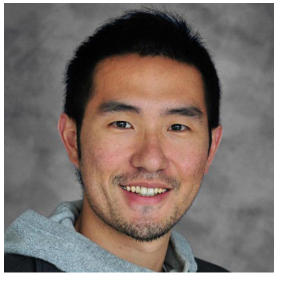
CNN^2: Viewpoint Generalization via a Binocular Vision



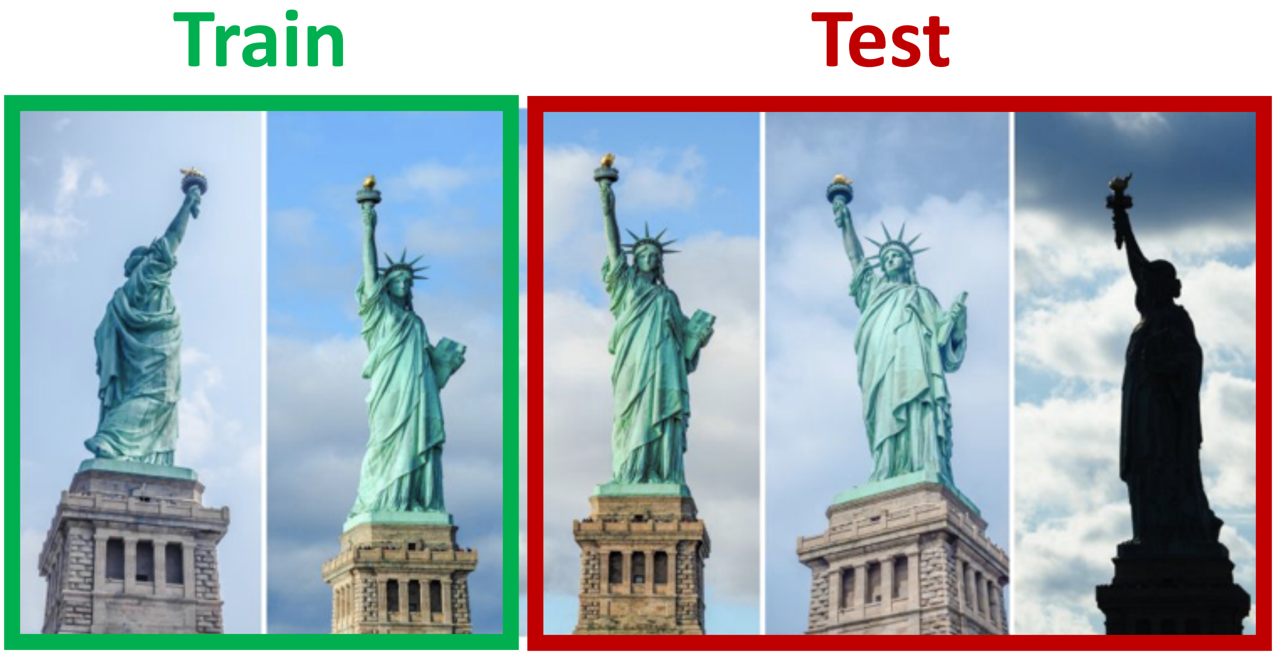
Wei-Da Chen and **Shan-Hung Wu**

CS Department, National Tsing-Hua University
Taiwan, R.O.C.
wdchen@datalab.cs.nthu.edu.tw, shwu@cs.nthu.edu.tw

On Generalizability of CNNs

- The Convolutional Neural Networks (CNNs) have laid the foundation for many techniques in various applications
- However, the ***3D viewpoint generalizability*** of CNNs is still far behind human's visual capabilities

3D Viewpoint Generalizability



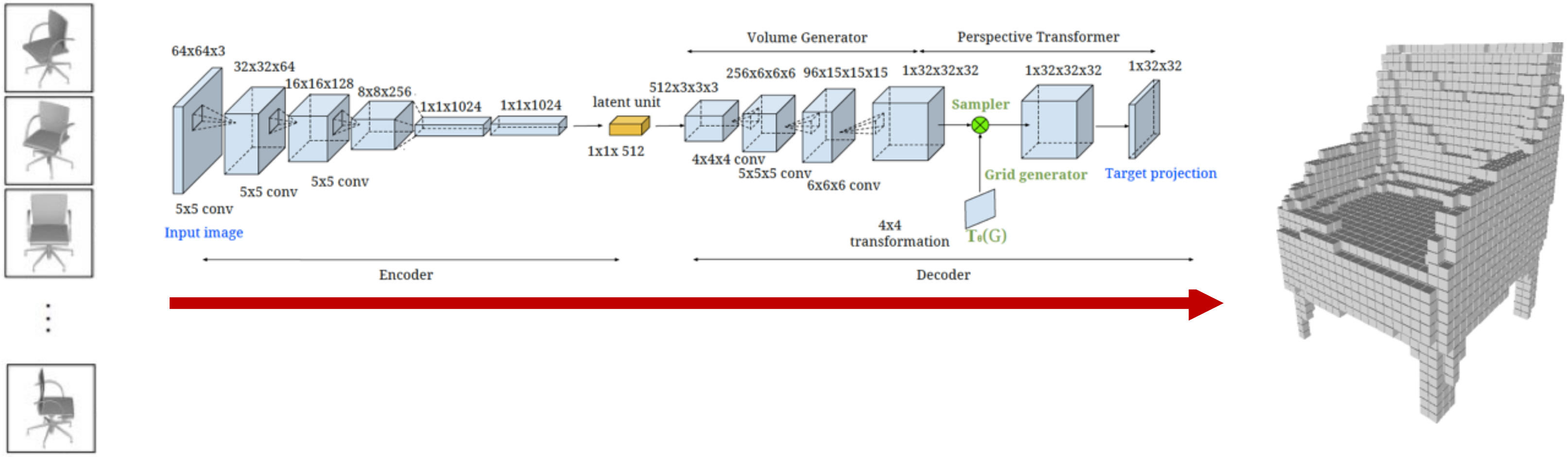
- Humans can recognize objects at ***unseen*** angles
- But CNNs cannot

Outline

- Related work
- CNN^2
  - Dual feedforward pathways
  - Dual parallax augmentation
  - Concentric Multiscale (CM) pooling
- Experiments

Voxel-Reconstruction Methods

- E.g., the Perspective Transformer Networks (PTNs) by Yan et al. 16

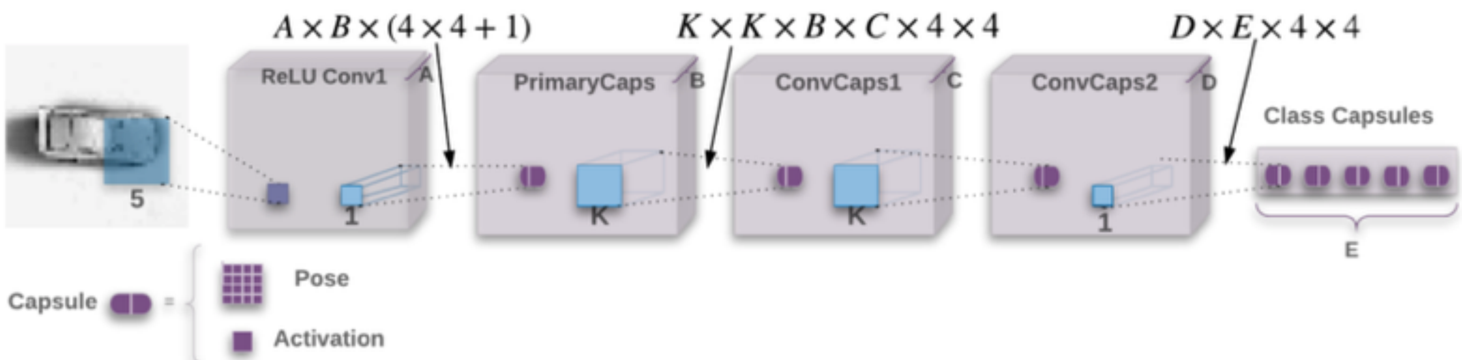


- Learn 3D models directly

Cons

- Require either
  - Voxel-level supervision, or
  - Omnidirectional images as input
- Both are expensive to collect in practice

CapsuleNets (Hinton et al. 17, 18)



- Different capsules are organized in a parse tree where lower-level capsules are dynamically routed to upper-level capsules using an agreement protocol
- When viewpoint changes, the “routes” will change in a coordinate way

But…

- People found that CapsuleNets are hard to train
  - Capsules increase the number of model parameters
  - Iterative routing-by-agreement algorithm is time-consuming
  - Does not ensure the emergence of a correct parse tree (Peer et al. 18)
- ***Not compatible*** with CNNs
  - and therefore cannot benefit the rich CNN ecosystem

Outline

- Related work
- CNN^2
  - Dual feedforward pathways
  - Dual parallax augmentation
  - Concentric Multiscale (CM) pooling
- Experiments

Our Goals

- A new model that
  - has improved 3D viewpoint generalizability
  - does not require expensive input and supervision
  - is CNN compatible

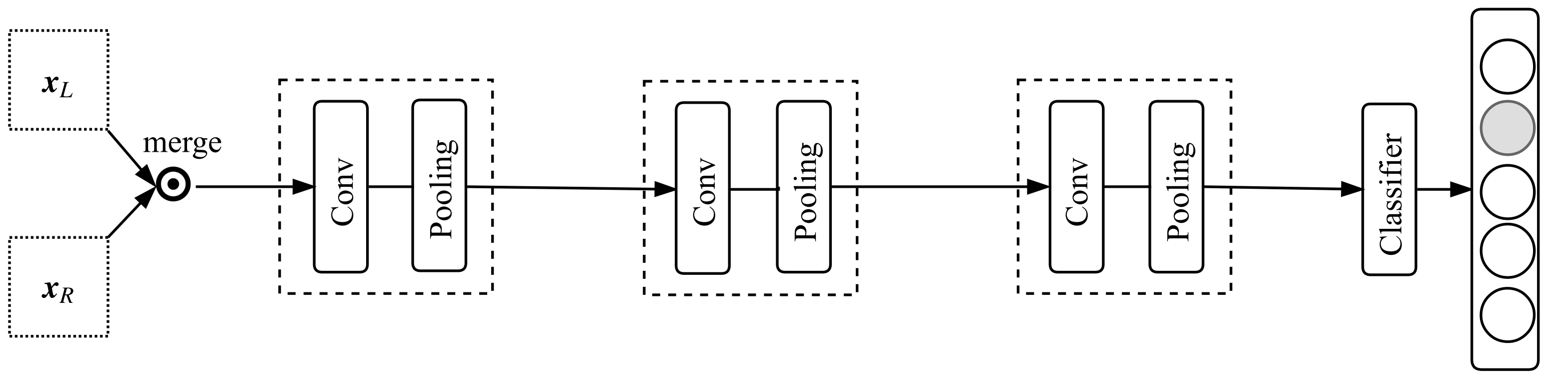
Observation:

Humans understand the world using two eyes!

Binocular Images

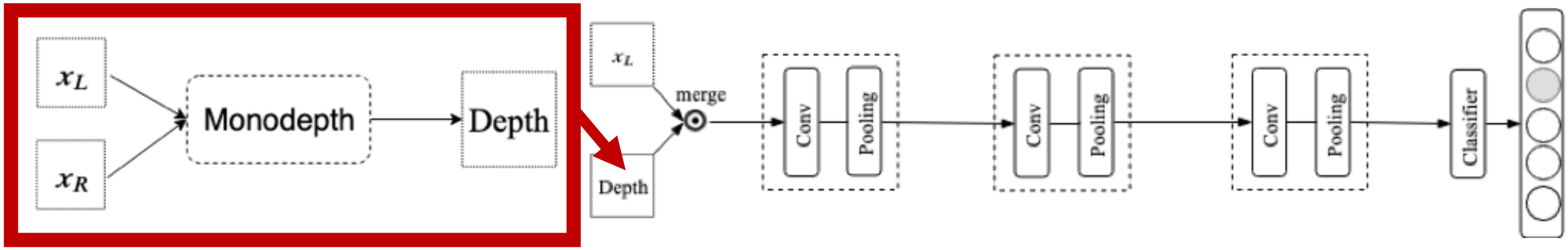
- Today, binocular images can be easily collected
- Majority of people are using their smartphones, which are now usually equipped with dual or more lens
- One can also extract two nearby frames in online videos to construct a large binocular image dataset

Binocular Solution 1 (LeCun et al. 14)



- Stacks up two binocular images along the channel dimension and then feeds them to a regular CNN
- But don’t model any prior of binocular vision

Binocular Solution 2:
Sol. 1 + Monodepth (Godard et al. 17)

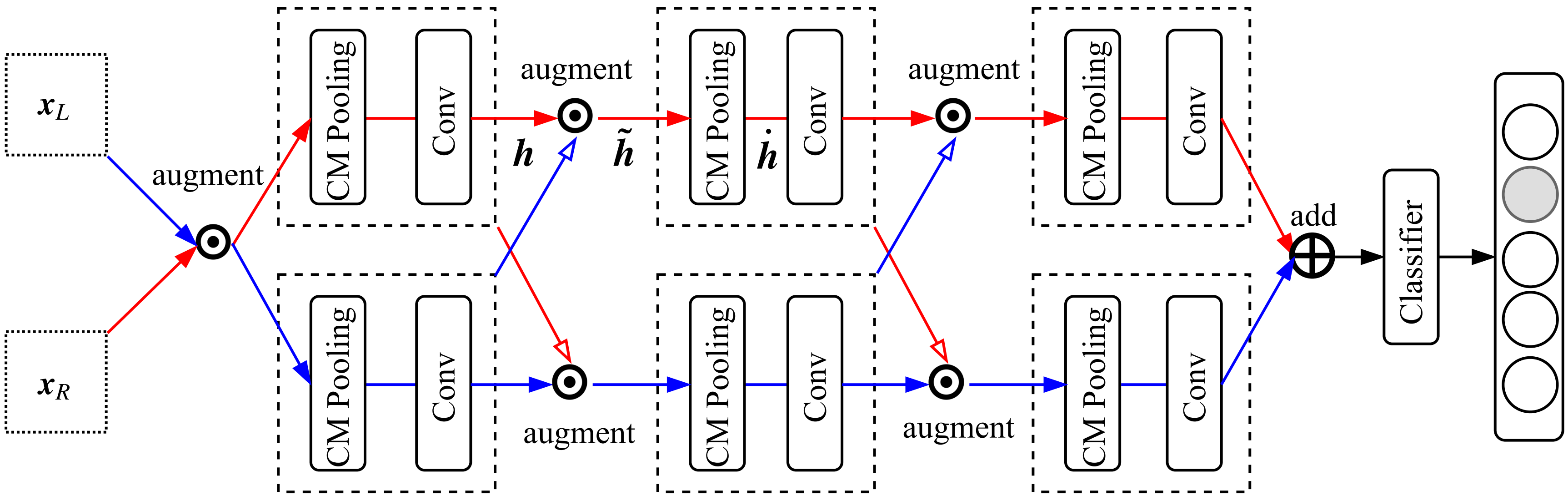


- Calculate the depth map explicitly, then add it as additional input channels

However…

- The depth information is only a subset of the knowledge that can be learned from binocular vision
- Studies in neuroscience have found out that human's visual system can detect
  - Stereoscopic edges (Von Der Heydt et al. 00)
  - Foreground and background (Qiu andVon Der Heydt 05; Maruko et al. 08)
  - Illusory contours of objects (Von der Heydt et al. 1984; Anzai et al. 07)

Our Solution: CNN^2

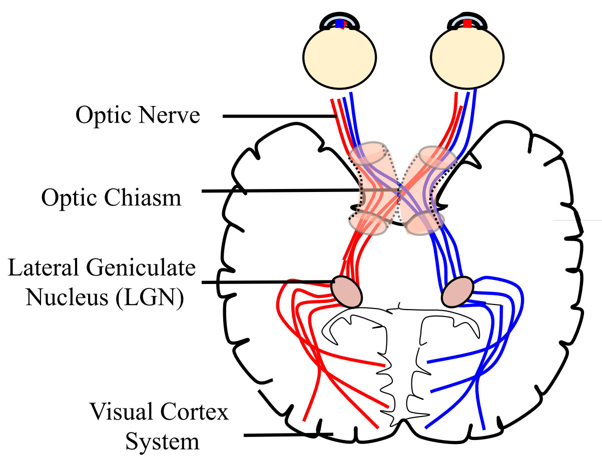


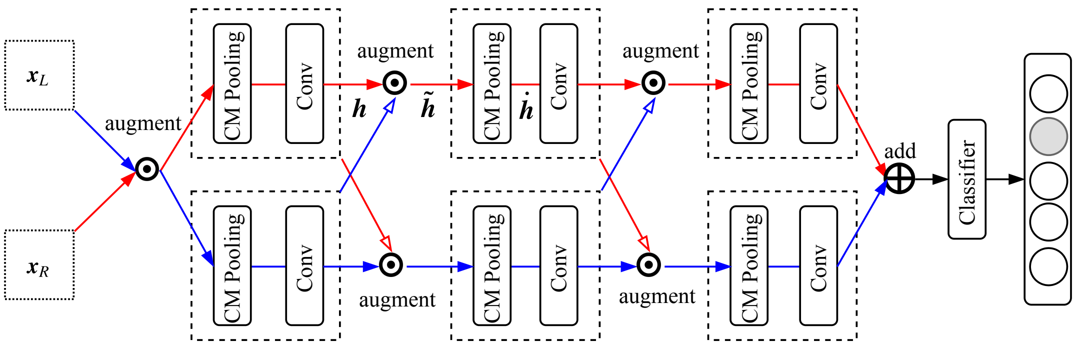
- Dual feedforward pathways
- Dual parallax augmentation
- Concentric Multiscale (CM) pooling

Outline

- Related work
- CNN^2
  - Dual feedforward pathways
  - Dual parallax augmentation
  - Concentric Multiscale (CM) pooling
- Experiments

Dual Feedforward Pathways

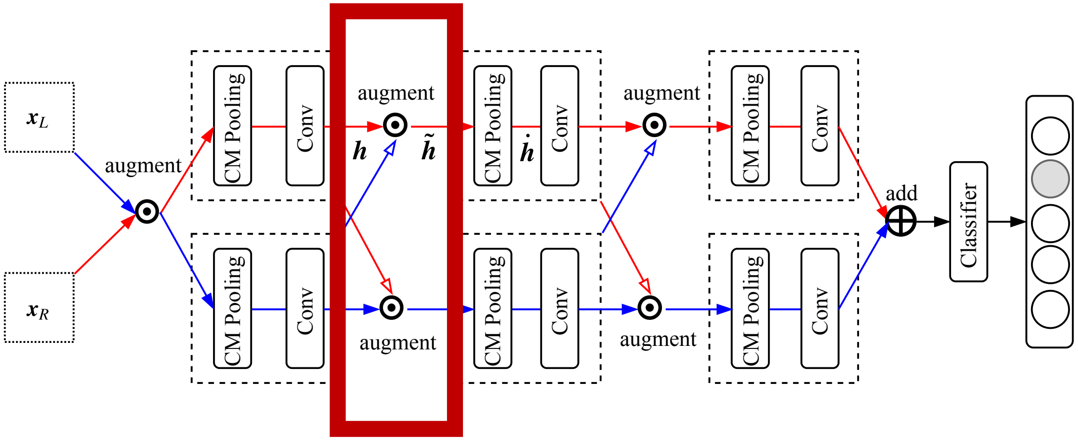




- Humans visual system at left and right sides of the brain are known to have bias (Gotts et al. 13)
- Filters/kernels in the left and right pathways can learn different (biased) features

Outline

- Related work
- CNN^2
  - Dual feedforward pathways
  - Dual parallax augmentation
  - Concentric Multiscale (CM) pooling
- Experiments



Dual Parallax Augmentation (1/2)

Left path:

W X H X C   W X H X C   W X H X C   W X H X ***2C***

$$h_L \ \textbf{concat}\left(h_R - h_L \right) = \tilde{h}_L$$

Right path:

W X H X C   W X H X C   W X H X C   W X H X ***2C***

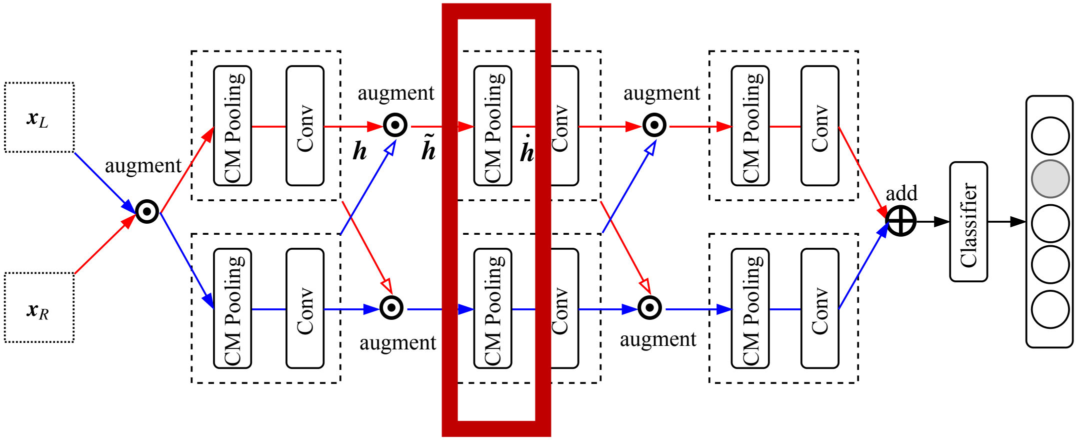
$$h_R \ \textbf{concat}\left(h_L - h_R \right) = \tilde{h}_R$$

Dual Parallax Augmentation (2/2)

- Allows the filters/kernels in convolutional layers to recursively detect ***stereoscopic features at different abstraction levels*** by looking into the parallax
- The small differences between the two input images at the pixel level and at shallow layers may add up to a big difference at a deeper layer

Outline

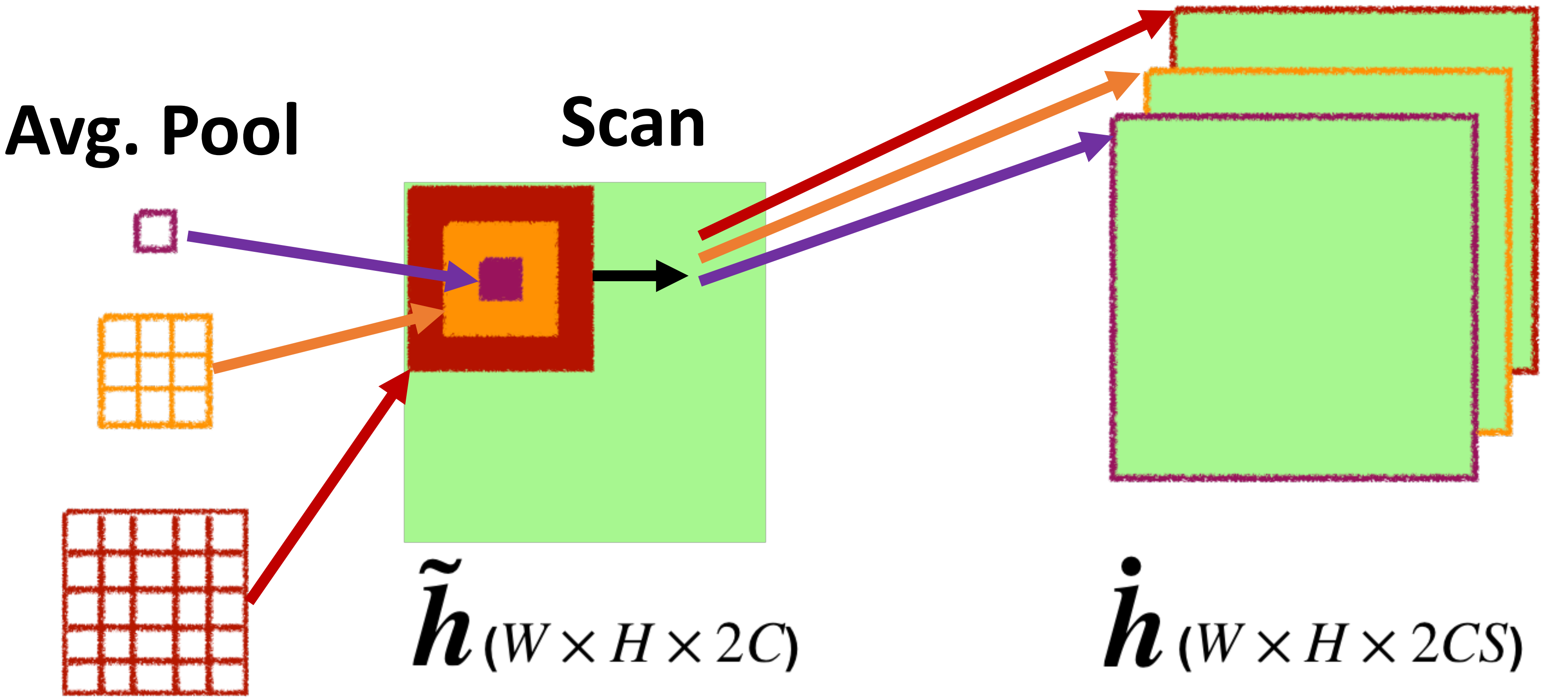
- Related work
- CNN^2
  - Dual feedforward pathways
  - Dual parallax augmentation
  - **Concentric Multiscale (CM) pooling**
- Experiments



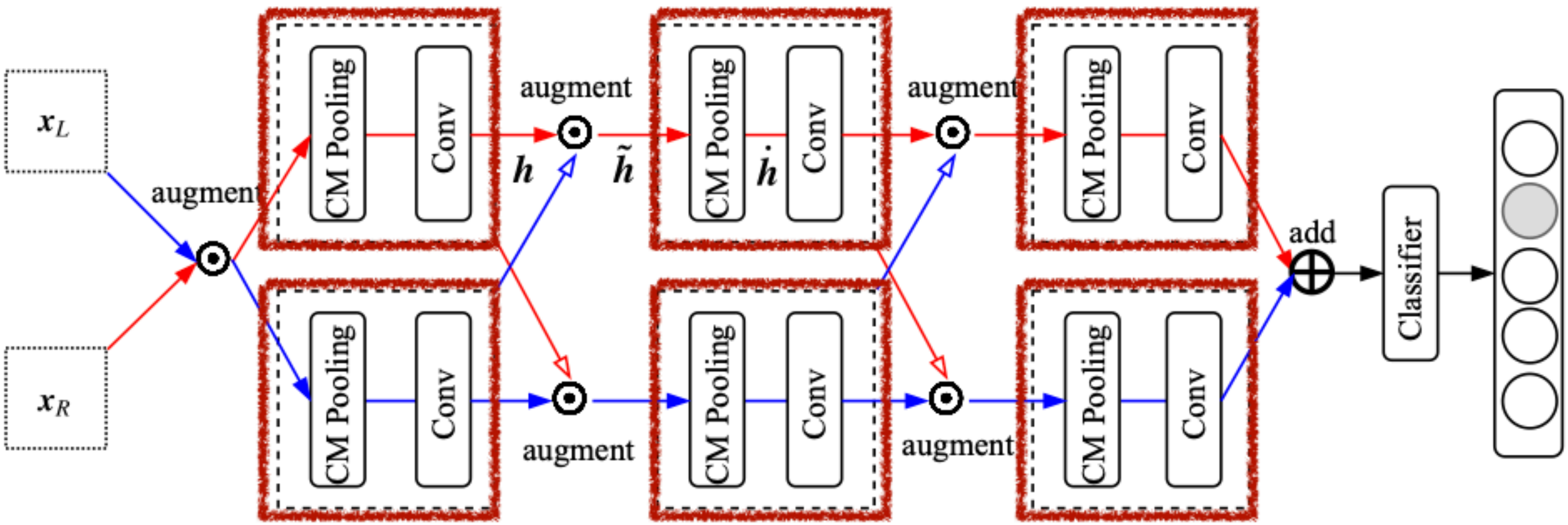
Concentric Multiscale (CM) Pooling (1/2)

- Areas that are out of focus are blurred

Concentric Multiscale (CM) Pooling (2/2)



Placed *Before* Convolution



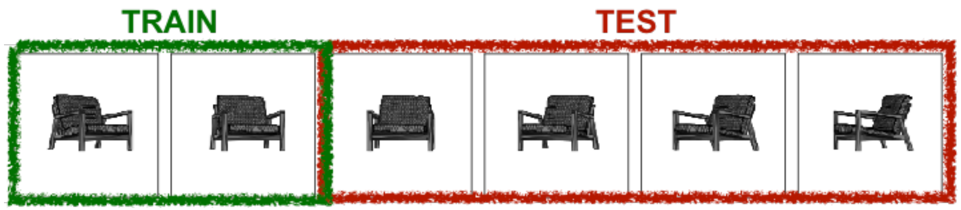
- Allows filters/kernels to contrast blurry features with clear features

Outline

- Related work
- CNN^2
  - Dual feedforward pathways
  - Dual parallax augmentation
  - Concentric Multiscale (CM) pooling
- Experiments

Datasets

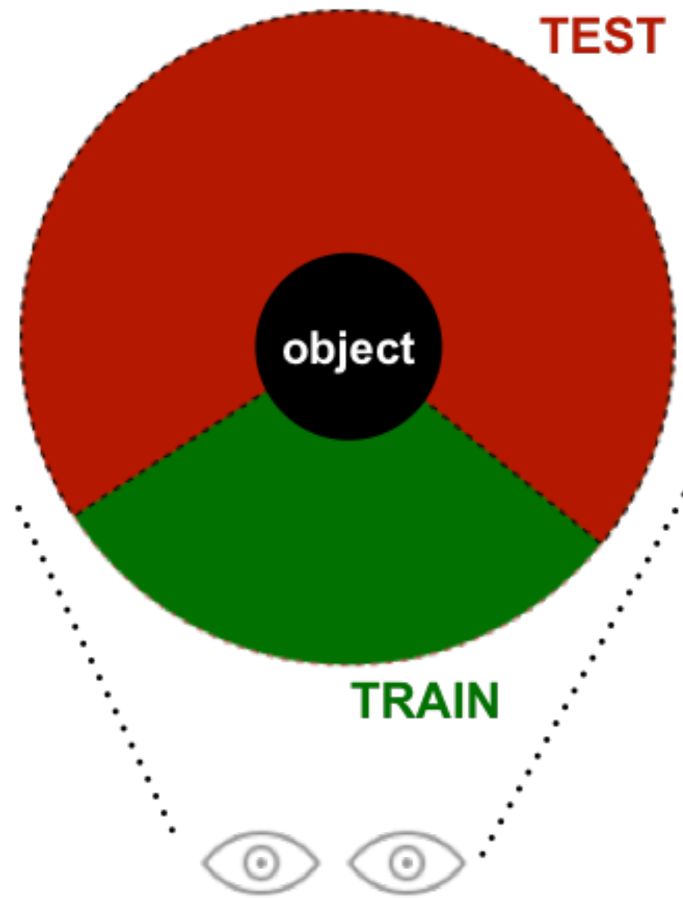
- ModelNet2D (gray scale)



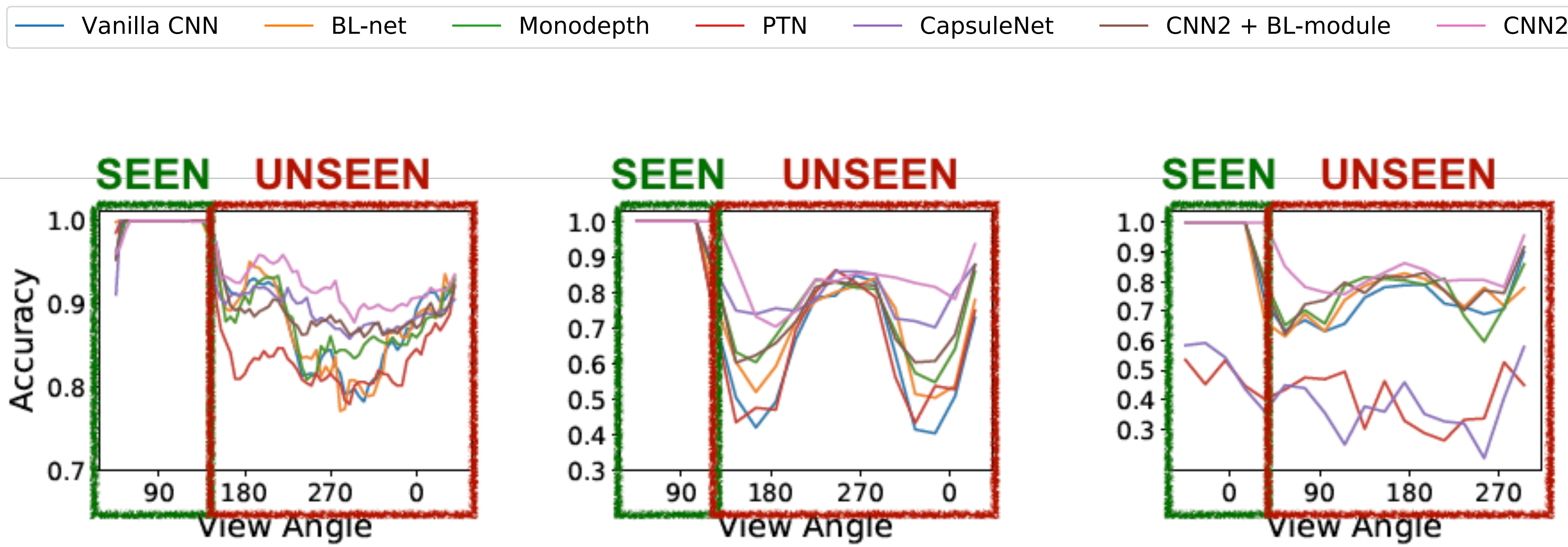
- SmallNORB (gray scale)

- RGBD-Object (RGB)

Train/Test Setting

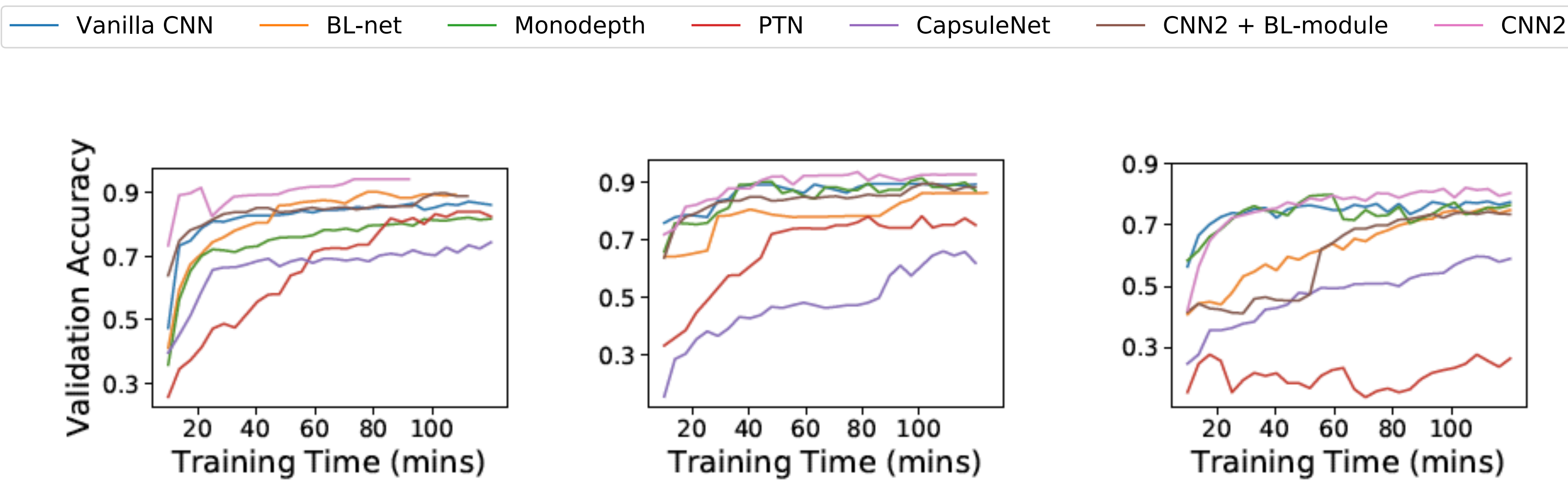


3D Viewpoint Generalization



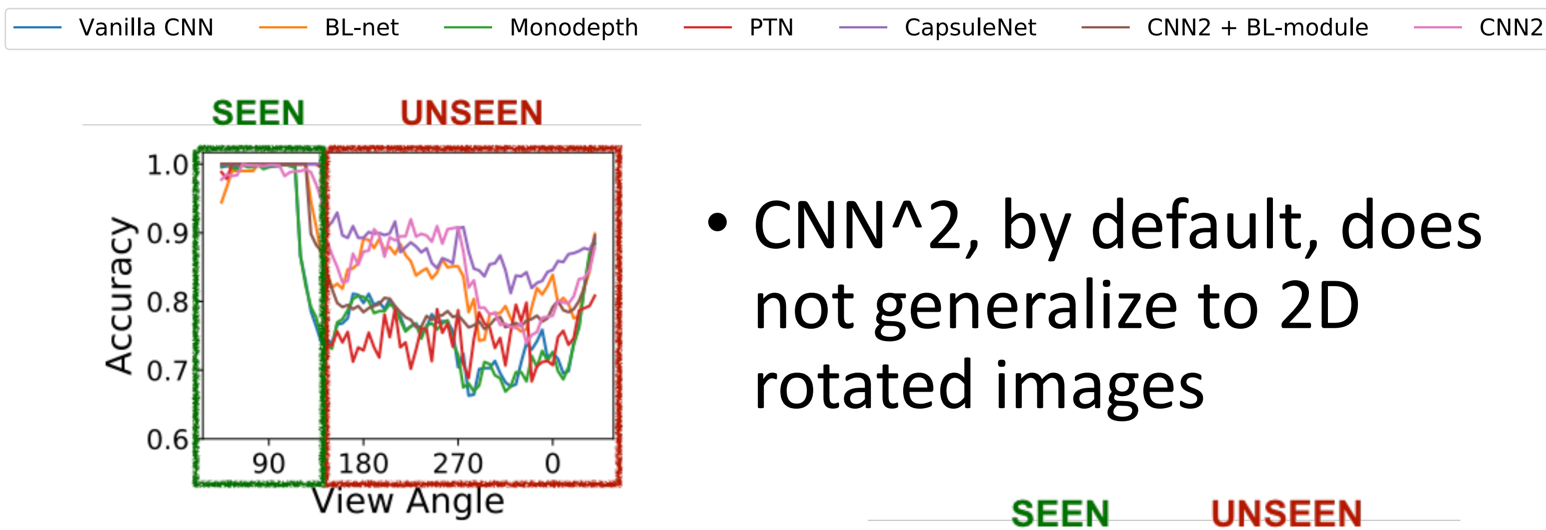
(a) ModelNet2D (b) SmallNORB (c) RGB-D Object

Learning Efficiency



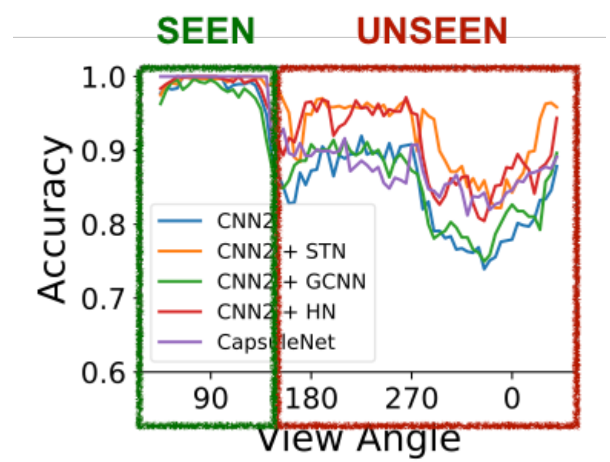
(d) ModelNet2D (e) SmallNORB (f) RGB-D Object

Backward Compatibility



- CNN^2, by default, does not generalize to 2D rotated images

- But can be enhanced by existing works on 2D rotation generalizability



Takwaways

- We propose CNN^2 that
  - gives improved 3D viewpoint generalizability
  - does not require expensive input or supervision
  - is compatible with CNNs and can benefit the rich CNN ecosystem
- Detects stereoscopic features beyond depth via:
  - Dual feedforward pathways
  - Dual parallax augmentation
  - Concentric Multiscale (CM) pooling

  from binocular images