Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification

Satoshi Iizuka, Edgar Simo-Serra, Hiroshi Ishikawa

Alper EMLEK

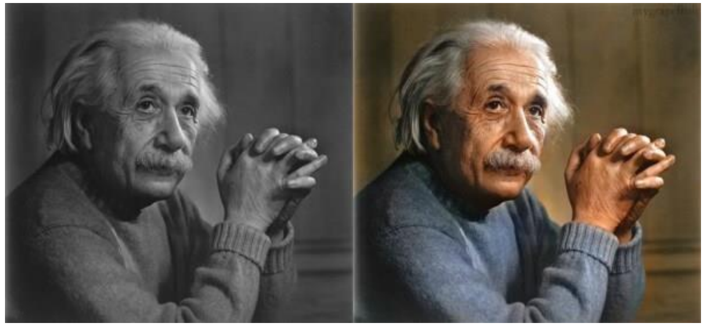
Fırat Coşkun DALGIÇ

Contents

Introduction

Image colorization assigns a color to each pixel of a target grayscale image

- Usually used for coloring of historical black and white photographs


- Q. What is any other usage area of image colorization ?

Introduction

- Traditional colorization techniques requires significant user interaction.
- In this paper, a fully automated data-driven approach proposed for colorization.
- This method requires neither pre-processing nor post-processing.
- This model consists of four main components:
  - A low-level features network
  - A mid-level features network
  - A global features network
  - A colorization network

Introduction

- A single network.

- This approach uses a combination of global image priors and local image features to colorize an image automatically.
  - Global priors
  - Local features

- It can also perform classification of the scene.

- This model to be run on input images of arbitrary resolutions, unlike most Convolutional Neural Networks.

Introduction

In summary, in this paper main contribution:

- A user-intervention-free approach to colorize grayscale images.
- A novel end-to-end network that jointly learns global and local features for an image.
- A learning approach that exploits classification labels to increase performance.
- A style transfer technique based on exploiting the global features.

Related works

- Colorization methods can be roughly divided into two categories.
  - Scribble-based colorization
  - Example-based colorization
  - Automatic colorization

Related works

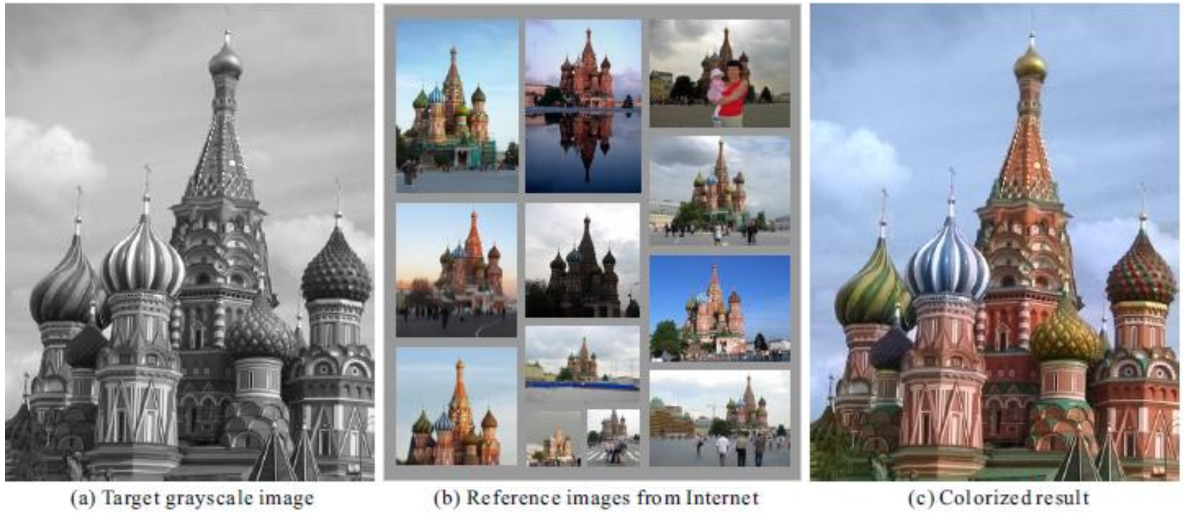
Levin+ 2004

- Scribbles-based
  - Levin et al. 2004
    - Simple colorization method that requires neither image segmentation, nor region tracking.
    - Based on a simple premise: neighboring pixels have similar intensities should have similar colors.
    - Formalize this premise using a quadratic cost function and obtain an optimization problem that can be solved efficiently using standard techniques.
  - Hunang et al. 2005
    - Imrove Levin's cost function for more sensetive to edge information, prevent the color bleeding over object boundaries

Related works

- Reference Image-based
  - Exploit the colors of a reference image .
  - Inspired by the color transfer techniques that are widely used for recoloring a color image.
  - Welsh et al. [2002]
    - Proposed a general technique to colorize grayscale images by matching the luminance and texture information between images.
    - Aim minimize the amount of human labor required for this task.
    - Further, the procedure is enhanced by allowing the user to match areas of the two images with rectangular swatches.
  - Gupta et al. [2012]
    - Matching superpixels between the input image and the reference image using feature matching
    - Space voting to perform the colorization

Related works



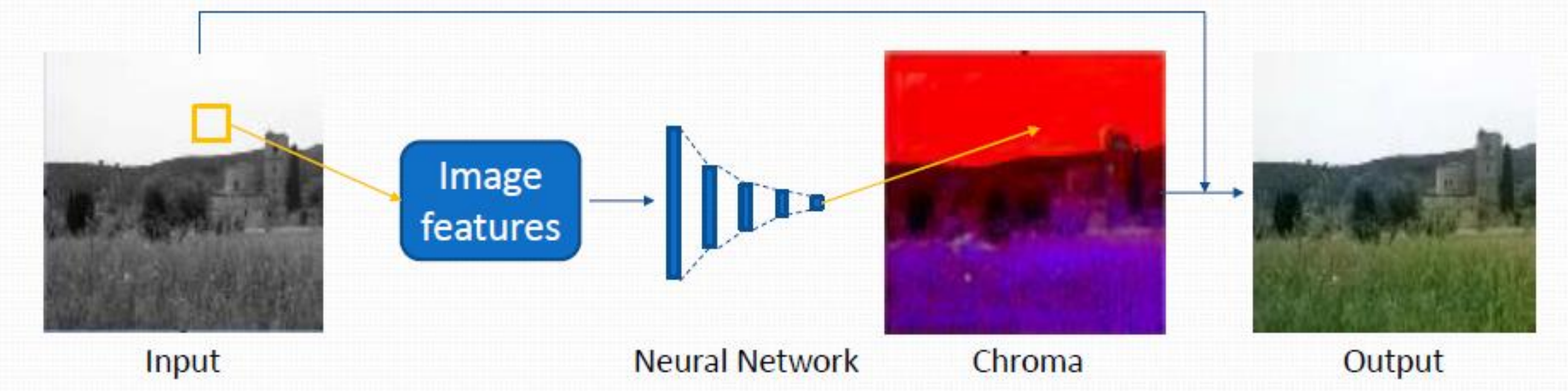
- Reference image-based
  - Liu et al. 2008
    - Reference images that are obtained directly from web search.
    - Its applicability is, however limited to famous landmarks where exact matches can be found.
  - Chia et al. 2011
    - Requires user to provide a semantic text label and segmentation cues for the foreground object.

Related works

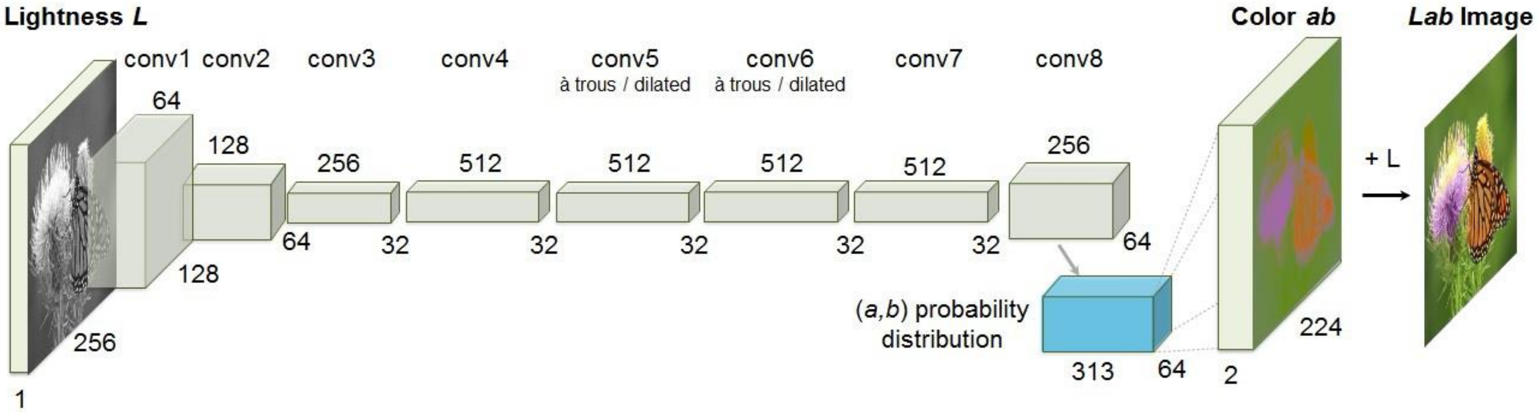
- Automatic colorization

  Aim to remove user interactiion.

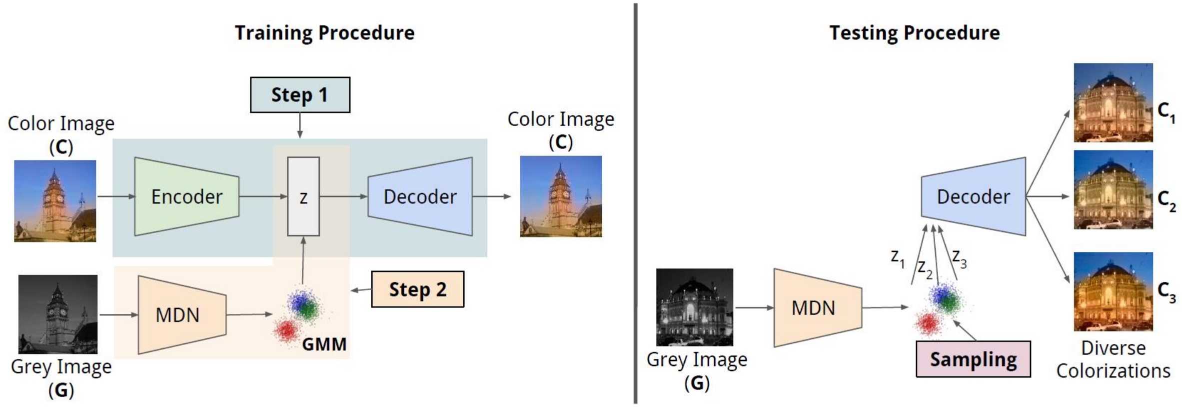
  - Cheng et al. 2015
    - Group these images into different clusters adaptively
    - Uses existing multiple image feature
    - Computes chrominance via shallow neural network
    - Depend on the performance of sematic segmentation
    - Only handles simple outdoor scenes



Related works

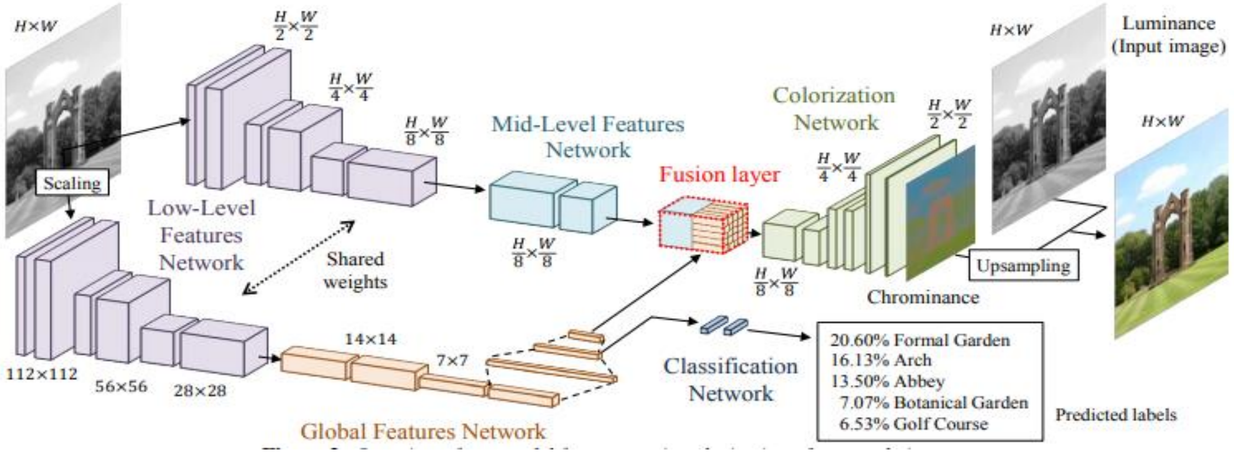


- Automatic colorization
  - Zhang et al. 2016
    - Given the lightness channel L, our system predicts the corresponding a and b color channels of the image in the CIE Lab colorspace.
    - Color prediction is inherently multimodal-many objects can take on several plausible colorizations.
    - To appropriately model the multimodal nature of the problem, we predict a distribution of possible colors for each pixel.
  - Deshpande et al. 2017
    - Previous methods only produce the single most probable colorization. Their goal is to model the diversity intrinsic to the problem of colorization and produce multiple colorizations.

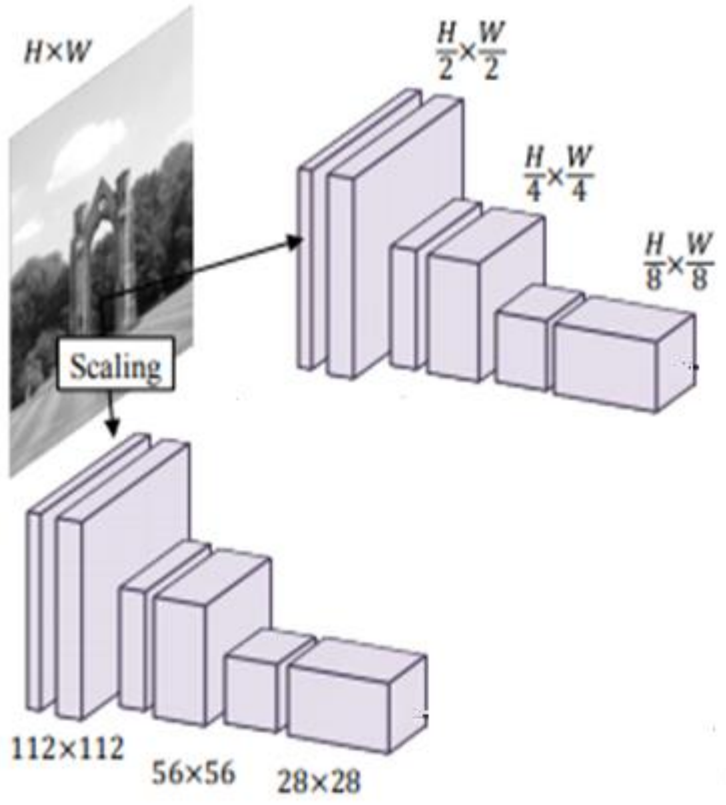


Analyzing Network Model

- In this section, first we will quickly overview the network model according to the subsection stated in article which are,
  - Low Level Features Network
  - High Level Features Network
  - Mid Level Features Network
  - Fusing Layer
  - Colorization Network
- Afterwards, we will examine the model by asking some questions. These questions will be stated later.

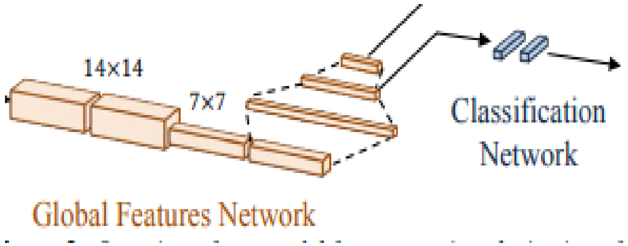


Low Level Features Network



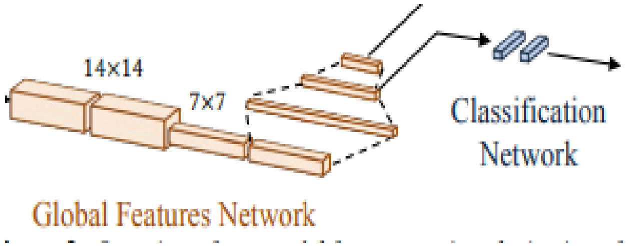
- Network properties are:
  - 6 layer CNN structure
  - Dimension reduction with **increasing stride,** NOT by using pooling!

Global Features Network



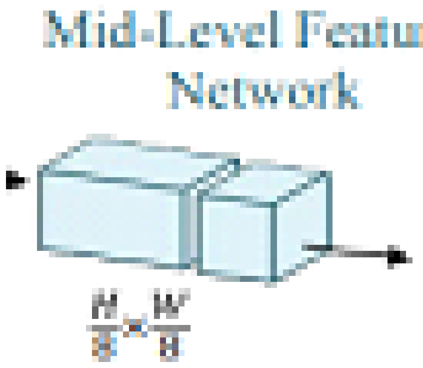
- Smaller network inside main network model. But WHY? What is the advantage of this smaller network?

Global Features Network



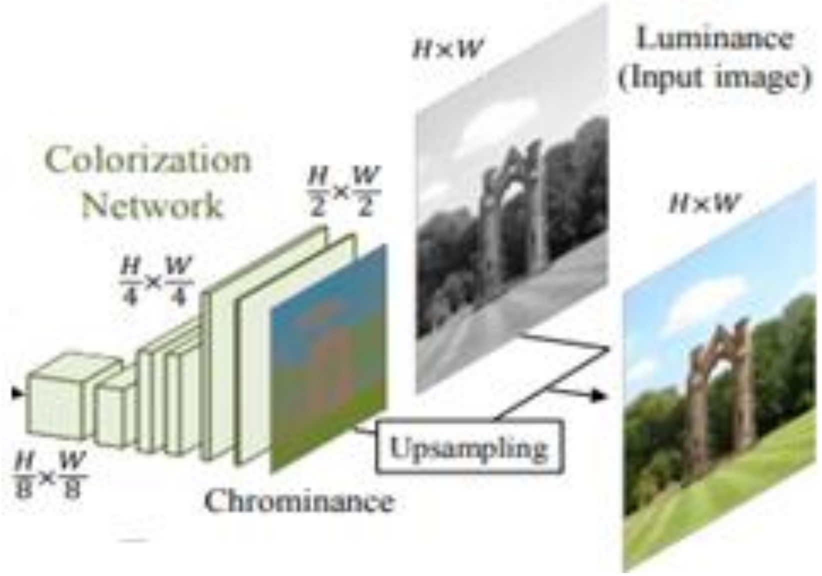
- Smaller network inside main network model. But WHY? What is the advantage of this smaller network?
  - Better understanding the context and scenery.
- How it worked?
  - Simply pretrained over for 205 different classes and specialized on training.

Mid Level Features Network



- ➢It is fully convolutional network which has 2 layer.
- ➢No dimension reduction

Colorization Network

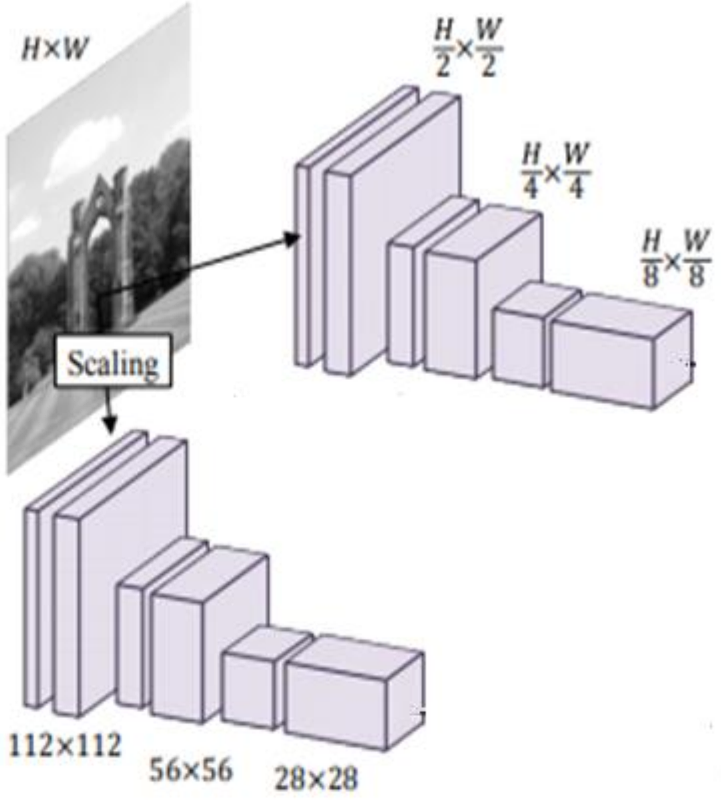


- ➢It is a deconvolution structure.
- ➢Upsamples till network width and height will be the same input size.
- ➢Combines deconvolution result with input intensities in order to construct colorfull image.

Question to Understanding Network Structure

- How they achived the process any image resolution?
- How they construct color image?
- How they reflect the content information in backpropogation?
- What activation function they used and why?
- What loss function they prefered?

How they achived the process any image resolution?

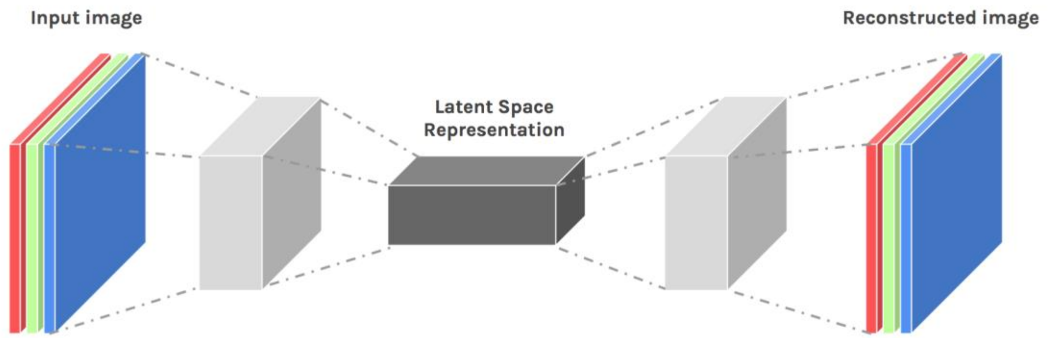


- Achieved by applying **scaling** on front of **Global Features Network**.
- However, this yields both **performance** and **accuracy** loss when we increase the input image size!

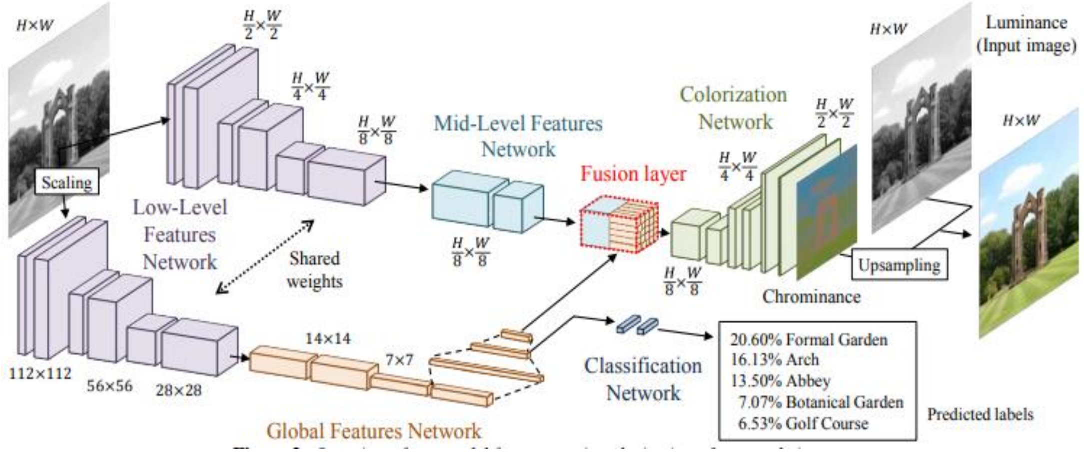
Question to Understanding Network Structure

- How they achived the process any image resolution?
- How they construct color image?
- How they reflect the content information in backpropogation?
- What activation function they used and why?
- What loss function they prefered?

How they construct color image?



https://hackernoon.com/autoencoders-deep-learning-bits-1-11731e200694

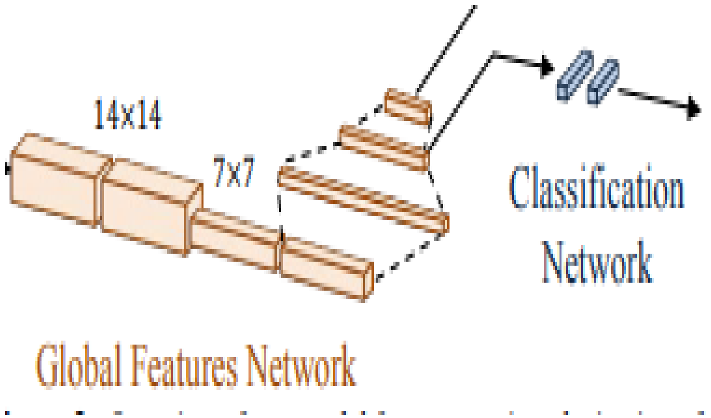


- By using **Autoencoder** strategy.
- **Fusing** global features at bottleneck.

Question to Understanding Network Structure

- How they achived the process any image resolution?
- How they construct color image?
- How they reflect the content information in backpropogation?
- What activation function they used and why?
- What loss function they prefered?

How they reflect the content information in back propagation?



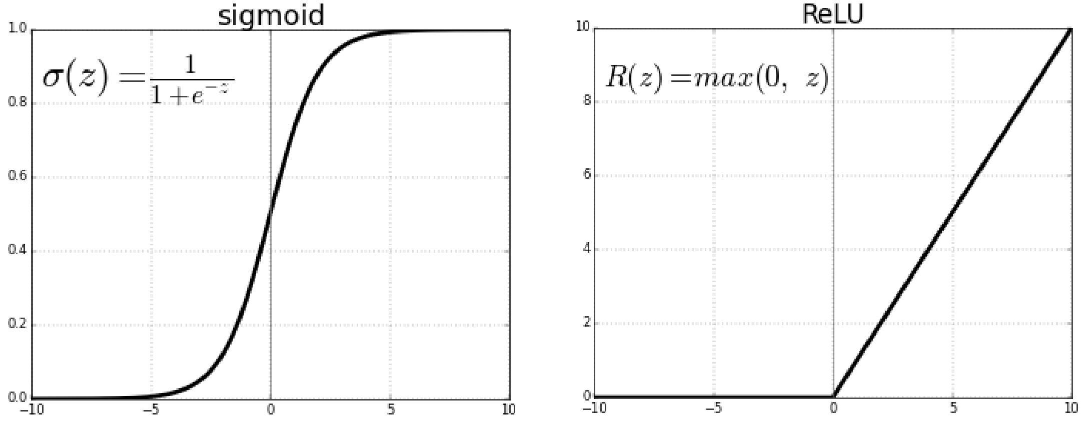
- By using **Classification Network loss** at back propogation.
- When they DID NOT use the classification loss, they realized that they still loose **content** information on **Global Features Network**.

Question to Understanding Network Structure

- How they achived the process any image resolution?
- How they construct color image?
- How they reflect the content information in backpropogation?
- What activation function they used and why?
- What loss function they prefered?

What activation function they used and why?

- They have tested network model with both ReLU and Sigmoid activation functions.
- After their experiments, they preferred to use Sigmoid function because:
    - ➢ Architecture is not so deep to cause harmful vanishing gradient problem.
    - ➢ In early stages, ReLU caused information loss especially at Global Features Network, therefore the Fusion layer became uneffective.

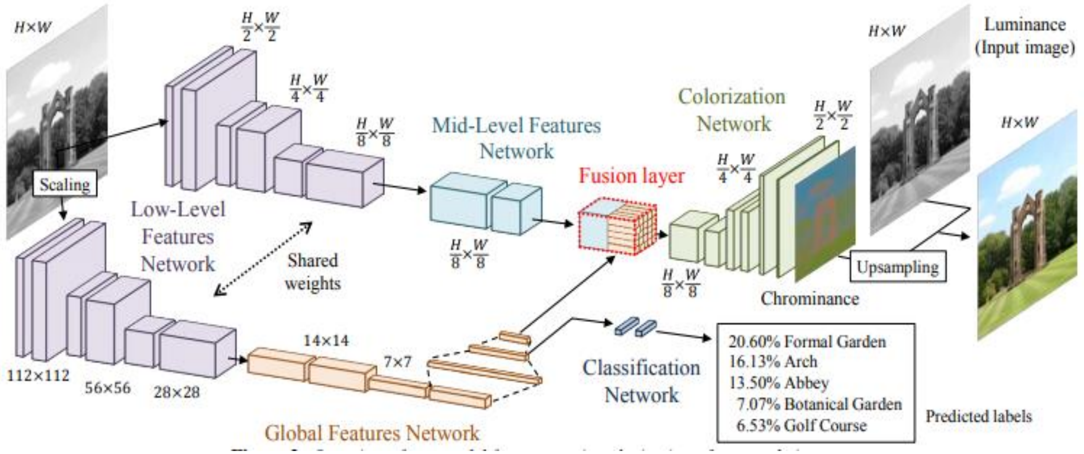


https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

Question to Understanding Network Structure

- How they achived the process any image resolution?
- How they construct color image?
- How they reflect the content information in backpropogation?
- What activation function they used and why?
- What loss function they prefered?

What loss function they prefered?

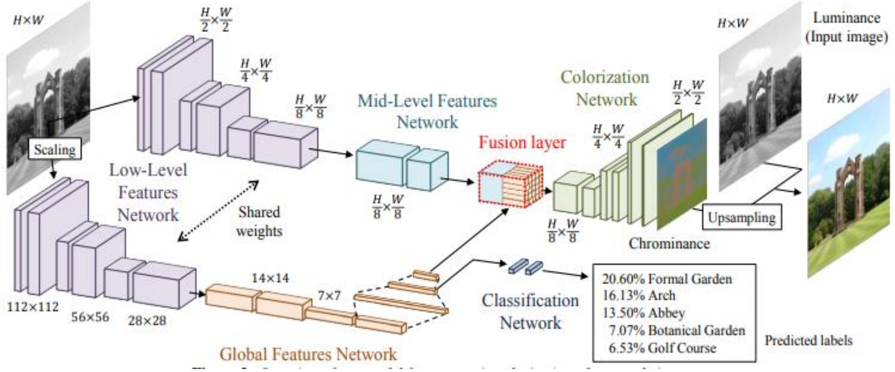


The global loss of network:

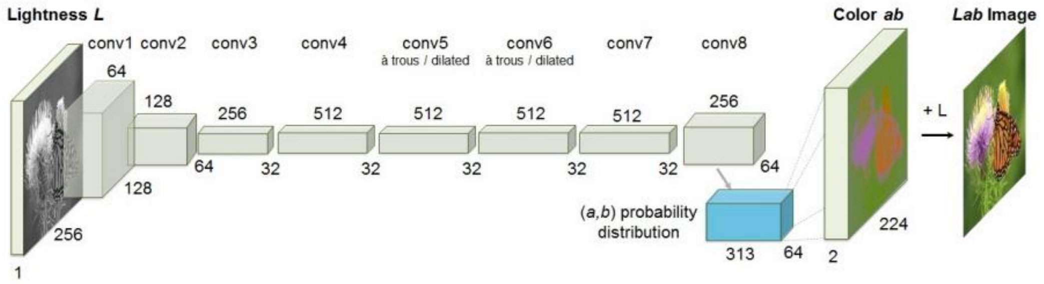
$$L(y^{\text{color}}, y^{\text{class}}) = \|y^{\text{color}} - y^{\text{color},*}\|^2_{\text{FRO}} - \alpha \left(y^{\text{class}}_{l^{\text{class}}} - \log \left(\sum_{i=0}^{N} \exp \left(y_i^{\text{class}} \right) \right) \right)$$

- The network has two main loss, classification loss and colorization loss.
- Colorization loss is the MSE between input and resultant image intensities.
- Classification loss is cross-entropy loss of classification network result.

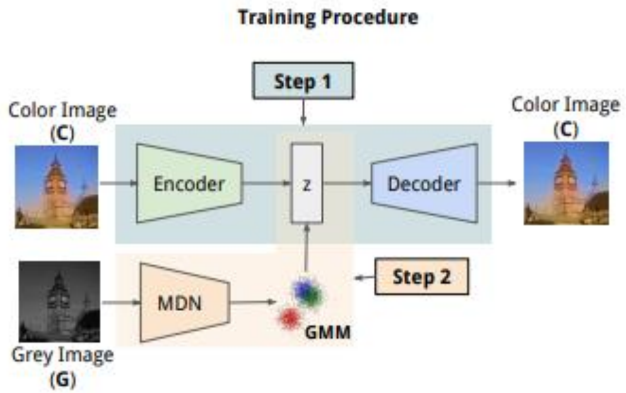
Comprasion with Modern State of Art Apporaches



Current Architecture

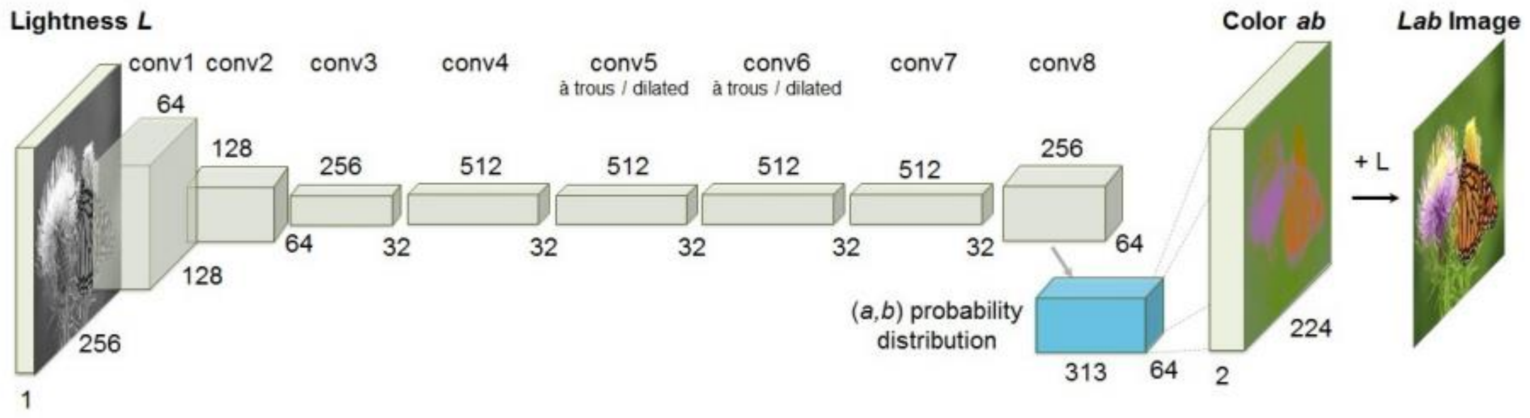


Colorful Image Colorization , Zhang et al. 2016



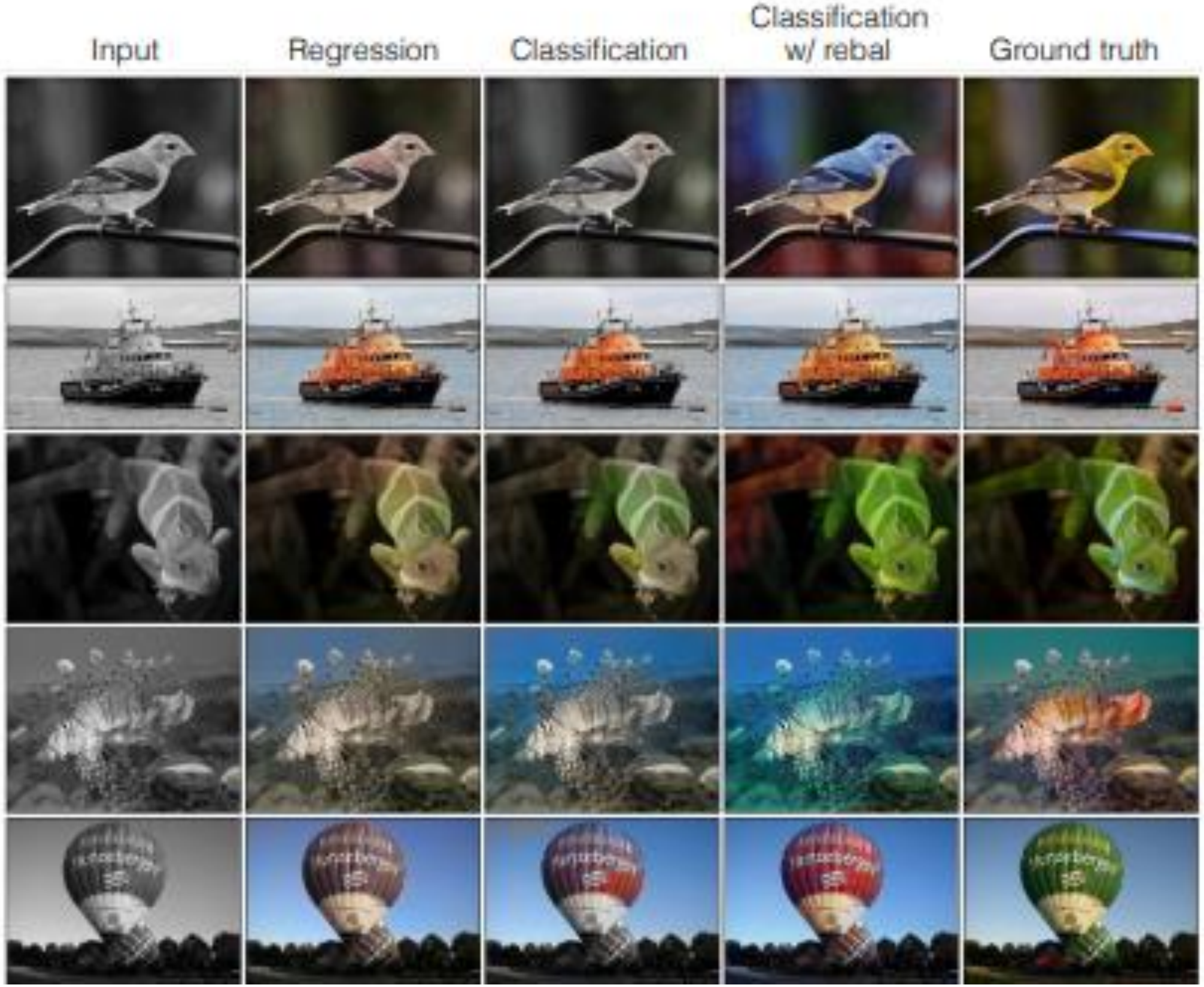
Learning Diverse Image Colorization , Deshpande et al. 2017

Colorful Image Colorization , Zhang et al. 2016

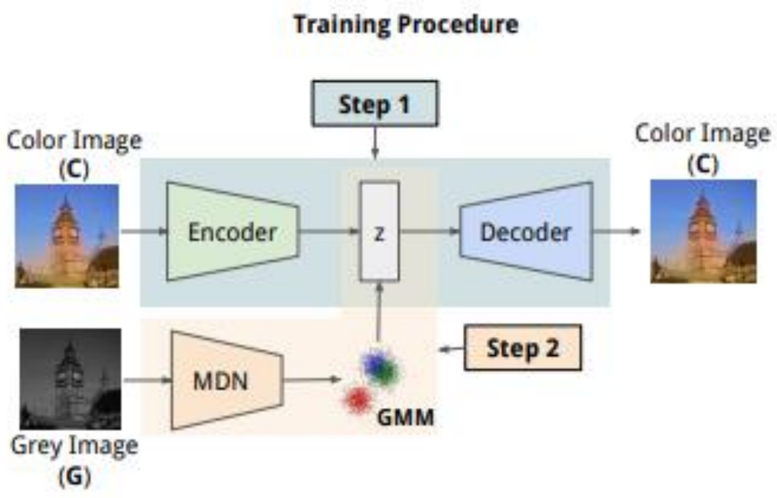


- Pros
- Any image can be used in training.
- Easly visualize the blackbox.
- Statistical preventing the overfitting problem (class re-balancing)
- Easy to apply transfer learnign.

- Cons
- Probability distrubition works not as excepted.
- Fixed image size.


Input
Regression
Classification
Classification w/ rebal
Ground truth

Learning Diverse Image Colorization , Deshpande et al. 2017



- Pros
- Single image, multiple possible outputs.
- Taking advantage of mixture density network
- Better statistical approach with using GMM.
- More accurate results.


Ours
Ours+Skip
cGAN
CVAE
GT
Ours
Ours+Skip
cGAN
CVAE
GT
Ours
Ours+Skip
cGAN
CVAE
GT
Ours
Ours+Skip

Dataset

- MIT Places Scene Dataset [Zhou+ 2014]
- 2.3 million training images with 205 scene labels
  - 256 x 256 pixels

Abbey

Airport terminal

Baseball field

Gift shop

Computational Time

CPU : Intel Core i7-5960X CPU @ 3.00 GHz with 8 cores

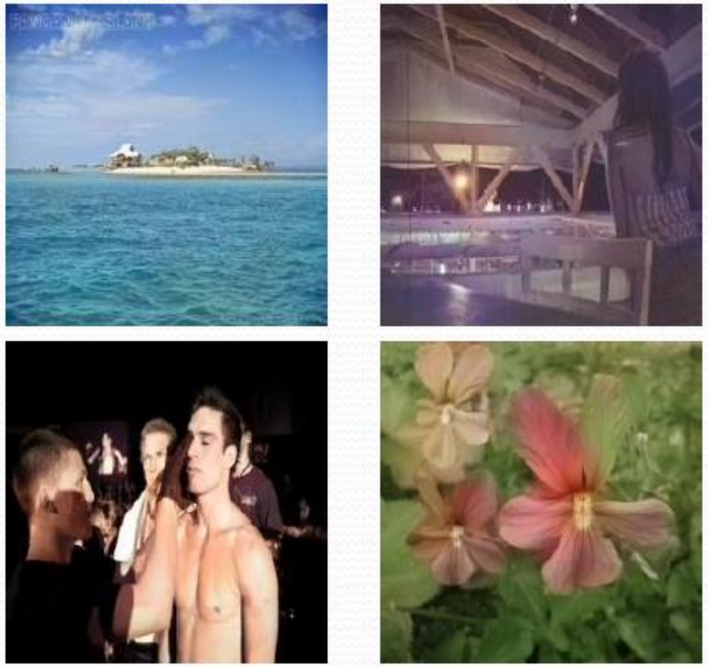
GPU : NVIDIA GeForce GTX TITAN X

Image Size	Pixels	CPU (s)	GPU (s)	Speedup
224x224	50,176	0.399	0.080	5.0 X
512x512	262,144	1.676	0.339	4.9 X
1024x1024	1,048,576	5.629	1.084	5.2 X
2048x2048	4,194,304	20.116	4.218	4.8 X

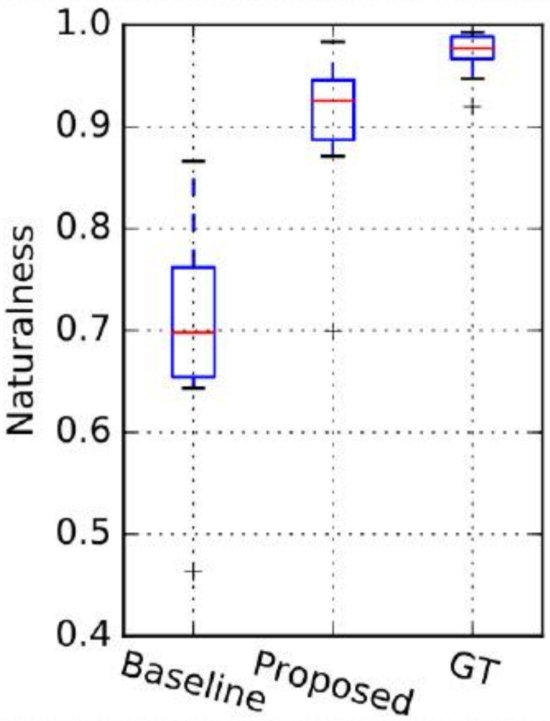
User Study

Does this image look natural to you?

- 10 users participated
- We show 500images of each type: total 1,500 images per user
- 90%of our results are considered “natural”



Natural Unnatural



Approach	Naturalness (median)
Ground Truth	97.7%
Proposed	92.6%
Baseline	69.8%

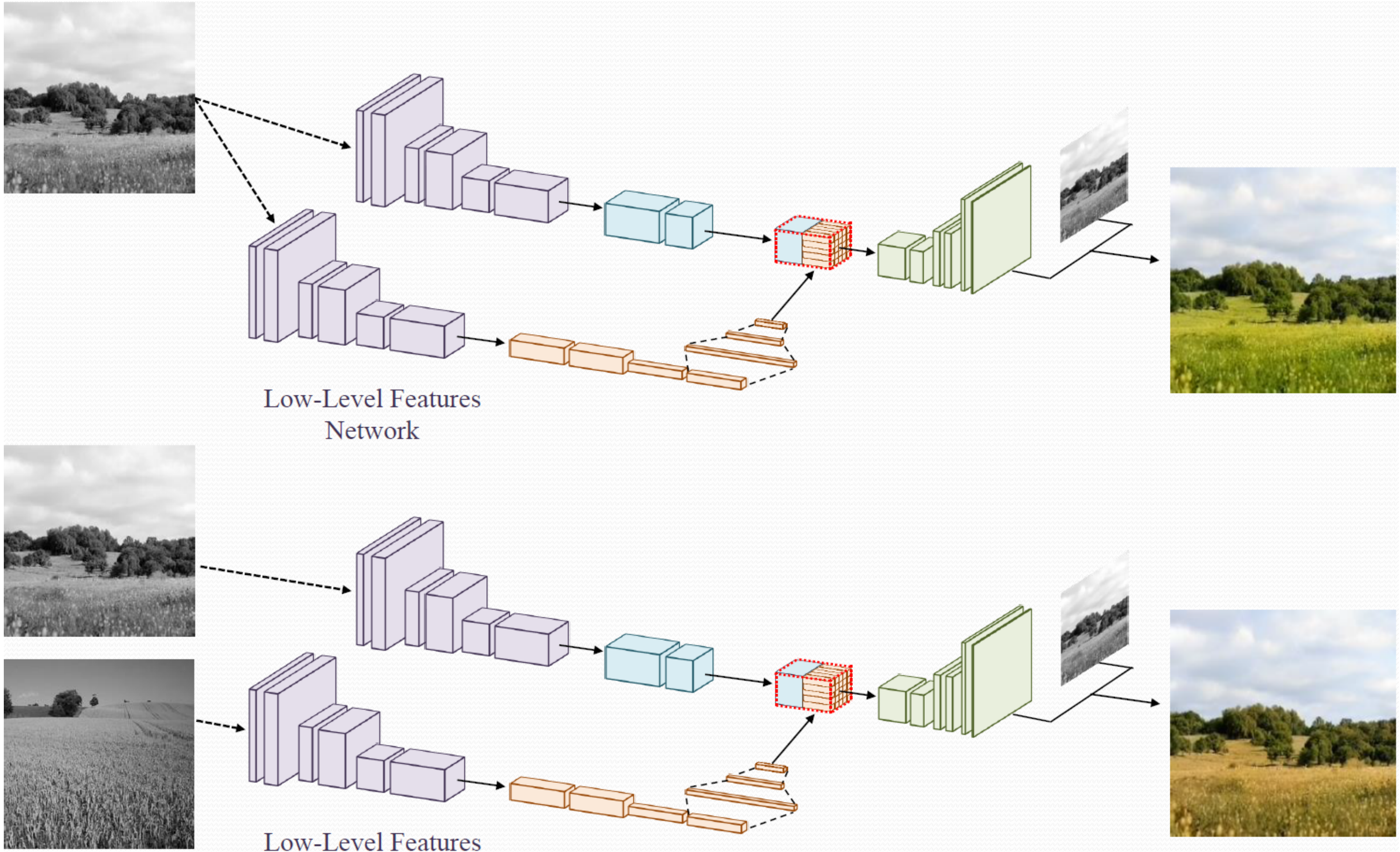
Results

Colorization of MIT Places dataset

Colorization of Historical Photographs

Mount Moran, 1941 | Scott's Run, 1937 | Youngsters, 1912 | Burns Basement, 1910

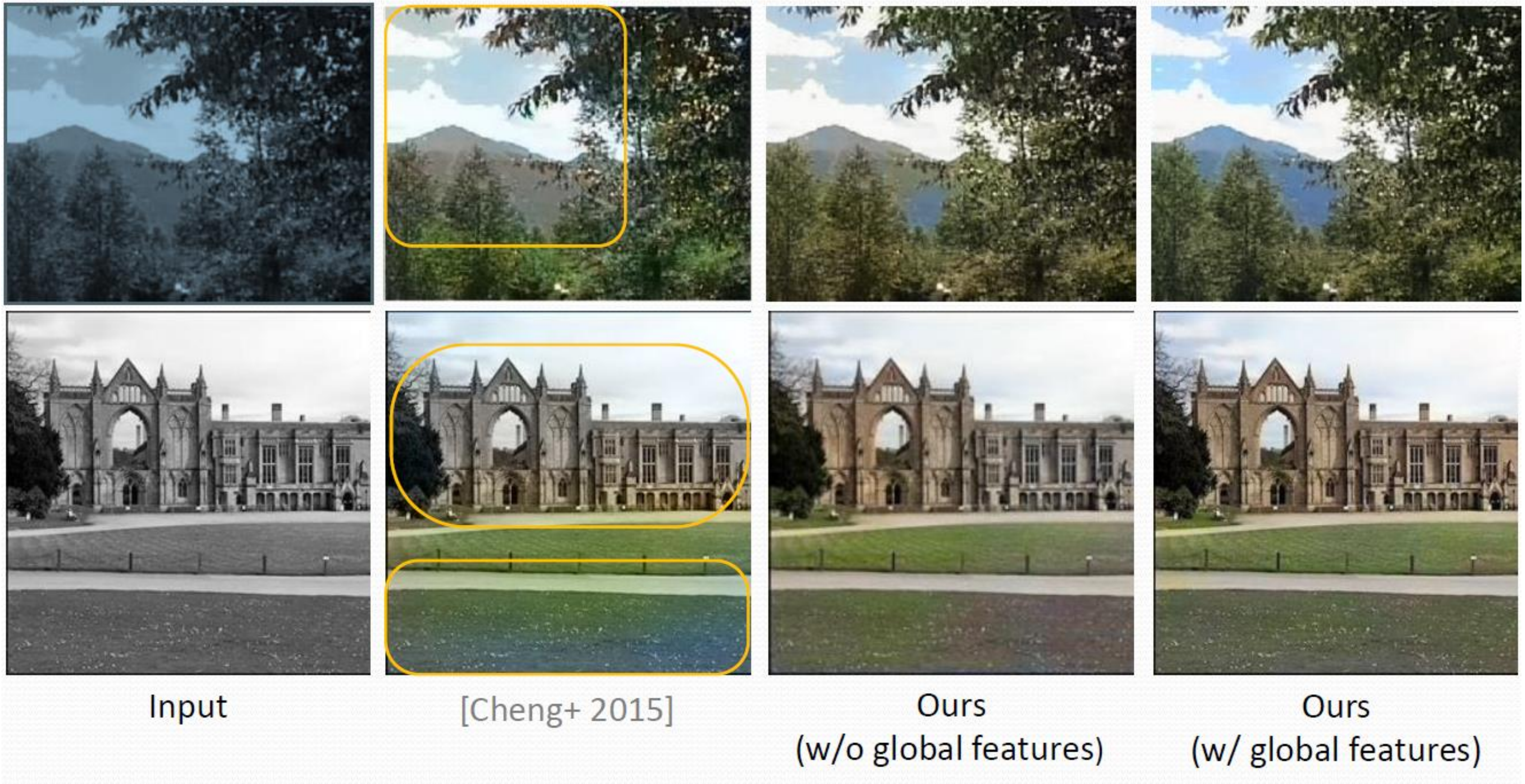
Style Transfer



Style Transfer

- Adapting the colorization of one image to the style of another

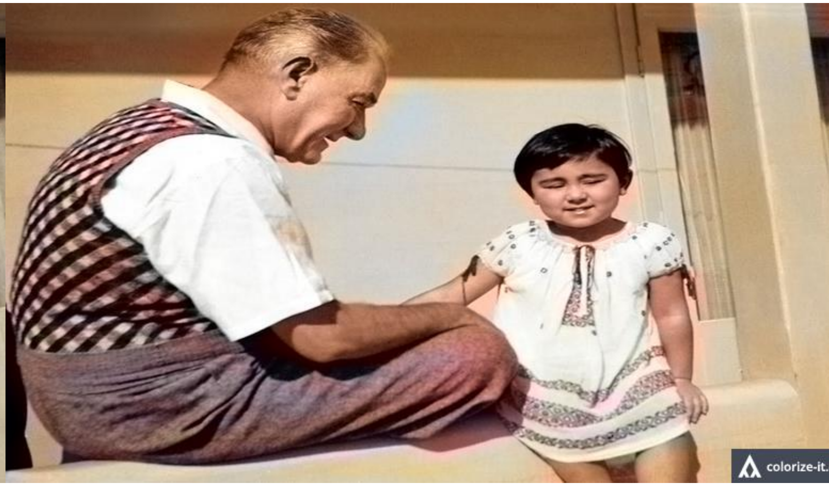
Comparisons



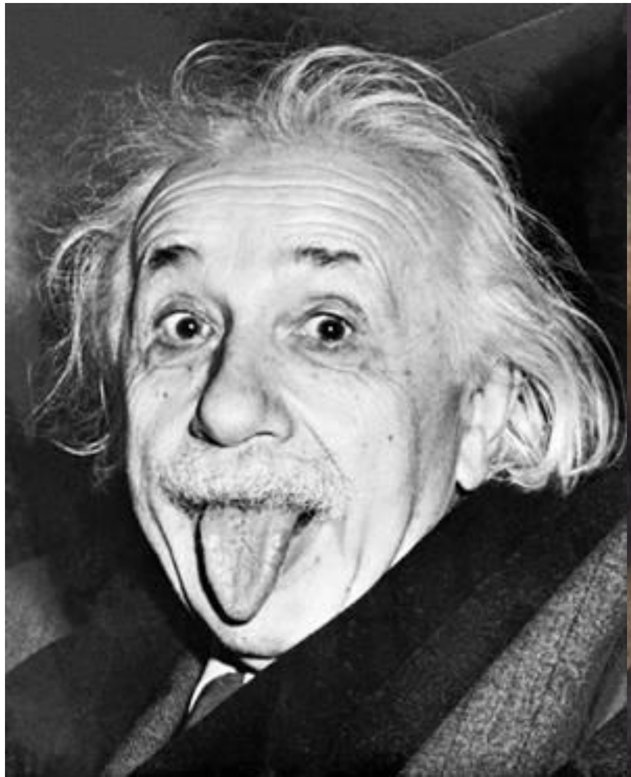
Grayscale

Lizuka et al.

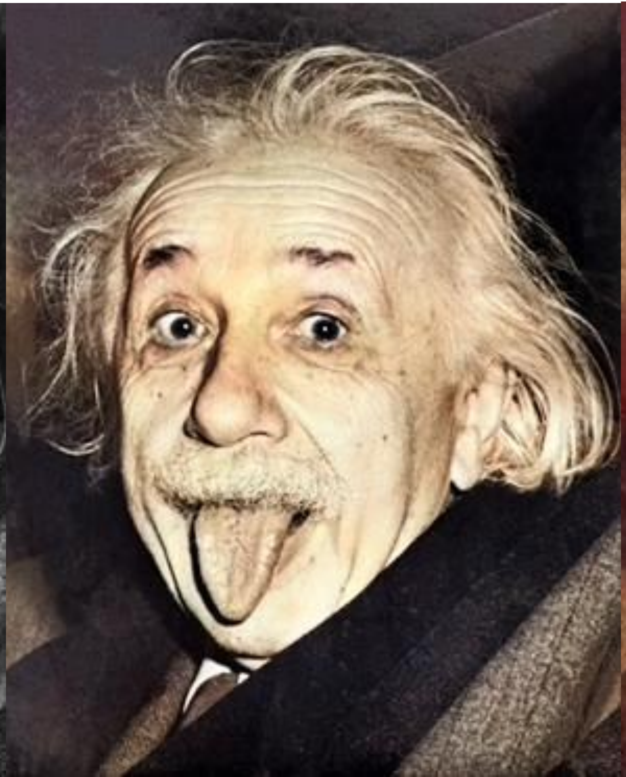
Zhang et al.



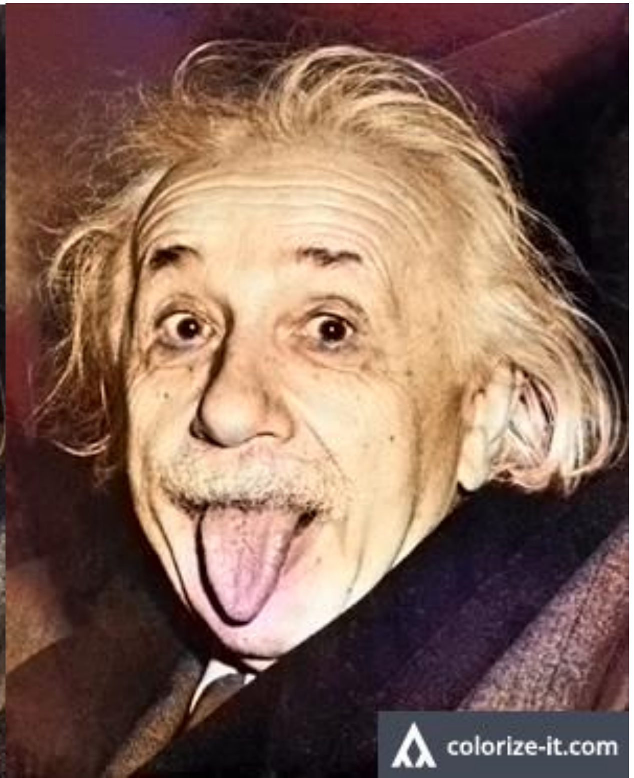
Grayscale



Lizuka et al.

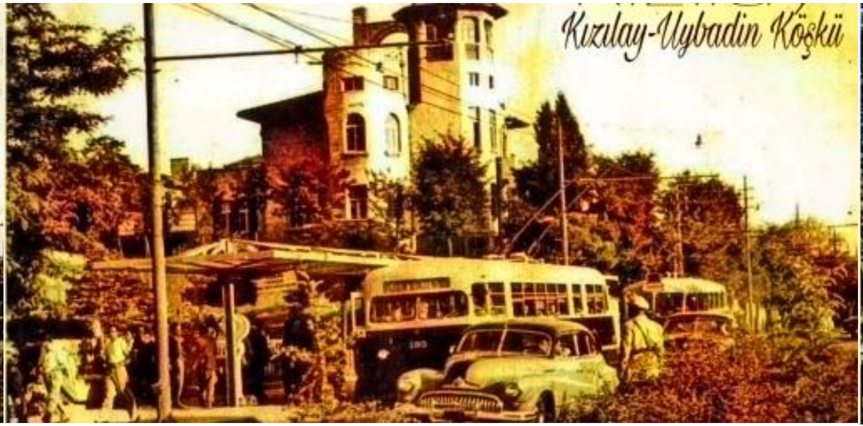


Zhang et al.



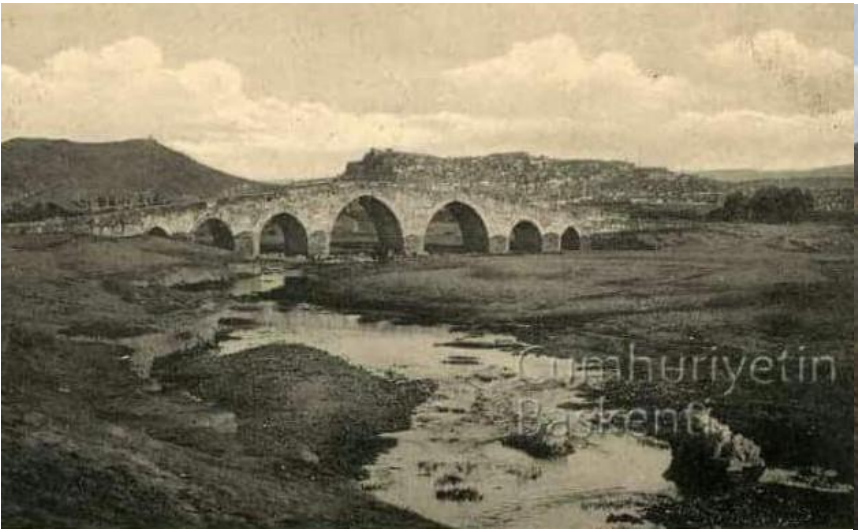
Grayscale

Lizuka et al.



Zhang et al.

Grayscale



Lizuka et al.

Zhang et al.

Grayscale	GT	Lizuka et al.	Zhang et al.

Grayscale

GT

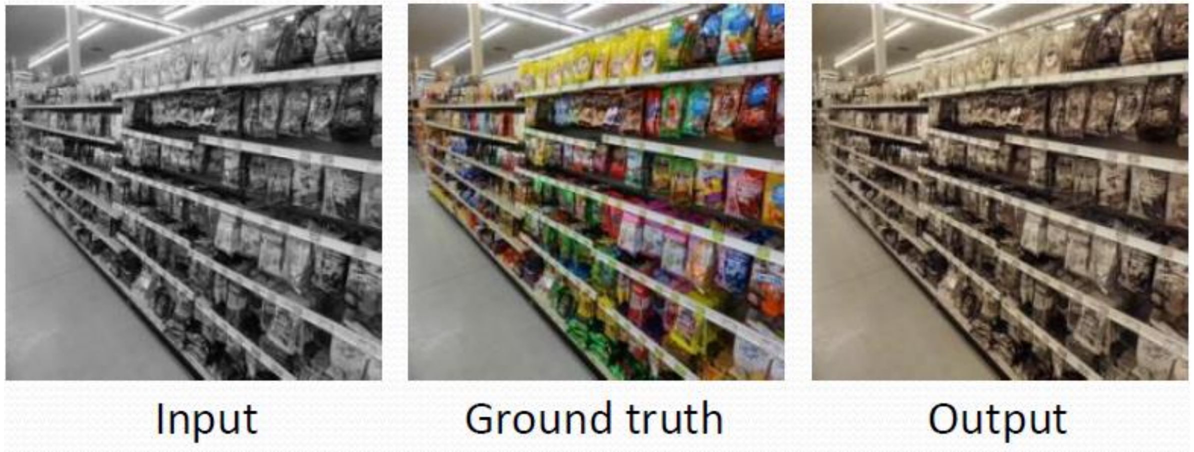
Lizuka et al.

Zhang et al.

Grayscale GT Lizuka et al. Zhang et al.

Limitations

- Difficult to output colorful images



- Cannot restore exact colors

Conclusion

- Novel approach for image colorization by fusing global and local information
  - Fusion layer
  - Joint training of colorization and clasification
  - Style taransfer
- Architecture allows us to process images of any resolution
- Using multi model CNN with adding conditional behavior after fusing layer
- Run in near real-time

Future Work

- If clasification layer performance improve, their result will be be more accuracy.
- However, this does not contain, for example, human-created images. If we wish to evaluate on significantly different types of images.
- Regularization with Dropout

Thank you!