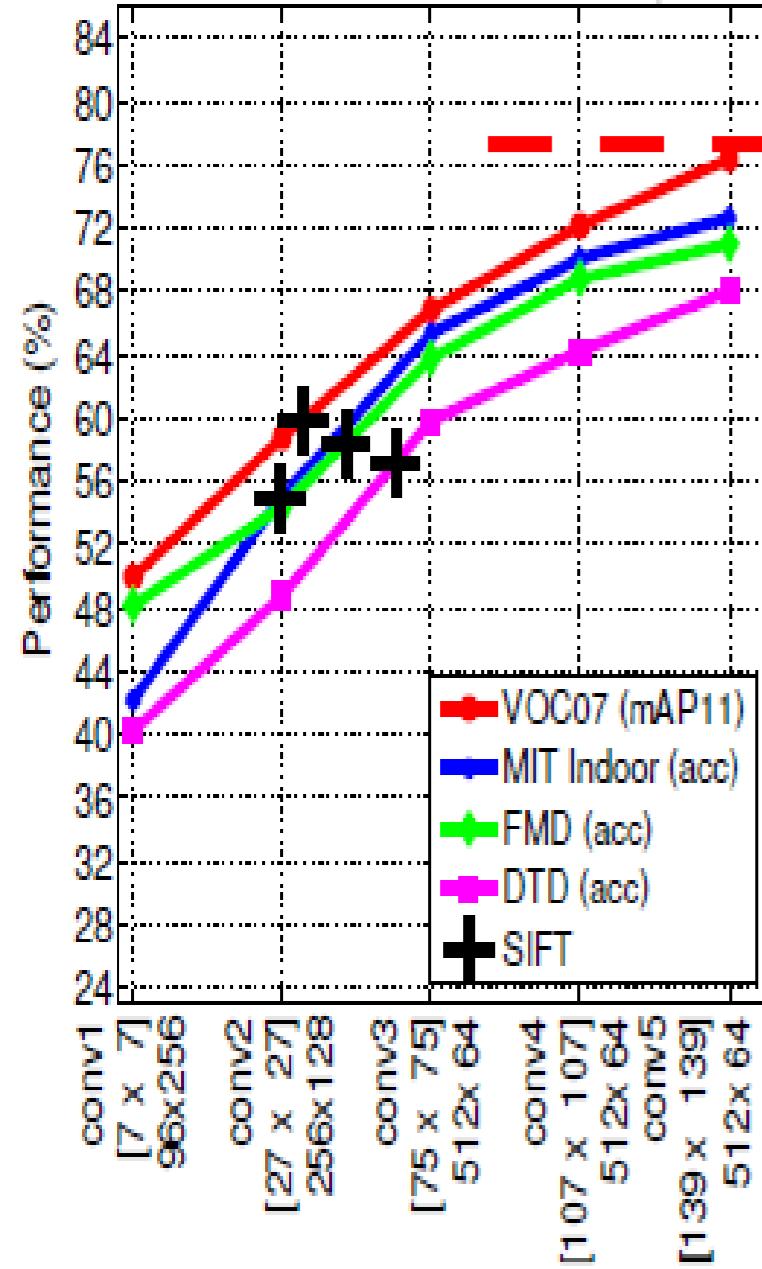
Number of Gaussians

53

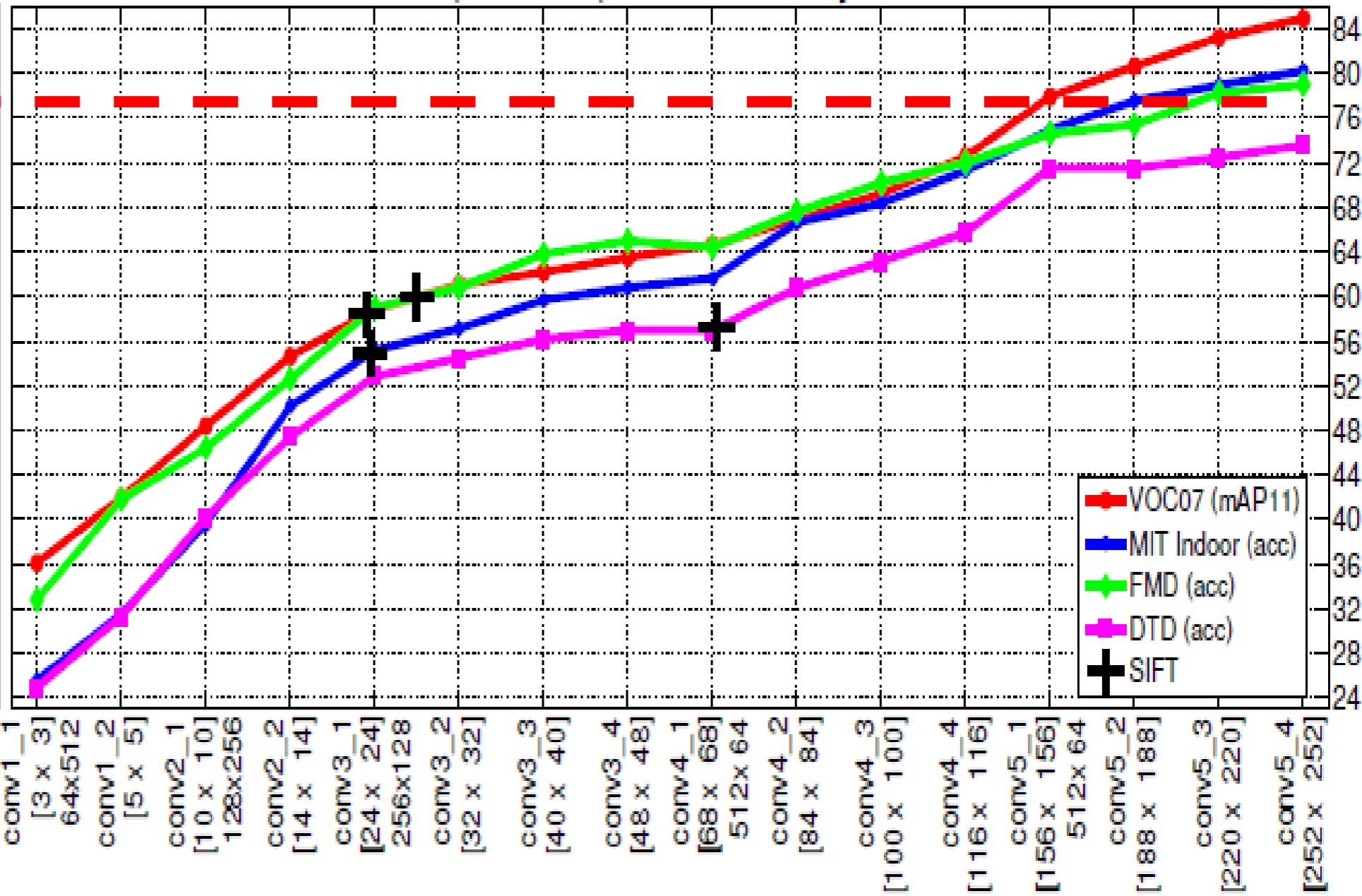


Effect of Depth on CNN Features

(VGG-M) Filterbank Analysis



(VGG-VD) Filterbank Analysis

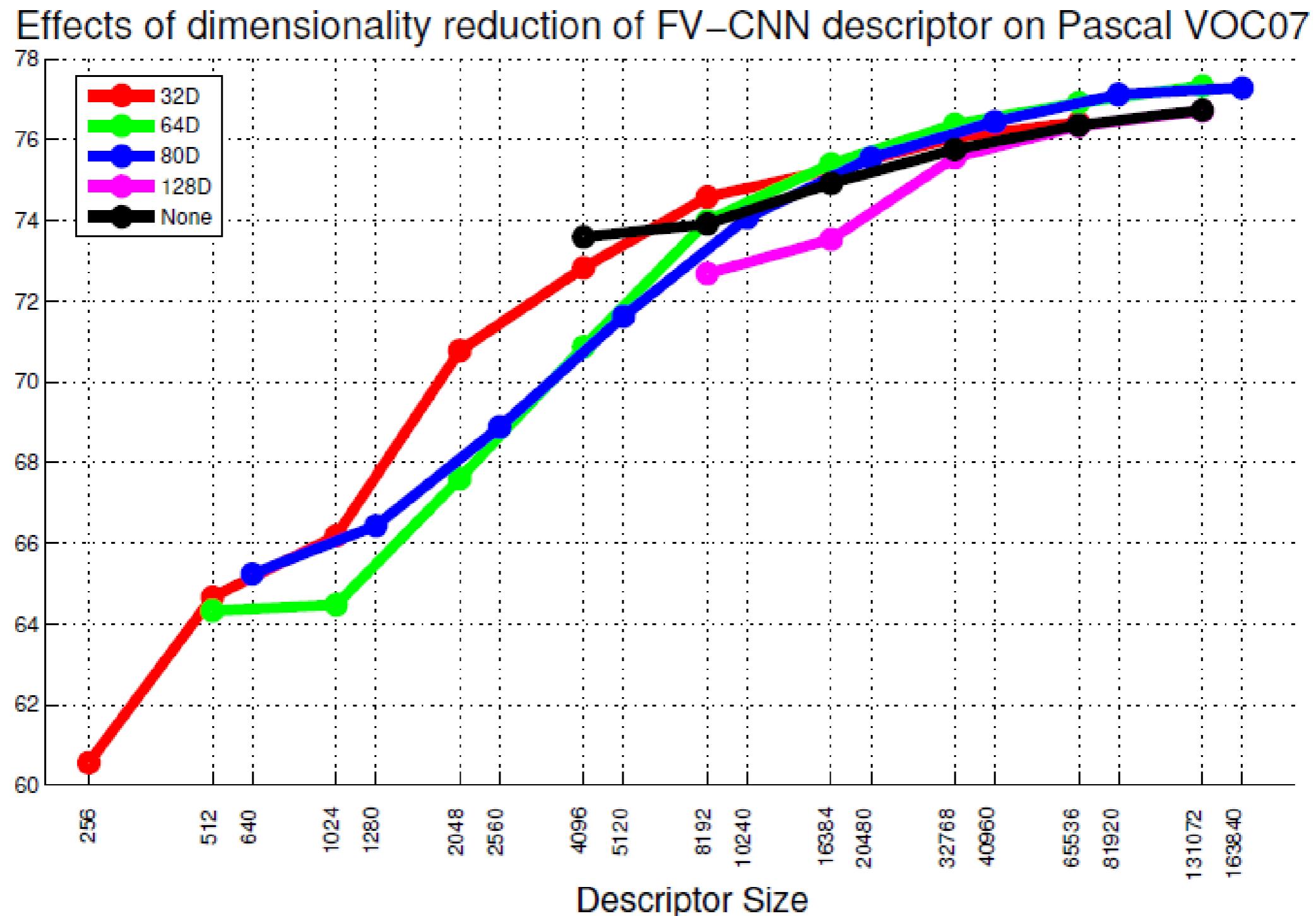


Conv5 for VGG-VD – extra 4%

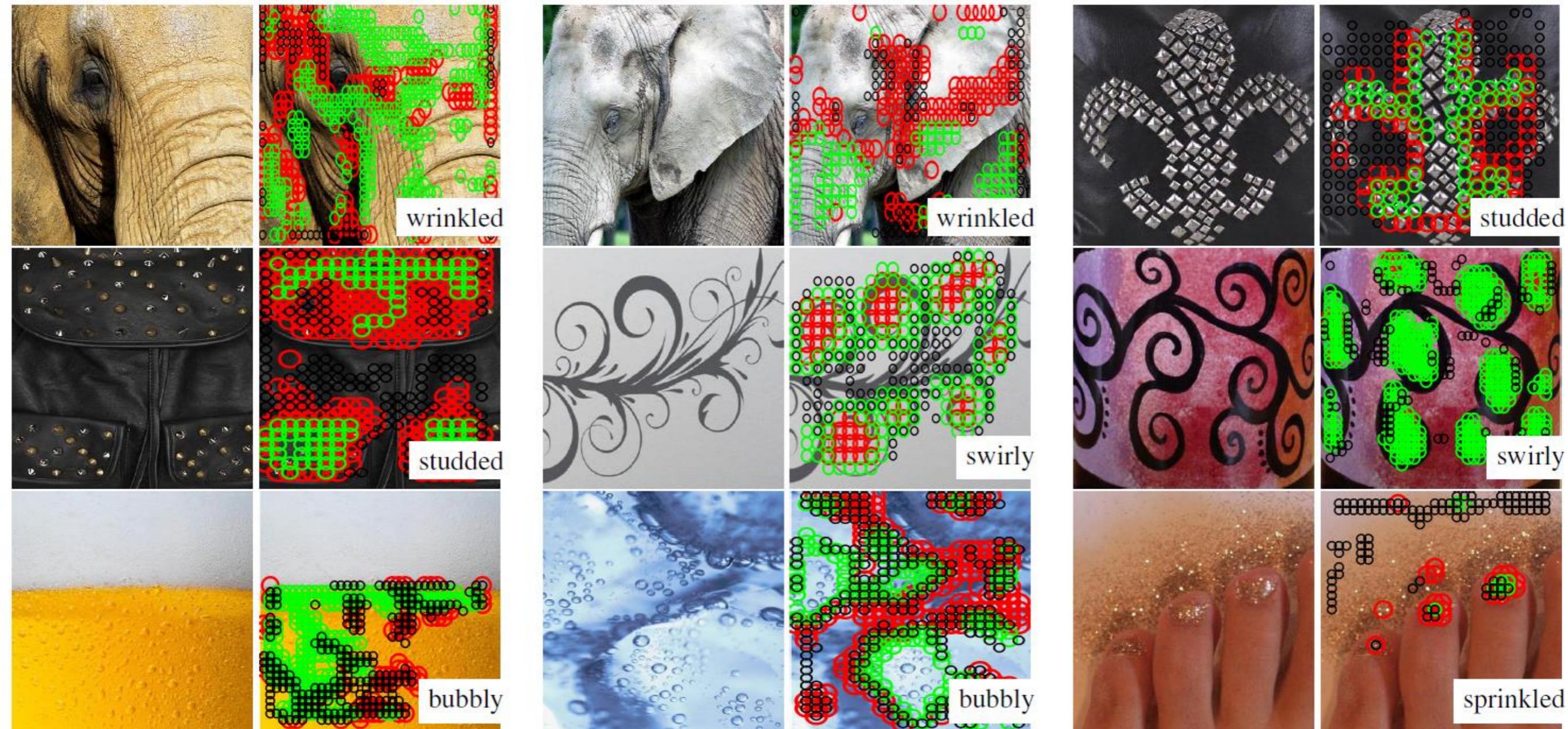
SIFT – same as Conv2 / Conv3

Dimensionality reduction and descriptor size

55



Visualizing top FV components



Locations of CNN descriptors that correspond to the FV-CNN components most strongly associated with the texture words (bubbly, studded, wrinkled ...)