Master Thesis Defense

Automatic Image Colorization Using Semantic Guides

Committee: Prof. Name 1
Prof. Name 2

Advisor: Prof. Name 3

Presenter: Name
Department of Artificial Intelligence Convergence,
Chonnam National University, Korea
email

December 1st, 2020

Agenda

1. INTRODUCTION

- **Problem: Fully Automatic Colorization**
  - Given the **grayscale image**, produce ***a plausible colorization to fool a human observer***.
  - **Input**: Grayscale image or L channel of image, output ab channel of image

1. INTRODUCTION

- **Challenges of Fully Automatic Colorization**:
  - **Averaging effect**: grayish, desaturated results due to 94% of the cells in our eyes determine brightness, only 6% for colors. Grayscale image is a lot sharper than the color layers.

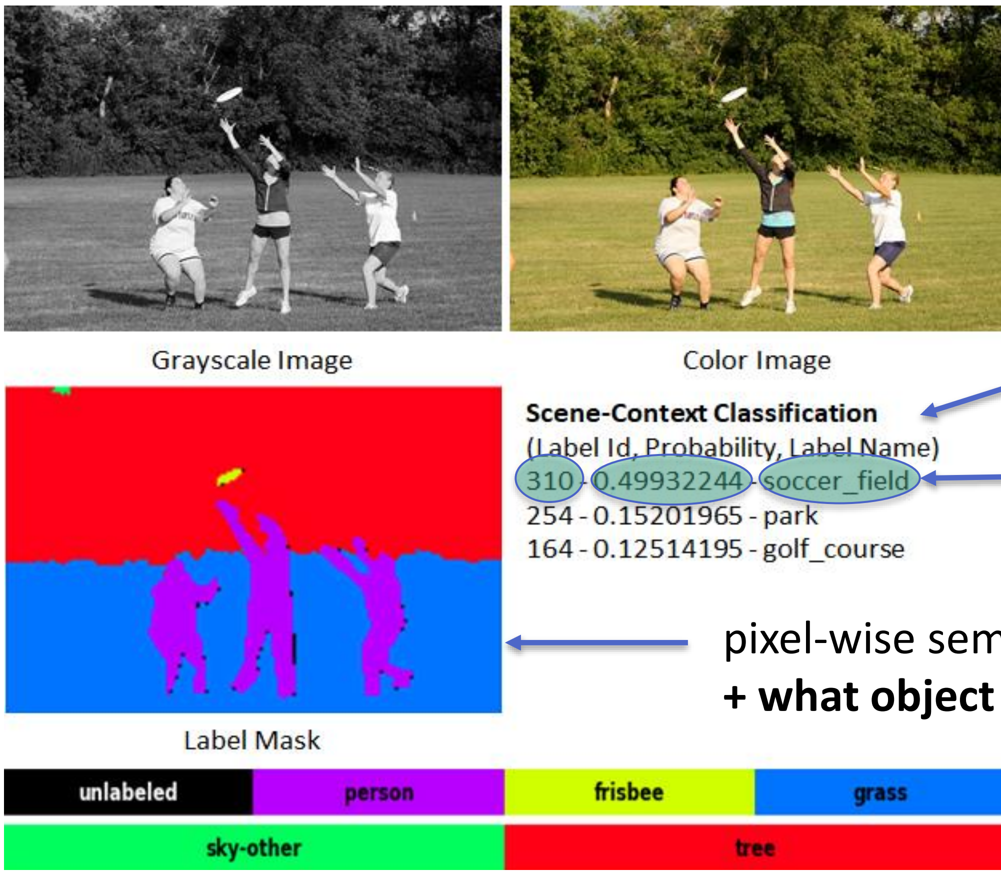
  - **Rare colors in images**: strongly biased due to the appearance of backgrounds such as clouds, pavement, dirt, and walls.

  - **Semantic information matters**: In order to colorize any kind of image, a system must interpret the semantic composition of the scene (what is in the image: faces, cars, plants, . . .) as well as localize objects (where things are).

GT: lagoon
top-1: balcony interior (0.136)
top-2: beach house (0.134)
top-3: boardwalk (0.123)
top-4: roof garden (0.103)
top-5: restaurant patio (0.068)

1. INTRODUCTION

- **Objectives**:
  - Integrate scene-context classification and pixel-wise semantic segmentation



scene-context classification
+ global scene information

Scene-context classes
(totally 365 classes)

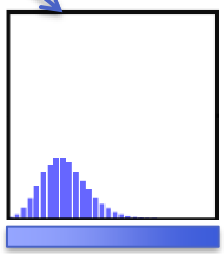
pixel-wise semantic segmentation
+ what object the pixels belong to

Segmentation classes in Coco-Stuff
(0: unlabeld, 1 – 182: objects & stuffes)

1. INTRODUCTION

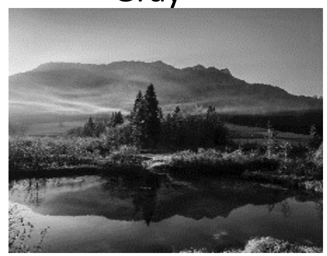
- **Objectives**:
  – Use ab color distribution ***to encourage rare color (rebalancing colors), and multi-modal in colorization***

With a pixel

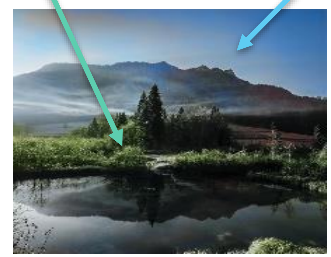


ab color distribution
vs.
ab color value

Gray

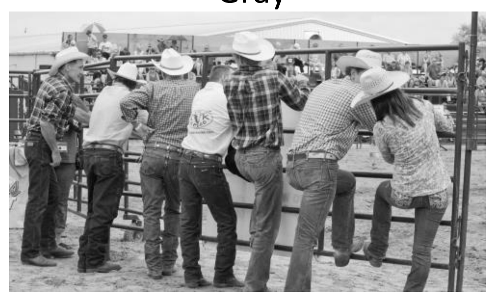


Tree
Sky (common colors)
Ours



GT

Gray



GT

Shirt (diversity colors, rare colors)

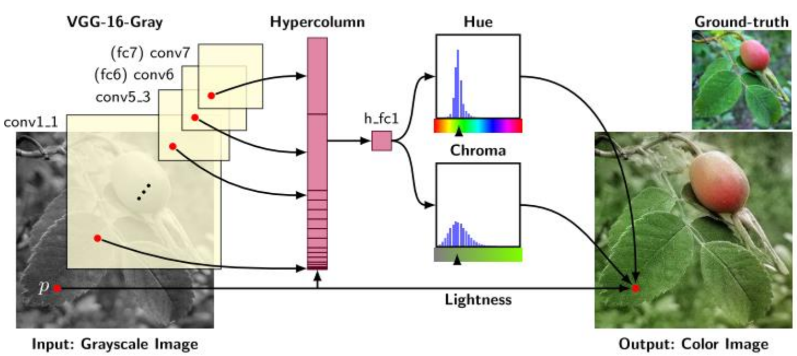
Grayish result

Multi-Modal Attribute or Bias
(many choice in colorization)
leading to
Grayish or Desaturated Effect

2. RELATED WORKS

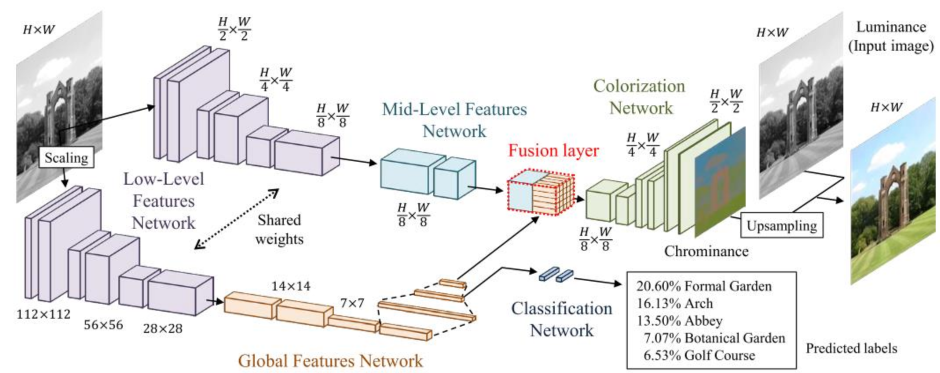
- **Larsson et al.**[1]: use un-rebalanced classification **loss**, build on hyper-columns on a VGG **network**, train on **ImageNet**, evaluate on **PSNR, RMSE**.
- **Iizuka et al.**[2]: use a regression **loss**, build a ***two-stream architecture*** fusing global and local features, train on ***Places365 scene dataset.***

Regression



Larsson et al.

Classification

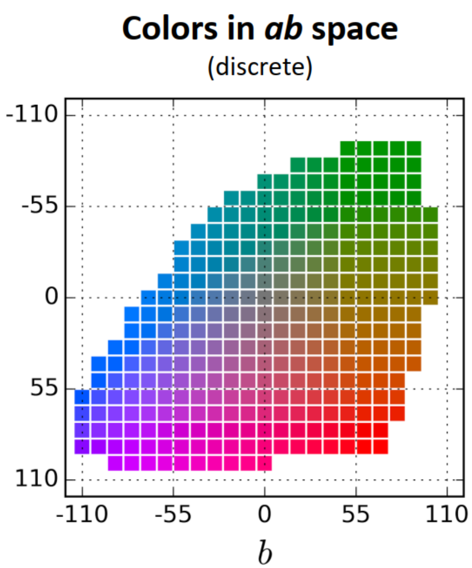


Iizuka et al.

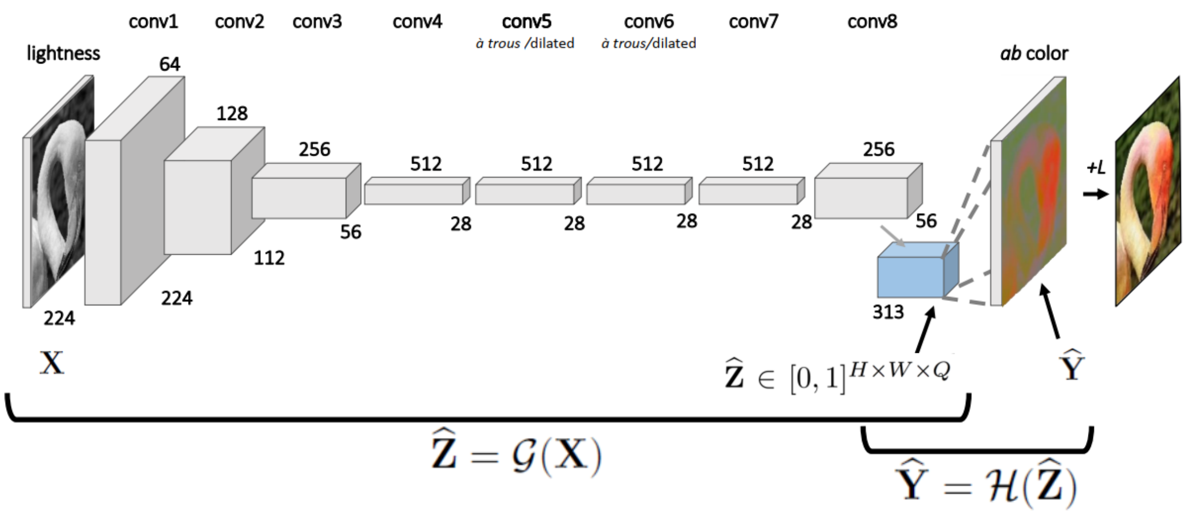
[1] G. Larsson, M. Maire, and G. Shakhnarovich, “***Learning Representations for Automatic Colorization***,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, 2016, pp. 577–593.
[2] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “***Let there be Color: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classificatio***,” *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–11, Jul. 2016.

2. RELATED WORKS

- **Zhang et. at**[1]
  - ***Multi-Class Classification*** problem by ***quantize ab space*** into grid size 10, keep 313 bins in gamut.
  - Category cross entropy loss with ***class rebalancing*** to encourage learning of rare colors.



quantize ab space with grid size 10 (313 bins)

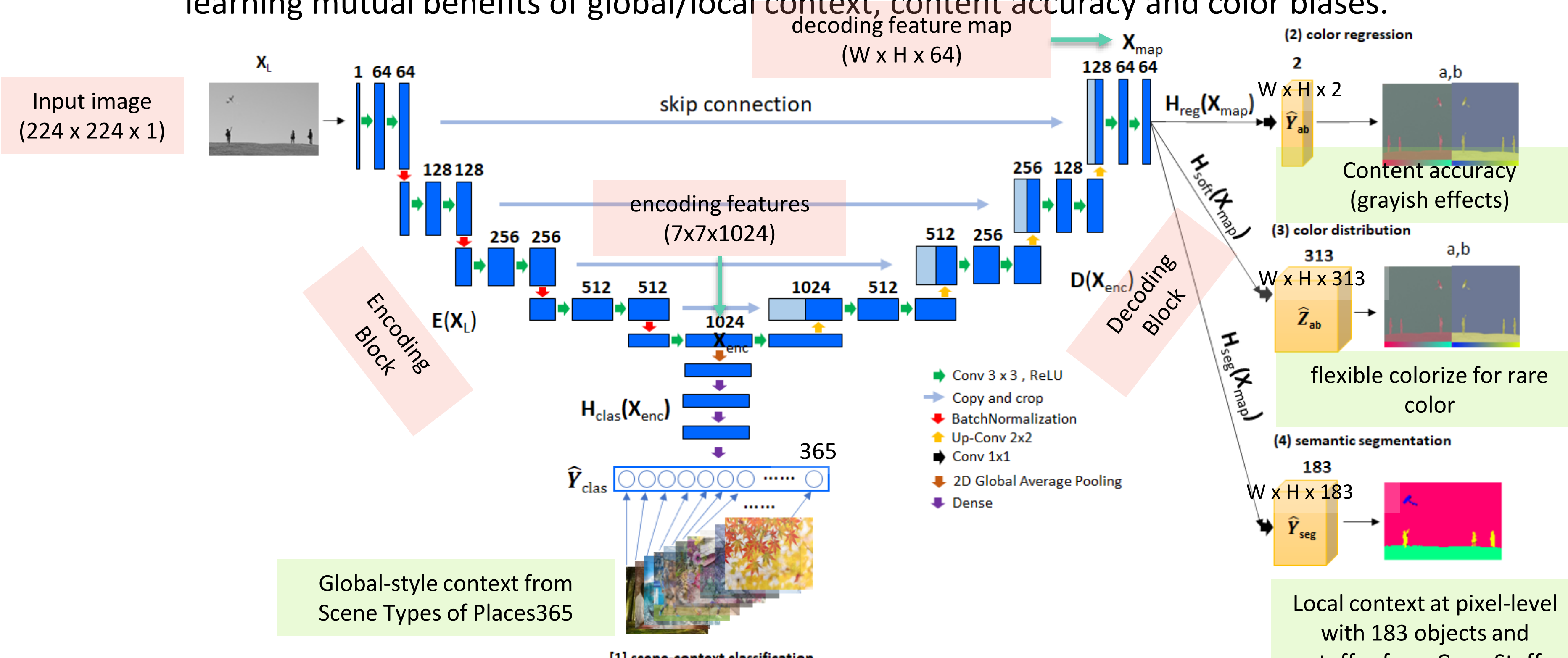


Deep network model

[1] R. Zhang, P. Isola, and A. A. Efros, **Colorful Image Colorization**, ECCV, pp. 649–666, 2016

3. PROPOSED METHOD

- **Our Model: Main Idea**
  - **Take advantage of skip connections** between the contracting and expanding path at the same depth level using U-Net model (prevent dying ReLU and vanishing problem[1])
  - **Use multi-task learning** with **end-end training** from **gray-scale image** to **four outputs** for learning mutual benefits of global/local context, content accuracy and color biases.

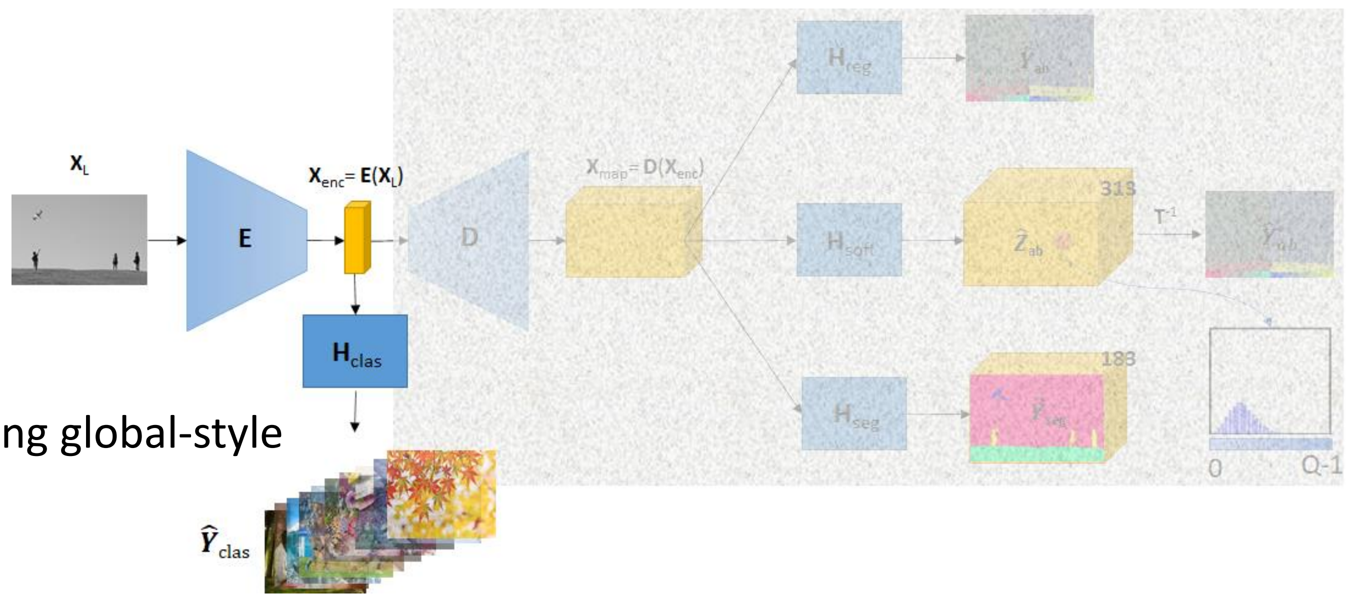


Semantic Image Colorization U-Net Model

[1] Douglas et al., "***Why RELU Units Sometimes Die: Analysis of Single-Unit Error Backpropagation in Neural Networks***," In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 864-868, 2018

3. PROPOSED METHOD

- **Explanation Details: Classification Branch**
  - Compute **backward gradient** of the classification loss to enhance encoding feature $\mathbf{X}_{enc}$ and **Encoder Block E** with scene global-style during training process
  - Create scene label ground-truth for training data:
    - **use pre-train weights** VGG16 on Places365 Dataset to predict scene labels
    - **apply label smoothing technique**

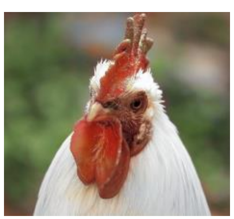


Learning global-style

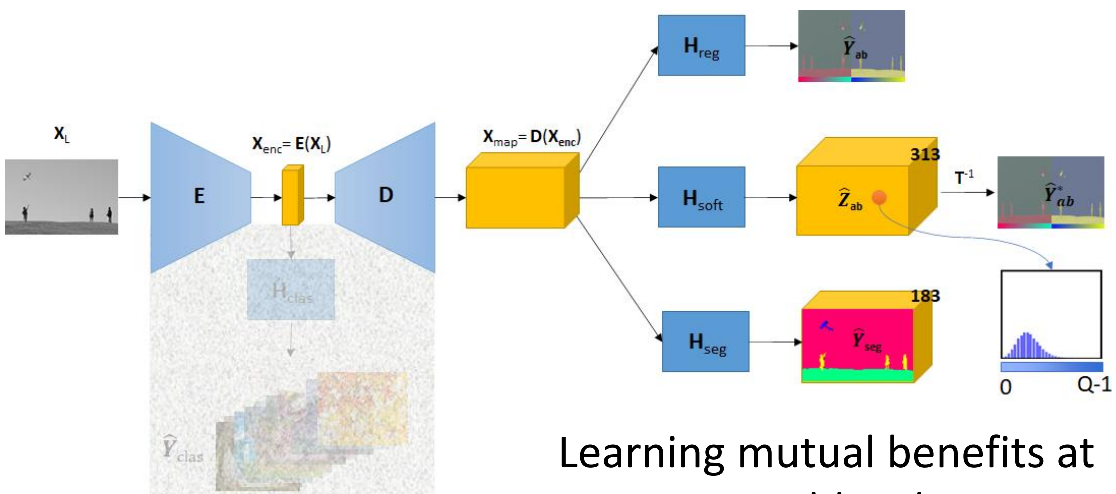
3. PROPOSED METHOD

- **Explanation Details: Regression/Color Distribution/Segmentation Branches**
  - Compute **backward gradients** of **three branches** to enhance decoding feature map $\mathbf{X}_{\mathbf{map}}$ and encoding feature $\mathbf{X}_{\mathbf{enc}}$
    - **regression branch** to keep the accuracy between prediction/ground-truth → **output results** with grayish and desaturated effects (not used as colorized result)
    - **color distribution branch** to encourage rare color (rebalancing colors) and multi-modal in colorization → output results with more vivid
    - **segmentation branch** to help the system understand what object the pixels belong to (with 183 object & stuff labels) → output results with more precise edge

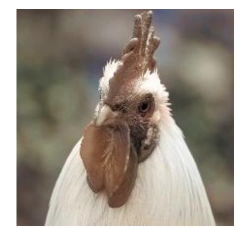
Gray



Ground-Truth

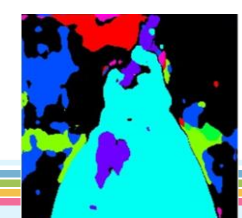


Learning mutual benefits at pixel-level



Reg

Soft colorize result



Seg

3. PROPOSED METHOD

- **More details: The scene-context classification:**
  - Extract the scene probabilities of training dataset (without scene-context ground-truth) based on **pre-trained model on Places365**[1].

Image

Pre-trained Model

0	1	2	3	...	n-1

Scene probability with n=365

  - **Label Smoothing**[2] **with top-5 prediction**: keep 5 highest probabilities, set all remain values to 0, and normalize the probabilities with sum 1.

Top-1: Cafeteria (0.179)
Top-2: Restaurant (0.167)
Top-3: dining_hall (0.091)
Top-4: coffe_shop (0.086)
Top-5: restaurant_patio (0.080)

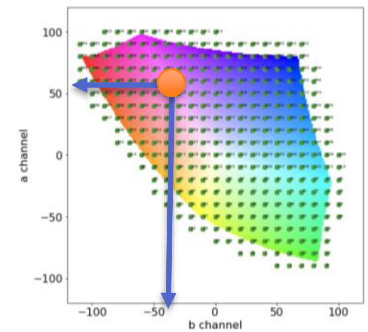
Top-1: Cafeteria (0.297)
Top-2: Restaurant (0.277)
Top-3: dining_hall (0.151)
Top-4: coffe_shop (0.143)
Top-5: restaurant_patio (0.132)

[1] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 Million Image Database for Scene Recognition,” IEEE transactions on pattern analysis and machine intelligence (TPAMI), vol. 40, no. 6, pp. 1452–1464, 2018
[2] R. Müller, S. Kornblith, and G. Hinton, “***When Does Label Smoothing Help?***,” In Advances in Neural Information Processing Systems (NeurIPS), pp.4696-4705, 2019.

3. PROPOSED METHOD

- **More details: The ab color distribution**
  - **Soft-Encoding Process:**
    - **Step 1**: For every pixel of image, convert from ab values to color index q (encoding) using K-Nearest

$I_{ab}(p) = (a, b) \longrightarrow q \in [0{,}312]$



quantize ab space with grid size 10 (313 bins)

- **Step 2**: Convert to one-hot encoding representation

0 ... **312**

	...	0	0	1	0	0	...	

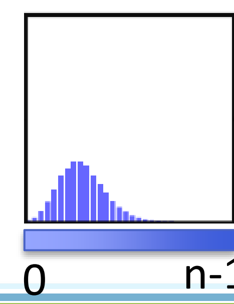
q

- **Step 3**: Apply label smoothing
  - Use K-Nearest neighbors to get 4 color indexes nearest q,
  - Generate 5 gaussian values, and normalize

0 ... **312**

	...	0.05	0.24	0.42	0.24	0.05	...	

q



3. PROPOSED METHOD

- **Multi-Task Losses:** $\mathcal{L}_{total} = w_{soft}\mathcal{L}_{soft} + w_{clas}\,\mathcal{L}_{clas} + w_{seg}\,\mathcal{L}_{seg} + w_{reg}\,\mathcal{L}_{reg}$
  - **Pixel Classification of ab color distribution**: Weighted Category Cross-Entropy Loss:

$$\mathcal{L}_{soft}(y,\hat{y}) = -\sum_{h,w} v(y_{h,w}) \sum_{i=1}^{N} y_{h,w,i} \log \hat{y}_{h,w,i}$$

Where N is the number of quantized colors of ab color distribution, $v(y_{h,w})$ is the weighted of color-classs at pixel (h,w) to encourage the rare-color, $y_{h,w,i}$ / $\hat{y}_{h,w,i}$ is the ground-truth/prediction probability of the soft-encoding color i at pixel (h,w).

  - **Scene-context classification:** Category Cross-Entropy (CCE) loss:

$$\mathcal{L}_{clas}(y,\hat{y}) = -\sum_{i=1}^{C} y_i \log \hat{y}_i$$

Where C is the number of scene, y_i/ $\hat{y}_i$ is the ground-truth/predicted scene probability.

  - **Pixel-wise sematic segmentation:** Dice loss:

$$\mathcal{L}_{seg}(y,\hat{y}) = 1 - \frac{2\sum_p y\hat{y}}{\left(\sum_p y\right)^2 + \left(\sum_p \hat{y}\right)^2}$$

  - **Regression ab channel**: Using Mean Square Error (MSE) Loss:

$$\mathcal{L}_{reg}(y,\hat{y}) = \frac{1}{2hw}\sum_{h,w} \|y-\hat{y}\|_2^2$$

4. EXPERIMENTS AND DISCUSSION

- **Training, Validation and Testing Datasets**:

Method	Training	Validation	Testing
COCO-Stuff [1]	118000	5000	1000 (ctest1k)
Places365 [2]			1000 (ctest1k)
DIV2K [3]			100 (high-resolution)
ImageNet [4]			1000 (ctest1k)

[1] Caesar, J. Uijlings, and V. Ferrari, "***COCO-Stuff: Thing and StuffClasses in Context***,"Proceedings of the IEEE Computer Society Confer-ence on Computer Vision and Pattern Recognition, pp. 1209–1218, 2018.

[2] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "***Places: A10 Million Image Database for Scene Recognition***,"IEEE Transactions onPattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452–1464,2018.

[3] E. Agustsson and R. Timofte, "***NTIRE 2017 Challenge on Single ImageSuper-Resolution: Dataset and Study***," inThe IEEE Conference on Com-puter Vision and Pattern Recognition (CVPR) Workshops, jul 2017.

[4] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei,"***ImageNet: A large-scale hierarchical image database***," in2009 IEEEConference on Computer Vision and Pattern Recognition.IEEE, jun2009, pp. 248–255.

4. EXPERIMENTS AND DISCUSSION

- **Quantitative comparisons**:

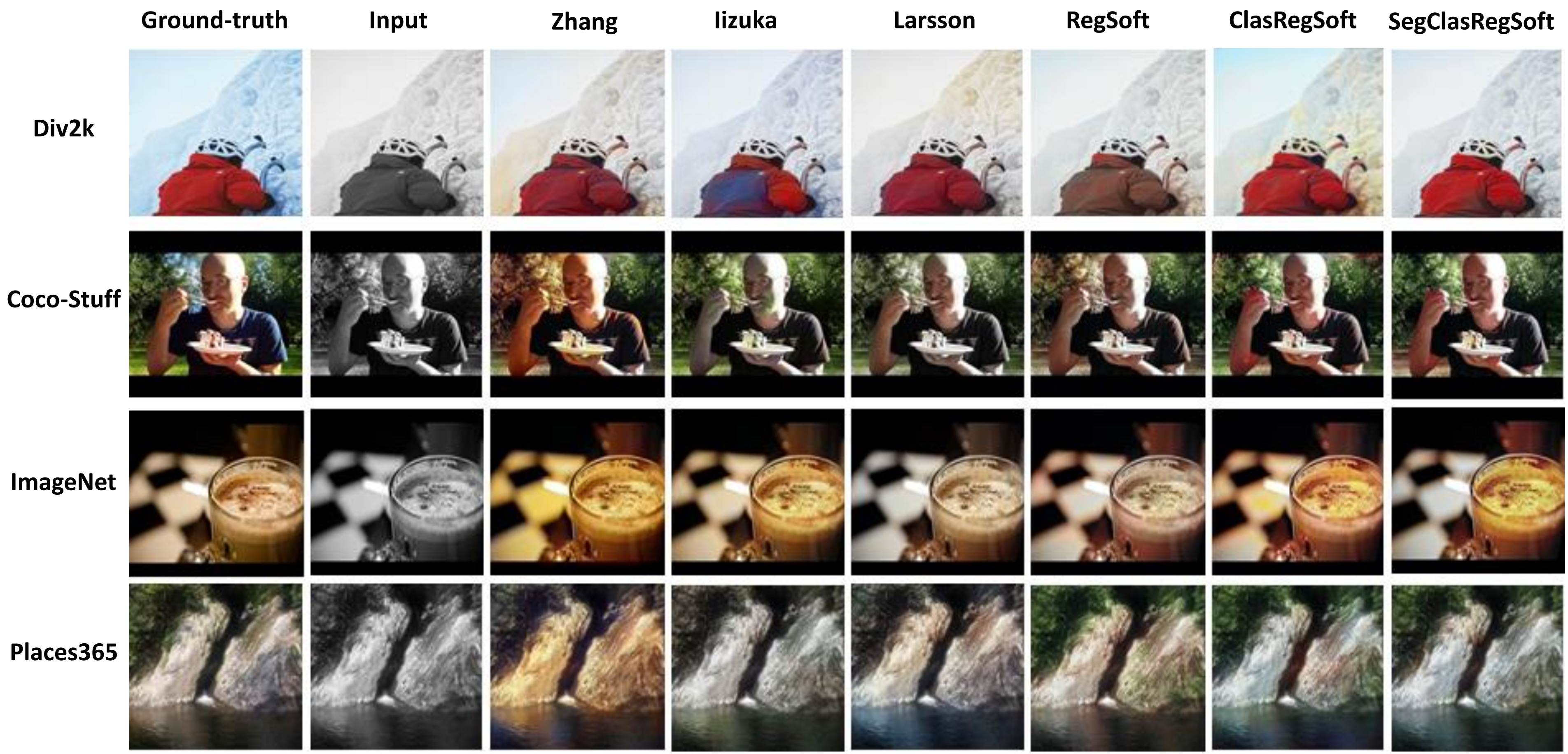
Method	ImageNet ctest1k			DIVK2K		
	PSNR ↑	SSIM ↑	$L2_{ab}$ ↓	PSNR ↑	SSIM ↑	$L2_{ab}$ ↓
Iizuka et al. [7]	22.841	0.865	0.277	22.981	0.919	0.079
Larsson et al. [8]	23.335	0.869	0.26	23.490	0.929	0.072
Zhang et al. [11]	21.297	0.848	0.286	20.929	0.896	0.079
Ours with RegSoft	22.102	0.896	0.269	22.026	0.914	0.071
Ours with ClassRegSoft	21.068	0.886	0.274	21.694	0.912	0.071
Ours with SegClassRegSoft	21.900	0.893	0.264	22.330	0.917	0.068

Method	Place365 ctest1k			COCO-Stuff ctest1k		
	PSNR ↑	SSIM ↑	$L2_{ab}$ ↓	PSNR ↑	SSIM ↑	$L2_{ab}$ ↓
Iizuka et al. [7]	25.572	0.948	0.481	23.541	0.871	0.242
Larsson et al. [8]	25.096	0.945	0.452	23.773	0.873	0.223
Zhang et al. [11]	23.076	0.928	0.484	21.502	0.851	0.245
Ours with RegSoft	23.599	0.932	0.474	22.872	0.912	0.23
Ours with ClassRegSoft	22.916	0.924	0.466	22.134	0.907	0.23
Ours with SegClassRegSoft	23.858	0.931	0.442	22.985	0.913	0.223

  – Larsson et al.: better on PSNR for ImageNet,DIV2K, and COCO-Stuff and on SSIM results for ImageNet and DIV2K.
  – Our methods: better on $L2_{ab}$ metric for DIV2K, Places365, and COCO-Stuff
  – Semantic segmentation played an important role in enhancing the colorization results, and it helped our method improve the accuracy of the ab channels.

4. EXPERIMENTS AND DISCUSSION

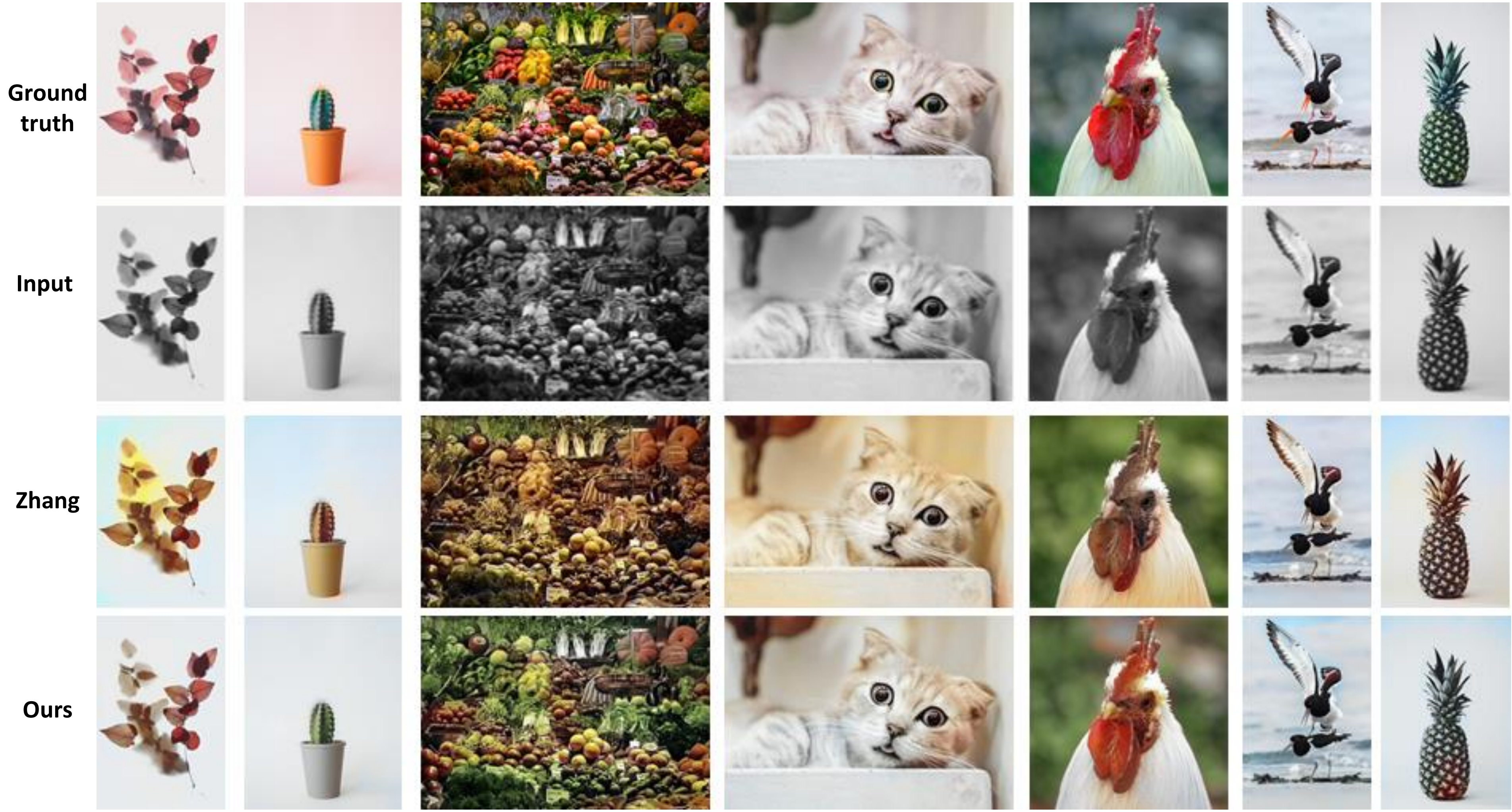
❖ SUCCESSFUL CASES



My results were more vibrant and had more precise edges than the other methods. Moreover, the yellow color noise also was reduced in our ClasRegSoft versions comparison on RegSoft version.

4. EXPERIMENTS AND DISCUSSION

❖ IMAGES FROM INTERNET



4. EXPERIMENTS AND DISCUSSION

❖ LEGACY IMAGES FROM INTERNET

Input

Ours

4. EXPERIMENTS AND DISCUSSION

❖ SOME FAIL CASES

Ground-truth Input Zhang Iizuka Larsson RegSoft ClasRegSoft SegClasRegSoft

My results met difficulties for colorization with incorrect colors, noise occurrences. These defects are similar to the results of Iizuka et al. and Larsson et al..

5. CONCLUSIONS

- I proposed the encoder-decoder architecture to deal with the global and local semantics in the colorization problem.

- Our colorization model is the result of the mutual benefit learning of
  - **scene-context classification** branch to bring the global image style,
  - **semantic segmentation** branch at pixel-level of objects in scenes,
  - **average colorization** branch in the regression branch and
  - **color distribution** branch in a soft-encoding branch.

- In the future, I will enhance the ***optimization process for multi-scale outputs*** to reduce noises. Moreover, I will use ***GANs model*** to colorize images better and get high resolution.

- Published Paper: "***Image colorization using the global scene-context style and pixel-wise semantic segmentation***, " IEEE Access 11/2020 (IF: 3.745), link: https://ieeexplore.ieee.org/document/9272287

THANK YOU FOR LISTENING