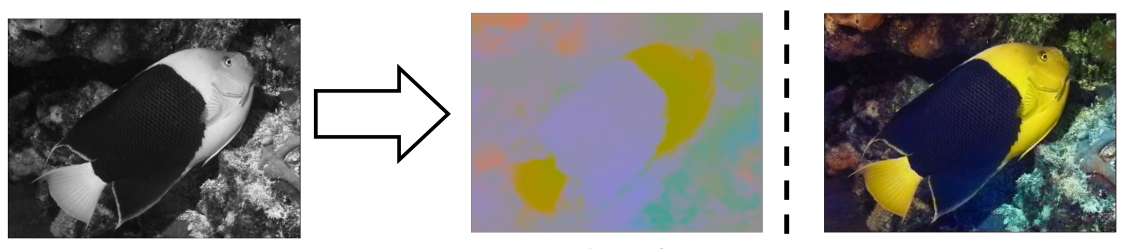
Colorful Image Colorization

Richard Zhang Phillip Isola Alexei A. Efros
Department of Electrical Engineering and Computer Sciences, UC Berkeley

Additional examples, Try our model!
richzhang.github.io/colorization

Problem Statement

Given a grayscale image, predict the color



Input: Grayscale image *L* channel — **Output:** Color information *ab* channels — Concatenate (*L*,*ab*) for plausible colorization

Our contributions

1) Graphics Task of Colorization
a) achieve state-of-the-art by training on 1M ImageNet photos
b) design an appropriate objective function that handles the multimodal uncertainty and captures a wide diversity
c) introduce a novel framework for testing colorization algorithms, potentially applicable to other image synthesis tasks

2) Colorization as Representation Learning
a) introduce colorization task as instance of ***cross-channel encoding***
b) evaluate colorization for representation learning, demonstrate competitive performance in self-supervision framework

Inherent Ambiguity

Multiple plausible colorizations may exist
➔ L2 loss is inadequate for this problem

Ground Truth — Grayscale — Our Output

Our Loss Function

Grayscale Image to color distribution
- ***multinomial classification*** problem
- quantize *ab* space into grid size 10, keep 313 bins in gamut
- cross entropy loss

$$L(\hat{\mathbf{Z}},\mathbf{Z}) = -\frac{1}{HW}\sum_{h,w} v(\mathbf{Z}_{h,w})\sum_q \mathbf{Z}_{h,w,q}\log(\hat{\mathbf{Z}}_{h,w,q})$$

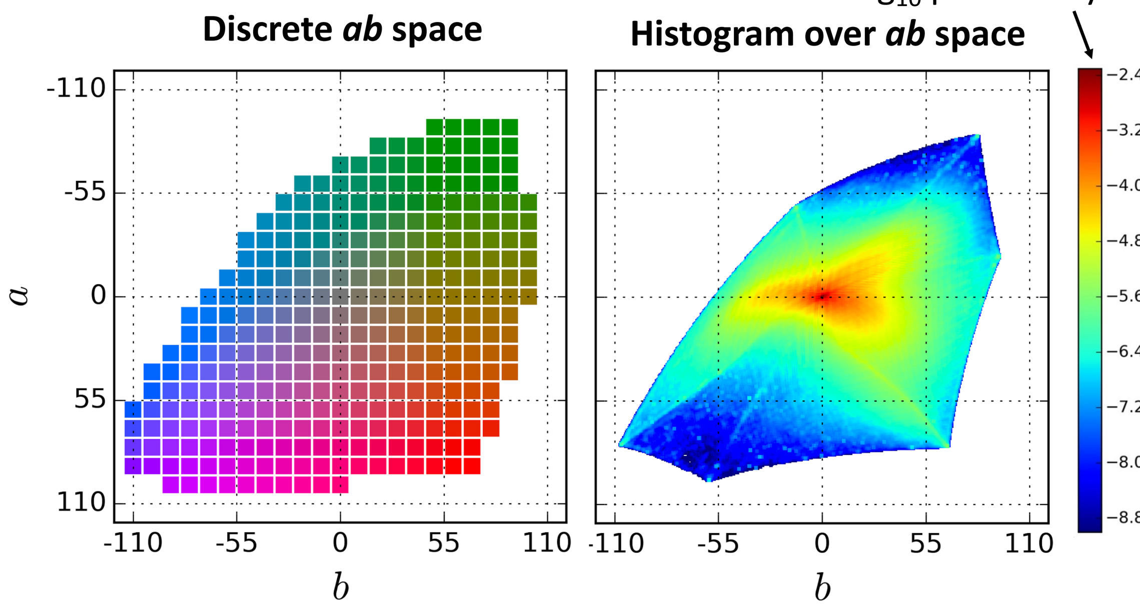
Rarity weighting — Target distribution — Predicted distribution

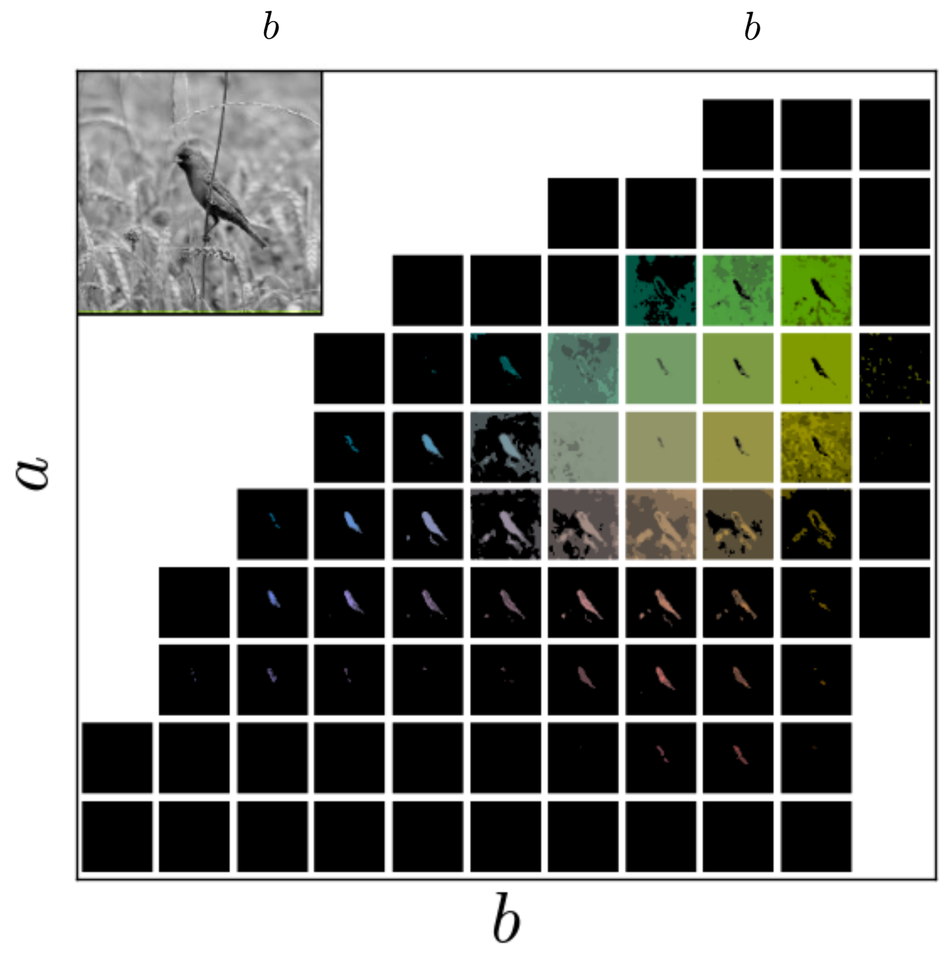
- ***Class rebalancing*** to encourage learning of *rare* colors

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

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

reweighting — empirical distribution — combine with uniform





Per-pixel Color Distribution to Single Point Estimate

- Mean is spatially coherent but desaturated
- Mode is vibrant but can have artifacts
- ***Interpolate*** between mean and mode with ***annealed-mean***

$$\mathcal{H}(\mathbf{Z}_{h,w}) = \mathbb{E}\big(f_T(\log \mathbf{Z}_{h,w})\big), \quad f_T(\mathbf{z}) = \frac{\exp(\mathbf{z}/T)}{\sum_q \exp(\mathbf{z}_q/T)}$$

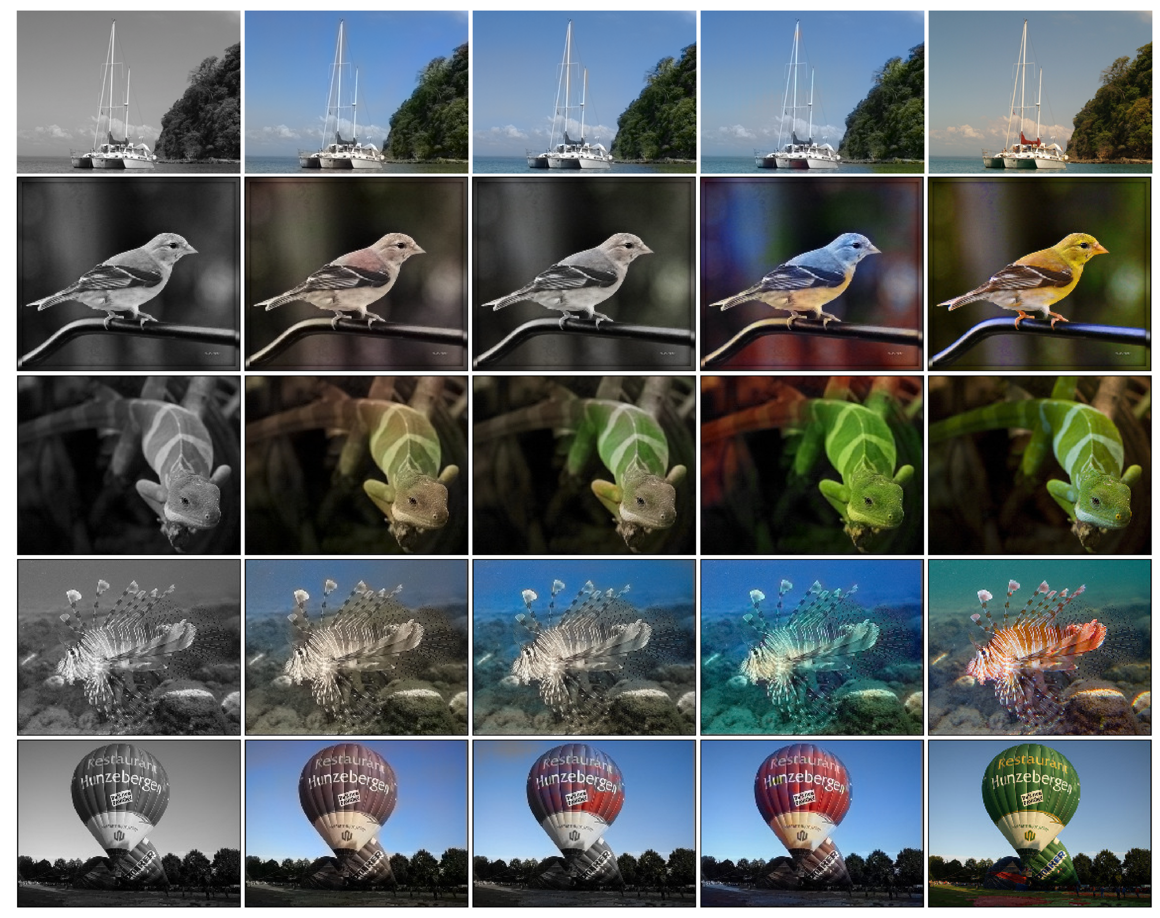
expectation over annealed distribution — annealed distribution

Mean (T=1) — **Annealed-Mean (T=.38)** — **Mode (T➔0)**

Qualitative Comparisons

Success Cases

Grayscale | L2 Regression | Classification | Classification w/ Rebal | Ground Truth

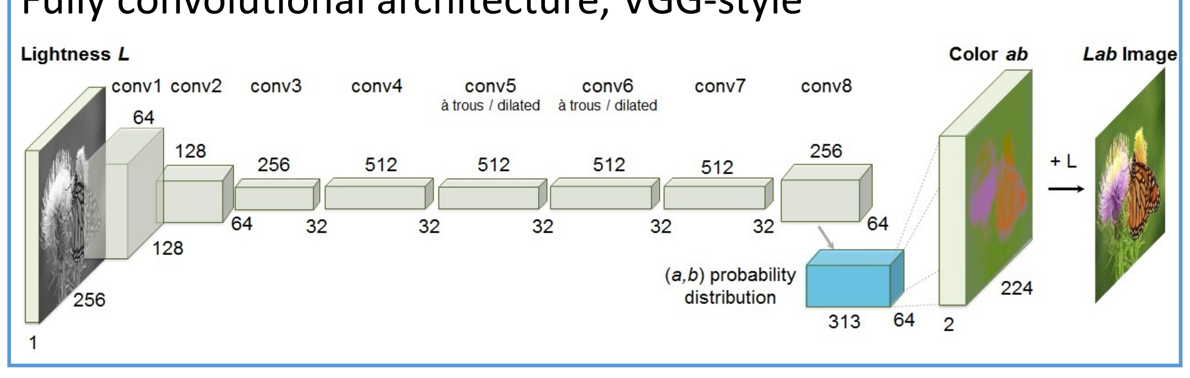


Failure Cases

Grayscale | L2 Regression | Classification | Classification w/ Rebal | Ground Truth

Network Architecture

Fully convolutional architecture, VGG-style



Quantitative Comparisons

Use 3 metrics of evaluation
(1) per-pixel accuracy (AuC CMF)
- commonly used metric for colorization
- does not evaluate plausibility, or joint interaction between pixels
(2) semantic interpretability (VGG)
(3) perceptual realism (AMT)

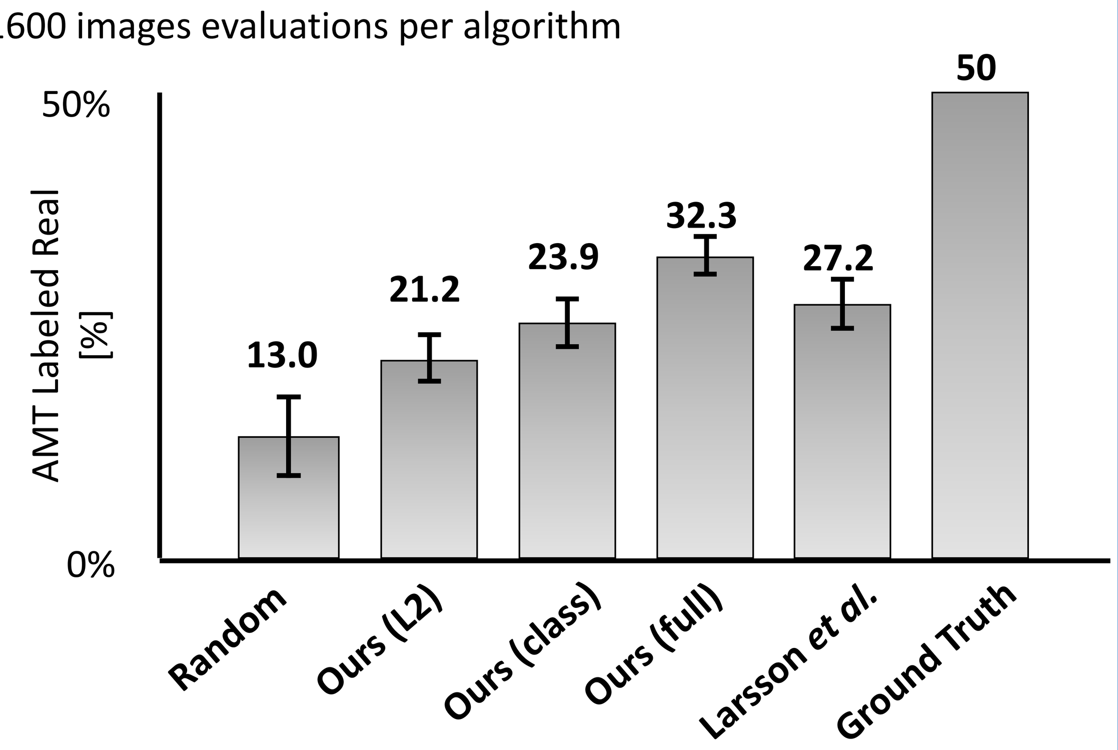
Colorization Results on ImageNet

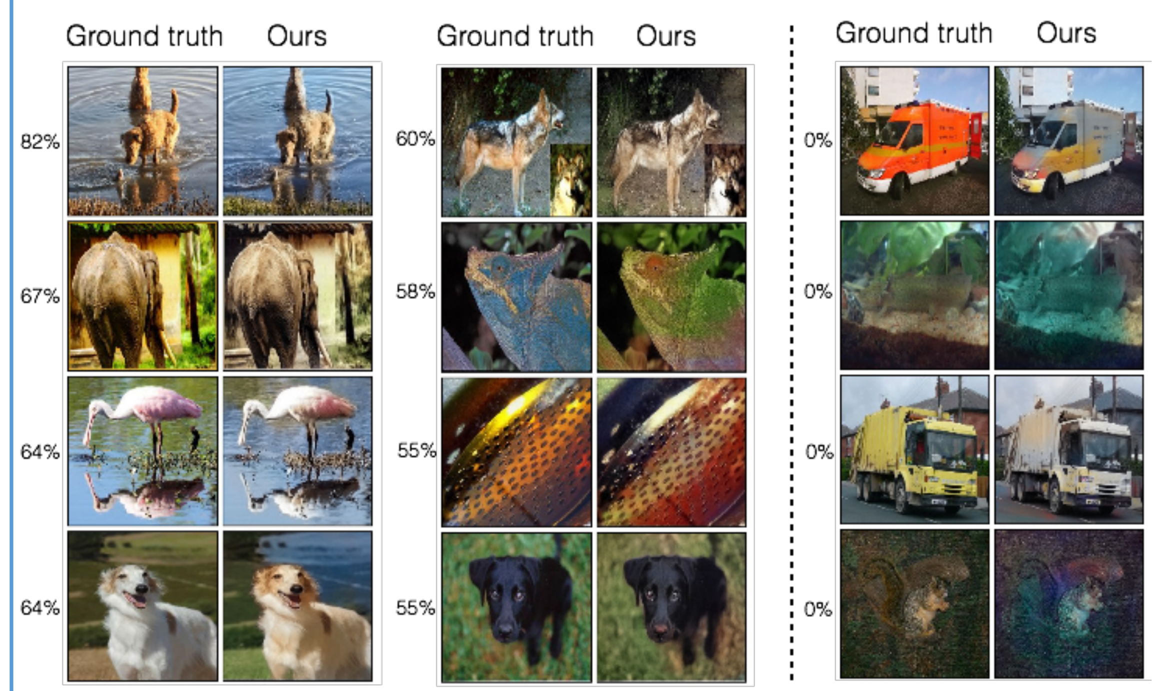
Method	Model Params (MB)	Model Feats (MB)	Model Runtime (ms)	AuC non-rebal (%)	AuC rebal (%)	VGG Top-1 Class Acc (%)	AMT Labeled Real (%)
Ground Truth	–	–	–	100	100	68.3	50
Gray	–	–	–	89.1	58.0	52.7	–
Random	–	–	–	84.2	57.3	41.0	13.0±4.4
Dahl [2]	–	–	–	90.4	58.9	48.7	18.3±2.8
Larsson et al. [23]	588	495	122.1	**91.7**	65.9	**59.4**	**27.2±2.7**
Ours (L2)	129	127	17.8	91.2	64.4	54.9	21.2±2.5
Ours (L2, ft)	129	127	17.8	91.5	66.2	56.5	23.9±2.8
Ours (class)	129	142	22.1	91.6	65.1	56.6	25.2±2.7
Ours (full)	129	142	22.1	89.5	**67.3**	56.0	**32.3±2.2**

Perceptual Realism Test (AMT Labeled Real)

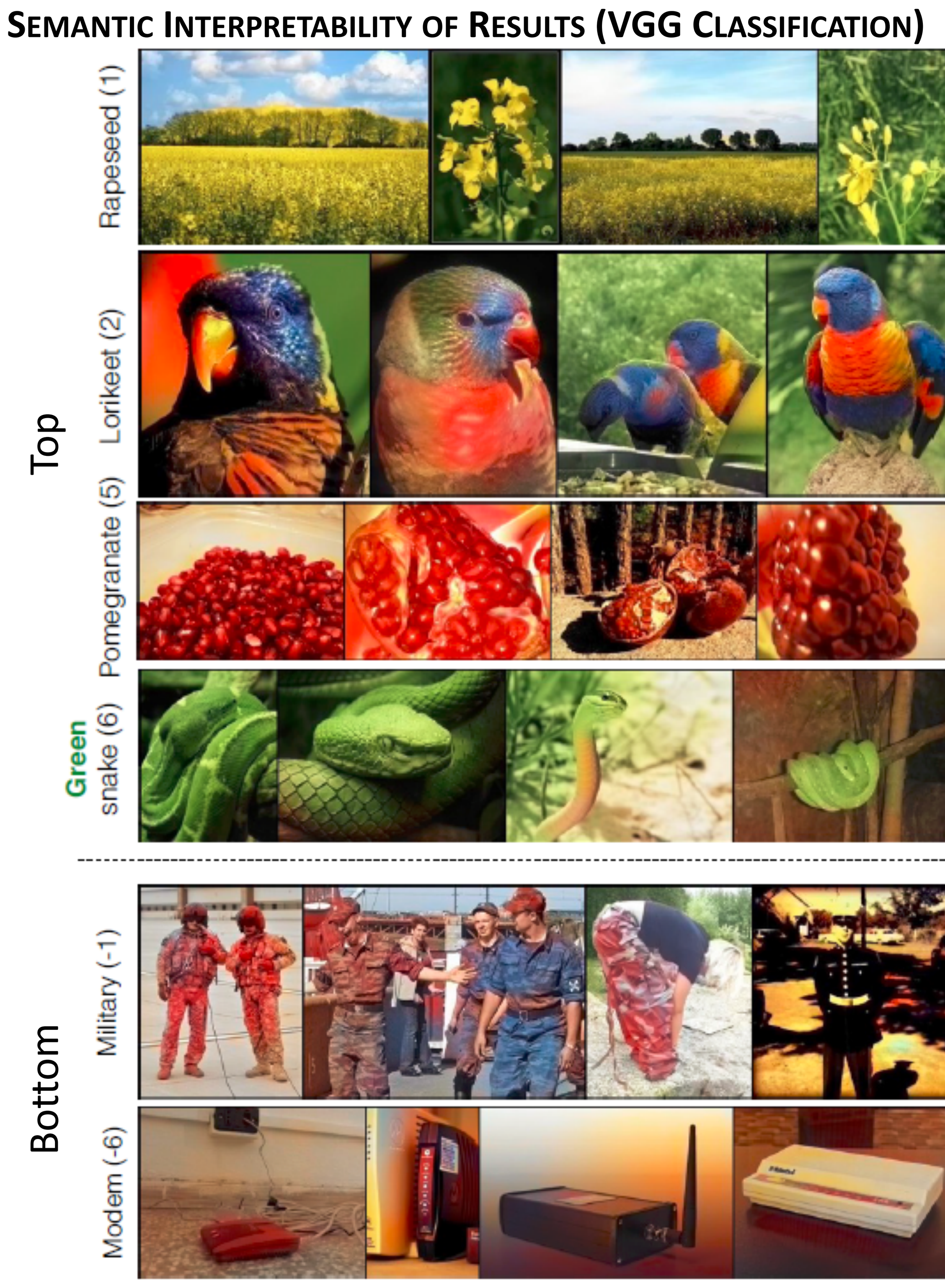
We introduce AMT as novel framework to evaluate ***visual plausibility*** of synthesized results

Test Procedure
- Participants asked to identify the generated vs ground truth image
- 1600 images evaluations per algorithm



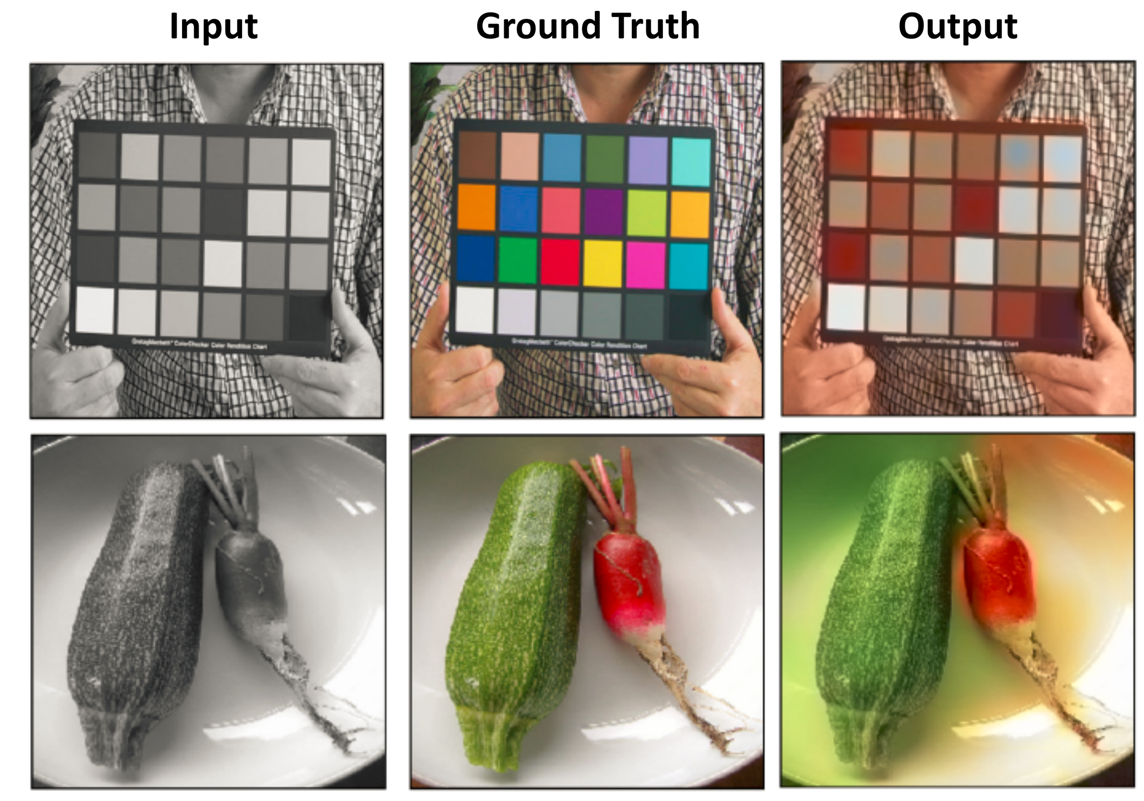


Semantic Interpretability of Results (VGG Classification)



Representation Learning via Cross-Channel Encoding

Input — **Ground Truth** — **Output**

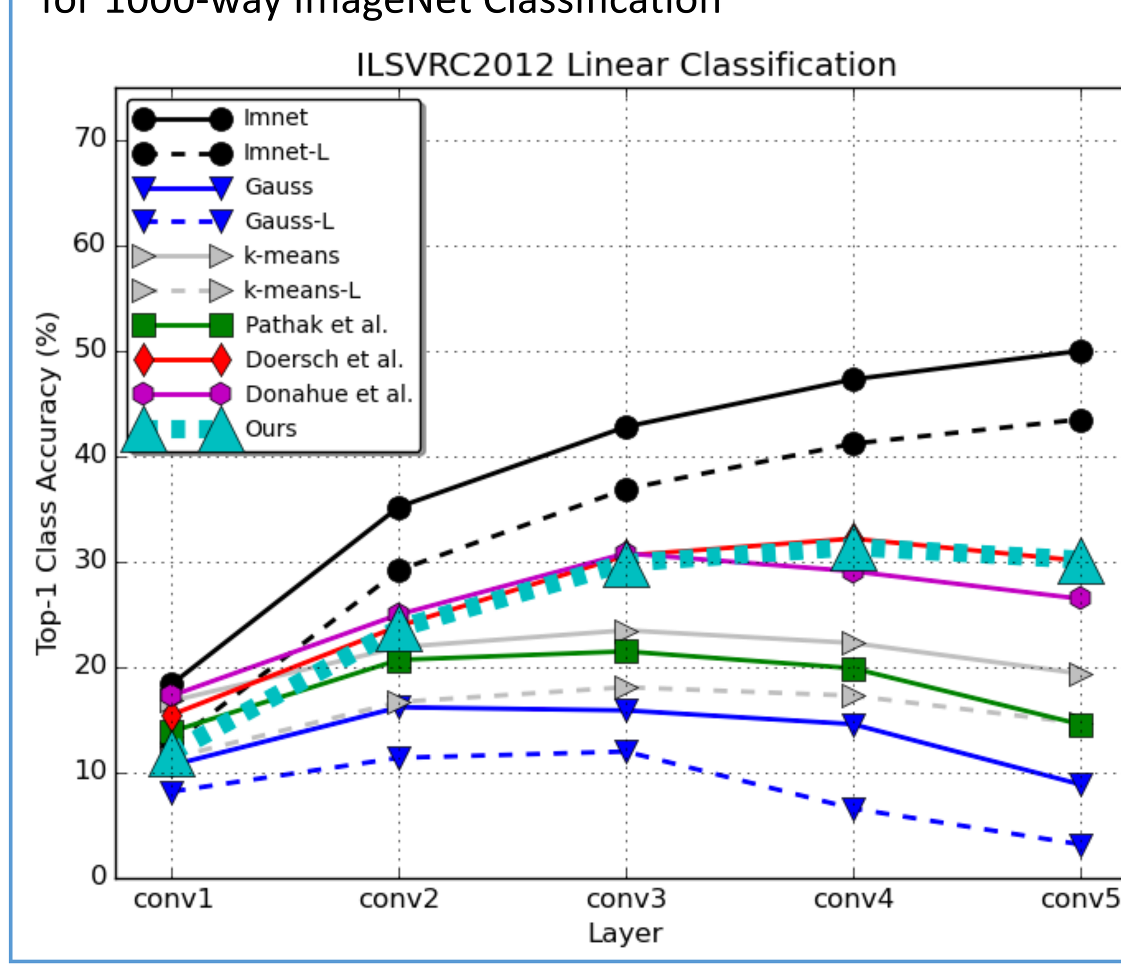


Supervised training: x_0 → CNN → Label y (with x_1)

Self-supervised training: x_0 → CNN → x_1

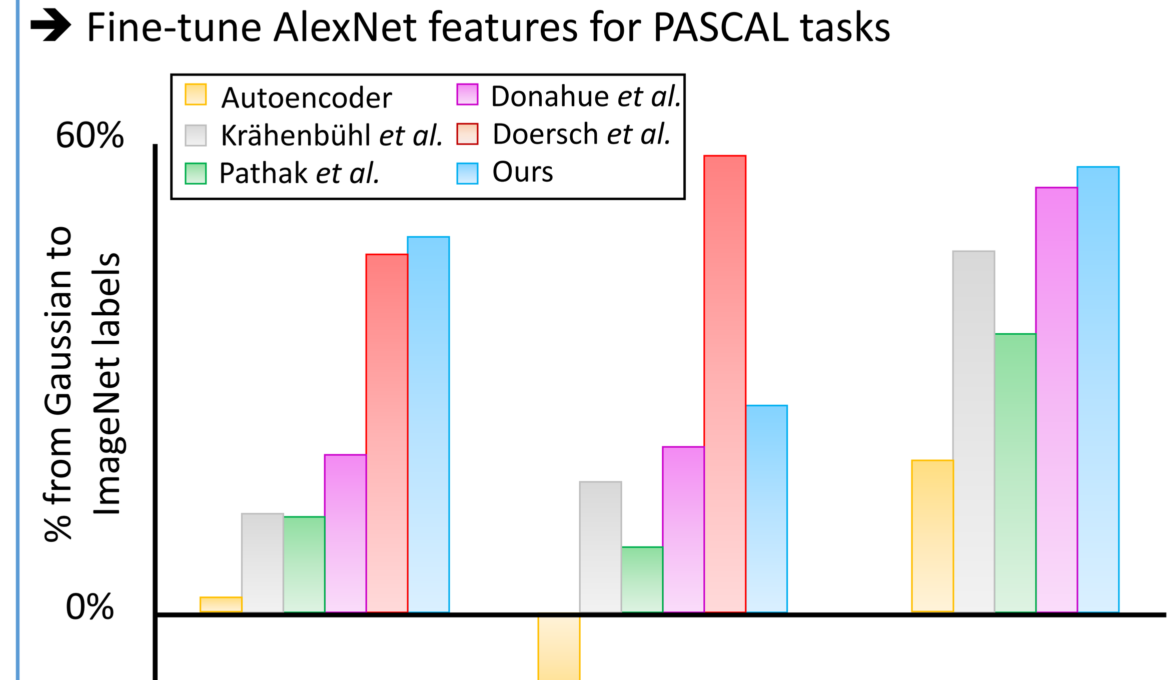
Task Generalization

How does colorization task generalize to semantics?
➔ Train linear classifiers on top of frozen AlexNet features for 1000-way ImageNet Classification



Dataset & Task Generalization

How does network generalize to *unseen data?*
➔ Fine-tune AlexNet features for PASCAL tasks

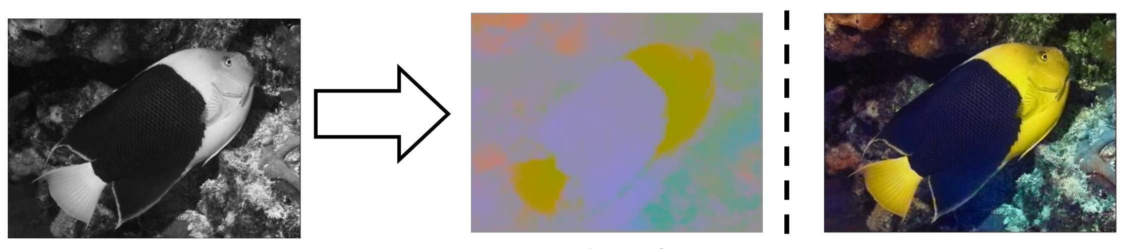


Colorful Image Colorization

Richard Zhang Phillip Isola Alexei A. Efros
Department of Electrical Engineering and Computer Sciences, UC Berkeley

Additional examples, Try our model! richzhang.github.io/colorization

Problem Statement
Given a grayscale image, predict the color



Input: Grayscale image *L* channel
Output: Color information *ab* channels
Concatenate (*L*,*ab*) for plausible colorization

Our contributions

1) Graphics Task of Colorization
a) set a new high-water mark on the task by training on 1M photos
b) design an appropriate objective function that handles the multimodal uncertainty and captures a wide diversity
c) introduce a novel framework for testing colorization algorithms, potentially applicable to other image synthesis tasks

2) Colorization as Representation Learning
a) introduce colorization task as instance of ***cross-channel encoding***
b) evaluate colorization for representation learning, demonstrate competitive performance in self-supervision framework

Inherent Ambiguity
Multiple plausible colorizations may exist
➔ L2 loss is inadequate for this problem

Ground Truth Grayscale Our Output

Our Loss Function
Grayscale Image to color distribution
- ***multinomial classification*** problem
- quantize *ab* space into grid size 10, keep 313 bins in gamut
- cross entropy loss

$$L(\widehat{\mathbf{Z}},\mathbf{Z}) = -\frac{1}{HW}\sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_q \mathbf{Z}_{h,w,q}\log(\widehat{\mathbf{Z}}_{h,w,q})$$

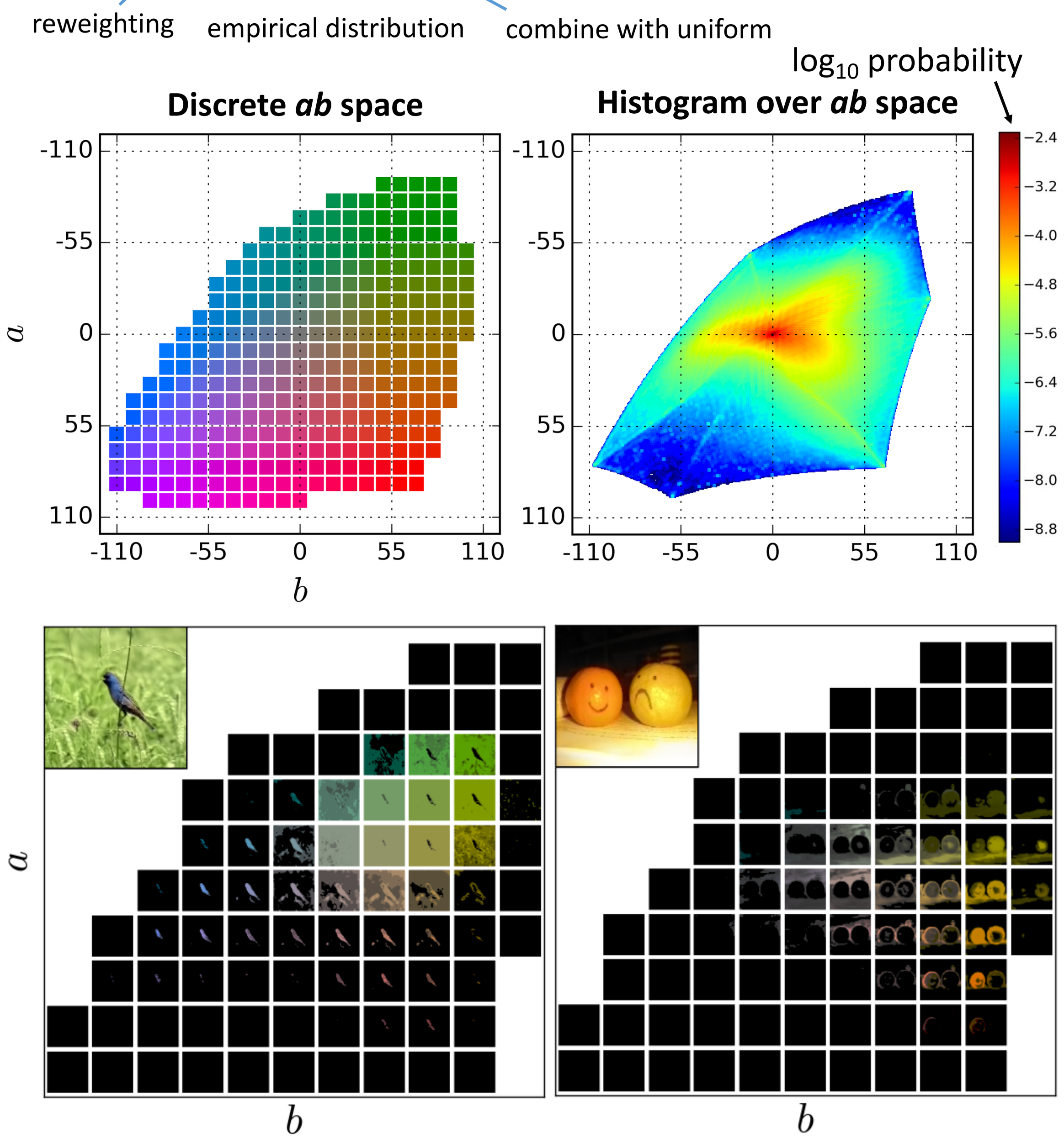
Rarity weighting Target distribution Predicted distribution

- ***Class rebalancing*** to encourage learning of *rare* colors

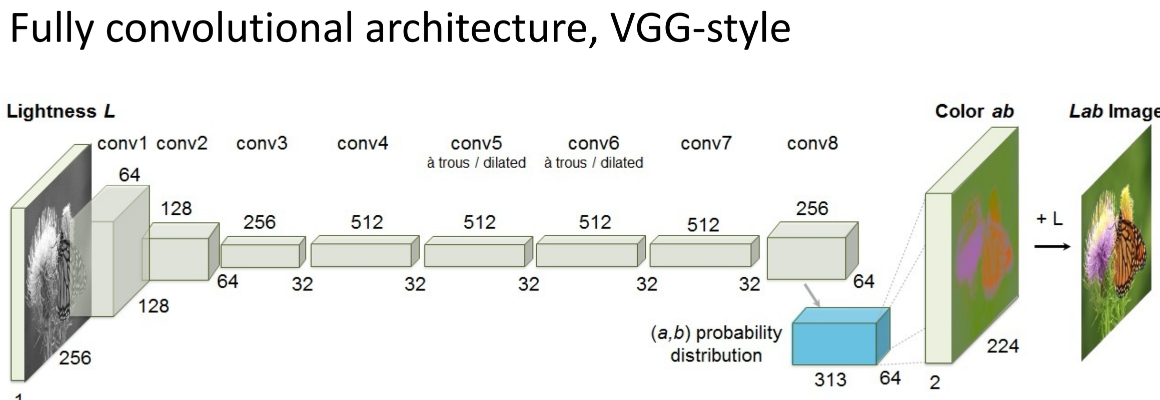
$$v(\mathbf{Z}_{h,w}) = \mathbf{w}_{q^*}, \text{ where } q^* = \arg\max_q \mathbf{Z}_{h,w,q}$$

$$\mathbf{w} \propto \left((1-\lambda)\widetilde{\mathbf{p}} + \frac{\lambda}{Q}\right)^{-1}, \quad \mathrm{E}[\mathbf{w}] = \sum_q \widetilde{\mathbf{p}}_q \mathbf{w}_q = 1$$

reweighting empirical distribution combine with uniform



Network Architecture
Fully convolutional architecture, VGG-style



Per-pixel Color Distribution to Single Point Estimate
- Mean is spatially coherent but desaturated
- Mode is vibrant but can have artifacts
- ***Interpolate*** between mean and mode with ***annealed-mean***

$$\mathcal{H}(\mathbf{Z}_{h,w}) = \mathrm{E}\big(f_T(\log \mathbf{Z}_{h,w})\big), \quad f_T(\mathbf{z}) = \frac{\exp(\mathbf{z}/T)}{\sum_q \exp(\mathbf{z}_q/T)}$$

expectation over annealed distribution annealed distribution

Mean (T=1) Annealed-Mean (T=.38) Mode (T→0)

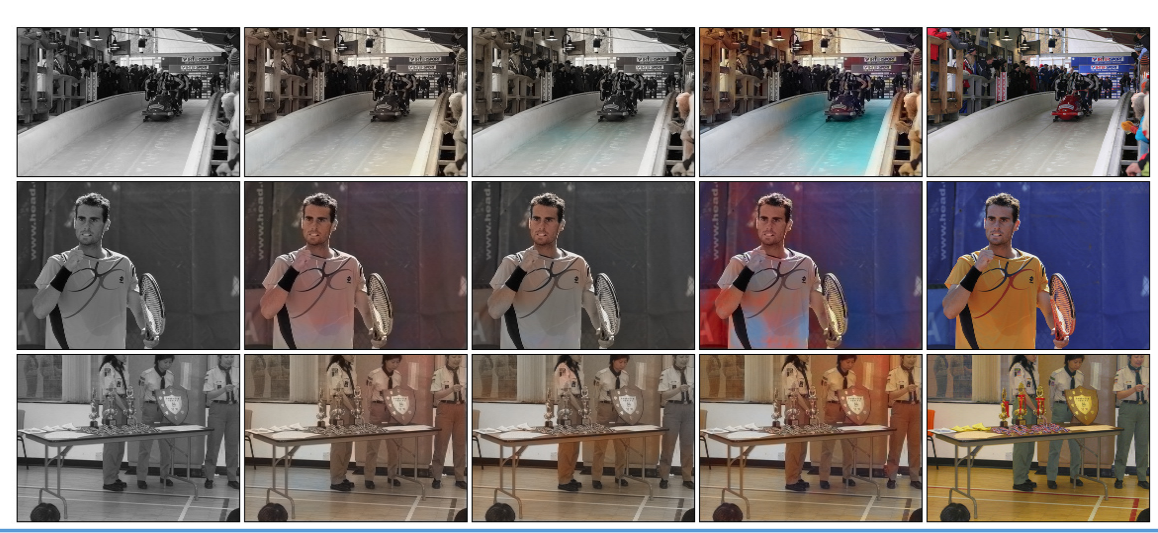
Selected ImageNet Results

Qualitative Comparisons
Success Cases

Grayscale L2 Regression Classification Classification w/ Rebal Ground Truth

Failure Cases

Grayscale L2 Regression Classification Classification w/ Rebal Ground Truth

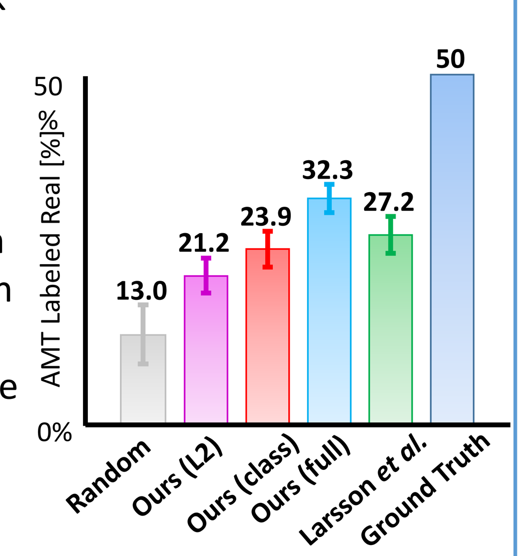


Quantitative Comparisons
Use 3 metrics of evaluation
(1) per-pixel accuracy (AuC CMF)
- commonly used metric for colorization
- does not evaluate plausibility, or joint interaction between pixels
- classification without rebalancing produces most "accurate" colors
(2) semantic interpretability (VGG)
(3) perceptual realism (AMT)

Colorization Results on ImageNet

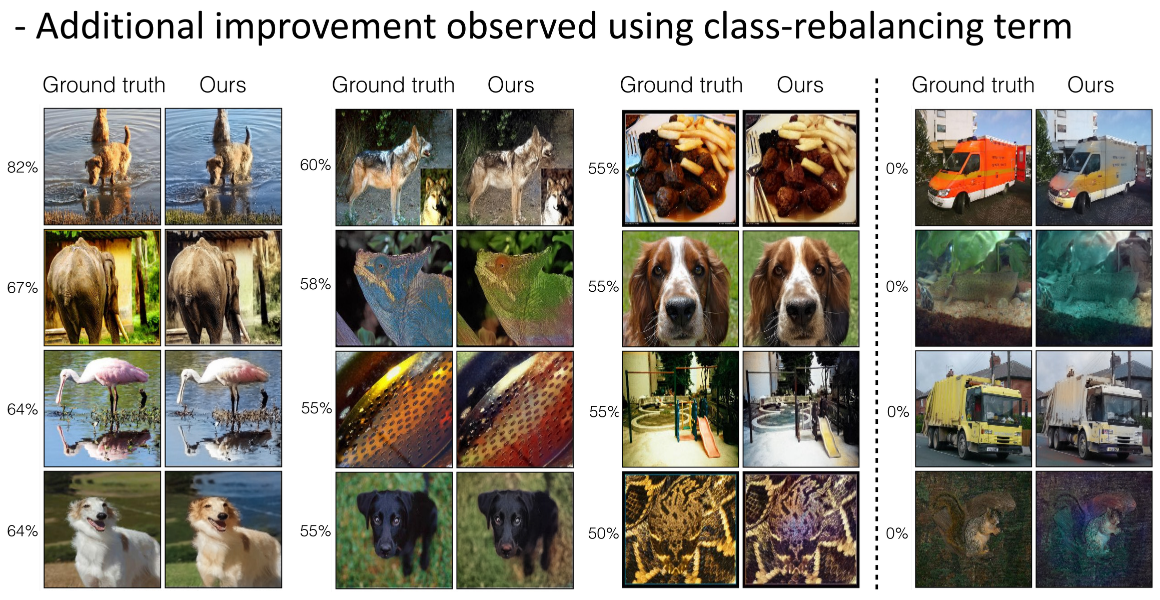
Method	Model Params (MB)	Model Feats (MB)	Runtime (ms)	AuC non-rebal (%)	AuC rebal (%)	VGG Top-1 Class Acc (%)	AMT Labeled Real (%)
Ground Truth	–	–	–	100	100	68.3	50
Gray	–	–	–	89.1	58.0	52.7	–
Random	–	–	–	84.2	57.3	41.0	13.0±4.4
Dahl [2]	–	–	–	90.4	58.9	48.7	18.3±2.8
Larsson et al. [23]	588	495	122.1	**91.7**	65.9	**59.4**	**27.2±2.7**
Ours (L2)	129	127	17.8	91.2	64.4	54.9	21.2±2.5
Ours (L2, ft)	129	127	17.8	91.5	66.2	56.5	23.9±2.8
Ours (class)	129	142	22.1	91.6	65.1	56.6	25.2±2.7
Ours (full)	129	142	22.1	89.5	**67.3**	56.0	**32.3±2.2**

Perceptual Realism Test (AMT Labeled Real)
We introduce AMT as novel framework to evaluate ***visual plausibility*** of synthesized results.



Test Procedure
- AMT participants shown ground truth and generated images, each for 1 sec in succession
- Participant asked to identify the image with "fake" colors
- 160 images evaluated for each algorithm, each evaluated ~10 times

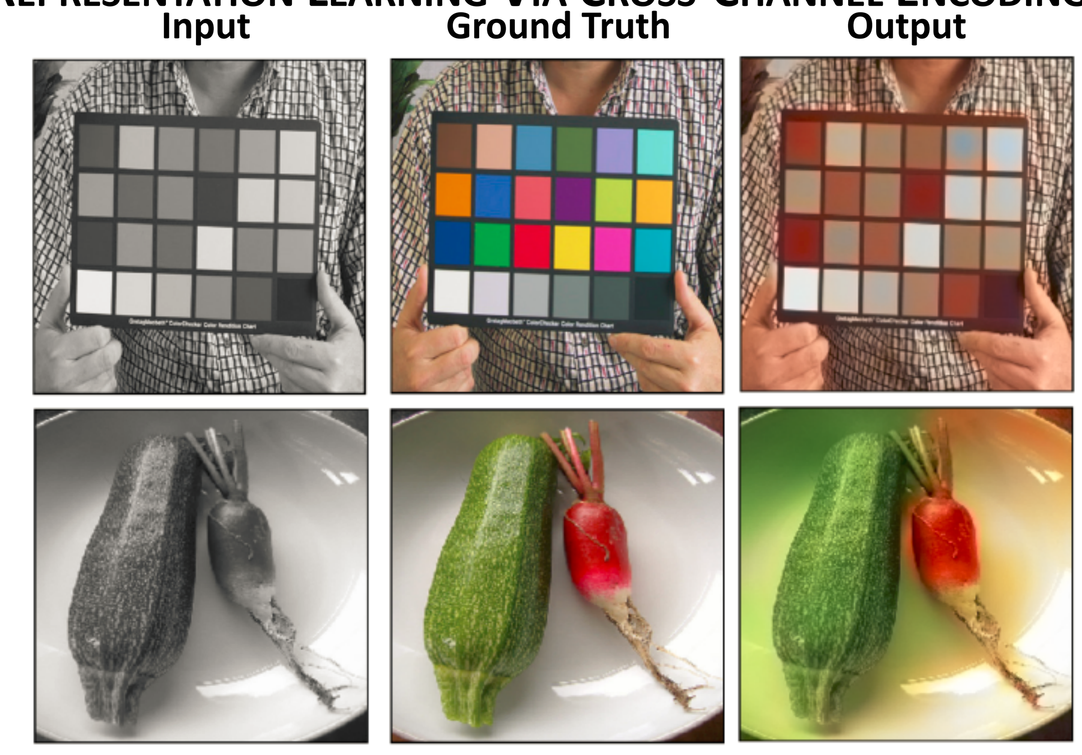
Conclusions
- Improvement in visual plausibility observed when using multinomial classification loss rather than L2 regression
- Additional improvement observed using class-rebalancing term



Semantic Interpretability of Results (VGG Classification)

Representation Learning via Cross-Channel Encoding

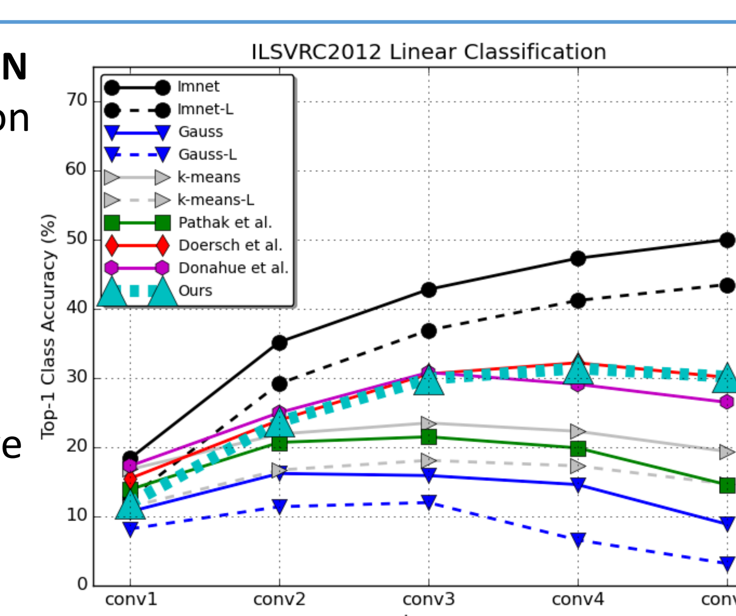
Input Ground Truth Output



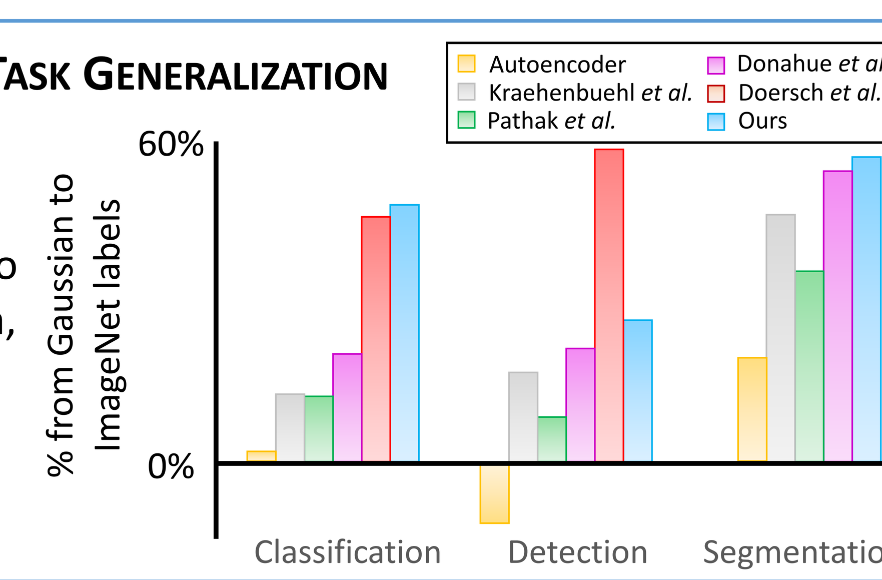
Supervised training: x_0, x_1 → CNN → Label y

Self-supervised training: x_0 → CNN → x_1

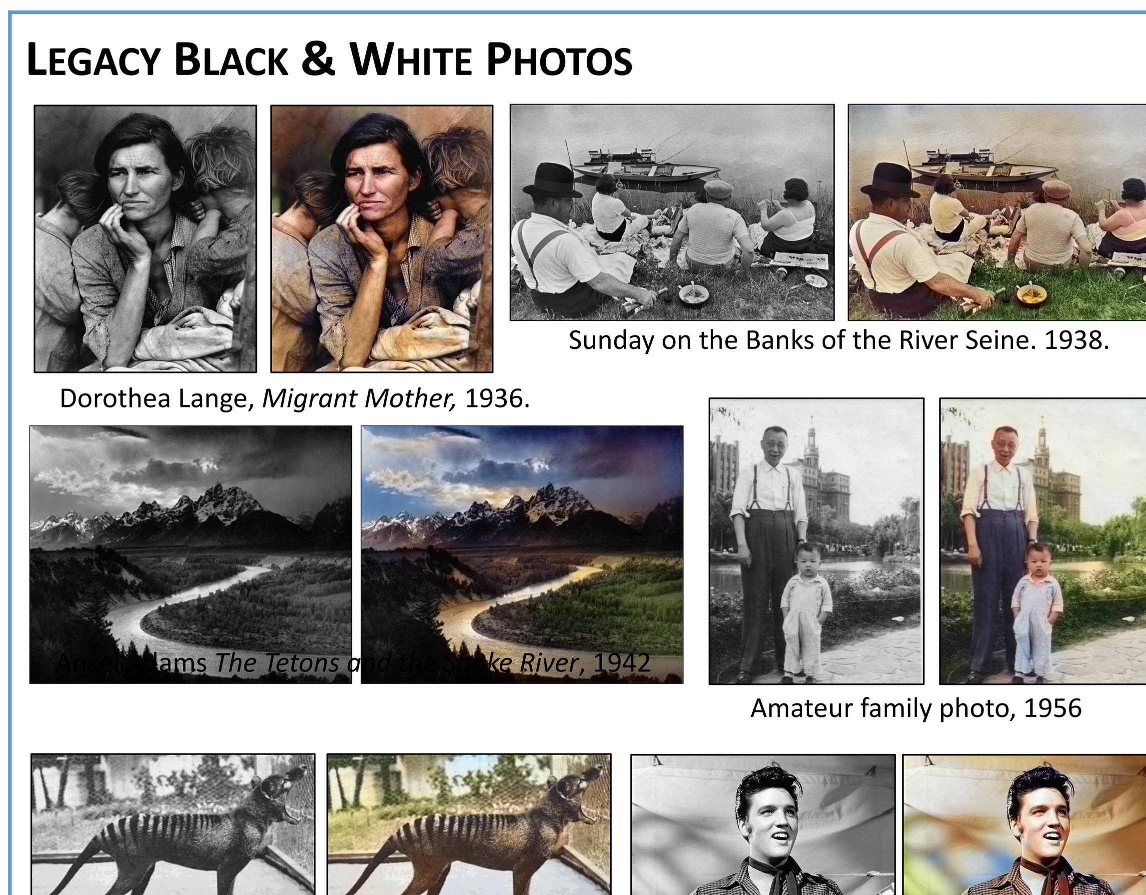
Task Generalization
How does colorization task generalize to semantics?
- conv1 low performance due to grayscale handicap
- conv2-5 competitive with self-supervised algorithms



Dataset & Task Generalization
How does network generalize to *unseen* data, along with unseen tasks?



Legacy Black & White Photos



Dorothea Lange, *Migrant Mother*, 1936.
Sunday on the Banks of the River Seine. 1938.
Amateur family photo, 1956
Dr. David Fleay, *Benjamin*, extinct in 1936
Photo of Elvis Presley