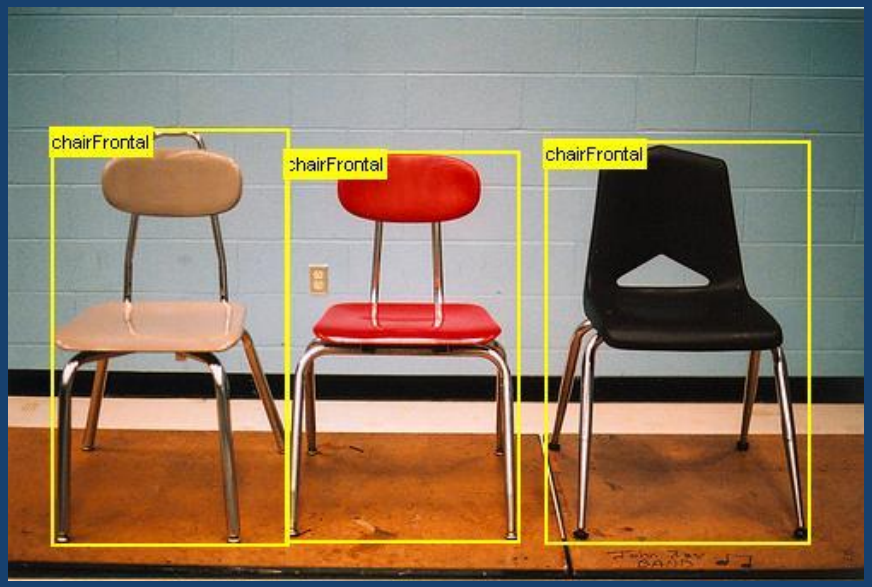
SELF-SUPERVISION

Nate Russell, Christian Followell, Pratik Lahiri

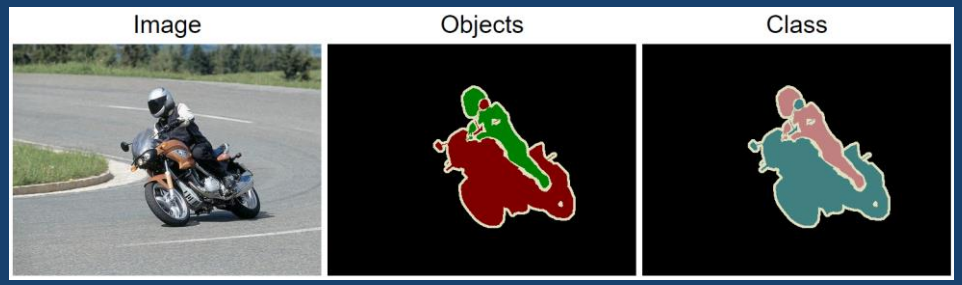
TASKS

Classification & Detection



Action Prediction

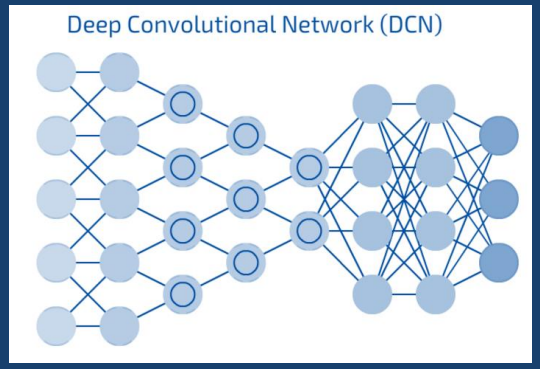
Segmentation

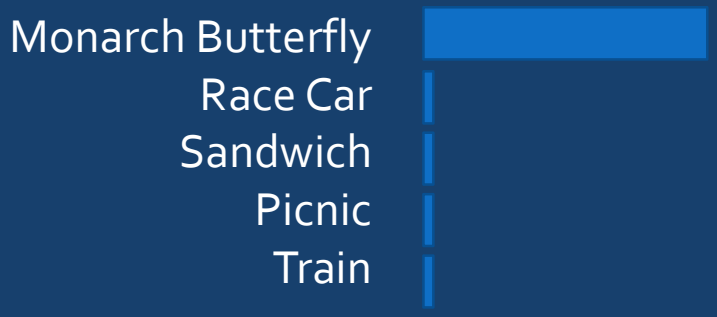


Retrieval

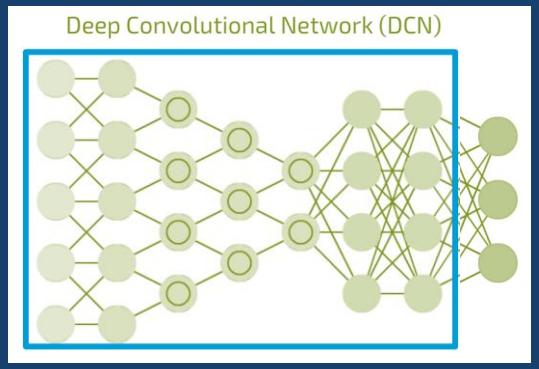
And Many More…

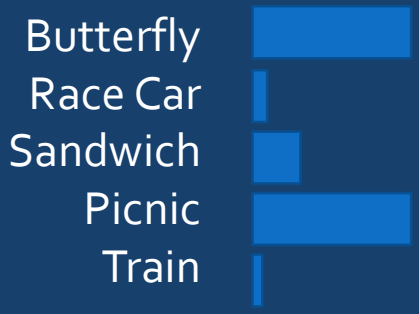
Strong Supervision



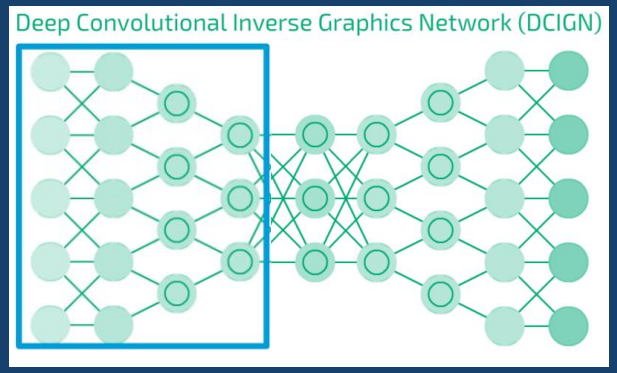


Weak Supervision



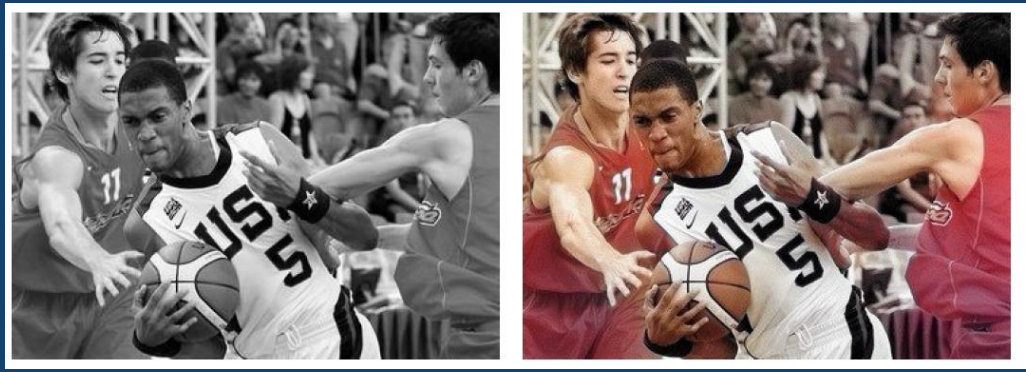


Self-Supervision

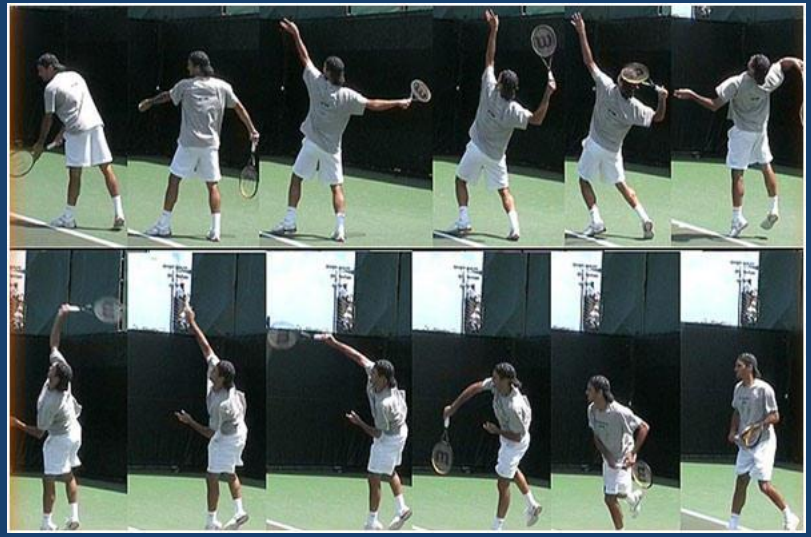


EXAMPLES OF SELF-SUPERVISION


I Context



II Color



III Motion

plane approaching zrh
avro regional jet rj

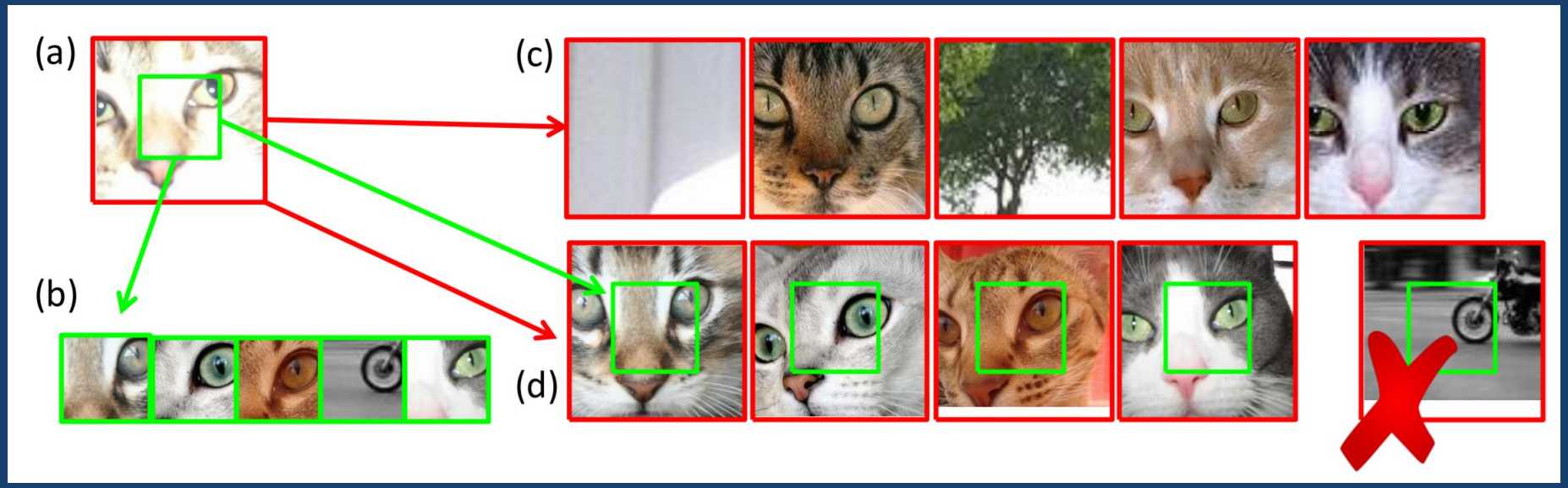
IV Ambient Noise & Noisy Labels

PIXEL CONTEXT

Patch Context & In-Painting

IMAGE - CONTEXT

Large Nearest Neighbors
Too Much Variance



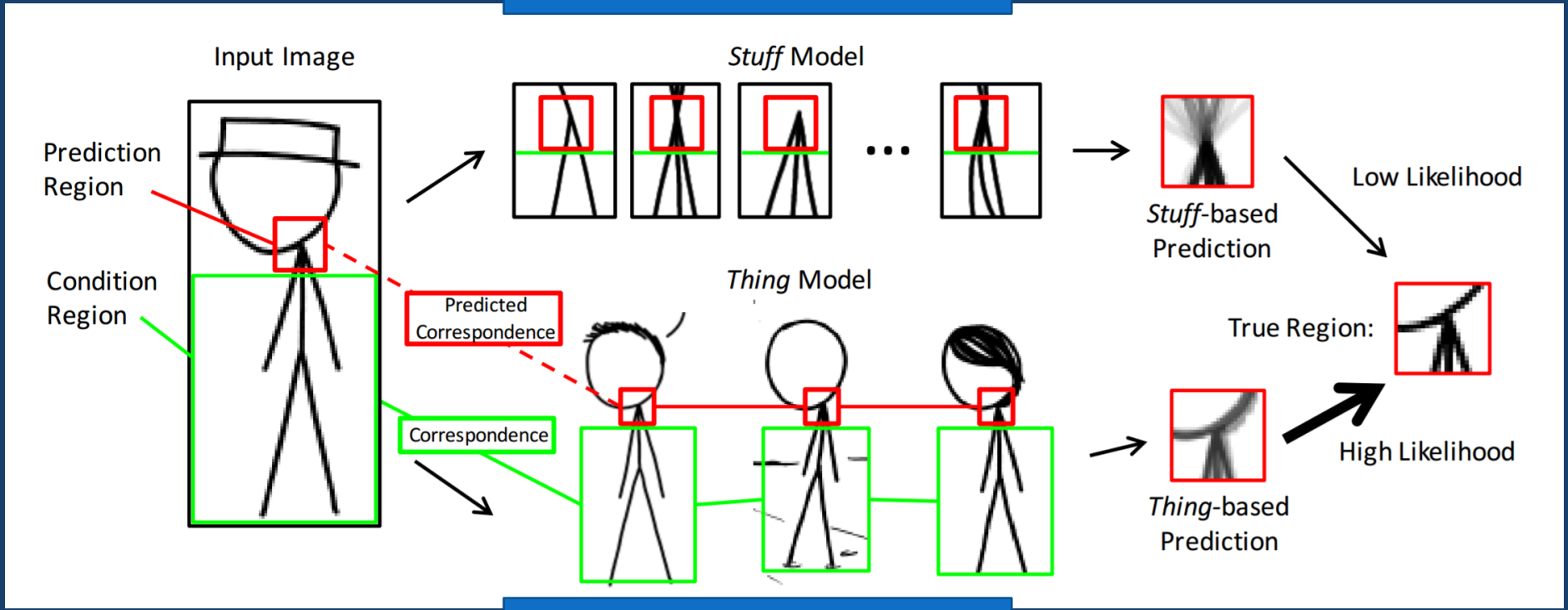
Small Nearest Neighbors
Not Enough Context

1. Get Small Nearest Neighbors
2. Prune Neighbors using Context

Carl Doersch, Abhinav Gupta, Alexei A. Efros, Context as Supervisory Signal: Discovering Objects with Predictable Context, ECCV 2014.

STUFF VS THING MODEL

Low Level Statistics

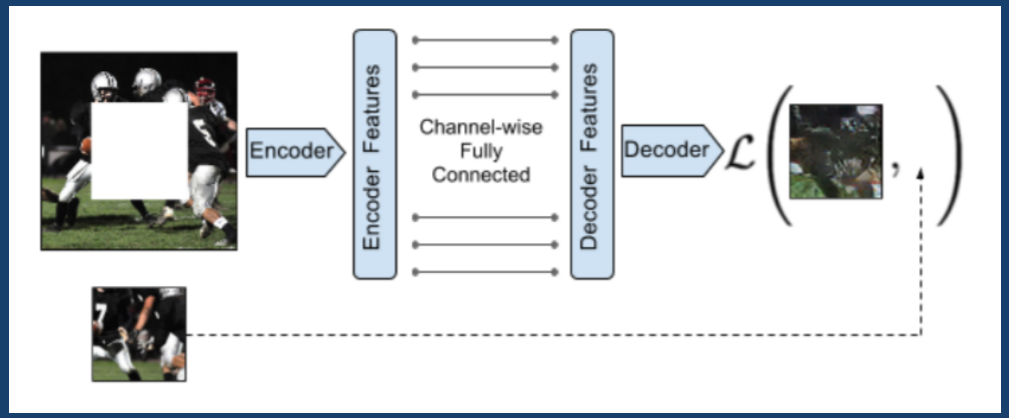


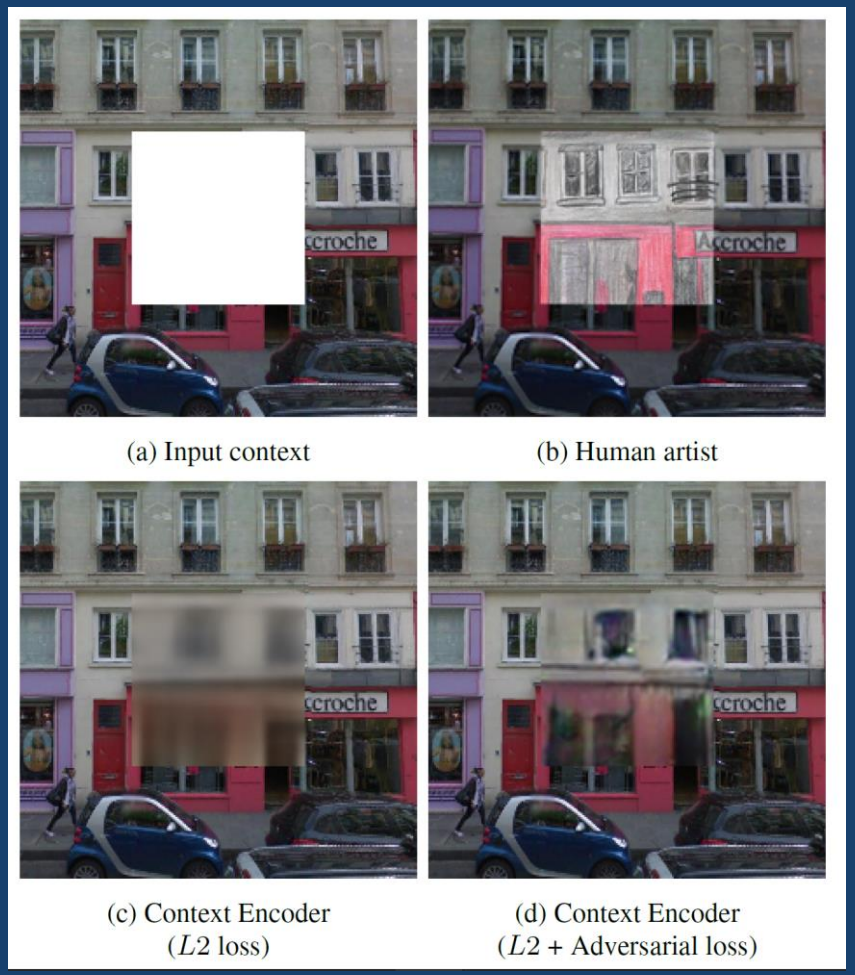
Cluster Correspondance

Carl Doersch, Abhinav Gupta, Alexei A. Efros, Context as Supervisory Signal: Discovering Objects with Predictable Context, ECCV 2014.

IMAGE IN-PAINTING

Correctly predicting large image patches in a photo realistic manner suggests that model has ability to generalize real world objects.



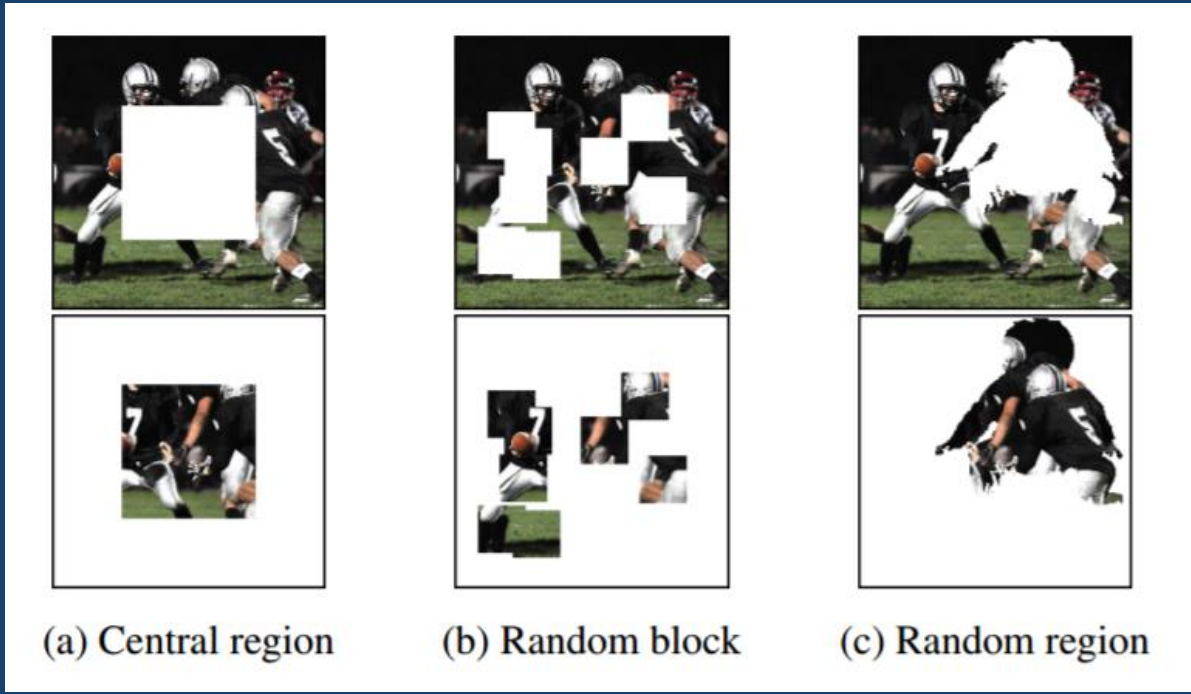


Context Encoders: Feature Learning by Inpainting Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

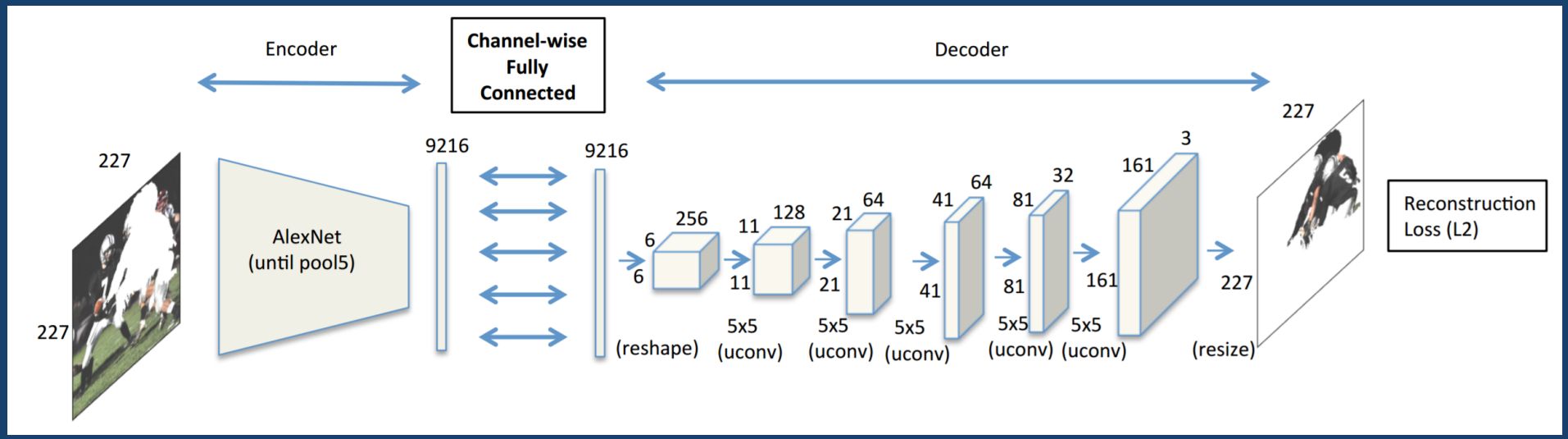
MASKS

Masks

- PASCAL VOC 2012 Shape Masks
- ¼ Image



MODEL ARCHITECTURE



Encoder

- *AlexNet Ending in Pool5*
- [227 x 227] -> [6 x 6 x 256] = 9216

Chanel-wise Fully Connected

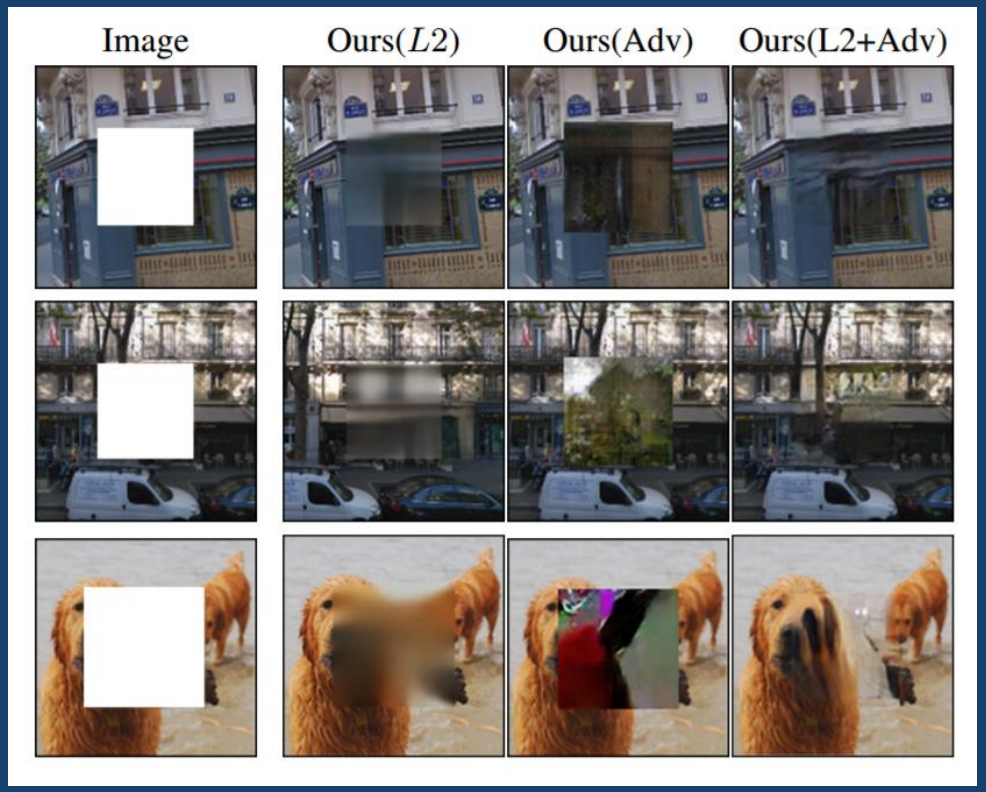
- $mn^4 < m^2n^4$
- No Inter Feature Map connections
- Needed to propagate global context

Decoder

- Stride 1 Convolution to propagate information across channels
- 5 up-convolutional layers

Context Encoders: Feature Learning by Inpainting Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

LOSS FUNCTION



Masked Loss

+ Contextual
- Blurry

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}$$

$$\mathcal{L}_{rec}(x) = \|\hat{M} \odot (x - F((1-\hat{M}) \odot x))\|_2^2$$

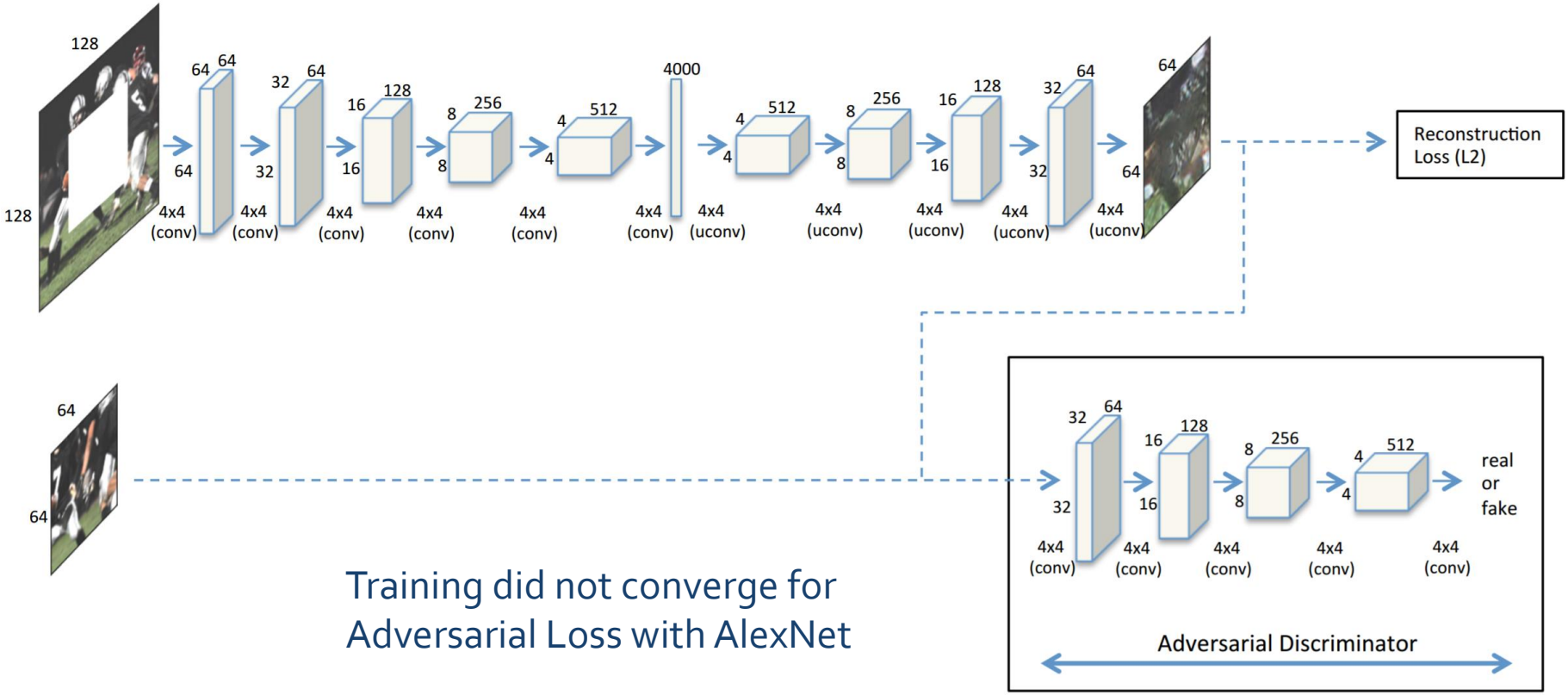
Adversarial Loss

+ In Focus
- Non-contextual

$$\mathcal{L}_{adv} = \max_{D} \quad \mathbb{E}_{x \in \mathcal{X}}[\log(D(x)) + \log(1 - D(F((1-\hat{M}) \odot x)))]$$

Context Encoders: Feature Learning by Inpainting Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

MODEL ARCHITECTURE



Training did not converge for Adversarial Loss with AlexNet

HOW TO TRANSFER?

Pretraining Initialization

1. Train Network on Task A
2. Replace last layer(s) for Task B
3. Begin training again on Task B. Allowing for all weights in network to update

Freeze and Fine Tune

1. Train Network on Task A
2. Replace last layer(s) for Task B
3. Train new layer weights on Task B but do NOT update weights learned from Task A

Feature Augmentation

1. Train Network on Task A
2. Train Network on Task B but use the activations from Network A and use them as Features

Joint Learning / Semi-Supervision / Multi-Task

1. Train One Network on Task A and Task B at the same time, controlling the tradeoff.

RESULTS

- Self-Supervision Pretraining -> PASCAL VOC
- Doersch et al. wins in detection likely due to local pixel patch wise methodology being better suited.

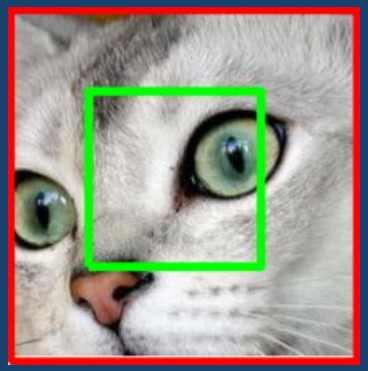
Pretraining Method	Supervision	Pretraining time	Classification	Detection	Segmentation
ImageNet	1000 class labels	3 days	72.8%	56.8%	48.0%
Wang *et.al.*	motion	1 week	58.7%	47.4%	-
Doersch *et.al.* *(First Method)*	Relative context	4 weeks	55.3%	46.6%	-
Pathak et al (In-Painting)	context	14 hours	56.5%	44.5%	30%

IMAGE CONTEXT PAPERS

Pathak et al.

- Predict Missing Patch
- Intended for Feature Learning
- Deep Model

Doersch et al.



- Predict Patch from local Context
- Intended for Unsupervised Clustering
- Non Deep Model based on HOG

- PASCAL VOC Benchmark

COLOR

Colorization & Depth

COLORIZE IMAGES

Correctly predicting color suggests understanding of latent object semantics and texture

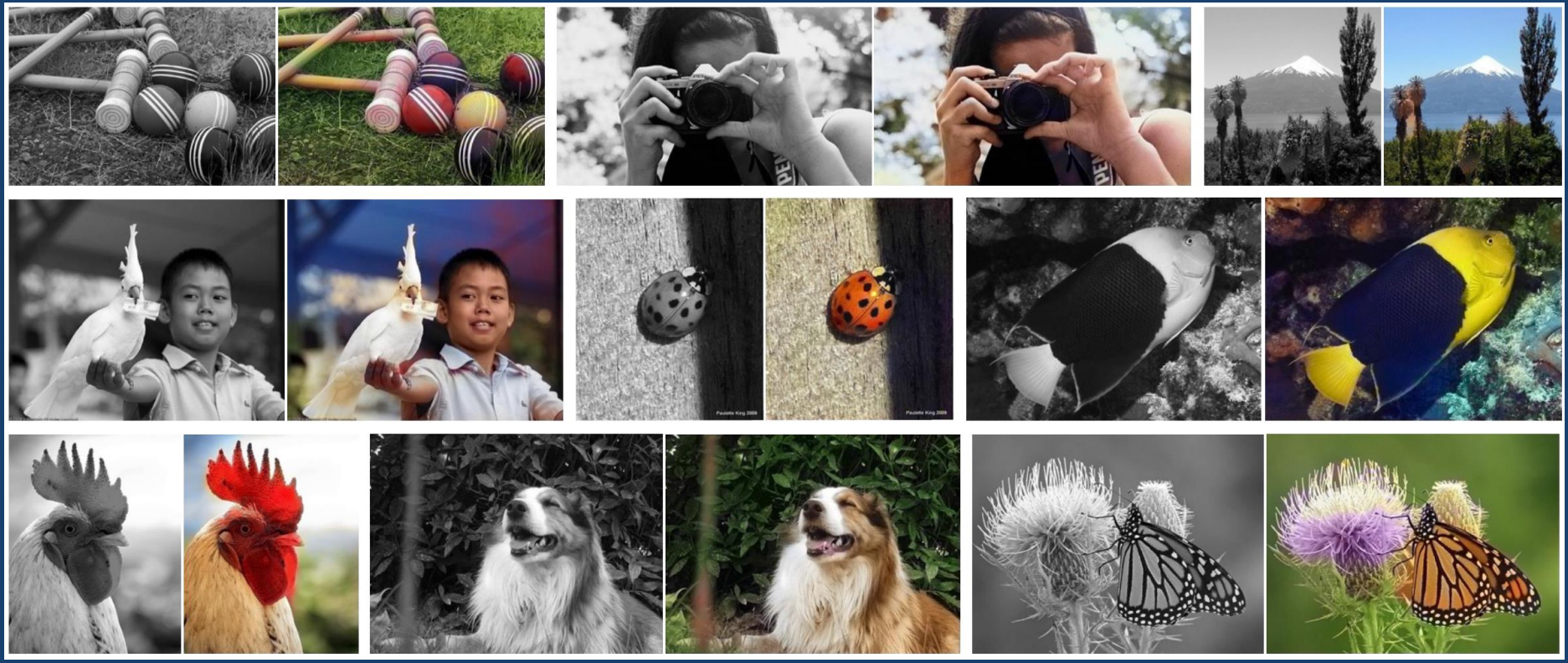
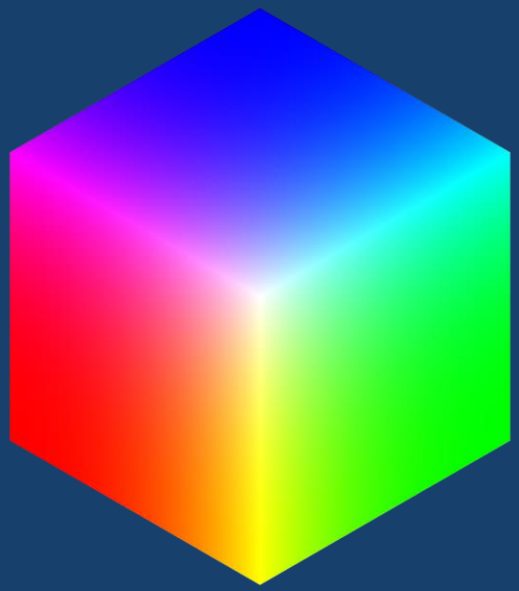


Figure from Colorful Image Colorization Richard Zhang, Phillip Isola, Alexei A. Efros

COLOR SPACE

RGB

X: Red
Y: Blue
Z: Green



CIELAB

L: White - Black
A: Red - Green
B: Blue - Yellow

HSL/HSV Bi-cone

L: White - Black
H: Hue
C: Chroma

ZHANG ET AL. NETWORK ARCHITECTURE

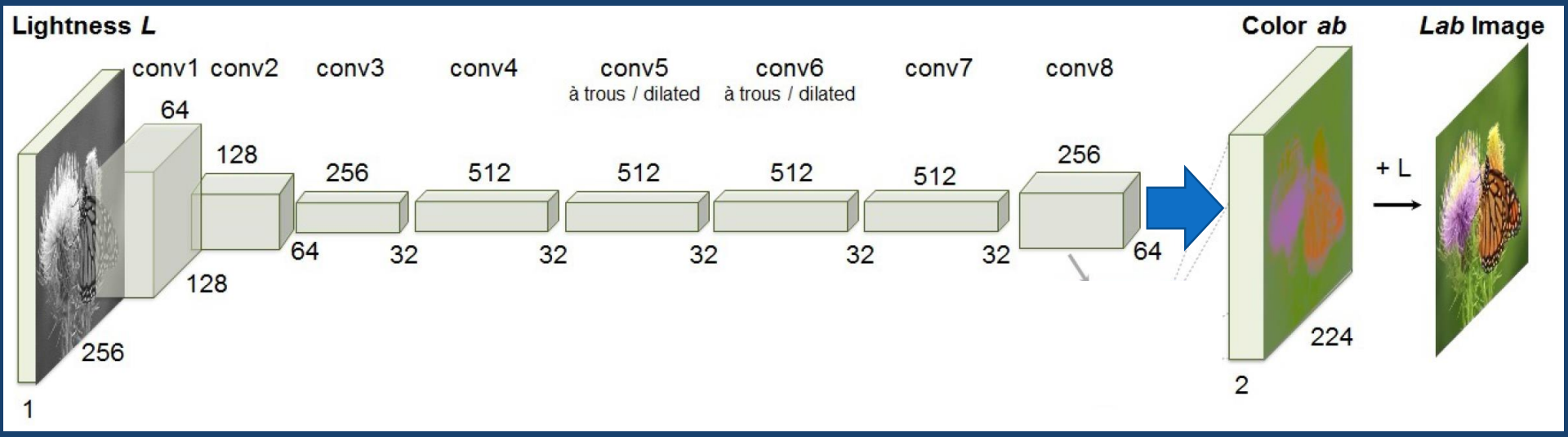


Figure from Colorful Image Colorization Richard Zhang, Phillip Isola, Alexei A. Efros

Each Block = 2 to 3 conv + ReLu layers followed by BatchNorm

No Pooling

Dilated Convolutions

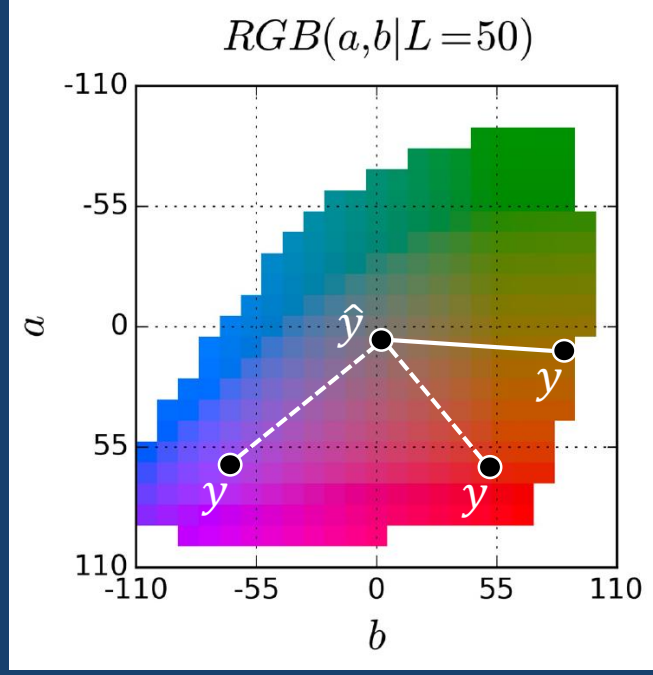
Regression Loss in LAB space

$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \| \mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w} \|_2^2$$

Sum over all pixels

Actual Pixel Color

Predicted Pixel Color

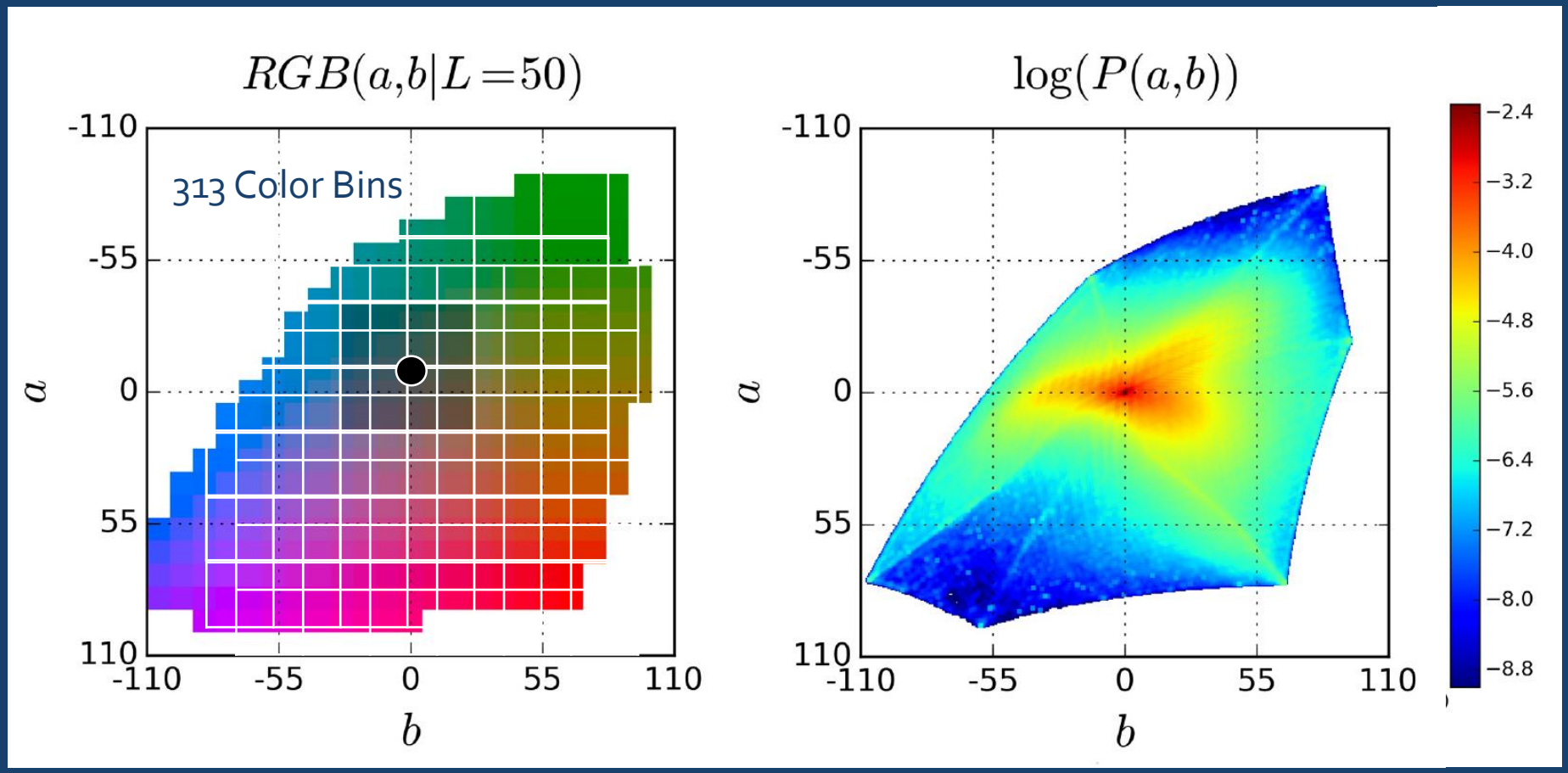


- Yields lots of Grey unsaturated images
- Color might not exist if color space is non-convex

Discretization and Rebalancing

Color Discretization &
Soft Encoding

Empirical Color
Distribution



ZHANG ET AL. NETWORK ARCHITECTURE

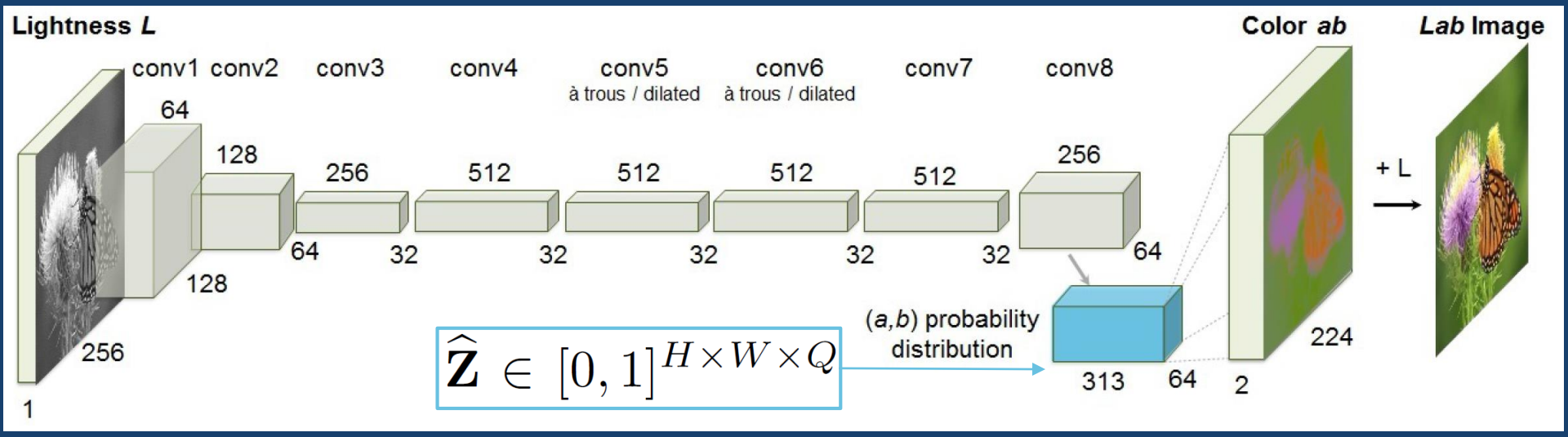


Figure from Colorful Image Colorization Richard Zhang, Phillip Isola, Alexei A. Efros

Each Block = 2 to 3 conv + ReLu layers followed by BatchNorm

No Pooling

Dilated Convolutions

Classification and Rebalancing

Soft Encoding: for each true sRGB color find 5 Nearest Quantized Neighbors. Assign probability mass to each of these bins proportional to their distance in AB space using Gaussian kernel. Helps learn relations between 313 bins

Multinomial Cross Entropy Loss

Sum over all pixels

$$\mathrm{L}_{cl}(\hat{\mathbf{Z}}, \mathbf{Z}) = -\sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_{q} \mathbf{Z}_{h,w,q} \log(\hat{\mathbf{Z}}_{h,w,q})$$

Actual Pixel Color Distribution

Predicted Pixel Color Distribution

$$v(\mathbf{Z}_{h,w}) = \mathbf{w}_{q^*}, \text{ where } q^* = \arg\max_{q} \mathbf{Z}_{h,w,q}$$

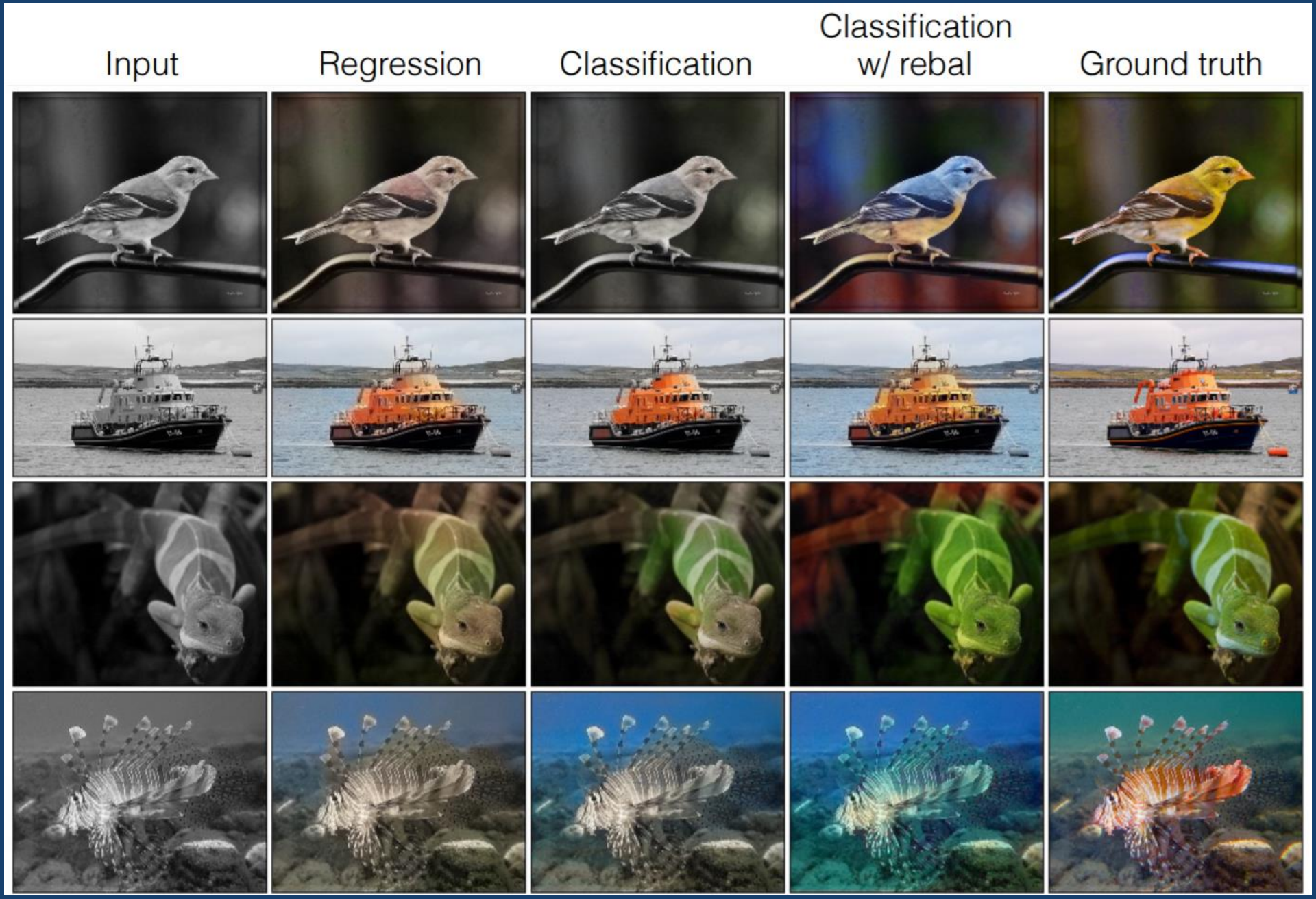
$$\mathbf{w} \propto \left((1-\lambda)\tilde{\mathbf{p}} + \frac{\lambda}{Q}\right)^{-1}, \quad \mathbb{E}[\mathbf{w}] = \sum_{q} \tilde{\mathbf{p}}_q \mathbf{w}_q = 1$$

Smooth Empirical Distribution with Gaussian and Mix it with Uniform distribution

$$\lambda = \tfrac{1}{2} \text{ and } \sigma = 5$$

Color Distribution -> Color

$$\mathcal{H}(\mathbf{Z}_{h,w}) = \mathbb{E}\left[f_T(\mathbf{Z}_{h,w})\right], \qquad f_T(\mathbf{z}) = \frac{\exp(\log(\mathbf{z})/T)}{\sum_q \exp(\log(\mathbf{z}_q)/T)}$$

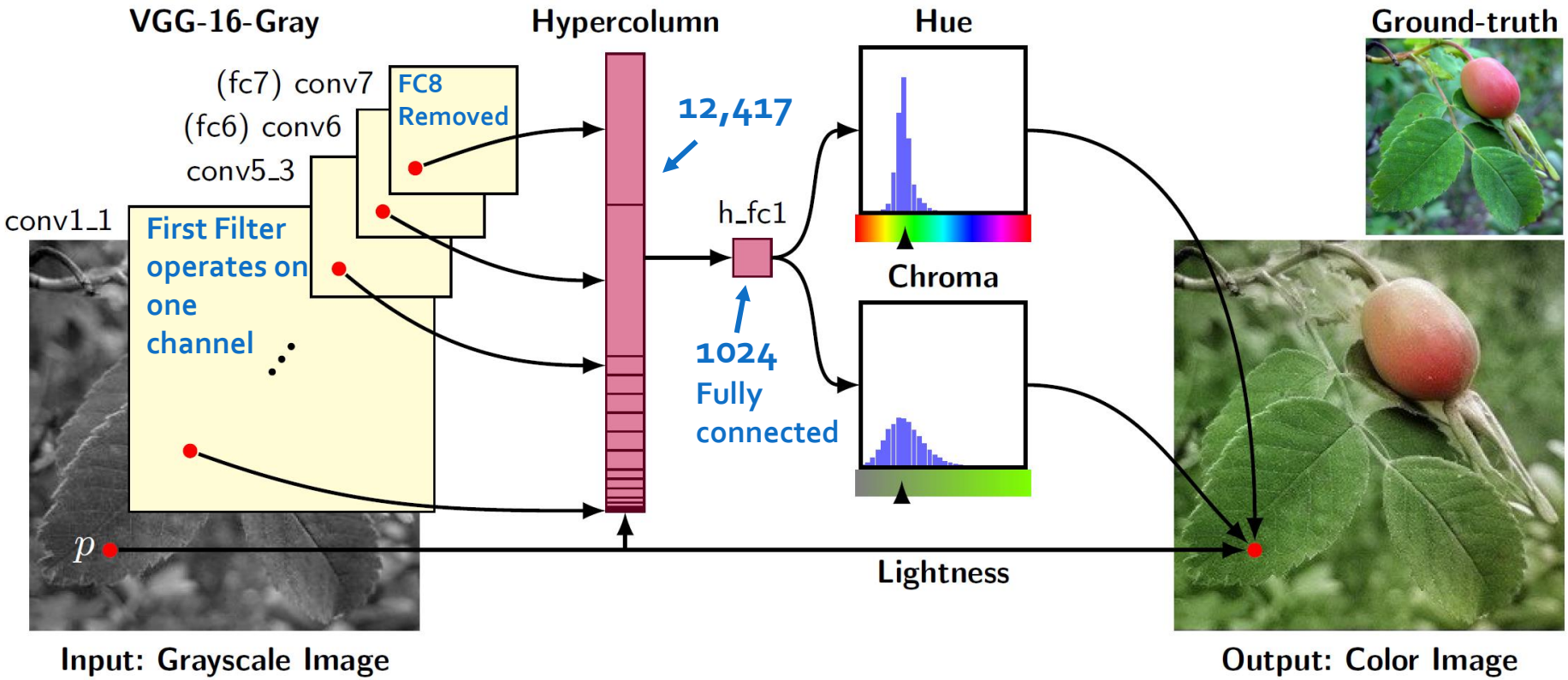


Colorful Image Colorization Richard Zhang, Phillip Isola, Alexei A. Efros

Modes of Failure

- Red-Blue Confusion
- Complex Indoor Scene -> Sepia Tones
- Color Consistency

LARSSON ET AL. NETWORK ARCHITECTURE

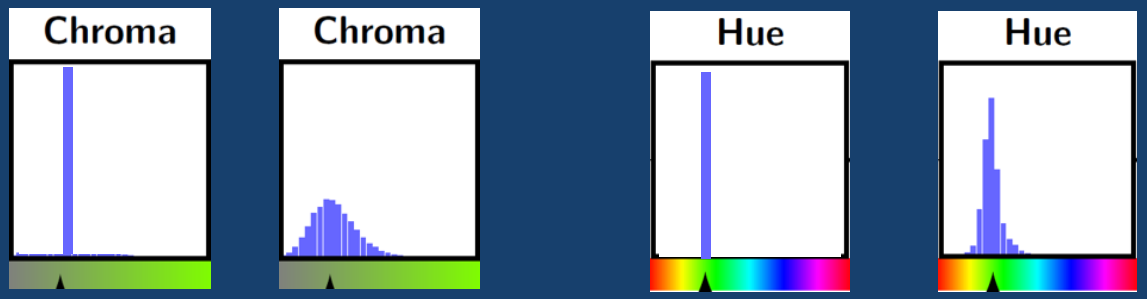


Loss

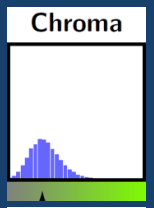
Balance Chroma and Hue so each is equally valuable to the loss

Per Pixel Loss

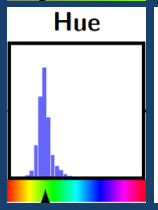
$$L_{\text{hue/chroma}}(\mathbf{x},\mathbf{y}) = D_{\text{KL}}(\mathbf{y}_{\text{C}} \| f_{\text{C}}(\mathbf{x})) + \lambda_H y_{\text{C}} D_{\text{KL}}(\mathbf{y}_{\text{H}} \| f_{\text{H}}(\mathbf{x}))$$



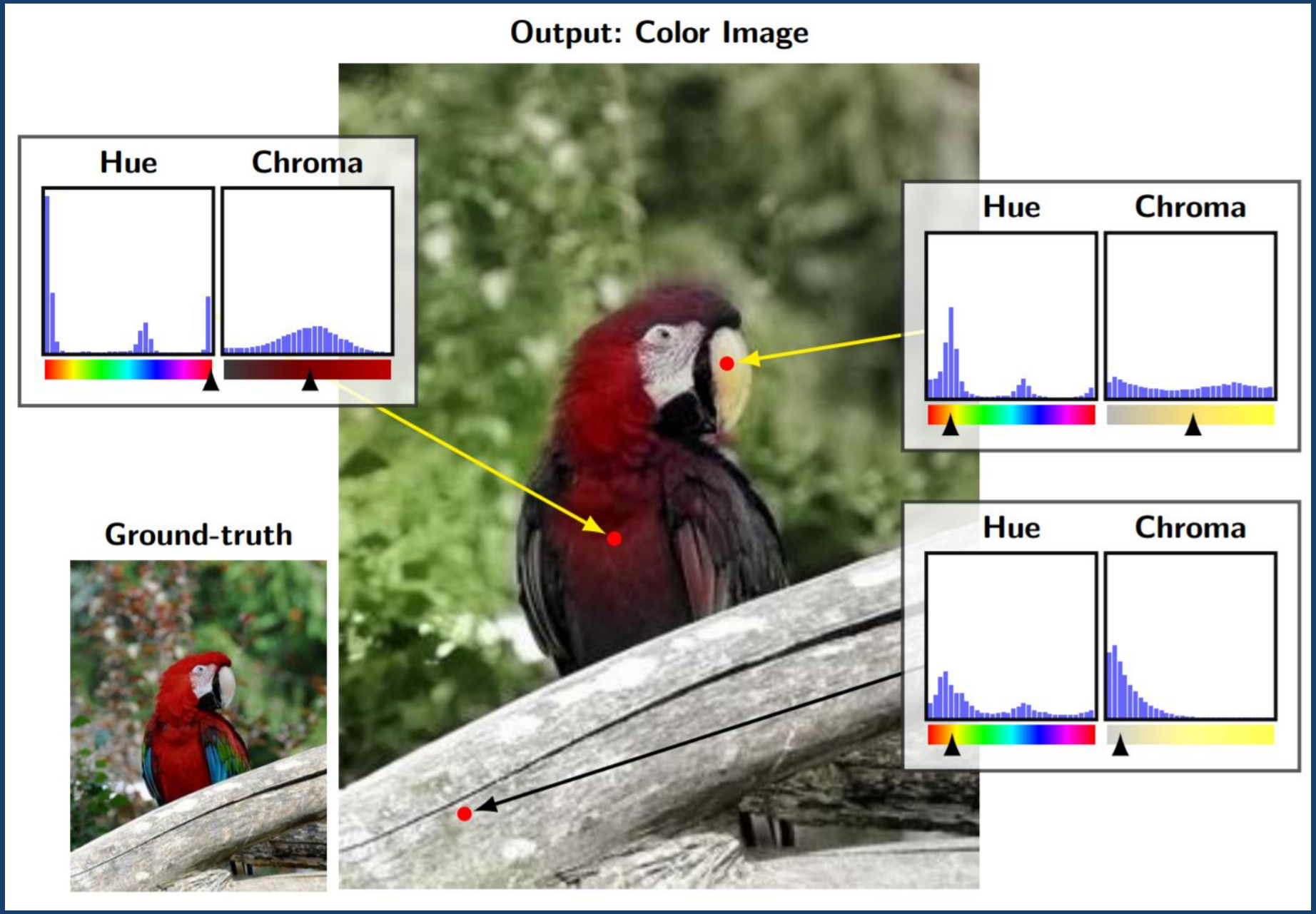
Color Distribution -> Color



Compute cumulative sum of $\hat{y}_n$ and use linear interpolation to find the value at the middle bin. That is the Z value.



$$z = \mathbb{E}_{H \sim f_h(\mathbf{x})}[H] \triangleq \frac{1}{K} \sum_k [f_h(x)]_k e^{i\theta_k}, \quad \theta_k = 2\pi \frac{k+0.5}{K}$$


Output: Color Image
Hue
Chroma
Hue
Chroma
Hue
Chroma
Ground-truth

MODES OF FAILURE

Too Desaturated

Edge Pollution

Inconsistent Chroma

Inconsistent Hue

Color Bleeding

HUMAN EVALUATION

Model **100%**

Real **0%**

HUMAN EVALUATION

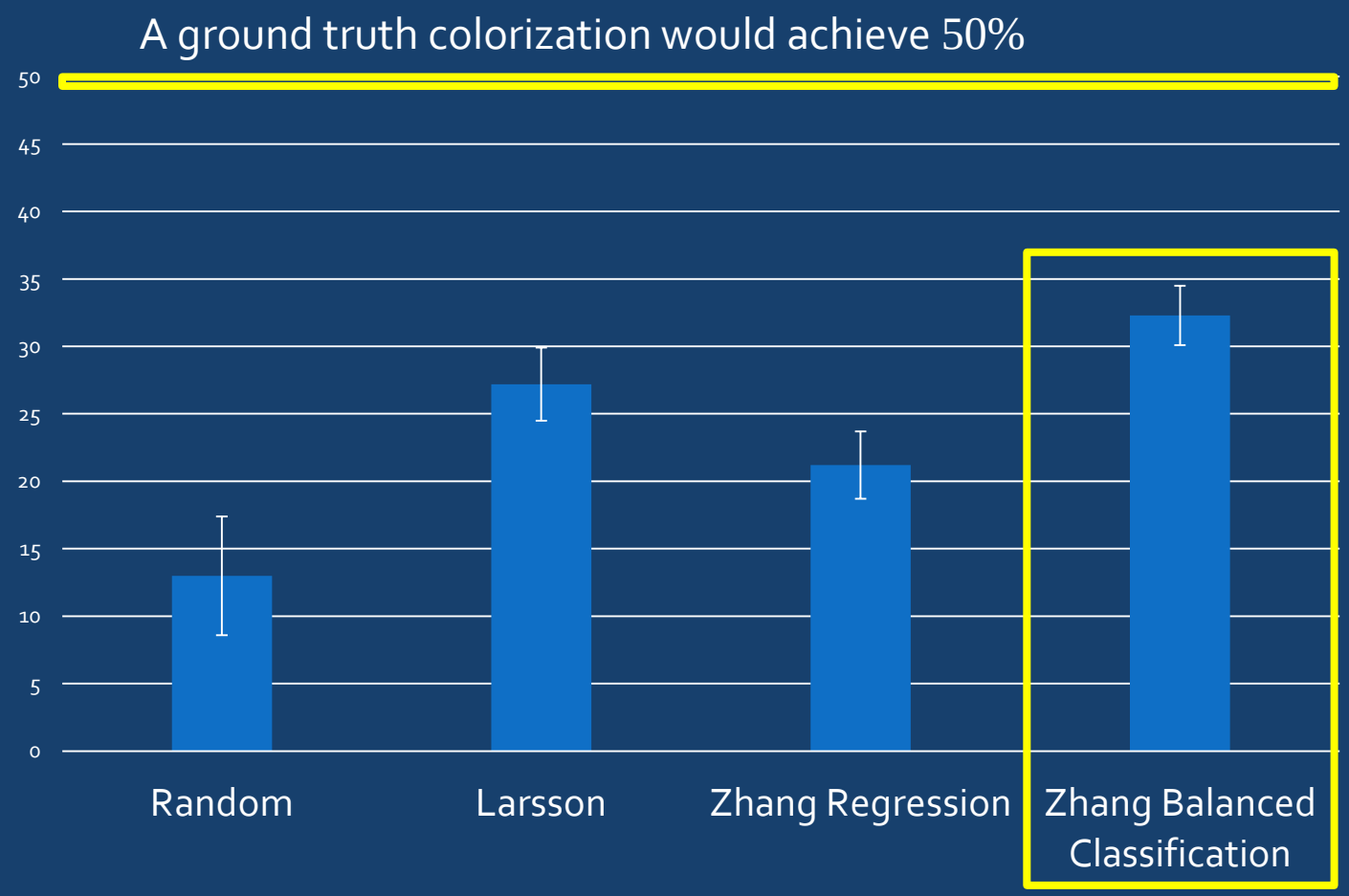
Real **18%**

Model **82%**

AMAZON MECHANICAL TURK

% of AMT workers that believed the Model generated image was the real Image

40 Participants
40 Pairs Each



IMAGENET

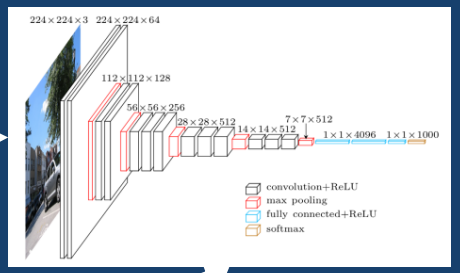
Train

VGG Net on Original Color Images

Test

VGG Net using Color

Original

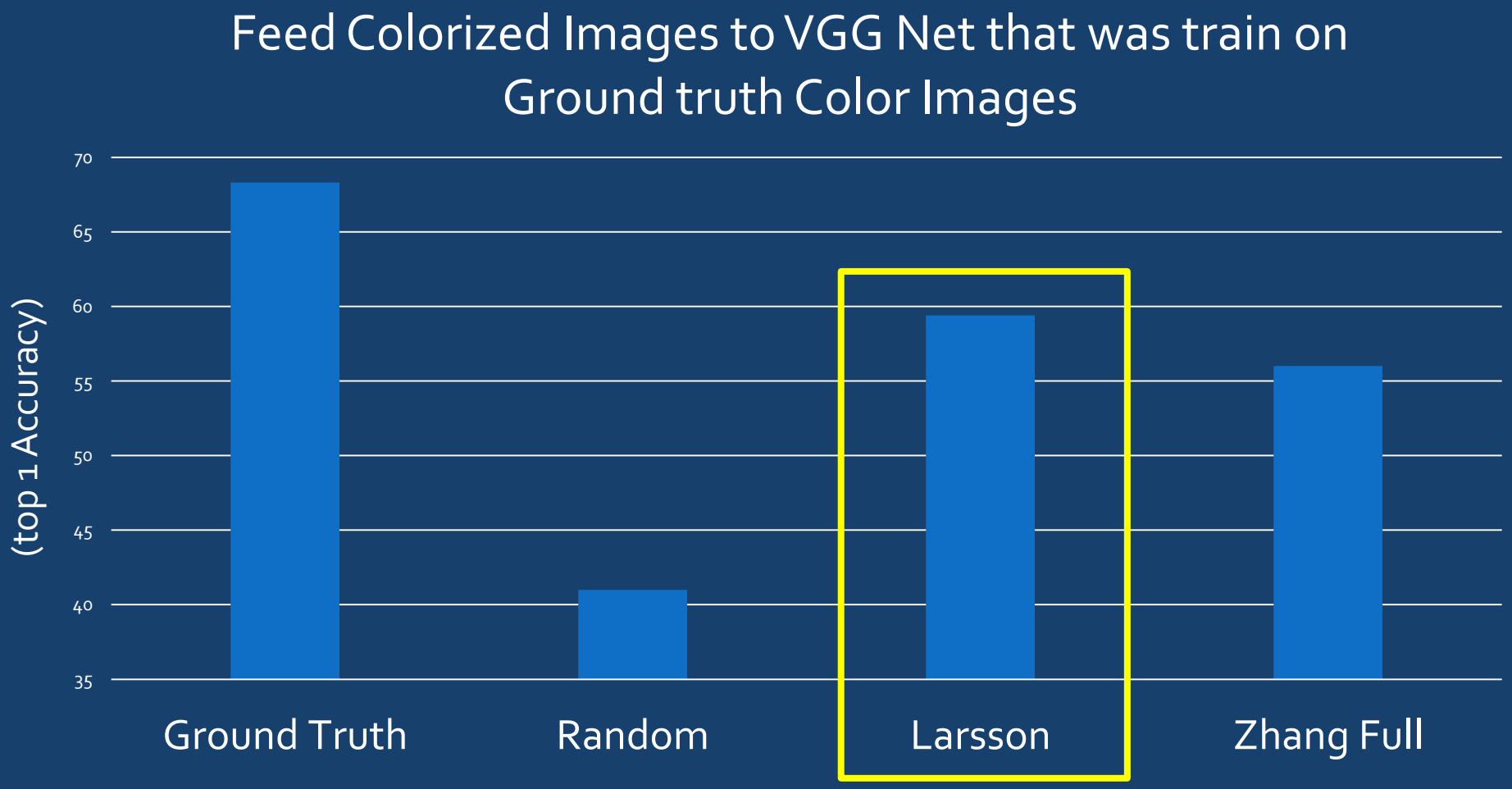


Larsson et al.

Zhang et al.

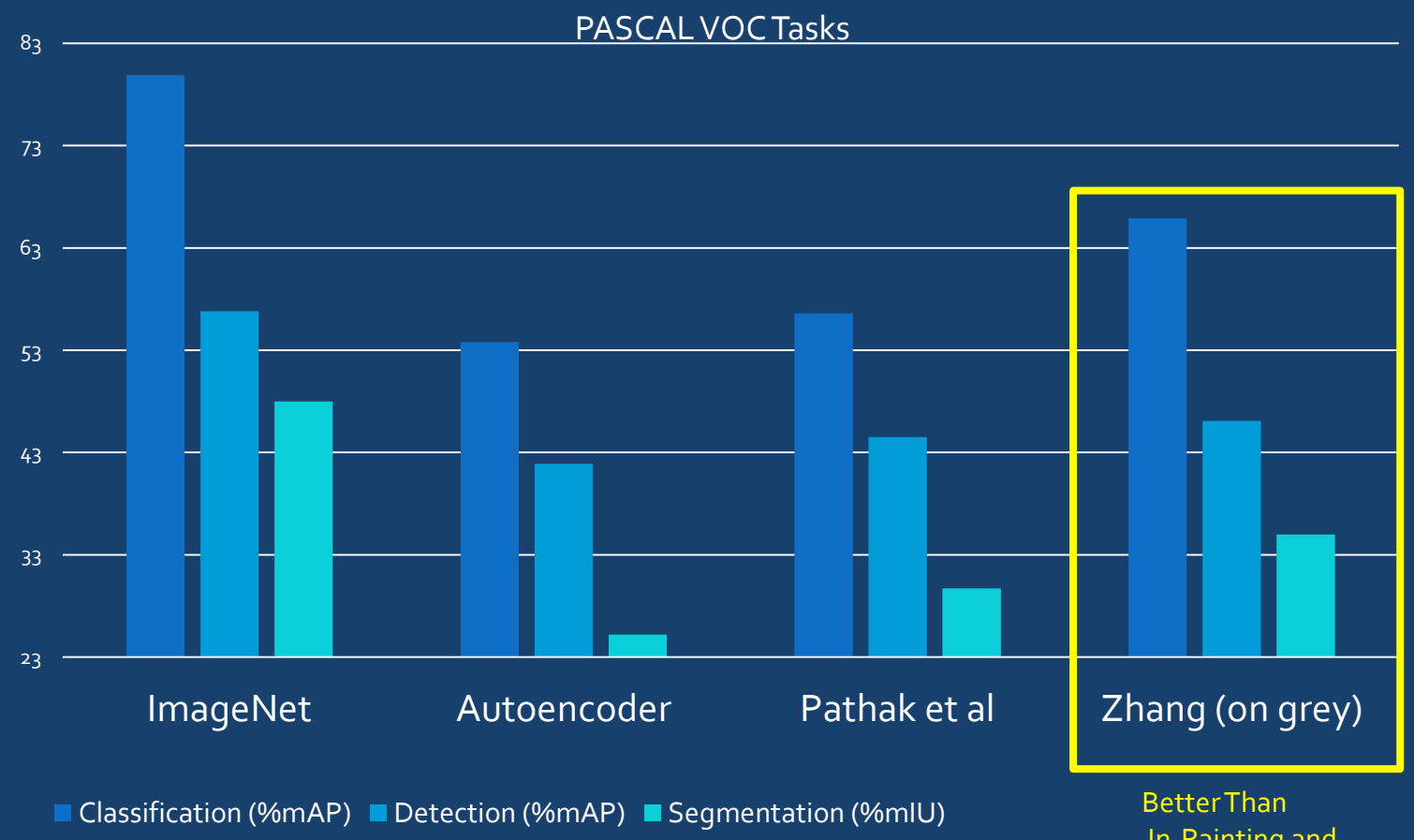
Colorful Image Colorization Richard Zhang, Phillip Isola, Alexei A. Efros

IMAGENET


Feed Colorized Images to VGG Net that was train on Ground truth Color Images
(top 1 Accuracy)
70
65
60
55
50
45
40
35
Ground Truth
Random
Larsson
Zhang Full

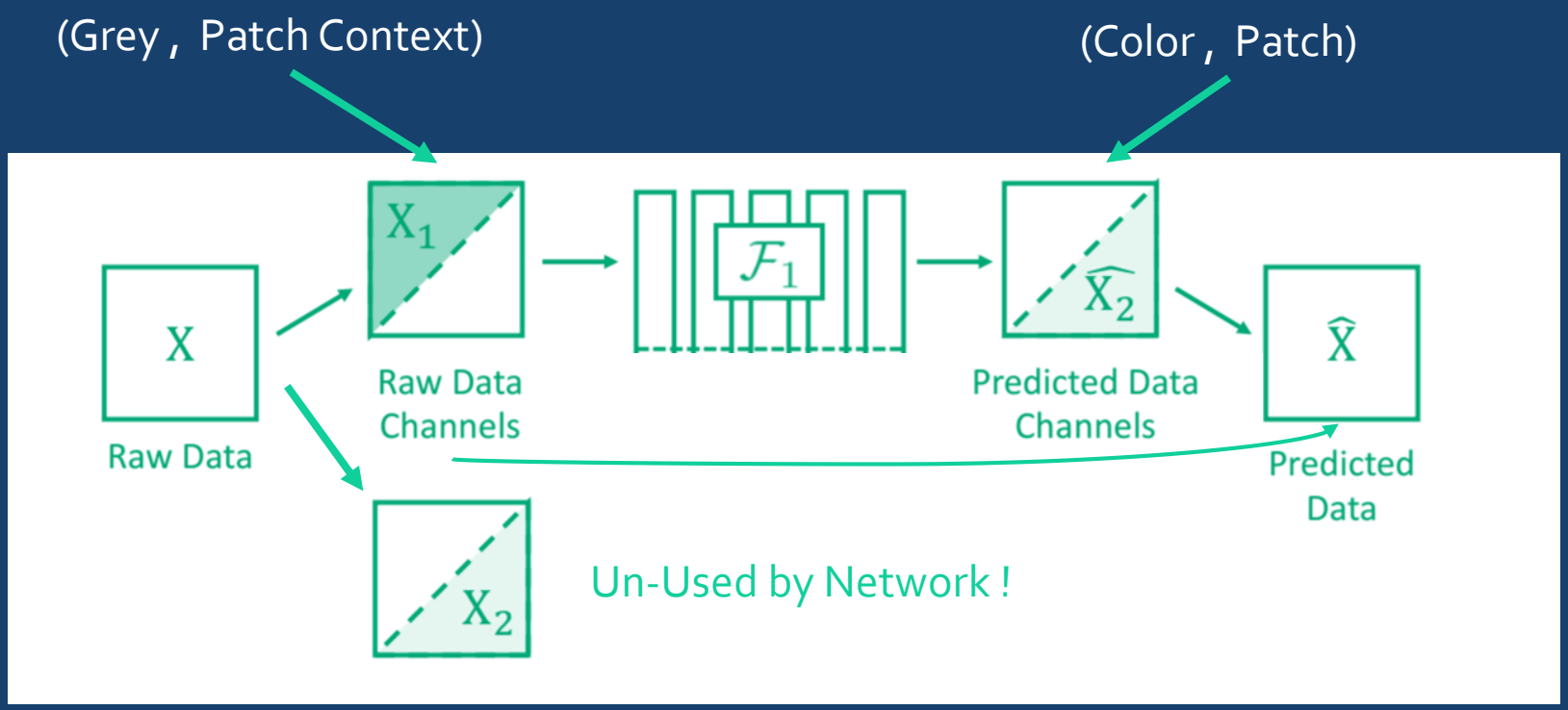
PASCAL EXPERIMENTS

- Frozen Weights with fine tuning at end
- Segmentation with FCN model
- Detection with R-CNN mode



Better Than In-Painting and Autoencoder

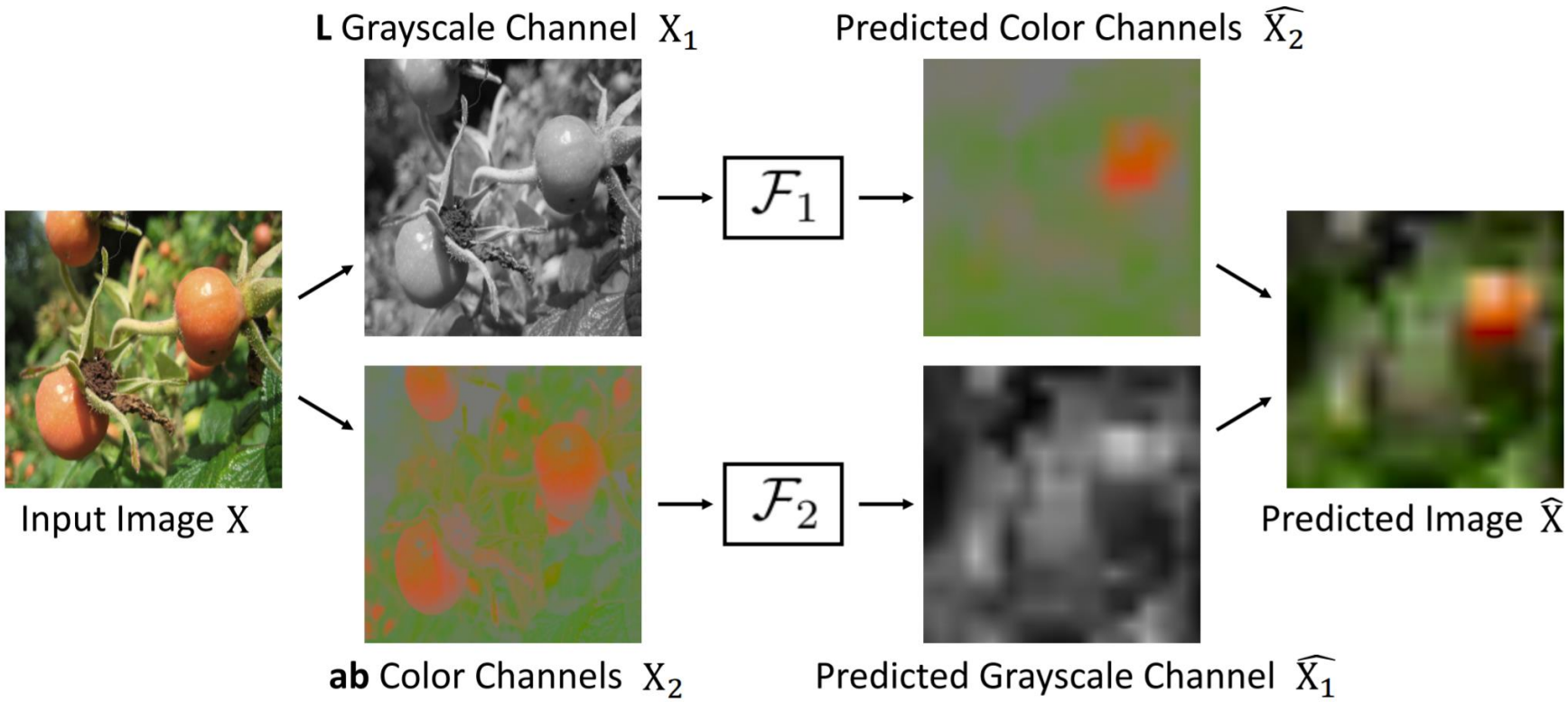
SPLIT BRAIN AUTOENCODER ARCHITECTURE



All Channels of Image Used !

Richard Zhang, Phillip Isola, Alexei A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017

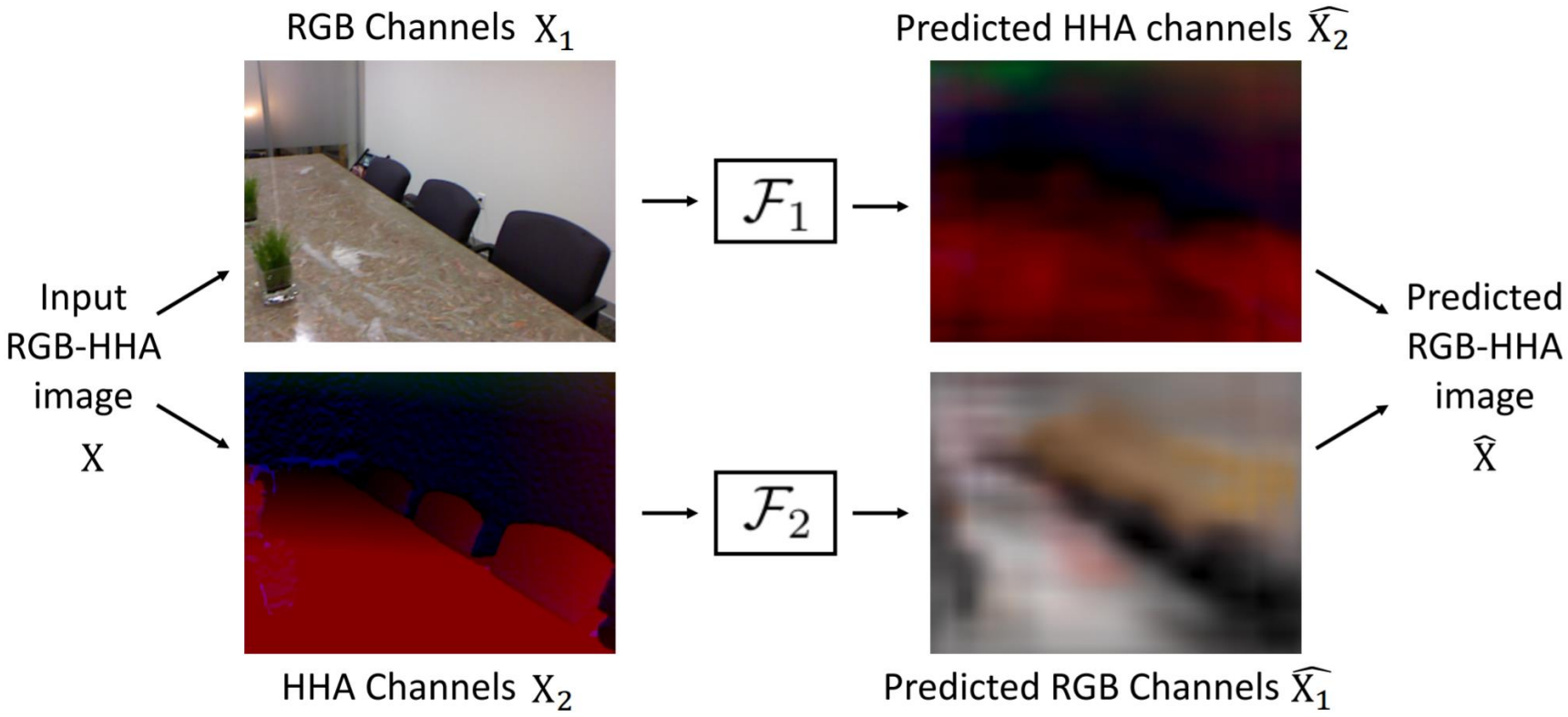
COLOR



(a) ***Lab* Images**

Richard Zhang, Phillip Isola, Alexei A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017

RGB-HHA

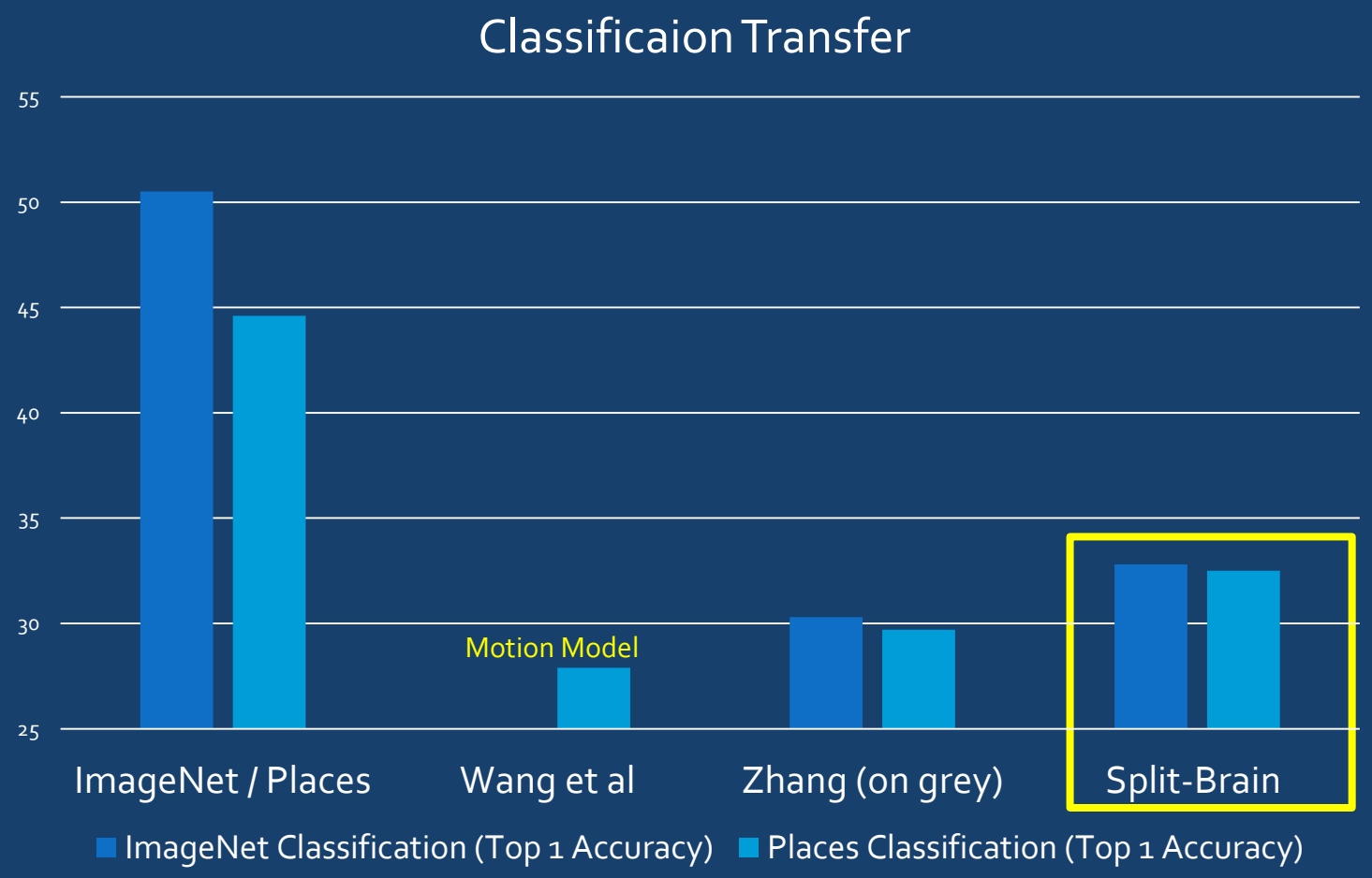


(b) **RGB-D Images**

Richard Zhang, Phillip Isola, Alexei A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017

IMAGENET AND PLACES

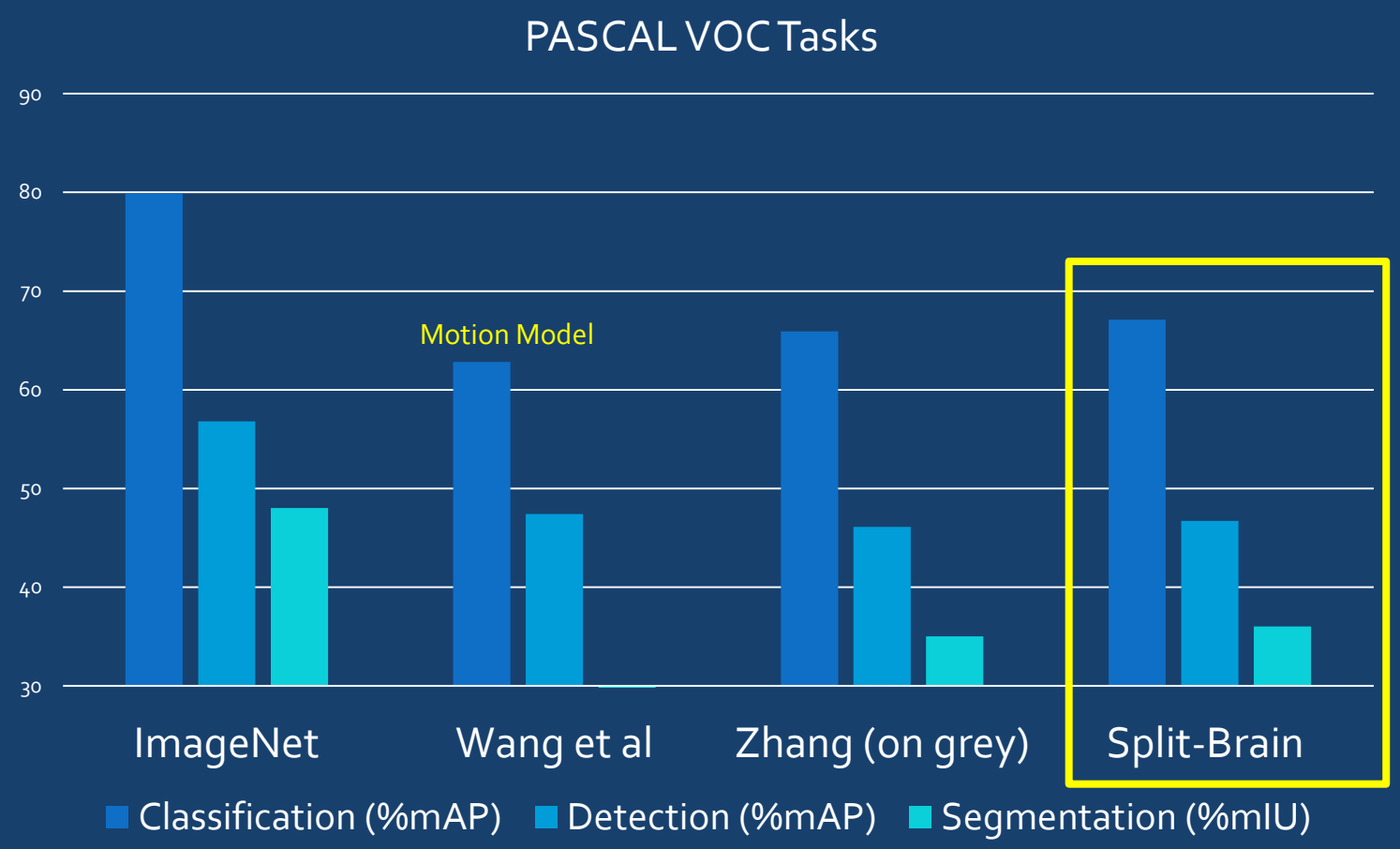
- Freeze network
- Train logistic layer after 5th Convolutional Layer



Richard Zhang, Phillip Isola, Alexei A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017

PASCAL VOC

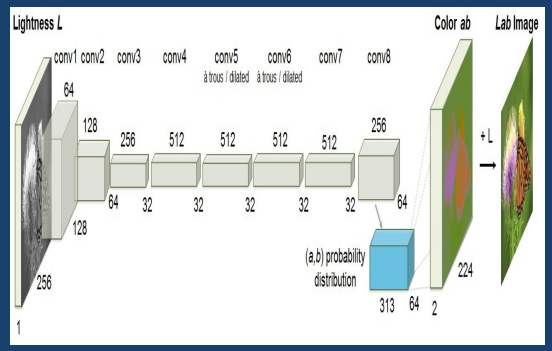
- Freeze network
- Train logistic layer after 5th Convolutional Layer



Richard Zhang, Phillip Isola, Alexei A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017

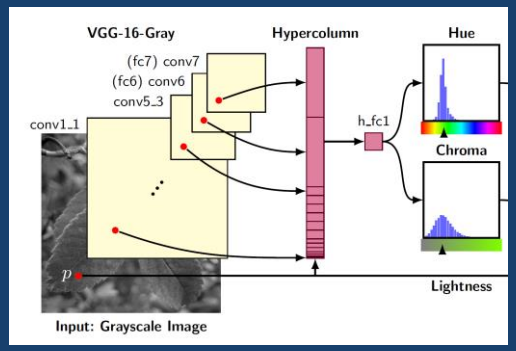
COLORIZATION PAPERS

Zhang et al. I



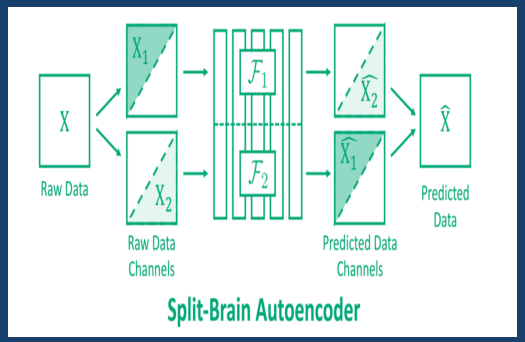
- Rebalanced Classification LAB Loss
- VGG Net + Extra Depth + Dilated Convolutions
- Better for Humans

Larsson et al.



- Un-Rebalanced Classification HSV/L Loss
- VGG Net + Hypercolumns
- Better for classification

Zhang et al. II



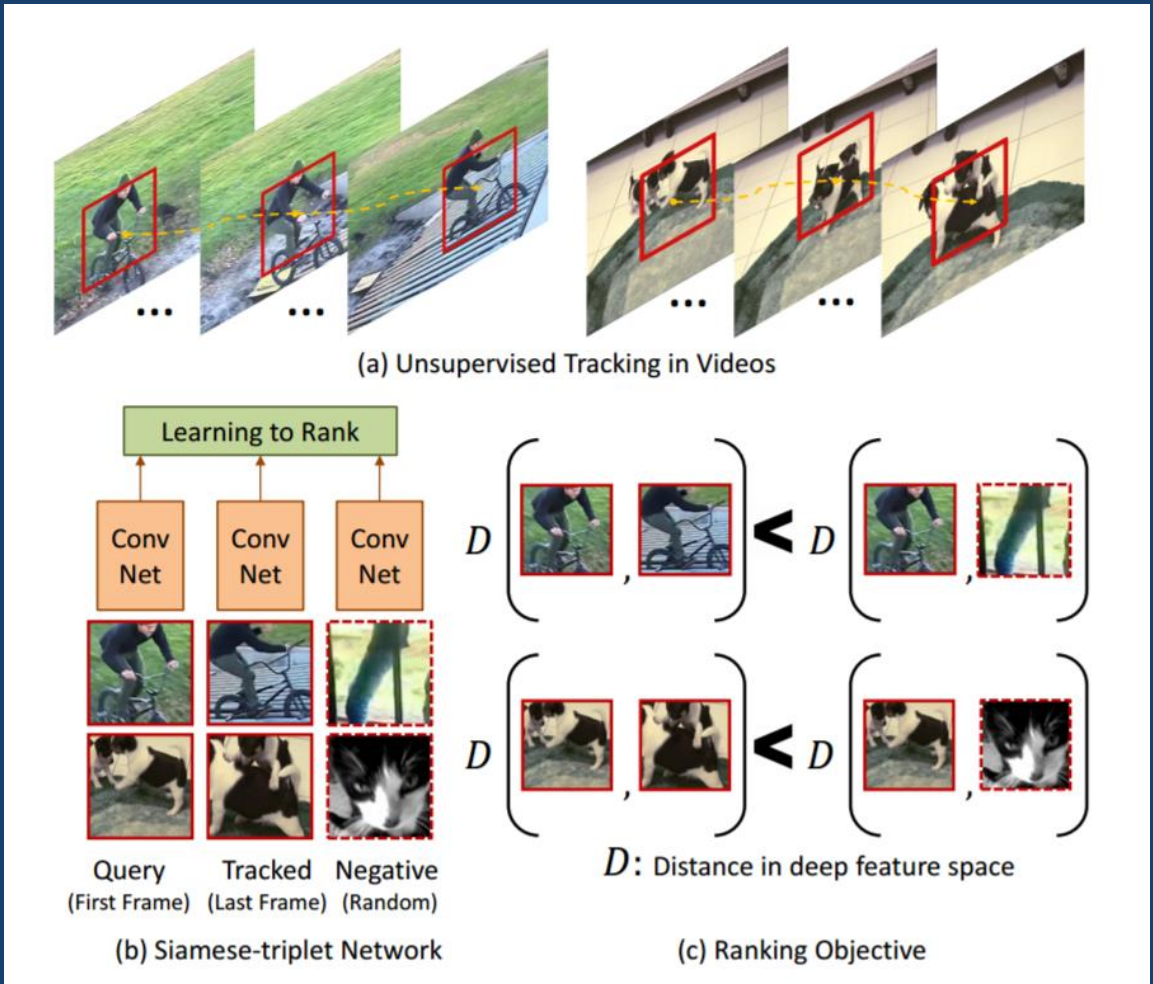
- More General Idea than just Colorization
- Nearly identical to Zhang et al. I
- Use color and grey

MOTION

Triplet Ranking, Contrastive, LSTM-Conv

MOTION

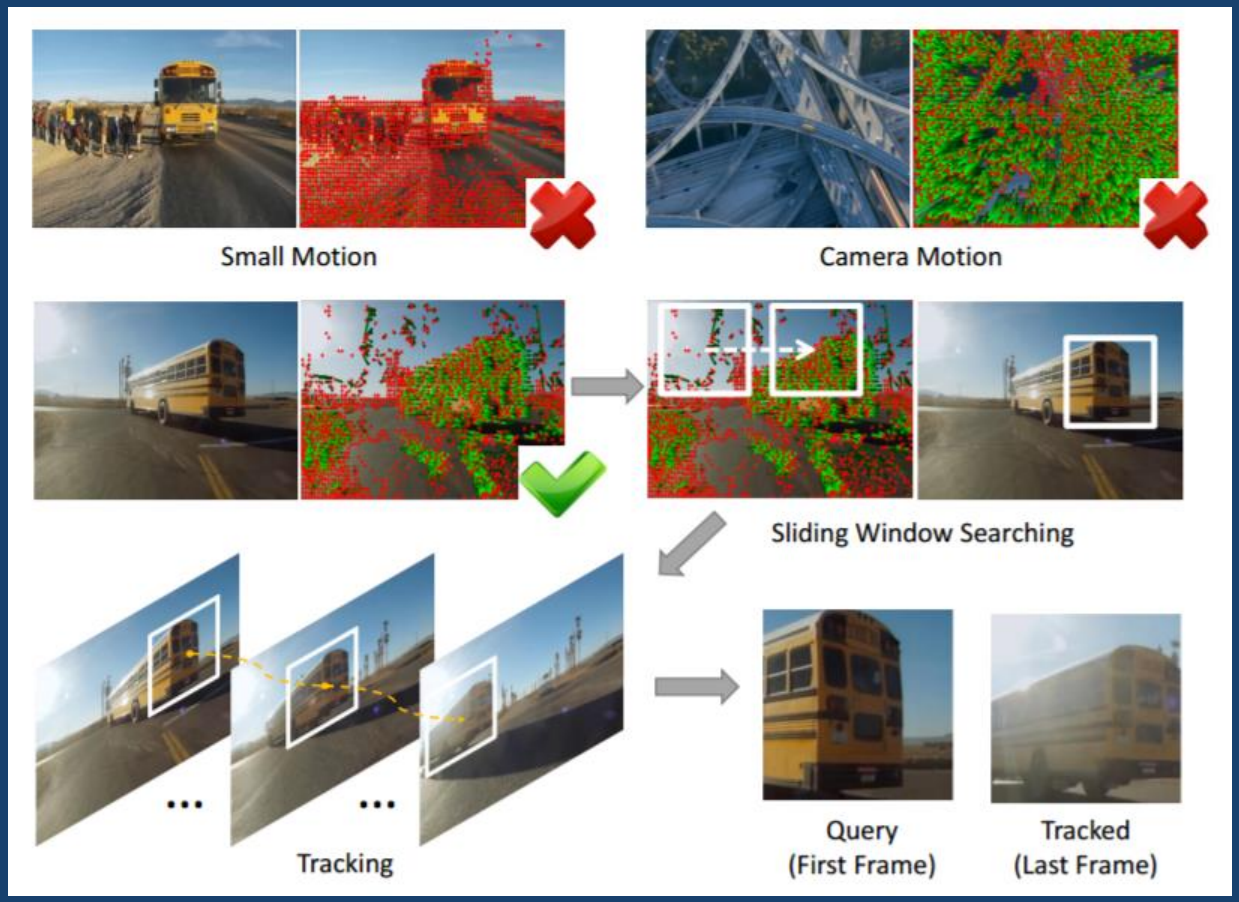
- Two patches connected by a track should have a similar representation in feature space
- Network learns invariance to scaling, lighting, transformation, etc.
- Analogous to how infants learn by tracking objects



Xiaolong Wang, Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015.

TRAINING DATA

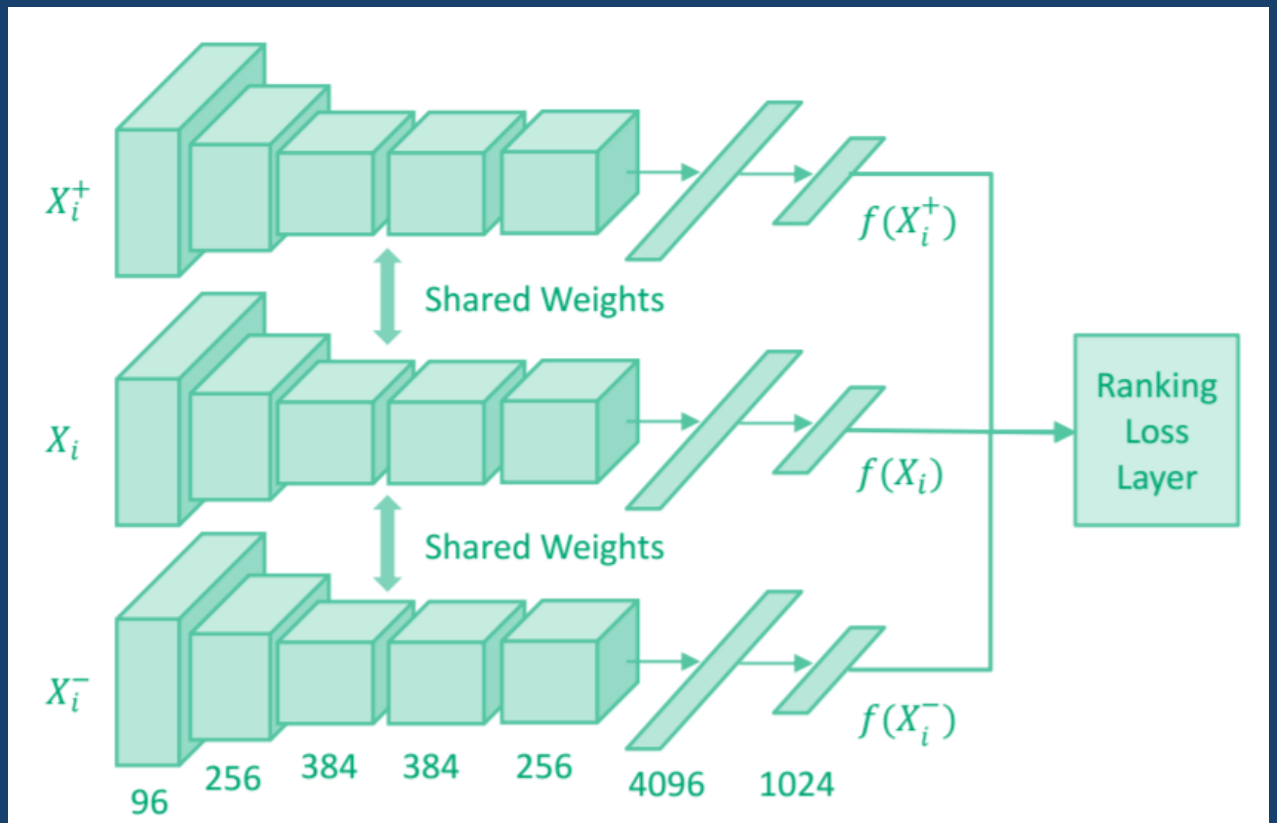
- Obtain feature points and classify point as moving if its trajectory > threshold
- Reject Frames with <25% (noise) or >75% (camera movement) moving points
- Find bounding box that contains greatest number of moving points as query patch
- Track box to obtain paired patch



Xiaolong Wang, Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015.

Architecture and Loss

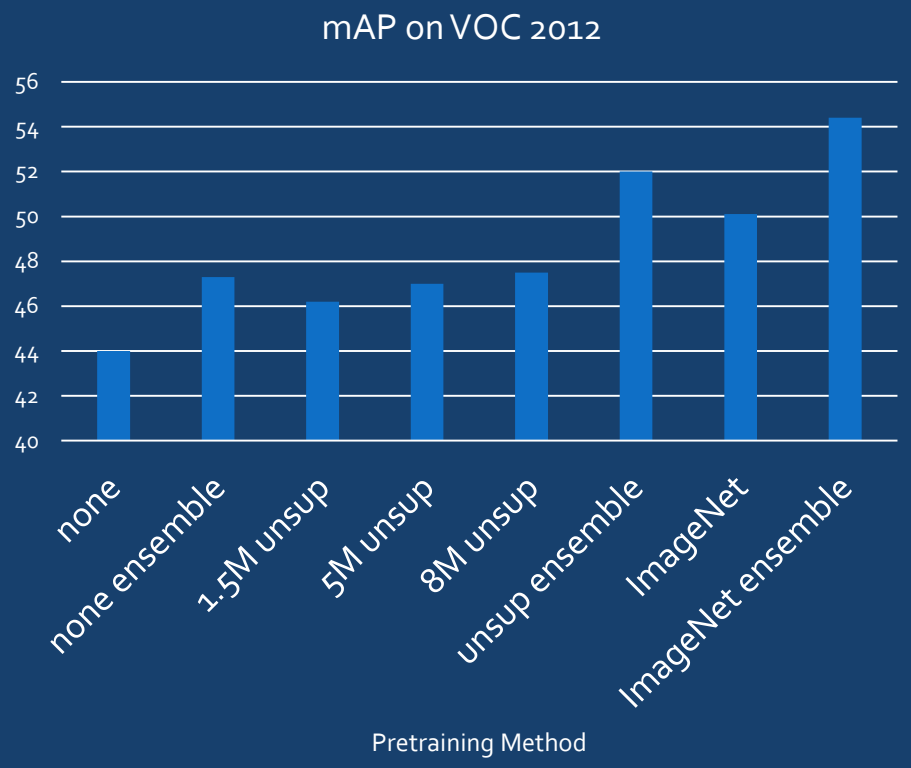
- Siamese Triplet Network based on AlexNet with two stacked fully connected layers on the pool5 outputs
- Triplet loss with margin on the 1024 dimensional feature space
- Hard negative mining to select negative patches

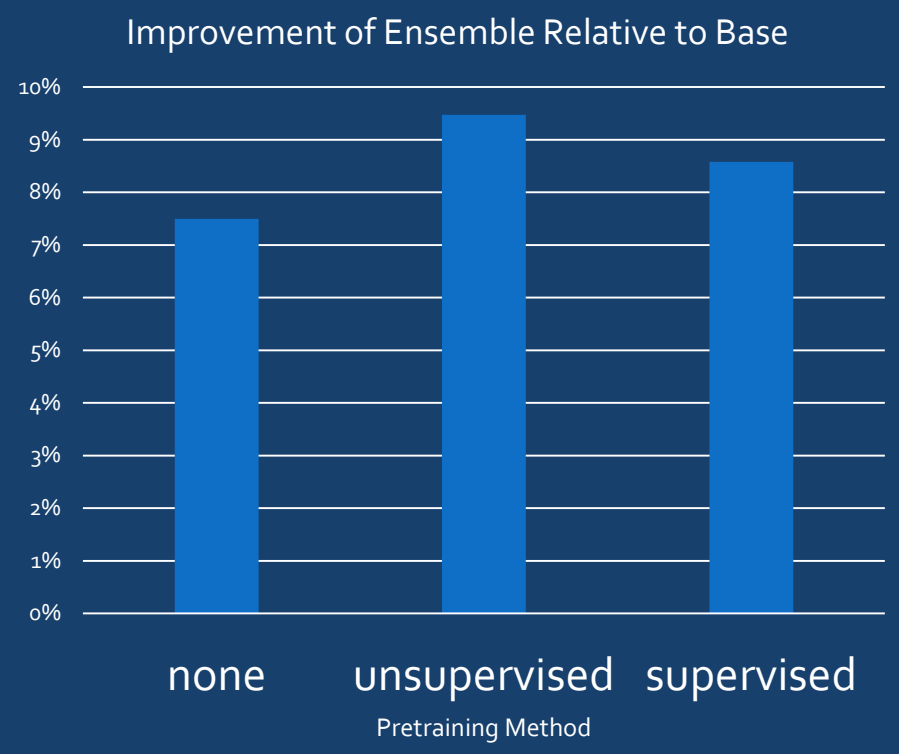


$$\min_W \frac{\lambda}{2} \| W \|_2^2 + \sum_{i=1}^{N} \max\{0, D(X_i, X_i^+) - D(X_i, X_i^-) + M\}$$

Xiaolong Wang, Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015.

PERFORMANCE



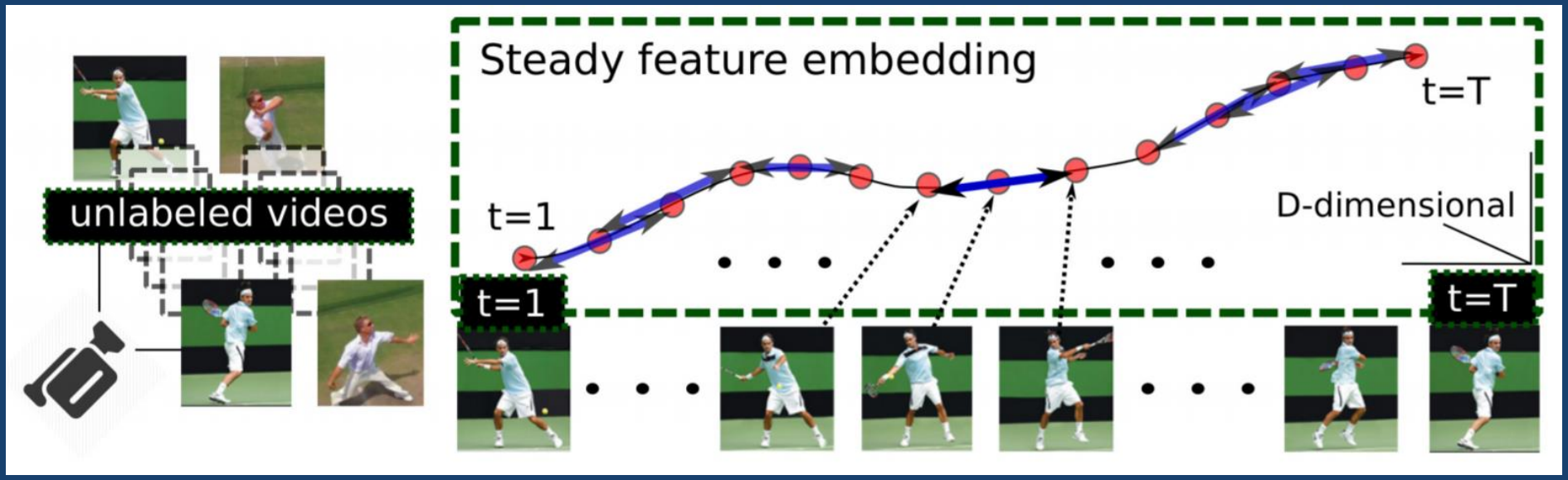


Supervised pretraining outperforms unsupervised pre-training, but adding more data (ensembling) greatly improves unsupervised pre-training

Xiaolong Wang, Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015.

MORE MOTION

- Previous paper explored 'slow' feature embedding
- We can explore 'steady' feature embedding in which changes in input should reflect similar changes in feature space

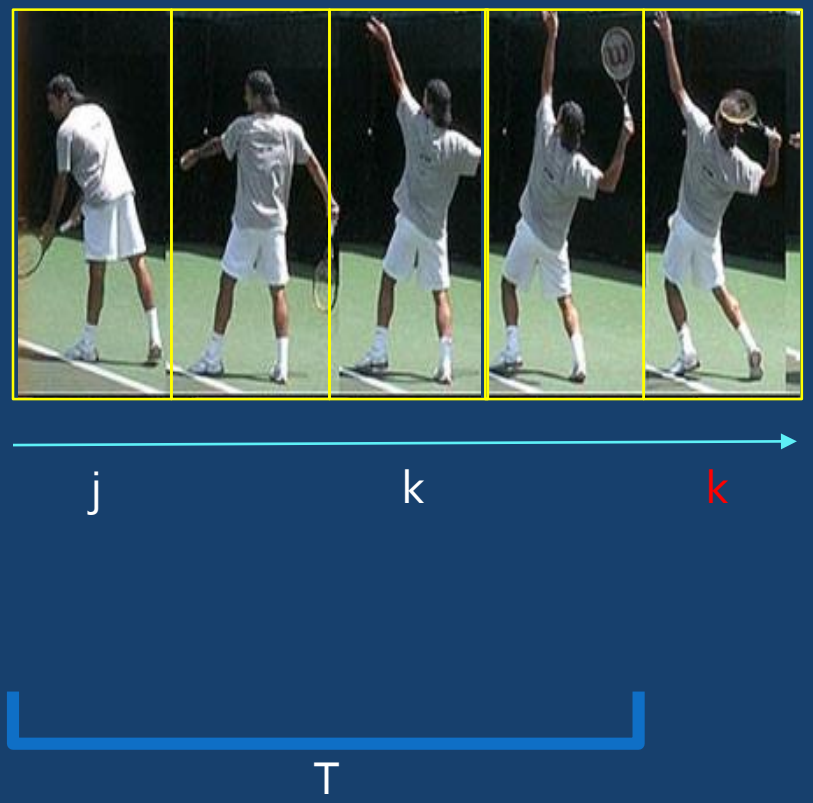


Dinesh Jayaraman, Kristen Grauman, Slow and steady feature analysis: higher order temporal coherence in video, CVPR 2016.

TRAINING DATA

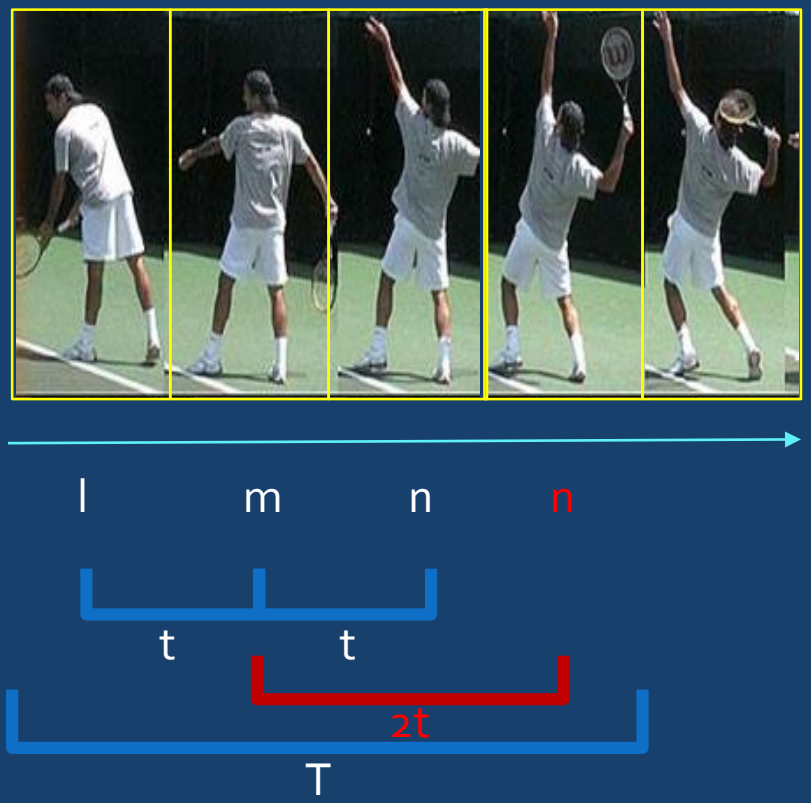
Slow Training Examples

Two frames within temporal window T



Steady Training Examples

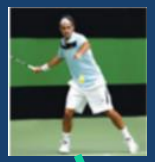
Three frames within temporal window T and equidistant in time from one another



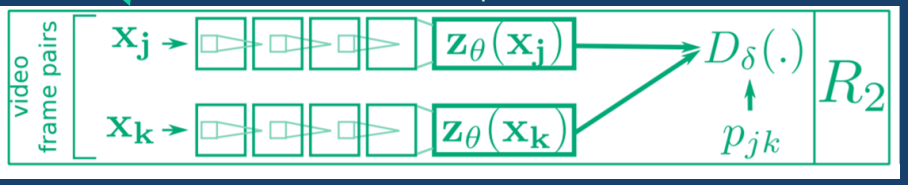
Dinesh Jayaraman, Kristen Grauman, Slow and steady feature analysis: higher order temporal coherence in video, CVPR 2016.

ARCHITECTURE AND LOSS

$$(\boldsymbol{\theta}^*, W^*) = \arg\min_{\boldsymbol{\theta}, W} L_s(\boldsymbol{\theta}, W, \mathcal{S}) + \lambda\left[R_2(\boldsymbol{\theta}, \mathcal{U}) + \lambda' R_3(\boldsymbol{\theta}, \mathcal{U})\right]$$



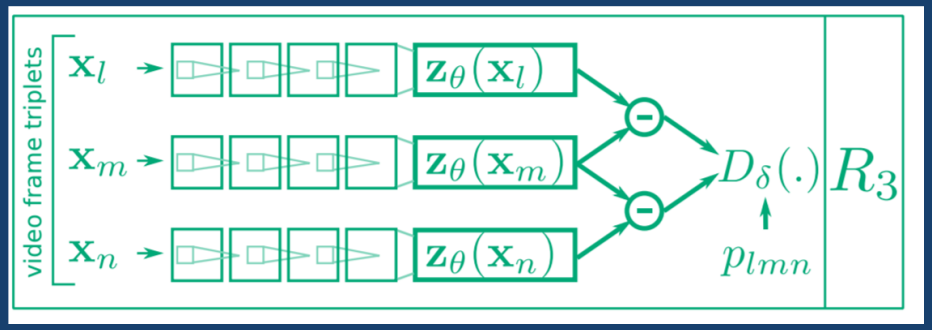
Difference of Images is 1st Order



$$R_2(\boldsymbol{\theta}, \mathcal{U}) = \sum_{(j,k)\in\mathcal{U}_2} D_\delta(\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_k), p_{jk})$$

$$= \sum_{(j,k)\in\mathcal{U}_2} \underbrace{p_{jk}\, d(\mathbf{z}_{\boldsymbol{\theta}j}, \mathbf{z}_{\boldsymbol{\theta}k})}_{\text{Neighbors Representations should be close}} + \underbrace{\overline{p_{jk}}\, \max(\delta - d(\mathbf{z}_{\boldsymbol{\theta}j}, \mathbf{z}_{\boldsymbol{\theta}k}), 0)}_{\text{Stranger Representations should be far (up to a margin)}}$$

Difference of differences is 2nd Order

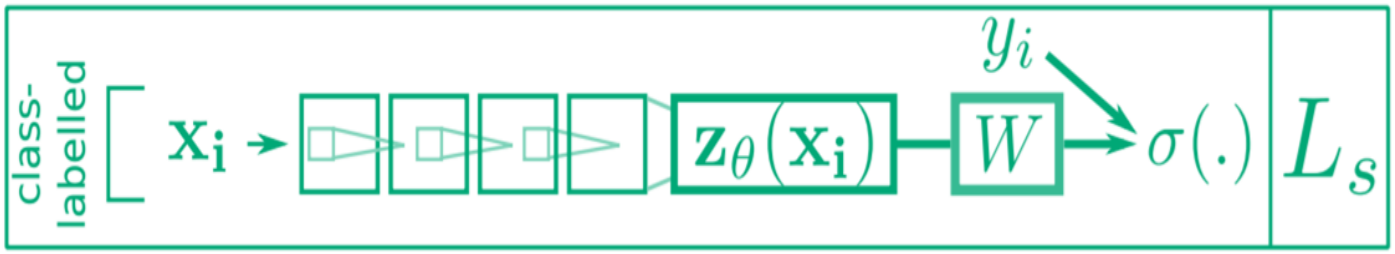


$$R_3(\boldsymbol{\theta}, \mathcal{U}) = \sum_{(l,m,n)\in\mathcal{U}_3} D_\delta(\mathbf{z}_{\boldsymbol{\theta}l} - \mathbf{z}_{\boldsymbol{\theta}m},\ \mathbf{z}_{\boldsymbol{\theta}m} - \mathbf{z}_{\boldsymbol{\theta}n},\ p_{lmn})$$

Dinesh Jayaraman, Kristen Grauman, Slow and steady feature analysis: higher order temporal coherence in video, CVPR 2016.

SUPERVISED PORTION

Included very small amounts of labeled data as part of the loss term

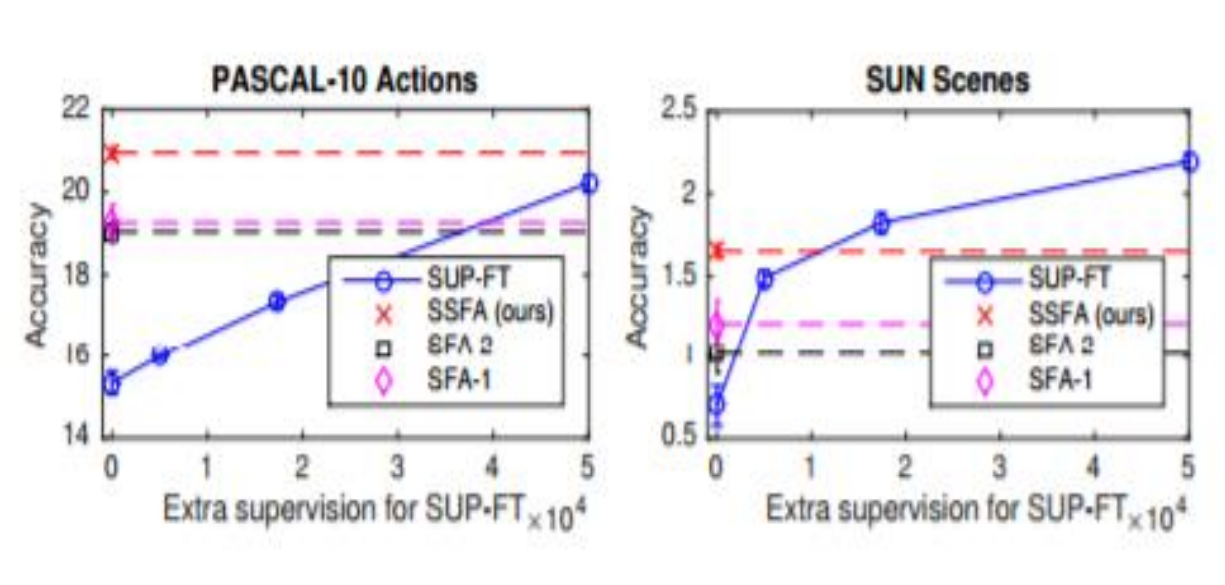


$$L_s(\theta, W, \mathcal{S}) = -\frac{1}{N_s}\sum_{i=1}^{N_s}\log(\sigma_{y_i}(W\mathbf{z}_\theta(\boldsymbol{x}_i)))$$

Dinesh Jayaraman, Kristen Grauman, Slow and steady feature analysis: higher order temporal coherence in video, CVPR 2016.

PERFORMANCE

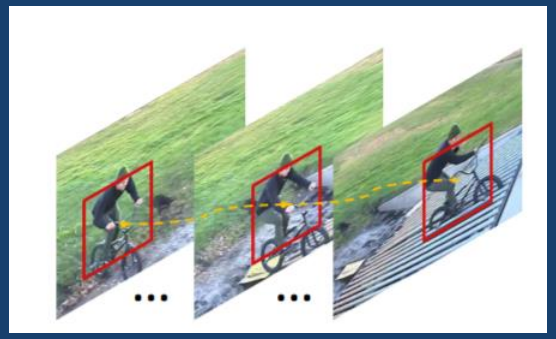
- Unsupervised pretraining method outperforms supervised pretraining method for PASCAL-10 dataset and is competitive with supervised pretraining for SUN scenes



Dinesh Jayaraman, Kristen Grauman, Slow and steady feature analysis: higher order temporal coherence in video, CVPR 2016.

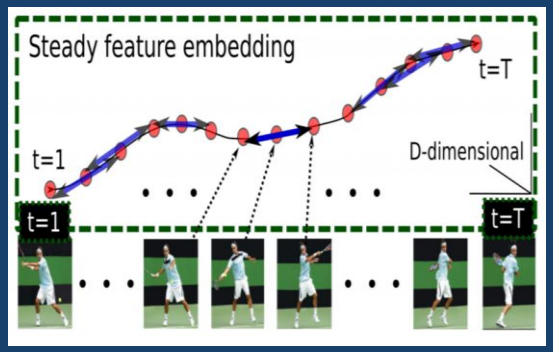
MOTION PAPERS

Wang et al.



- Motion Tracking
- Slow Feature Analysis
- Triplet Ranking Loss
- Completely Unsupervised

Jayaraman et al.



- No Tracking Needed
- Slow + Steady Feature Analysis
- Contrastive Loss
- Semi-Supervised

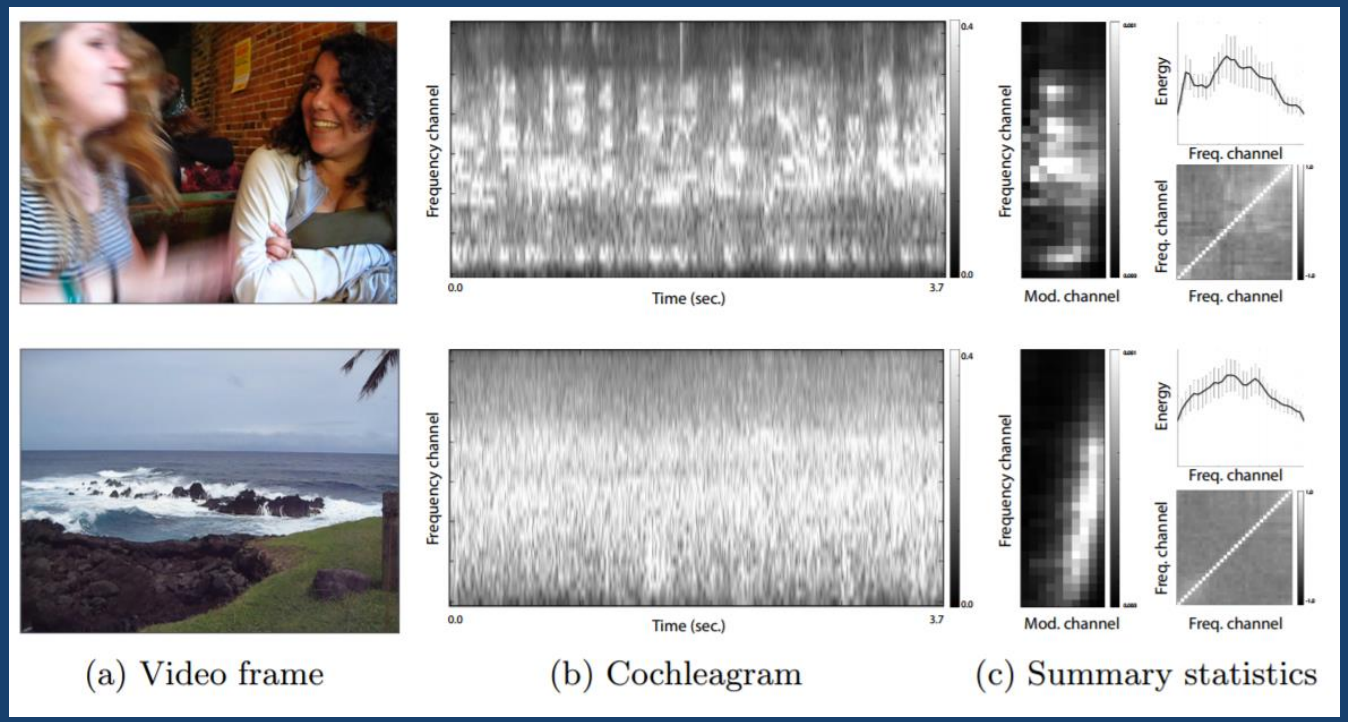
All rely on the concept of small localized semantically predictable motion in Video

WEAK SUPERVISION

Ambient Noise & Noisy Labels

AUDIO AS SUPERVISION

Audio is largely invariant to camera angle, lighting, scene composition and Carries a lot of information about the semantics of an image.

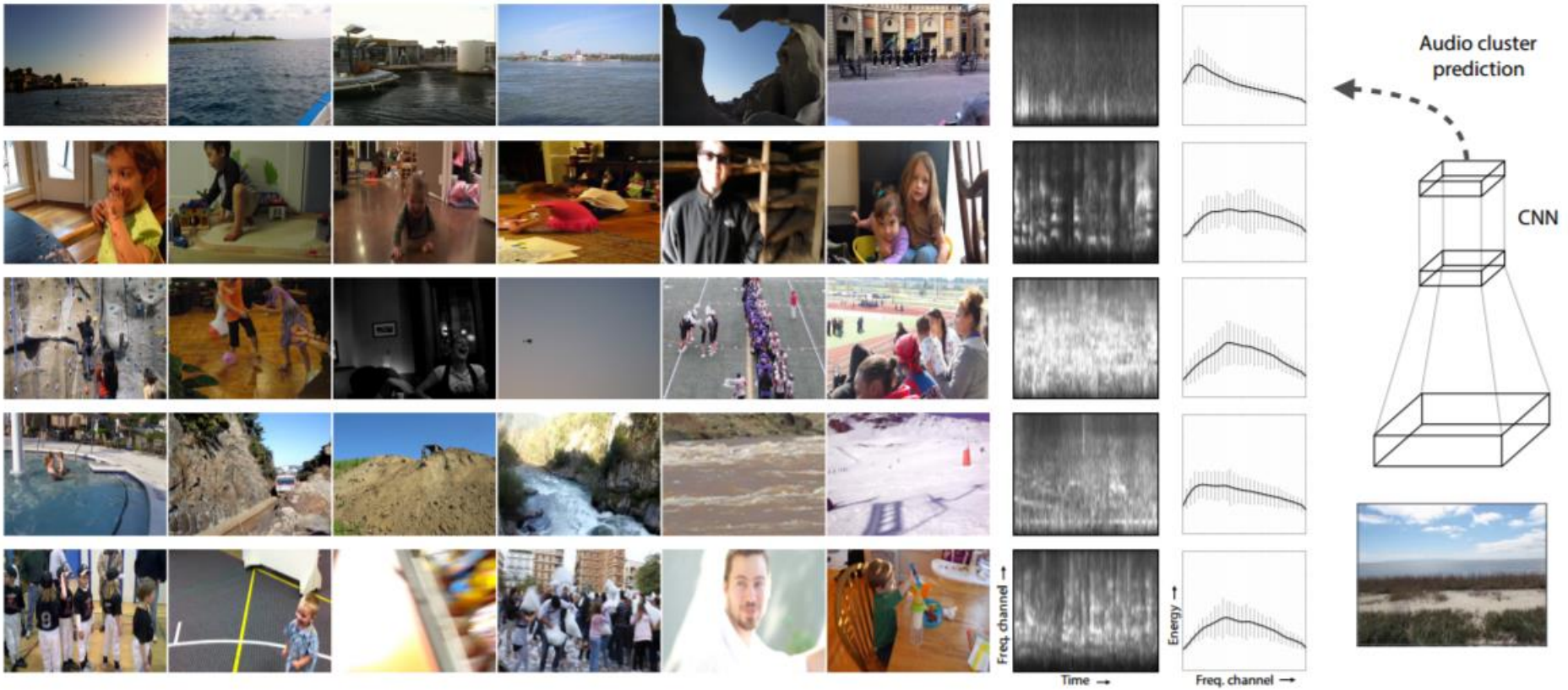


Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, Antonio Torralba, Ambient Sound Provides Supervision for Visual Learning, ECCV 2016.

ARCHITECTURE AND TRAINING

- CaffeNet + BatchNorm
- 360k videos from flicker
- 10 random frames from each video
- 1.8 million image frames in total
- Many were post processed

PREDICTION MODEL



(a) Images grouped by audio cluster (b) Clustered audio stats. (c) CNN model

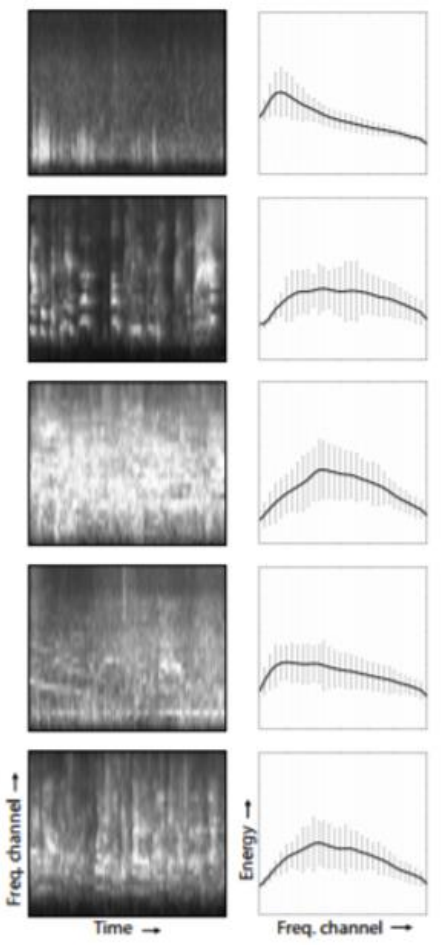
Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, Antonio Torralba, Ambient Sound Provides Supervision for Visual Learning, ECCV 2016.

SOUND REPRESENTATION

Audio Texture Summarization

McDermott, J.H., Simoncelli, E.P.: Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71(5), 926–940 (2011)

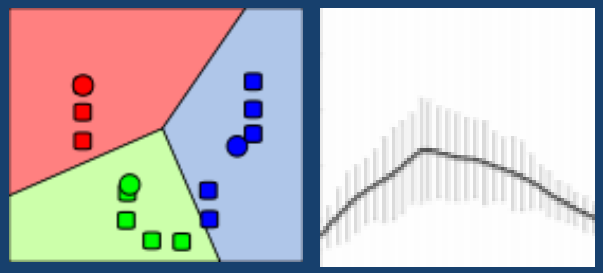
- Timing is very hard to predict given static image
- Audio Averaged over multiple seconds
- Audio Decomposed and amplified to mimic human hearing conditions
- 512 Dimensional vector



Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, Antonio Torralba, Ambient Sound Provides Supervision for Visual Learning, ECCV 2016.

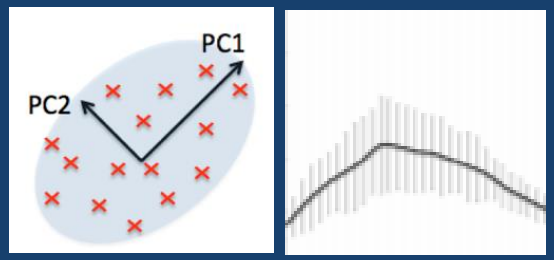
SOUND REPRESENTATION

Clusters



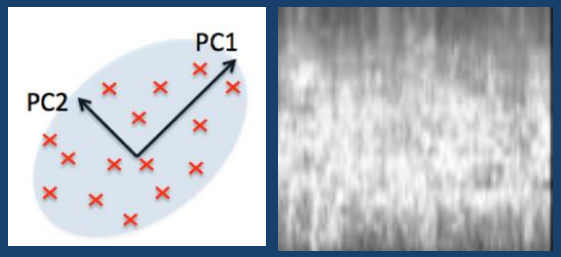
- Audio Texture
- K-means, single cluster allocation (unlike colorization)
- Toss out examples more than median distance from cluster

Binary Encodings



- Audio Texture
- PCA + Binary thresholding on each Dimension
- Cross-Entropy Loss

Spectrum



- Frequency Representation
- PCA + Binary thresholding on each Dimension
- Cross-Entropy Loss

Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, Antonio Torralba, Ambient Sound Provides Supervision for Visual Learning, ECCV 2016.

NEURON ACTIVATIONS REFLECT OBJECTS WITH SOUND CLUSTER TRAINING

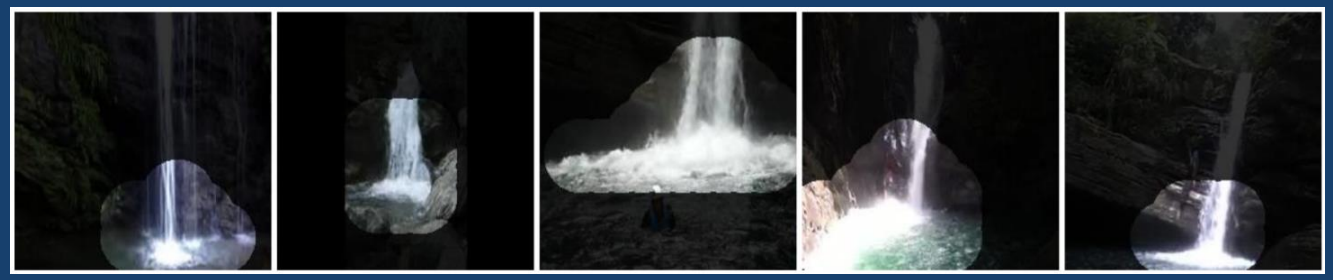
- Sample 200k Images from Test Set
- Find top 5 that spike a Neuron in 5th Convolutional Layer
- Use "synthetic visualization" to find receptive field.
- 91/256 units

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: ICLR 2015 (2014)

Neuron A = Baby

Neuron B = Sky

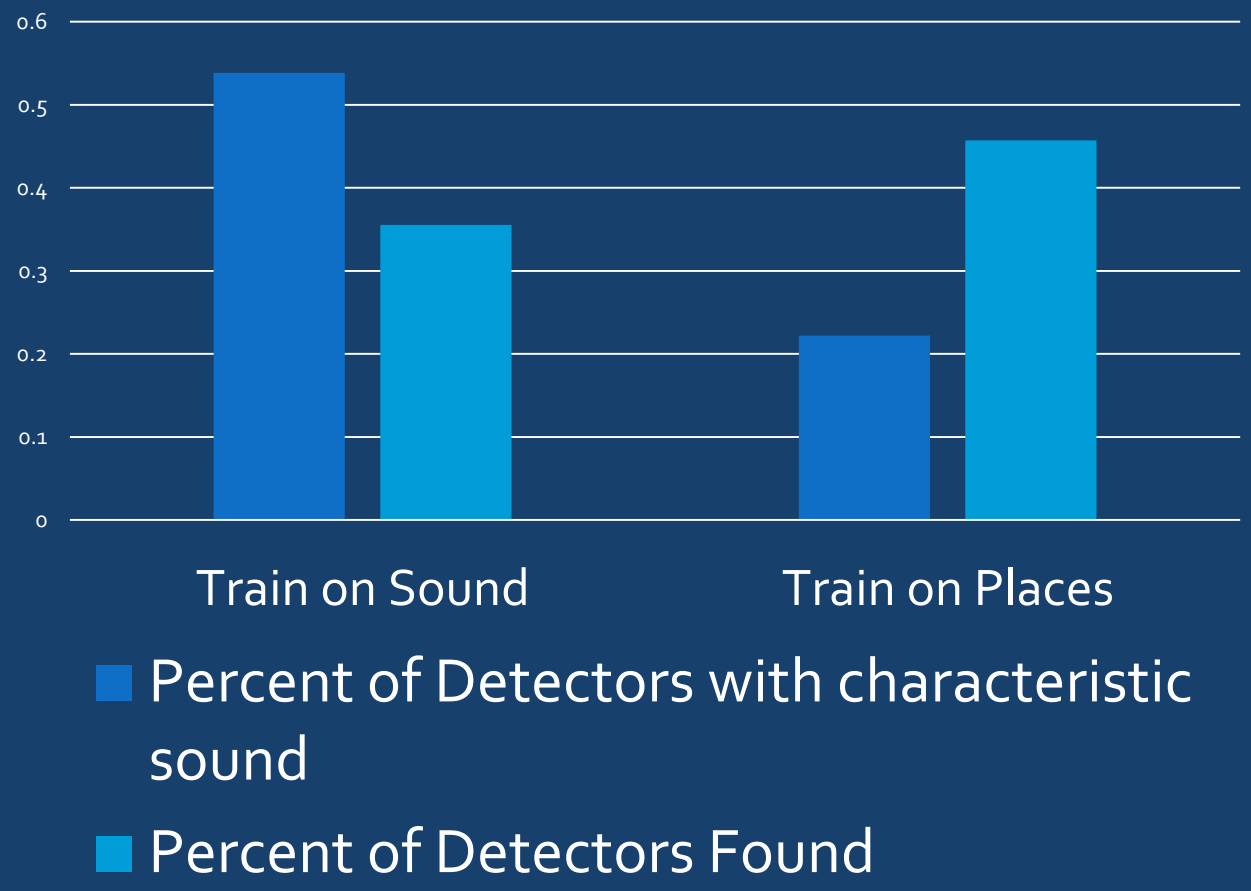
Neuron = Waterfall



SOUND VS PLACES

Places and Sound Learn different Detectors

Sound can Boost overall performance with an ensemble approach



MOTION VS SOUND TRANSFER

Classification
Train SVM weights at end of network. Tune with grid search

Detection
Fine Tuning with Fast-RCNN

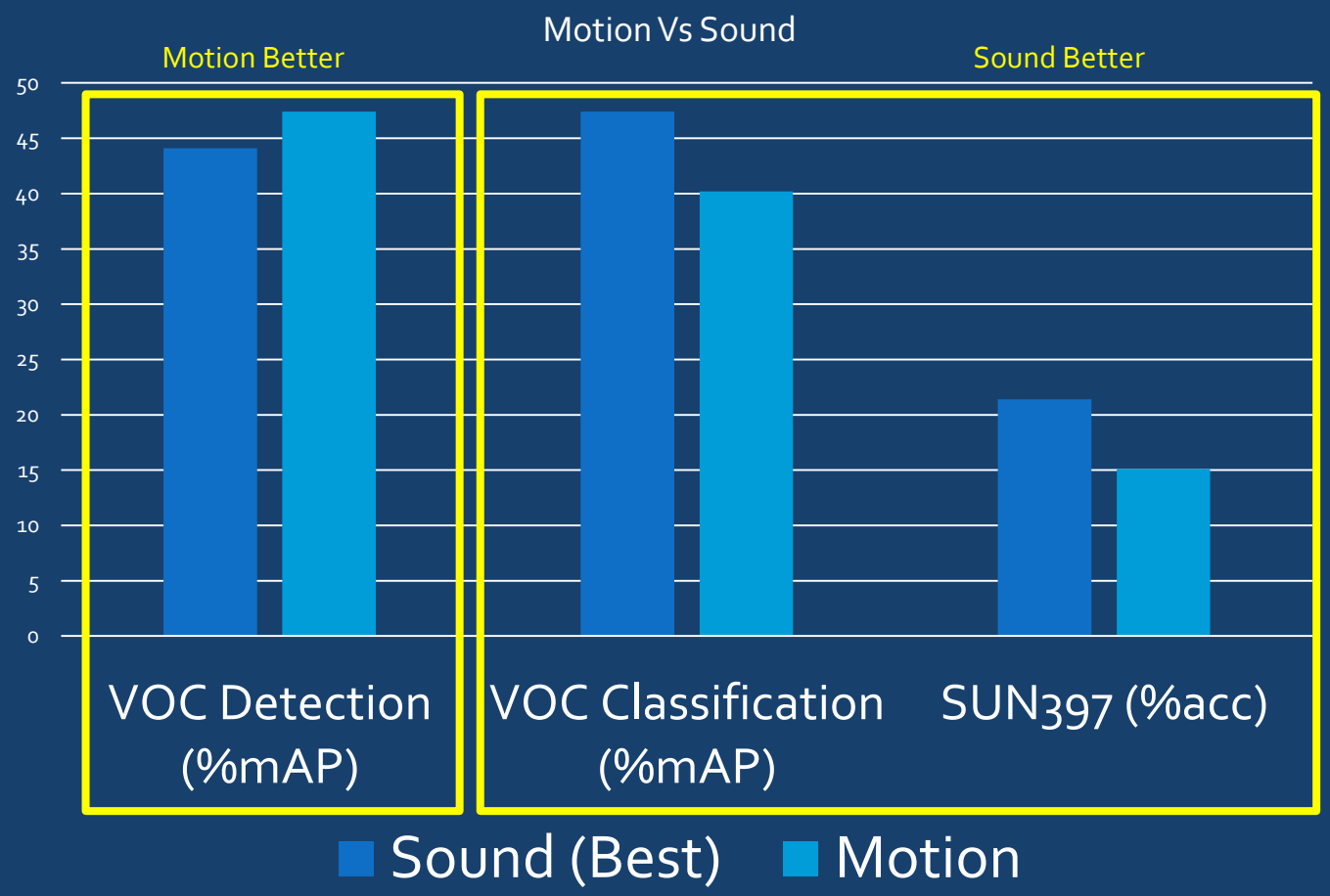
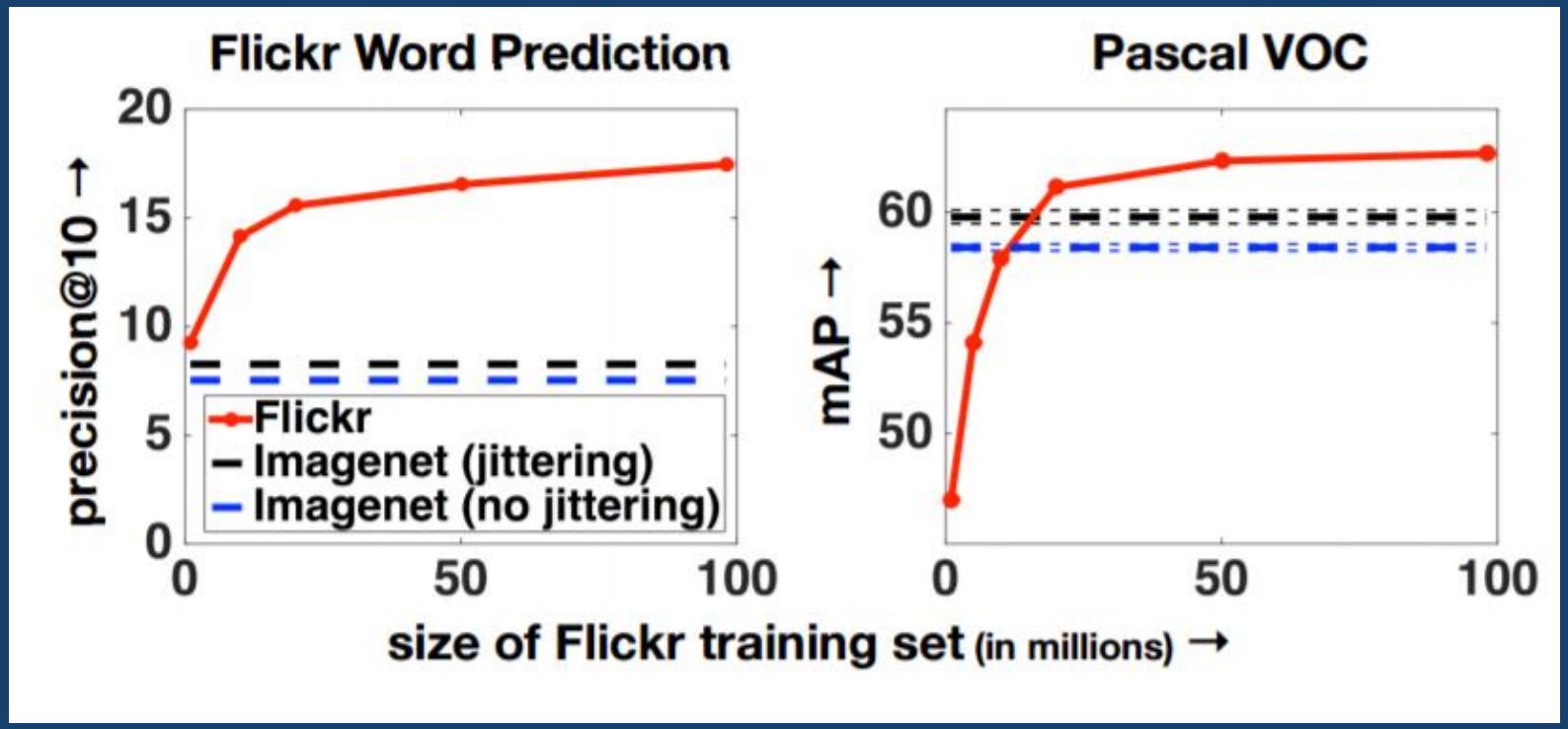


IMAGE ANNOTATION

On average, Bag of words should relate to semantic content of image. Predicting the bag of words suggests the model understands human focus and labels.

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

RESULTS



Pretrained

- AlexNet on ImageNet features
- Logistic classifier

End-to-end

- Multi-class logistic

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

WEAK SUPERVISION PAPERS

Joulin et al.

- Flicker image captions instead of imagenet labels or pascal segments
- Multilabel Loss function.
- Massive data will find signal in noisy captions

Owens et al.

- Flicker video provides audio data to supervise image learning
- Audio statistic prediction
- Massive data will find signal in noisy captions

SUMMARY

What did we learn?

SUPERVISION SPECTRUM

Labels/Values	Strong Supervision	Weak Supervision	Self Supervision
Source	Experts / Experiments	Humans / Somewhere	Data Itself
Quality	Low Noise	High Noise	Variable Noise
Relevance	Good-Great	Poor-Good	???
Cost	High	Low-Free	Free
Amount	Low	High	Massive
Examples	ImageNet , Pascal VOC	Flicker, Snapchat	Color, Motion, Audio

RELATIONS TO OTHER TOPICS

Self Supervision

Is compatible with

- Convolutional Networks
- Recurrent Networks

Can improve convergence and performance

- Training Techniques

Is framed as and useful For

- General Classification / Regression
- Pixel Wise Generation / Classification
- Ranking & Similarity Learning

Is an umbrella category including

- VAEs / GANS

Image Content

- Carl Doersch, Abhinav Gupta, Alexei A. Efros, Context as Supervisory Signal: Discovering Objects with Predictable Context, ECCV 2014.
- Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros, Context Encoders: Feature Learning by Inpainting, CVPR 2016.

Colorization

- Richard Zhang, Phillip Isola, Alexei A. Efros, Colorful Image Colorization, ECCV 2016.
- Gustav Larsson, Michael Maire, Gregory Shakhnarovich, Learning Representations for Automatic Colorization, ECCV 2016.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction, CVPR 2017

Video

- Xiaolong Wang, Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015.
- Chelsea Finn, Ian Goodfellow, Sergey Levine, Unsupervised Learning for Physical Interaction through Video Prediction, NIPS 2016.
- Dinesh Jayaraman, Kristen Grauman, Slow and steady feature analysis: higher order temporal coherence in video, CVPR 2016.

Weak Supervision

- Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.
- Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, Antonio Torralba, Ambient Sound Provides Supervision for Visual Learning, ECCV 2016.

QUESTIONS?

Thank You

START OF SUPPLEMENT

(Refuse Too)

DENOISING AUTOENCODER

General Purpose Feature Representation Learning

Add Noise $\varphi: X \rightarrow \tilde{X}$

Project into latent representation $\phi: \tilde{X}, X \rightarrow F$

$$X \in R^x \quad F \in R^f \quad x \ll f$$

Project back into original representation $\psi: F \rightarrow X$

Minimize Squared Loss $\underset{\psi,\phi}{\operatorname{argmin}} \|X - (\psi \circ \phi)X\|^2$

PREPROCESSING

- 100 Million Flicker images and caption
- Images cropped to central 224 x 224
- Drop 500 most frequent tokens (The, a, it, etc.)
- Predict Multi-label bag of words size 1k, 10k, and 100k

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

TWO LOSSES

One vs All Loss
- Sensitive to Class Imbalance

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{y_{nk}}{N_k}\log\sigma(f(\mathbf{x}_n;\theta))+\frac{1-y_{nk}}{N-N_k}\log(1-\sigma(f(\mathbf{x}_n,\theta))),$$

Multi-Class Logistic Loss
+ Performs better empirically in all experiments

$$\ell(\theta,\mathbf{W};\mathcal{D})=\frac{-1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}y_{nk}\log\left[\frac{\exp(\mathbf{w}_k^\top f(\mathbf{x}_n;\theta))}{\sum_{k'=1}^{K}\exp(\mathbf{w}_{k'}^\top f(\mathbf{x}_n;\theta))}\right]$$

Ranking Loss
- Too Slow

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

LARGE SOFTMAX WOES

Stochastic Gradient over Targets

100K Softmax is huge and slow. Can reduce training from months to weeks by only updating weights for the potential 128 words per image.

$$s_k = \exp\left(\mathbf{w}_k^\top f(\mathbf{x}_n;\theta)\right)$$

$$\mathbb{E}\left[\log \sum_{c\in\mathcal{C}} s_c\right] \leq \log\left(\sum_{k=1}^{K} s_k\right) = \log(Z).$$

Approximate Loss does NOT overestimate True Loss

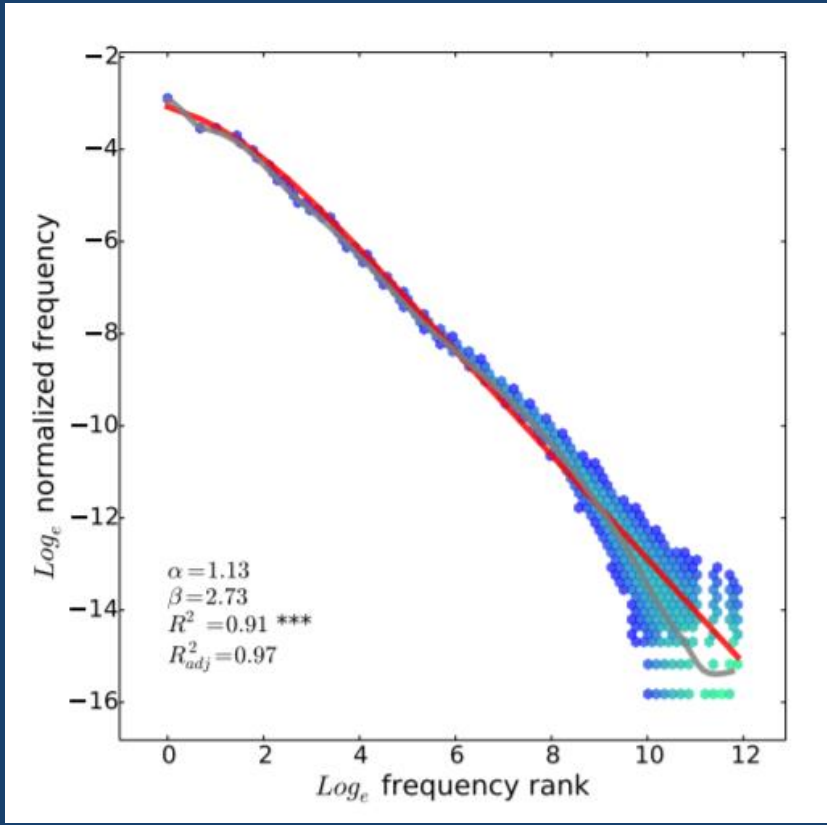
$$\mathbb{E}\left[\log \sum_{c\in\mathcal{C}} s_c\right] \geq P\left(\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} s_c \geq \frac{1}{K}Z\right)\left(\log\frac{|\mathcal{C}|}{K} + \log Z\right)$$

Found Lower Bound on expected loss

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

SAMPLING AND BALANCE

- Zipf Distribution of words
  - Many Infrequent Words
  - Few, Highly Frequent Words
- Sampling
  - Pick a Word uniformly at random
  - Pick image with that word. All other words for that image are considered noisy
  - Noisier Gradients but Fast and Empirically strong

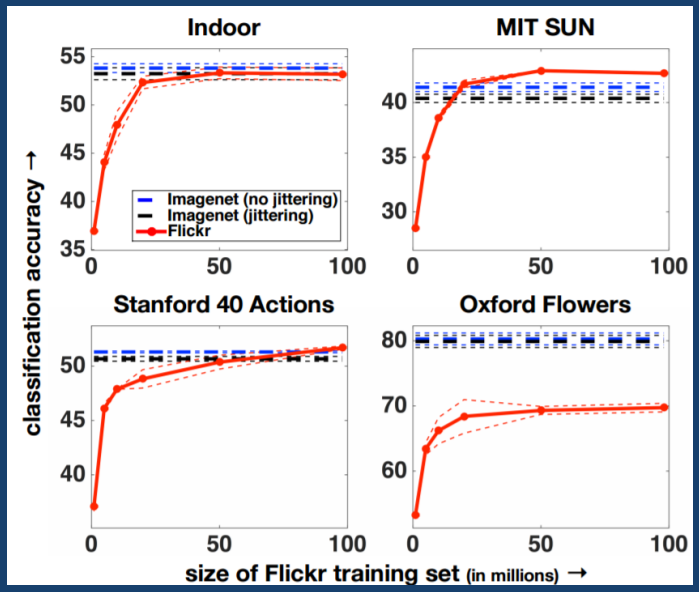


Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

TRANSFER LEARNING

Week Supervision can be great in combination with regular supervision but underperforms on its own.

Dataset	Model	Indoor	SUN	Action	Flower	Sports	ImNet
Imagenet	AlexNet	53.82	41.40	51.27	80.28	86.07	53.63
	GoogLeNet	64.00	48.76	67.10	79.05	95.91	69.89
Flickr	AlexNet	53.19	42.67	51.69	69.72	86.79	34.93
	GoogLeNet	55.56	44.43	52.84	65.80	87.40	33.62
Combined	AlexNet	58.76	47.27	56.35	83.28	87.50	–
	GoogLeNet	67.87	55.04	69.19	83.74	95.79	–

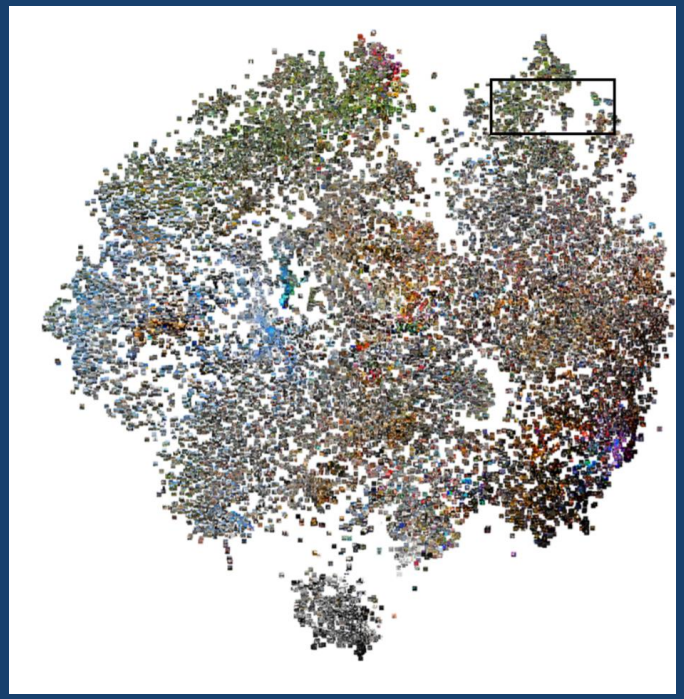


Results suggest GoogLeNet has too small capacity for Flicker

Dataset	Model																						mAP
Imagenet	AlexNet	75.7	61.9	66.9	66.5	29.3	56.1	73.5	68.0	47.1	40.9	57.4	60.0	74.0	63.2	86.2	38.8	57.9	45.5	75.7	51.1	59.8
	GoogLeNet	91.3	84.0	88.4	87.2	42.4	79.6	87.3	85.0	59.1	66.5	69.5	83.3	86.6	82.9	88.4	57.5	75.8	64.6	89.5	73.8	77.1
Flickr	AlexNet	84.0	72.2	70.2	77.0	29.5	60.8	79.3	69.5	49.2	40.5	54.0	57.1	79.2	64.6	90.2	43.0	47.5	44.1	85.0	50.7	62.4
	GoogLeNet	91.5	83.7	84.1	88.5	41.7	78.0	86.8	84.0	54.7	55.5	63.3	78.5.	86.0	77.4	91.1	51.3	60.8	52.7	91.9	60.9	73.2
Combined	AlexNet	82.96	70.32	73.28	76.29	32.21	61.84	79.81	72.91	51.56	43.82	60.77	63.32	78.63	67.72	90.26	45.45	53.15	49.14	84.8	55.8	64.7
	GoogLeNet	94.09	85.03	89.71	88.47	49.35	81.47	88.1	85.2	60.51	68.37	71.65	85.81	88.87	85.22	88.69	60.45	77.26	66.61	90.71	74.49	79.0

VISUALIZATION

T-SNE based on last layers of Alex Net with 1k Words



Visually and Semantically similar !

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

WORD EMBEDDINGS

Last layer of ALexNet can be interpreted as word embedding's

Surprising that vector space operations are preserved in deep model

Analogical Reasoning Tests

Model	K = 1,000	K = 10,000	K = 100,000
AlexNet	67.91	29.29	0.85
GoogLeNet	71.92	24.06	–
word2vec	71.92	61.35	47.24
AlexNet + word2vec	74.79	57.26	44.35
GoogLeNet + word2vec	75.36	56.05	–

Word Similarity Rank Statistics

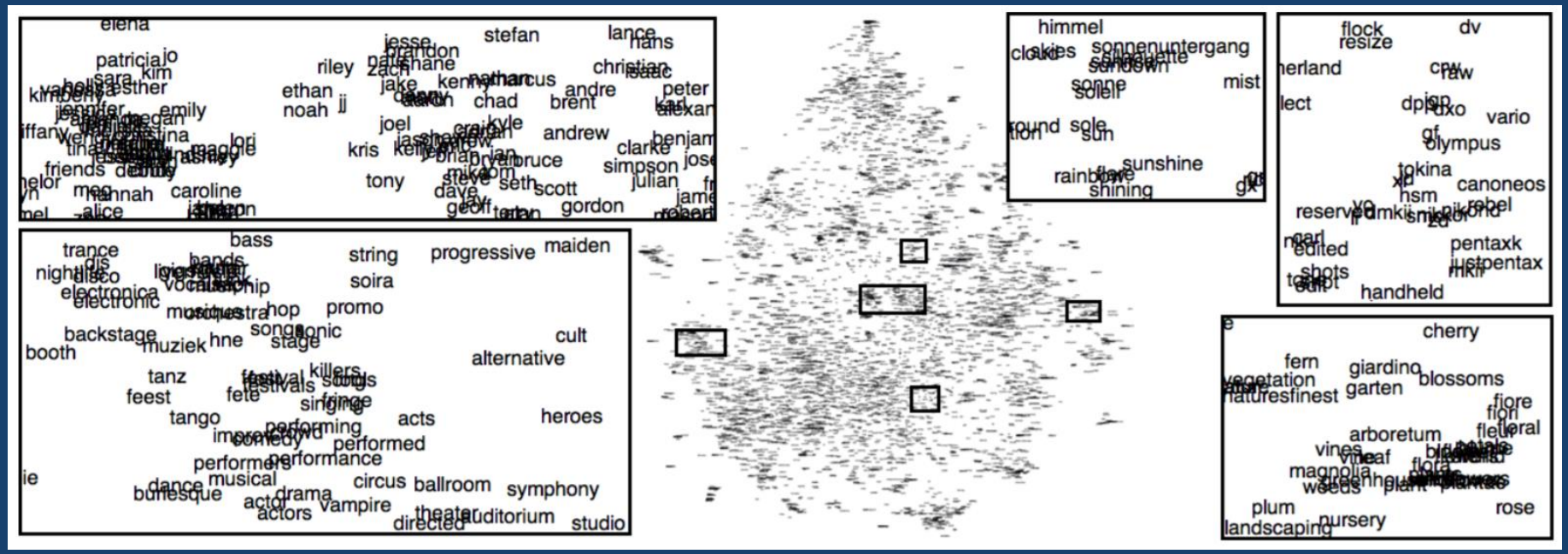
Model	K = 1,000	K = 10,000	K = 100,000
AlexNet	73.77	75.73	67.35
GoogLeNet	75.72	75.89	–
word2vec	75.25	77.53	77.91
AlexNet + word2vec	78.17	79.24	78.57
GoogLeNet + word2vec	78.75	79.11	–

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

WORD EMBEDDING VISUALIZATION

Gendered Names

Strange Results



Roughly Music/ Arts

Garden

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

DISCUSSION

- Bag of Words loses valuable relations between words.
- Joint embedding of text and images we will see later improves upon this
- Stochastic Gradient Descent over Targets is generally useful for large SoftMax Multi-label prediction

GENERAL STRATEGY

1. Sample Random patches from PASCAL VOC
2. Using HOG Space, Find top few Nearest Neighbors for each patch. These are clusters proposals.
3. Discard Patches that are inconsistent with other patches
4. Rank the clusters by sum of patch scores
5. Discard clusters that do not contain visually consistent objects

Score Nth Patch by predicting its context using the the N-1 other patches. But, Must account for difficulty of prediction.

ARCHITECTURE

AlexNet

- 15M – 415M Parameters
- 7 Layers
- 2 weeks to train

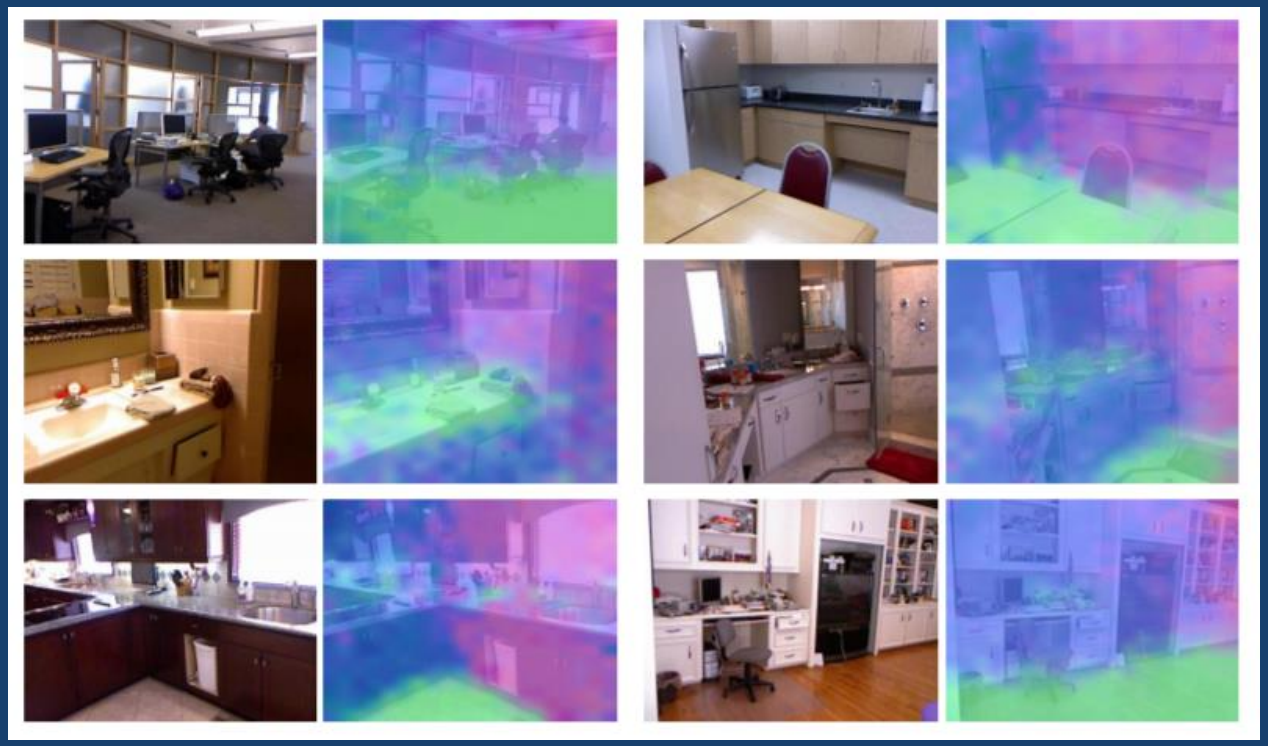
GoogleLeNet

- 4M – 404M Parameters
- 12 layers
- Auxiliary classifier
- 3 weeks to train

Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache, Learning Visual Features from Large Weakly Supervised Data, ECCV 2016.

SURFACE NORMAL PREDICTION

	(Lower Better)		(Higher Better)		
	Mean	Median	11.25°	22.5°	30°
scratch	38.6	26.5	33.1	46.8	52.5
unsup + ft	34.2	21.9	35.7	50.6	57.0
ImageNet + ft	33.3	20.8	36.7	51.7	58.1
UNFOLD [13]	35.1	19.2	37.6	53.3	58.9
Discr. [25]	32.5	22.4	27.4	50.2	60.2
3DP (MW) [12]	36.0	20.5	35.9	52.0	57.8



Xiaolong Wang, Abhinav Gupta, Unsupervised Learning of Visual Representations using Videos, ICCV 2015.