Colorful Image Colorization

Richard Zhang, Phillip Isola, Alexei A. Efros

Presenters: Aditya Sankar and Bindita Chaudhuri

Introduction

- ❖ Fully automatic approach (self-supervised deep learning algorithm)
- ❖ Aim: estimate the 2 unknown color dimensions from the known color dimension
- ❖ Under-constrained problem; goal is **not to match ground truth** but produce vibrant and **plausible colorization**

- ❖ “Colorization Turing test” to evaluate the algorithm

Related Work

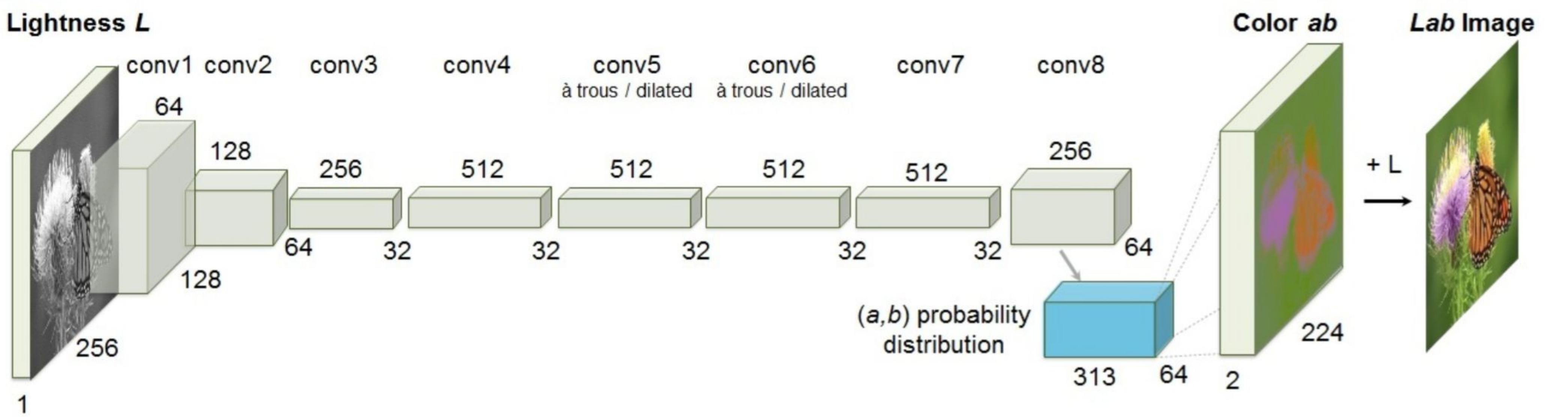
- **Non-parametric methods:**
  - Use one or more color reference images provided by user based on input grayscale image
  - Transfer color to input image from analogous regions of reference image(s)
- **Parametric methods:**
  - Learn mapping functions for color prediction
  - Generally on smaller datasets and using smaller models
- **Concurrent methods:**
  - Iizuka et. al.[1] – Two-stream architecture; regression loss; different database
  - Larsson et. al.[2] – Un-rebalanced classification loss; use of hypercolumns

[1] Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. ACM Transactions on Graphics (Proc. of SIGGRAPH 2016) 35(4) (2016)

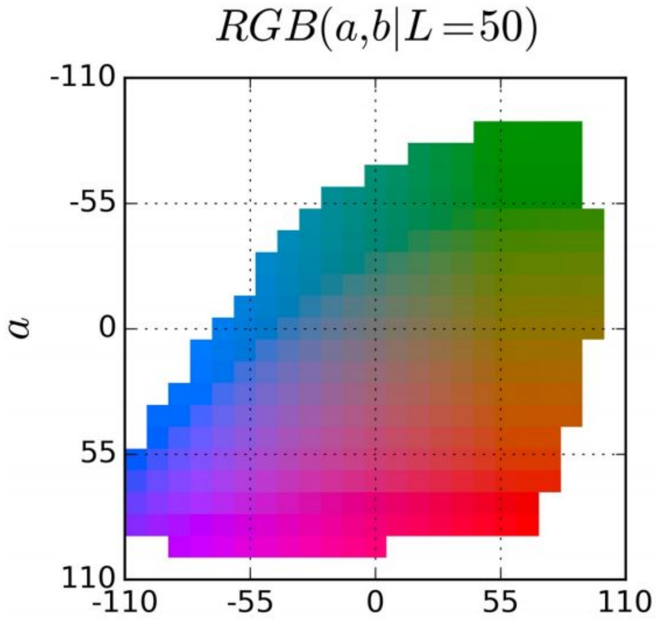
[2] Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. European Conference on Computer Vision (2016)

Network architecture

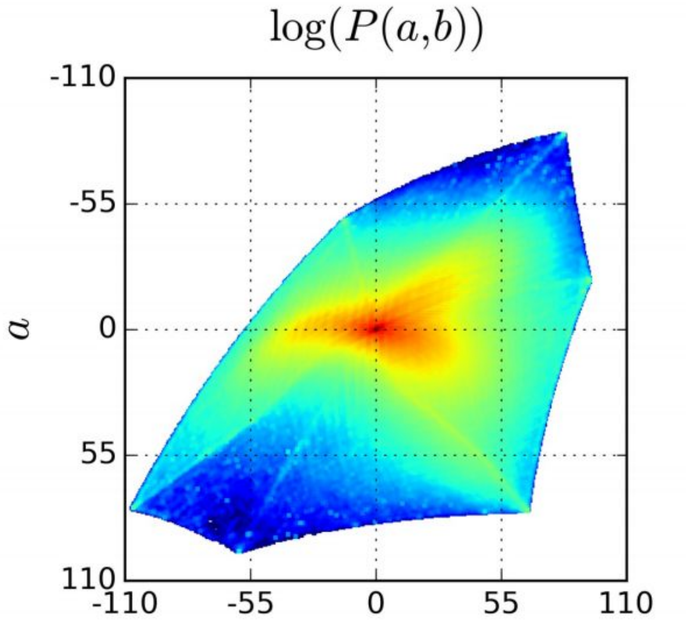
- CIE *Lab* color space used for perceptual similarity to human vision
- Input: $\mathbf{X} \in \mathbb{R}^{H\times W\times 1}$; H,W – image dimensions
- Intermediate result: $\widehat{\mathbf{Z}} = \mathcal{G}(\mathbf{X}) \in [0,1]^{H\times W\times Q}$; Q = 313 quantized *ab* values
- Output $\widehat{\mathbf{Y}} = \mathcal{H}(\widehat{\mathbf{Z}}) \in \mathbb{R}^{H\times W\times 2}$



ab – Space and Need for Rebalancing



Quantized *ab* Color Space



Empirical probability distribution within *ab* space

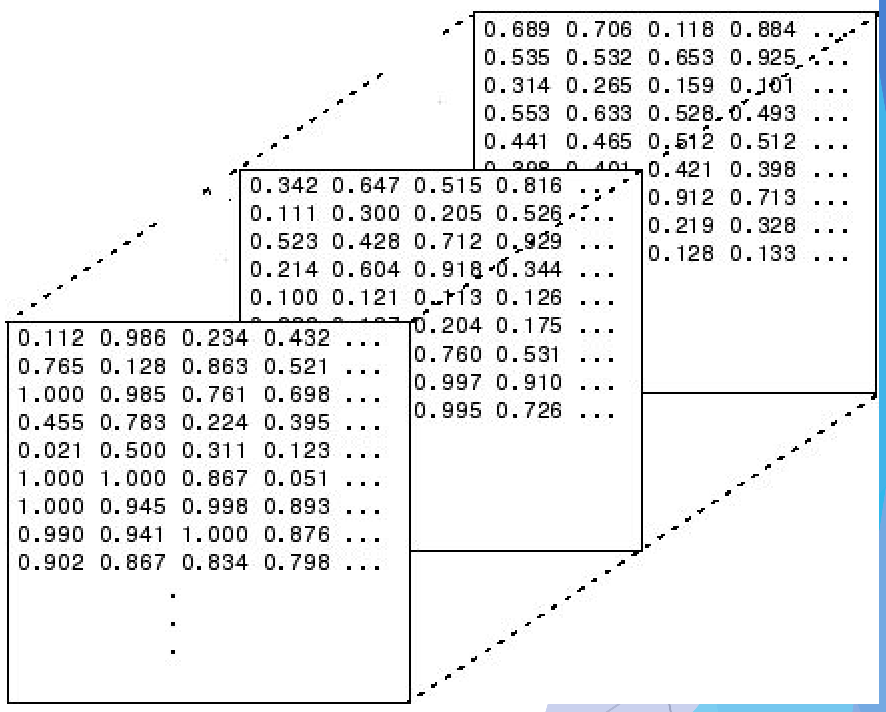


Illustration of $\mathbf{Z}$

Methodology

- CNN maps $\mathbf{X}$ to $\widehat{\mathbf{Z}}$
- Ground truth $\mathbf{Y}$ is mapped to $\mathbf{Z}$ using a soft-encoding scheme
- CNN is trained to minimize the following multinomial cross-entropy loss:

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

- Weights v are added to take care of class imbalance

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

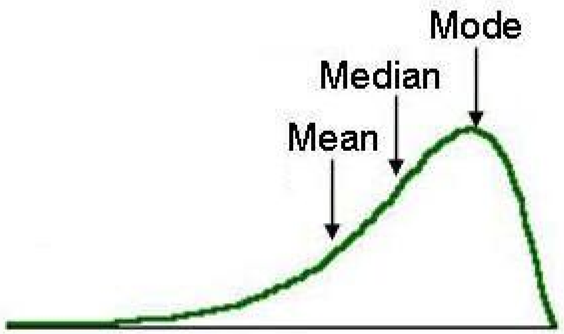
$$\mathbf{w} \propto \left((1-\lambda)\widetilde{\mathbf{p}} + \frac{\lambda}{Q}\right)^{-1}, \quad \mathbb{E}[\mathbf{w}] = \sum_{q} \widetilde{\mathbf{p}}_q \mathbf{w}_q = 1$$

Methodology (Cont.)

$\widehat{\mathbf{Z}}$ finally mapped to $\mathbf{Y}$ using the **annealed mean** of the color distribution.

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

Mean of distribution produce spatially consistent but desaturated results
Mode of distribution produce vibrant but spatially inconsistent results

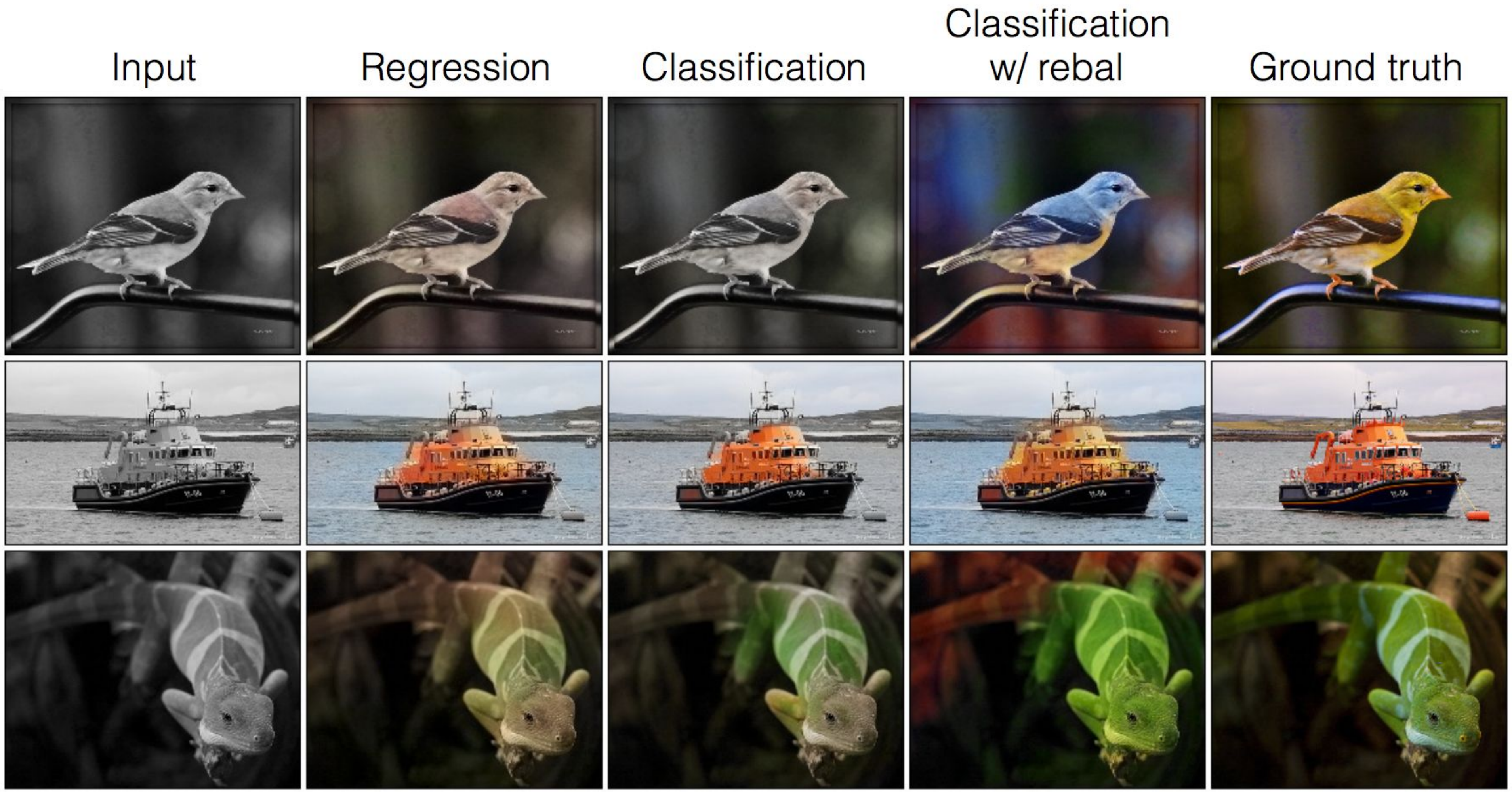


Experimental Details

- Data used:
  - 1.3 million training images from ImageNet training set
  - First 10K images for validation from ImageNet validation set
  - A separate set of 10k images for testing from ImageNet validation set

- CNN trained on various loss functions
  - Regression (L2-loss)
  - Classification, without rebalancing
  - Classification, with rebalancing (Full method)
  - Larsson, Dahl methods
  - Random colors and gray scale images

Qualitative Results



Qualitative Results (contd..)

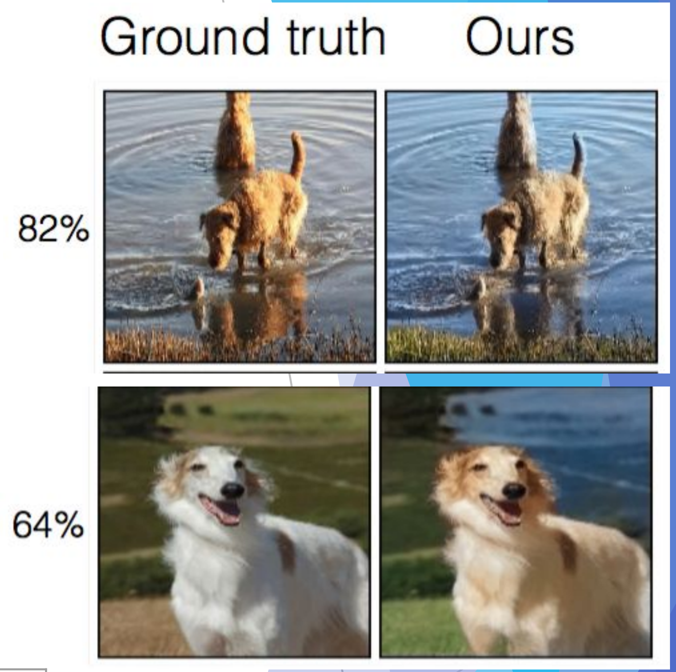
Failure Cases

Legacy Images

Results with legacy black and white photos

Quantitative Results

- Measure of 'Perceptual Realism' via Amazon Mechanical Turk
  - Real v/s Fake two-alternate choice experiment
  - 256x256 image pairs shown for 1 second
  - Turkers select the 'real' image for 40 pairs
  - Ground Truth v/s Ground Truth will have expected result of 50%
  - Random baseline produced 13% error (seems high)



"Better than Ground Truth results"

	Ground Truth	Random	Dahl [2]	Larrson [23]	Ours [L2]	Ours [L2, ft]	Ours (Class)	Ours (Full)
Labeled Real	50	13.0 ± 4.4	18.3 ± 2.8	**27.2 ± 2.7**	21.2 ± 2.5	23.9 ± 2.8	25.2 ± 2.7	**32.3 ± 2.2**

Other Observations

- **Semantic Interpretability**:
  - How does the colorization effect object detection?
  - VGG Object detection on ground truth images: 68.30%
  - VGG Object detection on desaturated images: 52.70%
  - VGG Object detection on (their) re-colorized images: 56.00%
  - VGG Object detection on Larsson re-colorized images: 59.40%

- **Raw Accuracy:**
  - L2-distance from ground truth ab values
  - Predicting grey values actually performs quite well for L2 and Larsson outperforms them in this metric
  - They rebalance color weights by frequency of occurrence and in this rebalanced metric outperform Larsson and Grey scale.

Conclusion and Discussion

- Deep learning and a well-chosen objective function produce results similar to real color photos.

- Network learns a representation; can be extended to object detection, classification and segmentation

- Visual results are great. Quantitative metrics and other observations are just OK..

- Need to consider global consistency and contextual information for complex scene colorizations

THANK YOU