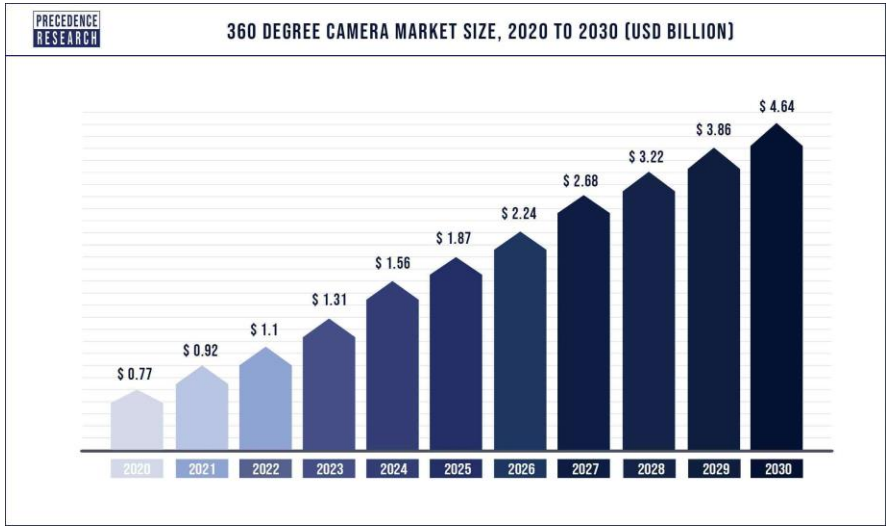
Learning for 360° Vision

Yu-Chuan Su
Research Scientist
Google

Google

Growing Popularity of 360° Media

Expected market growth



[Precedence Research]

Applications

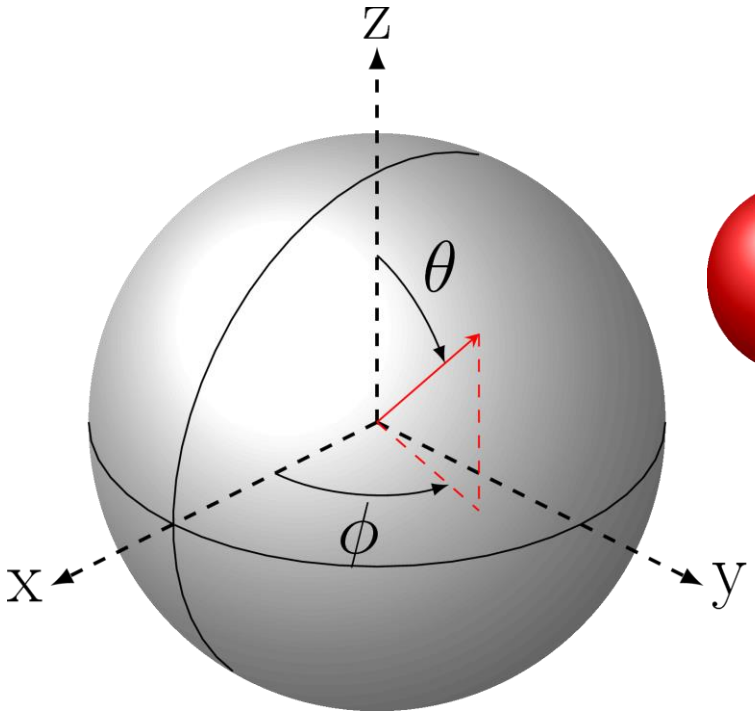
Virtual / Augmented Reality

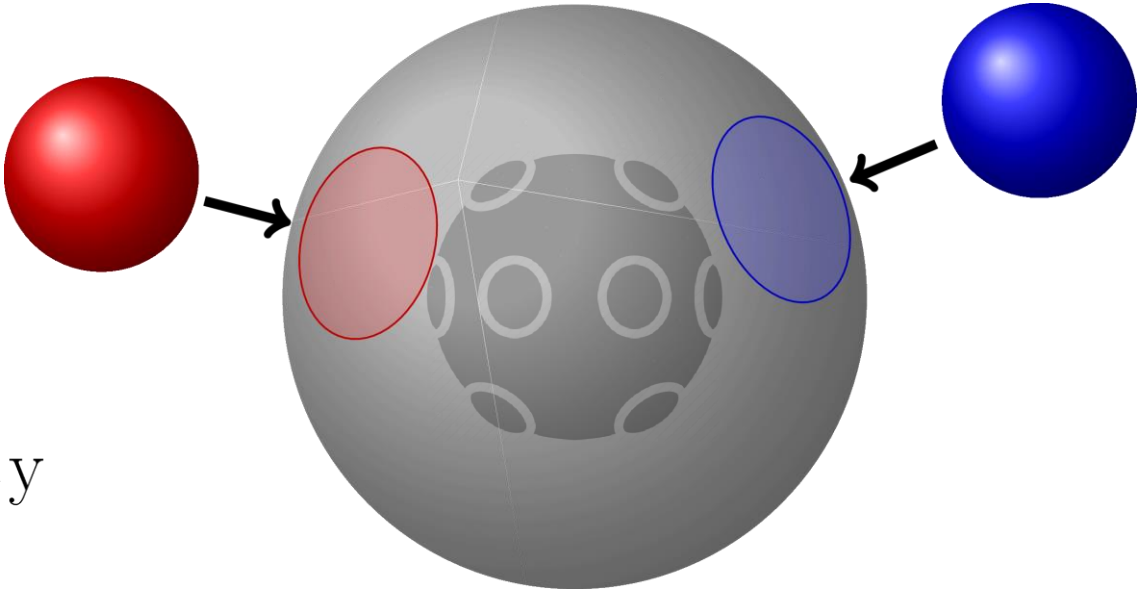
Photography

Drone

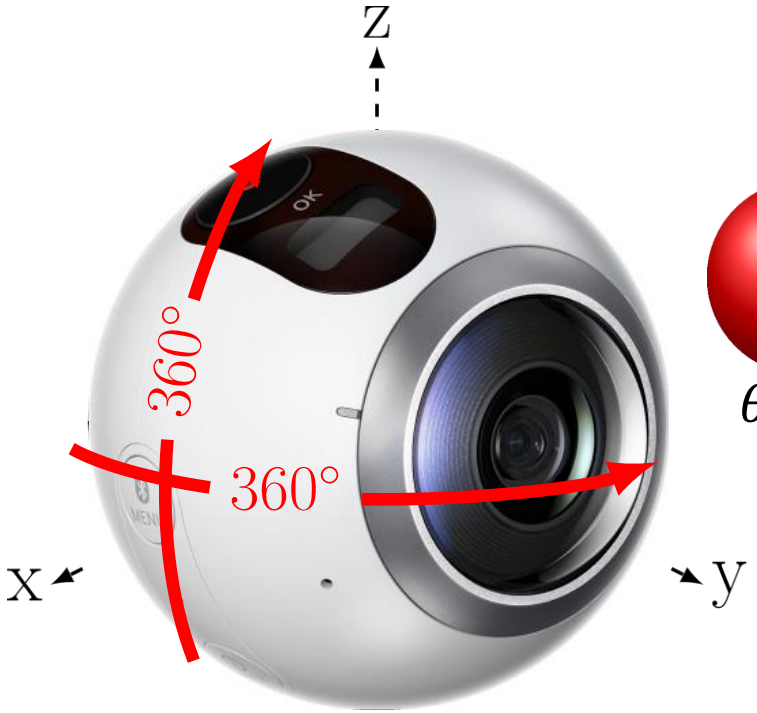
Surveillance

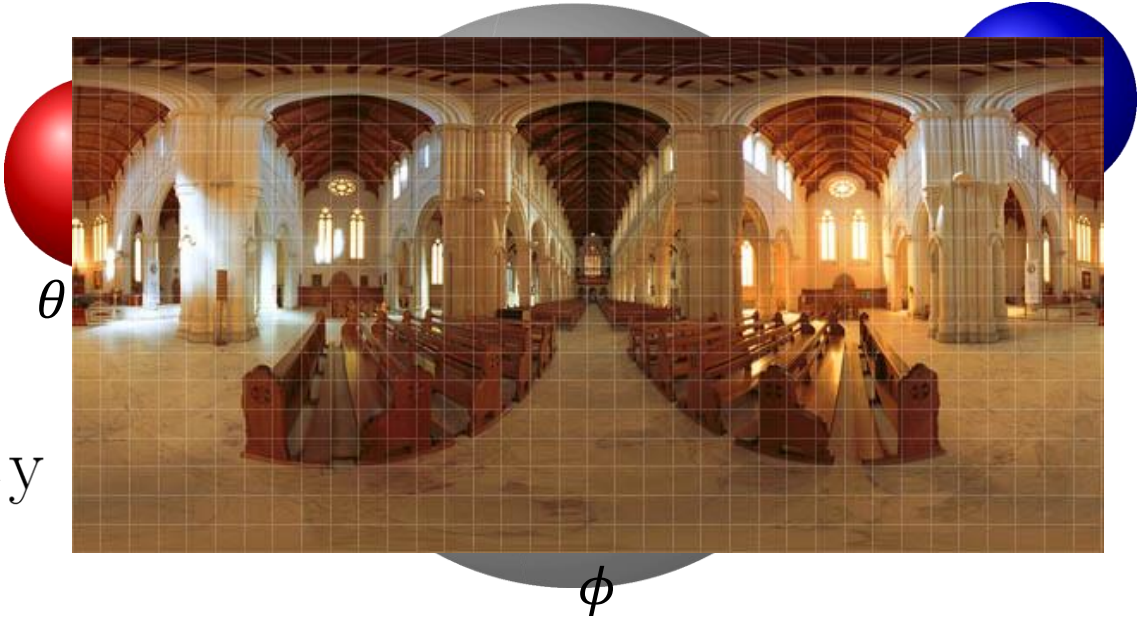
360° Image vs. Perspective Image





360° Image vs. Perspective Image





Visual Recognition on 360° Imagery

- Applications depend on visual recognition

  [Zhang et al., ECCV 2014], [Hu et al., CVPR 2017], [Lai et al., TOG 2017], [Chou et al., AAAI 2018], [Yu et al., AAAI 2018], [Lee et al., CVPR 2018]

- CNNs for spherical data

  [Boomsma and Frellsen, NeurIPS 2017], [Cohen et al., ICLR 2018], [Cheng et al., CVPR 2018], [Esteves et al., ECCV 2018], [Coors et al., ECCV 2018], [Zhang et al., ECCV 2018], [Khasanova and Frossard, ICML 2019]

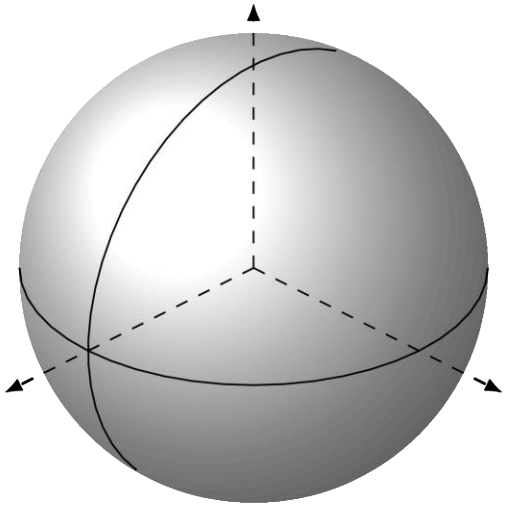
  - New architectures designed specifically for spherical data
  - **Not data efficient** – require annotations in spherical formats

Applying CNNs to 360° Imagery

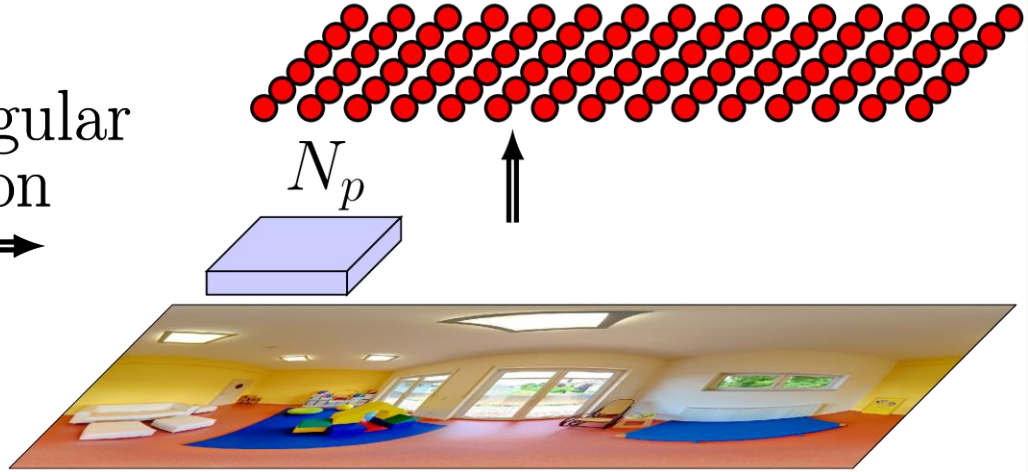
Strategy I (Equirectangular)

[Hu CVPR 2017], [Lai TOG 2017]

- **Fast** – single projection
- **Inaccurate** – distortion in projection



Equirectangular Projection



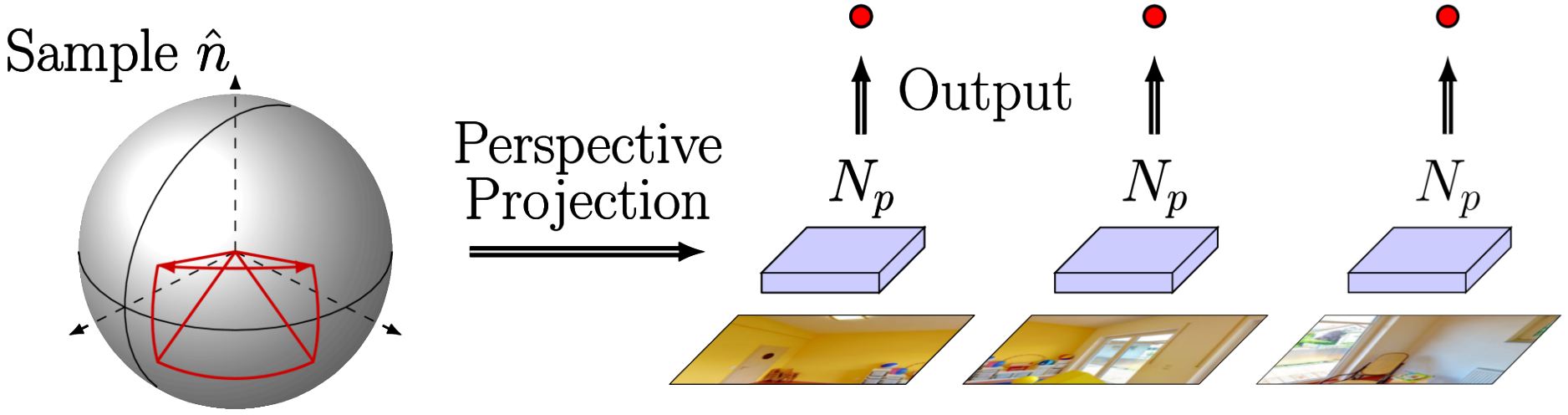
Google

Applying CNNs to 360° Imagery

Strategy II (Perspective)

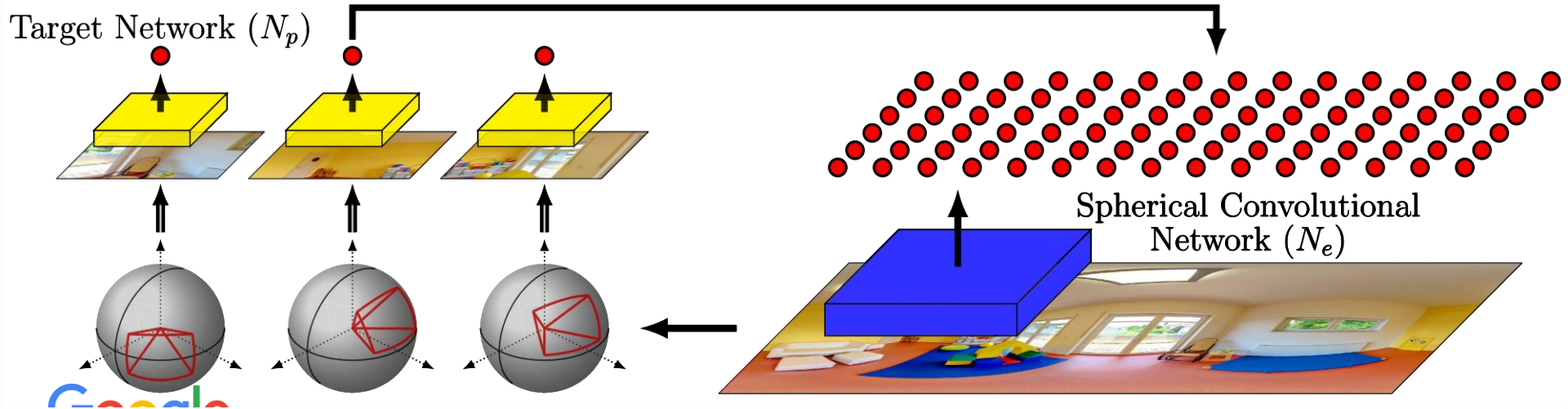
[Zhang ECCV 2014], [Chou AAAI 2018], [Yu AAAI 2018]

- **Accurate** – exact convolution output
- **Slow** – repeated projection



Spherical Convolution

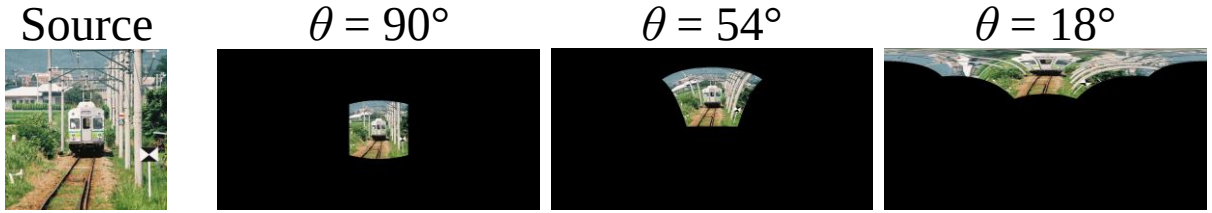
- **Fast** – single equirectangular projection
- **Accurate** – simulate exact convolution output
- **Data efficient** – does not require additional annotations



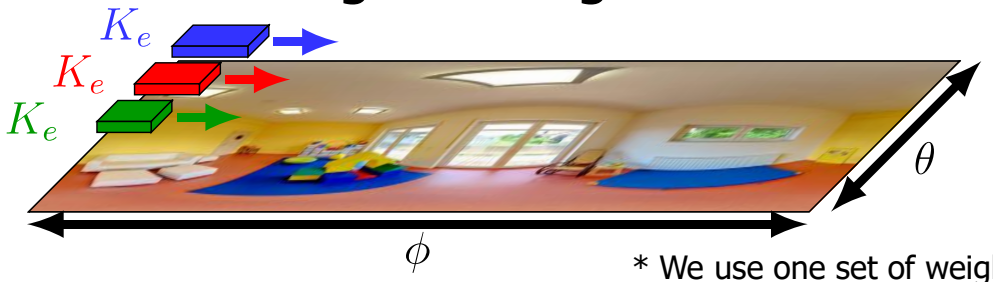
Google

Spherical Convolution Model Architecture

- Goal – account for distortion
- Problem – distortion is polar angle *(θ)* / row dependent



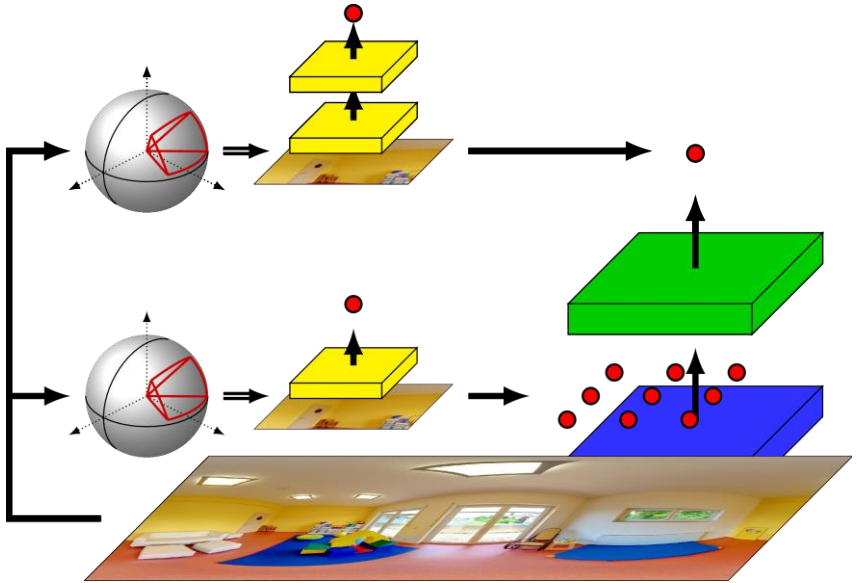
- Solution – untie kernel weights along the rows*



* We use one set of weights every 5 rows in practice

Layer-wise Training

- Goal – accelerate training
- Idea – require N_e to reproduce all intermediate features

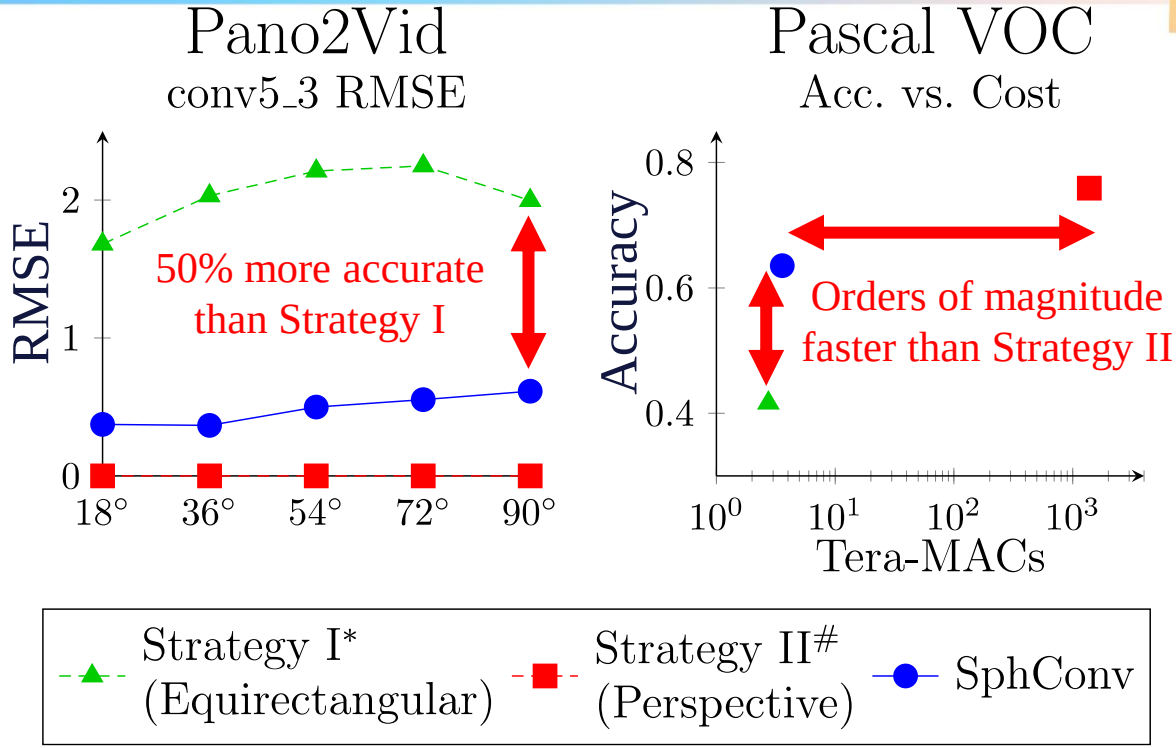


Experiment: Datasets

Apply SphConv to Faster R-CNN

- Pano2Vid (Train)
  - 360° video dataset
  - Sample **1,056** frames for training
- Pascal VOC 2007 (Test)
  - Perspective object detection dataset
  - Project bounding boxes to spheres at five polar angle ($\boldsymbol{\theta}$)

SphConv Evaluation



* [Hu et al., CVPR 2017], [Lai et al., TOG 2017]

\# [Zhang et al., ECCV 2014], [Chou et al., AAAI 2018], [Yu et al., AAAI 2018]

Problem of Spherical Convolution

- Model size increases linearly w.r.t. input resolution (H)

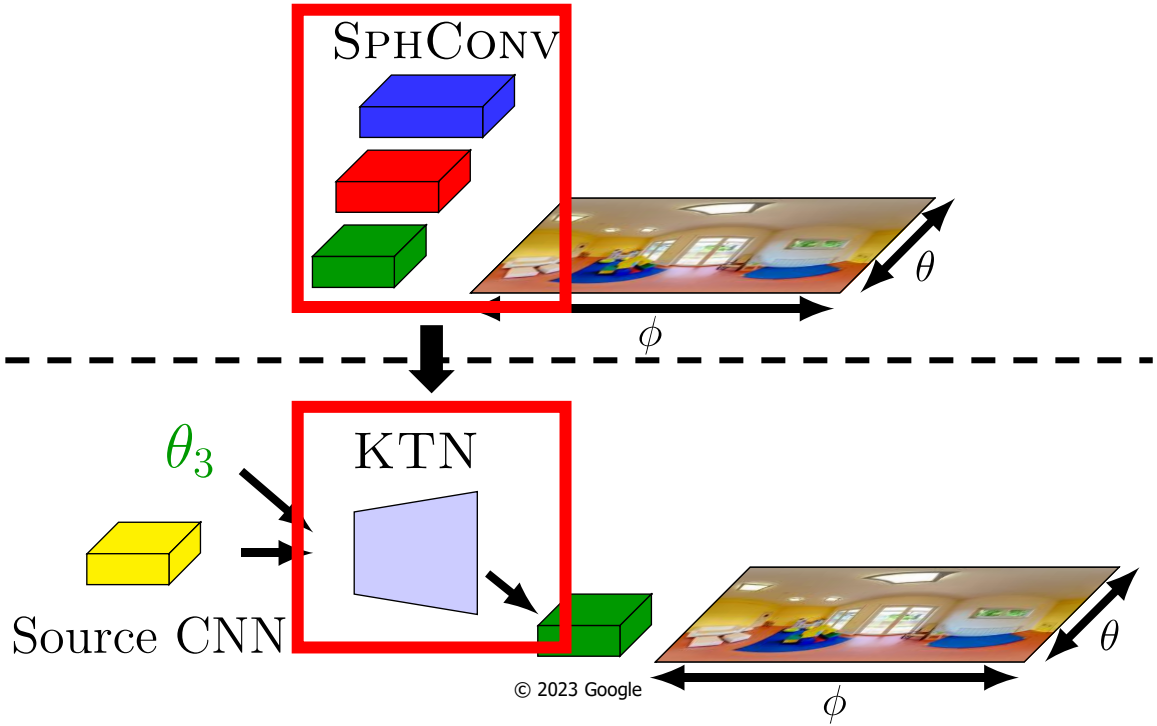
	CNN	SphConv
Asymptotic	c^2k^2	c^2k^2H
VGG*	56 MB	29 GB

* Over 640x320 input images

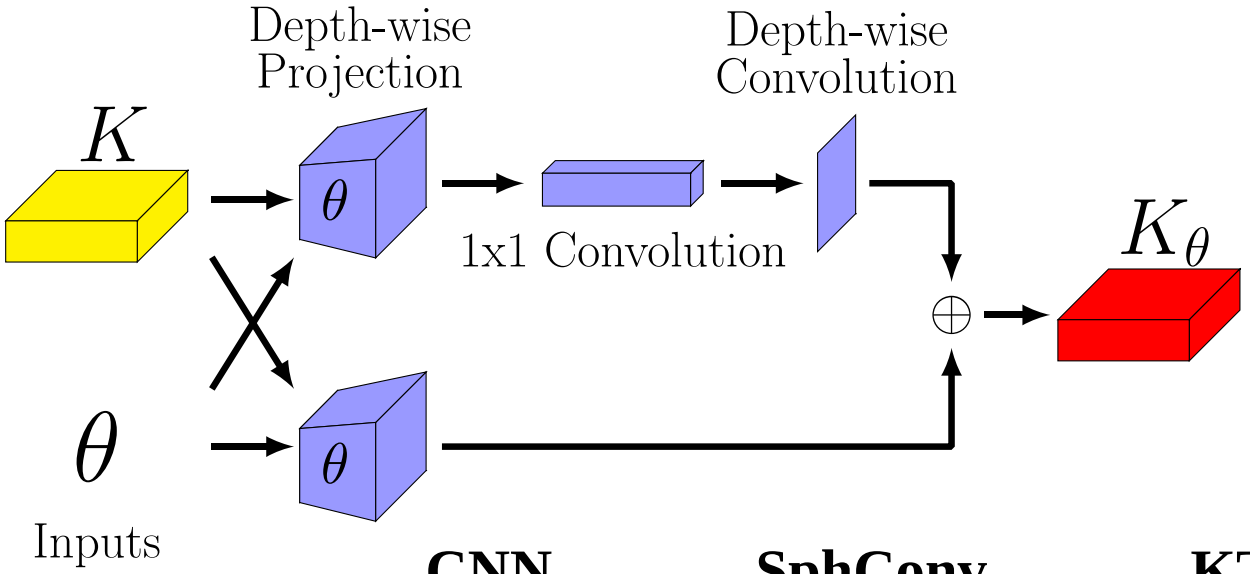
- SphConv kernels are highly correlated → weights are redundant

Kernel Transformer Network

Idea – generates SphConv kernels from source kernels

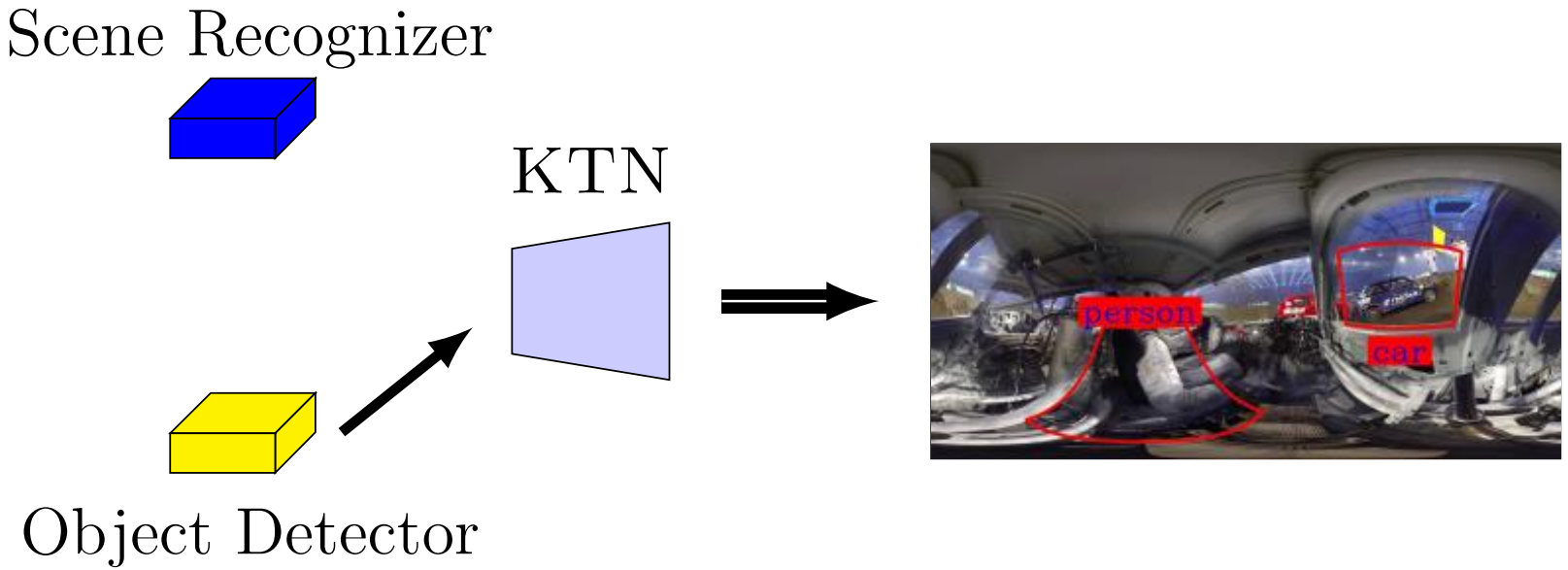


Kernel Transformer Network Architecture



	CNN	SphConv	KTN
Asymptotic	c^2k^2	c^2k^2H	$c^2k^2+c^2$
VGG	56 MB	29 GB	70 MB

Transferability of Kernel Transformer Network



We need only one KTN for different tasks!

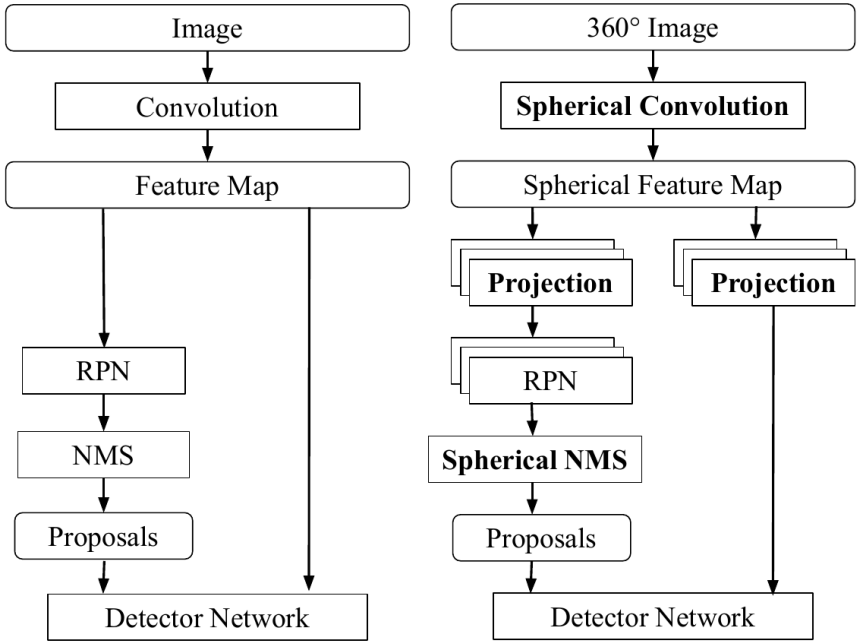
Application: 360° Object Detection

- Appy KTN to Faster R-CNN

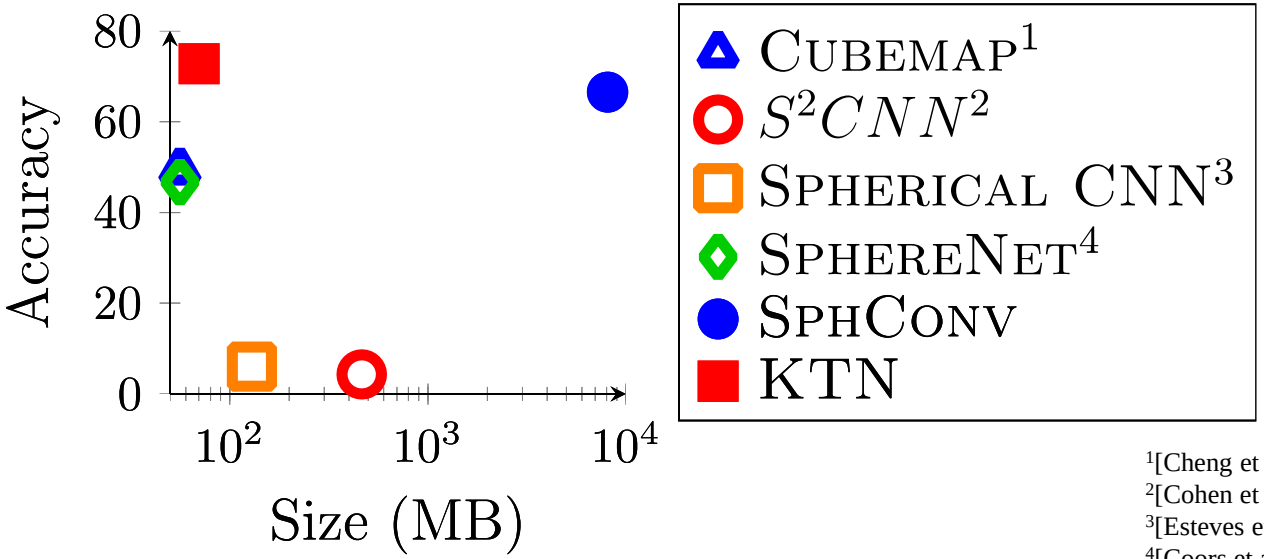
  [Ren et al., NIPS 2015]

  - Extract equirectangular feature map
  - Project features to tangent planes
  - Spherical non-maximum suppression

- Does not need new annotation



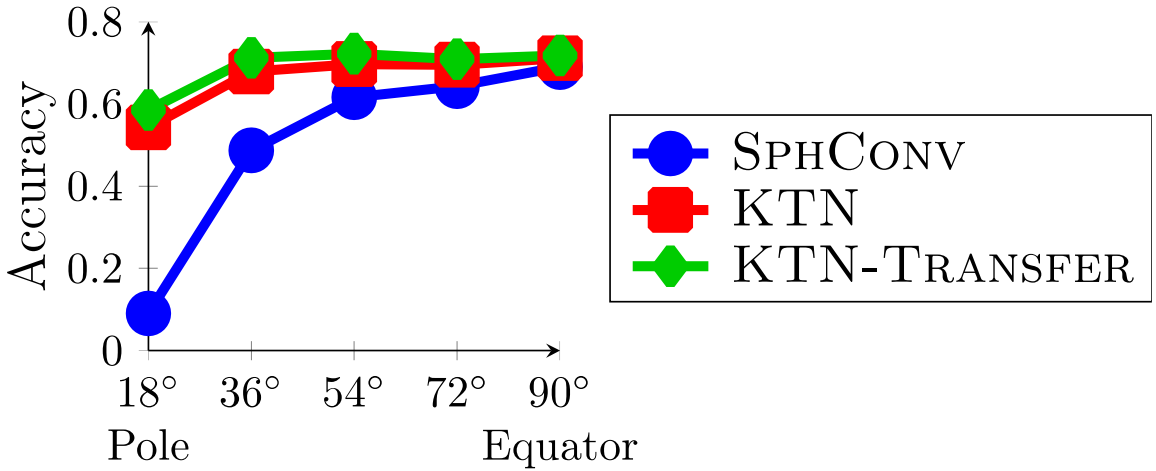
Experiment: Object Detection Accuracy

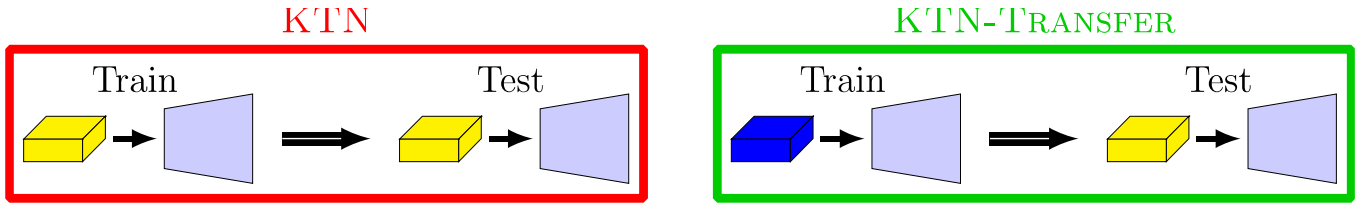


[1][Cheng et al., CVPR 2018],
[2][Cohen et al., ICLR 2018],
[3][Esteves et al., ECCV 2018],
[4][Coors et al., ECCV 2018]

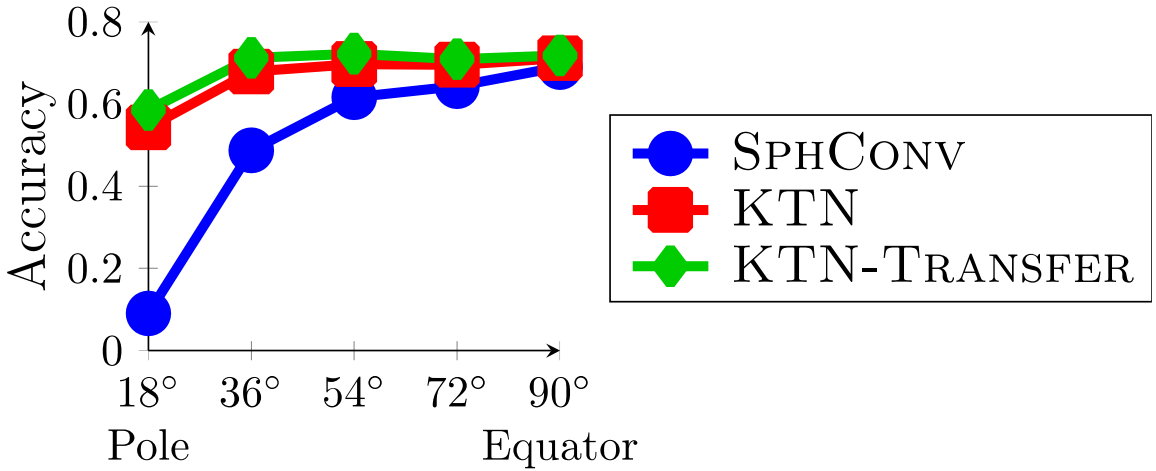
- KTN is much more compact than SphConv
- KTN outperforms recent CNNs on spherical data

Experiment: KTN Transferability





Experiment: KTN Transferability



KTN performs almost identically regardless of the training source!

Conclusion

	Translation Invariance	Supervised Training	Model Size	Transferable Across Models
EQUIRECTANGULAR	No	No	c^2k^2	No
CUBEMAP[1]	No	No	c^2k^2	No
S^2CNN[2]	Yes	Yes	c^2H	No
SPHERICAL CNN[3]	Yes	Yes	c^2H	No
SPHERENET[4]	Yes	Yes	c^2k^2	No
SPHCONV	Yes	No	c^2k^2H	No
KTN	Yes	No	$c^2k^2 + c^2$	Yes

- Accounts for distortion in 360° images
- Does not require annotated 360° images
- Is transferable across different source CNNs

[1][Cheng et al., CVPR 2018],
[2][Cohen et al., ICLR 2018],
[3][Esteves et al., ECCV 2018],
[4][Coors et al., ECCV 2018]

Open door to off-the-shelf CNN recognition on 360° images

Resources

- Su and Grauman, Learning Spherical Convolution for Fast Features from 360° Imagery, NeurIPS 2017
- Su and Grauman, Kernel Transformer Networks for Compact Spherical Convolution, CVPR 2018
- Su and Grauman, Learning Spherical Convolution for 360 Recognition, PAMI 2021
- Zhang et al., PanoContext: A Whole-Room 3D Context Model for Panoramic Scene Understanding, ECCV 2014
- Hu et al., Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Video, CVPR 2017
- Lai et al., Semantic-driven Generation of Hyperlapse from 360° Video, TOG 2017
- Boomsma and Frellsen, Spherical convolutions and their application in molecular modelling, NeurIPS 2017

Resources

- Cheng et al., Cube Padding for Weakly-Supervised Saliency Prediction in 360° Videos, CVPR 2018
- Lee et al., A Memory Network Approach for Story-Based Temporal Summarization of 360° Videos, CVPR 2018
- Chou et al., Self-view Grounding Given a Narrated 360° Video, AAAI 2018
- Yu et al., A Deep Ranking Model for Spatio-Temporal Highlight Detection from a 360 Video, AAAI 2018
- Cohen et al., Spherical CNNs, ICLR 2018
- Esteves et al., Learning SO(3) Equivariant Representations with Spherical CNNs, ECCV 2018
- Coors et al., SphereNet: Learning Spherical Representations for Detection and Classification in Omnidirectional Images, ECCV 2018
- Zhang et al., Saliency Detection in 360° Videos, ECCV 2018
- Khasanova and Frossard, Geometry Aware Convolutional Filters for Omnidirectional Images Representation, ICML 2019