Image Colorization using CNNs and Inception-Resnet-v2

Federico Baldassarre, Lucas Rodés-Guirao, Diego González Morín

KTH Royal Institute of Technology

{*fedbal, lucasrg, diegogm*} *@kth.se*

May 23, 2017

TL;DR

- Overview of several approaches to the colorization process
- Combining a deep CNN with high-level feature from a pre-trained Inception-ResNet-v2 to enhance the coloring process
- Training and results on ImageNet
- User study to assess public acceptance
- Colorization of historical pictures

Goal

Black-and-White image colorization using a deep Convolutional neural Network and Inception-Resnet-v2 pre-trained network.

Background - First steps

Scribble-based methods - Welsh (2002)

- Transfers color from a related reference image

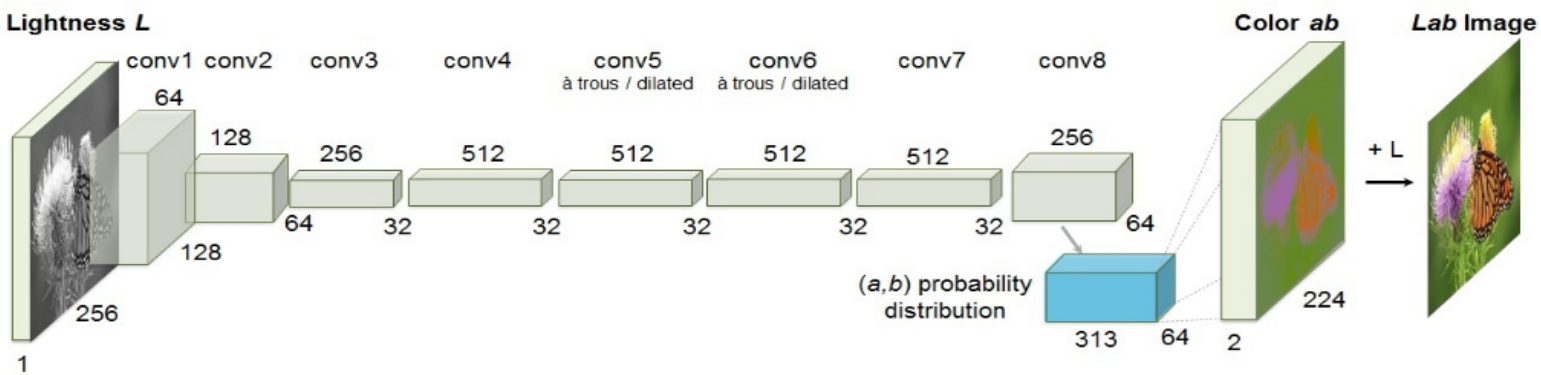
Based in reference images - Levin(2004)

- User-guided color using scribbles

Background - Convolutional networks

End-to-end Convolutional Neural Network colorization approaches

- Iizuka, S.(2016) - proposed a Convolutional Neural Network structure in which global-level and mid-level features were used to encode and colorize the images and colorize them.
- Zhang, R.(2016) - multi-modal Convolutional Neural Network scheme, where each pixel was given a probability value for each possible color.



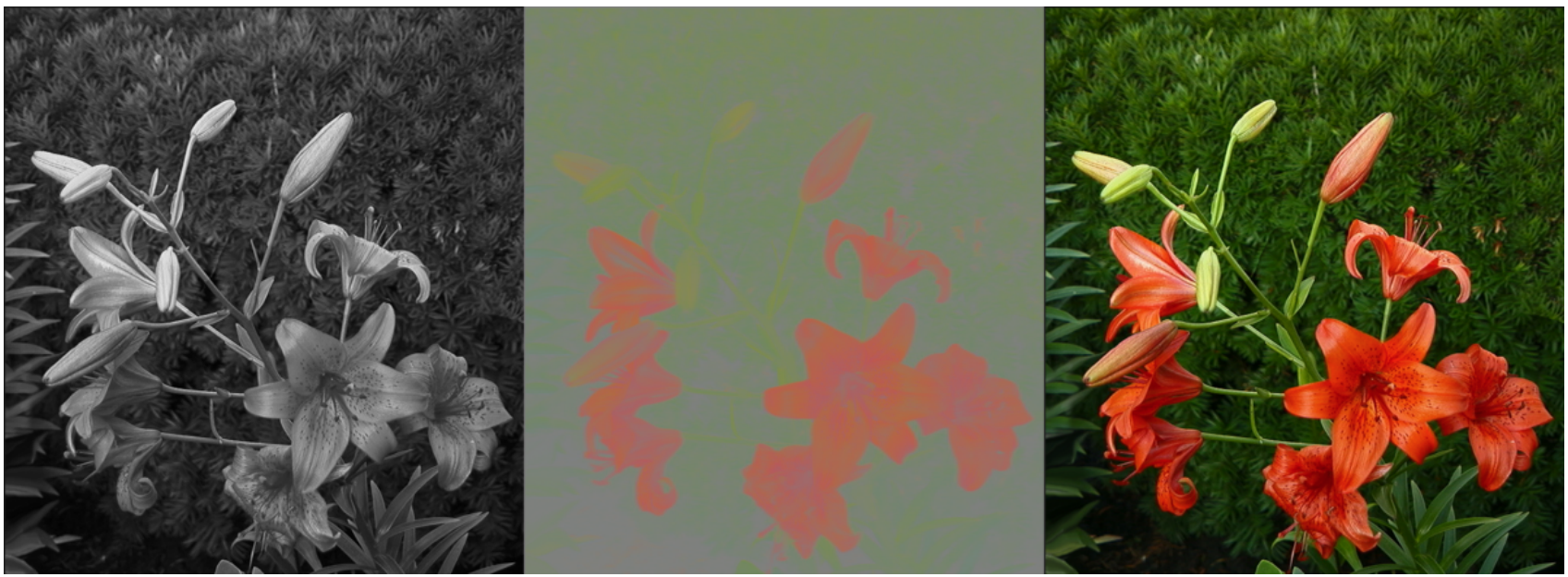
Background - Convolutional networks

End-to-end Convolutional Neural Network colorization approaches

- Zhang, R.(2017) - presented an an end-to-end CNN approach incorporating user "hints" in the spirit of scribble based methods allowing real-time use.

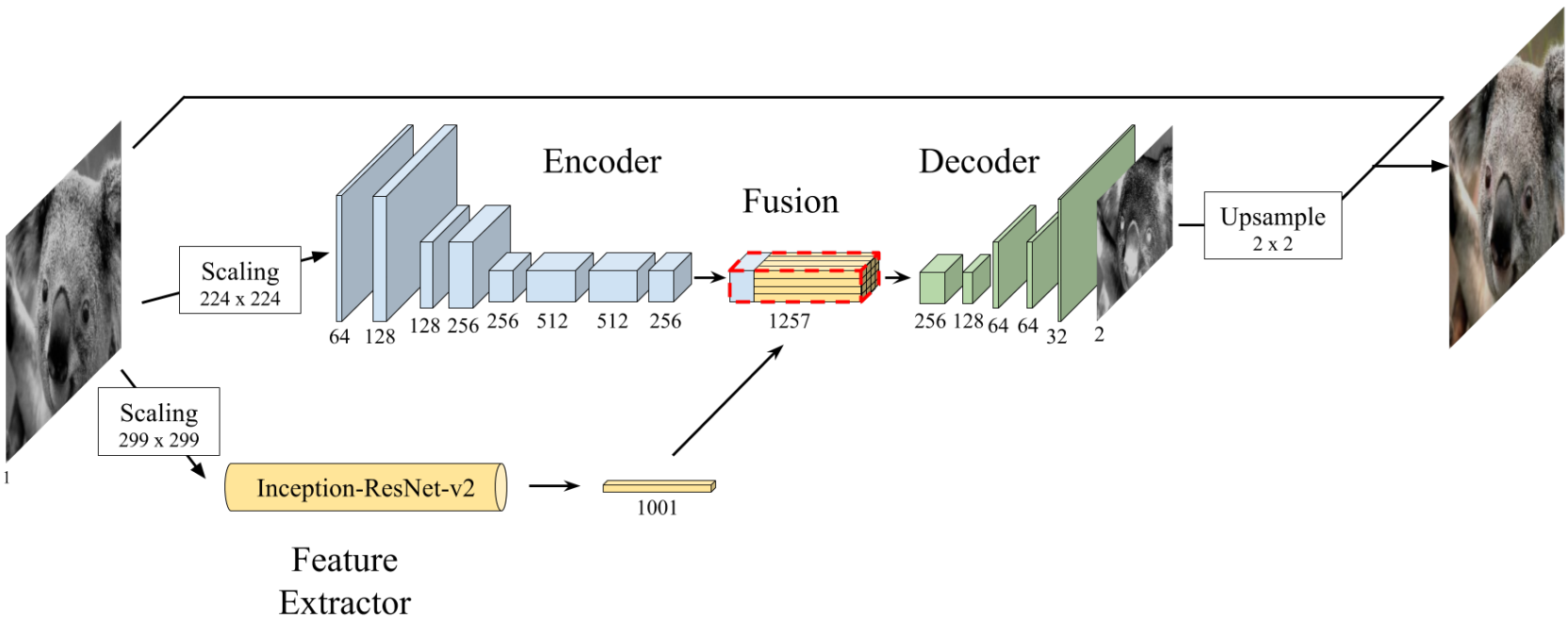
Color representation

- CIE L*a*b* colorspace for color representation:
  - Luminance dimension comes from the grayscale image
  - The network predicts the a*b* color components
- $$\mathcal{F} : \mathbf{X_L} \rightarrow (\tilde{\mathbf{X}}_\mathbf{a}, \tilde{\mathbf{X}}_\mathbf{b}),$$



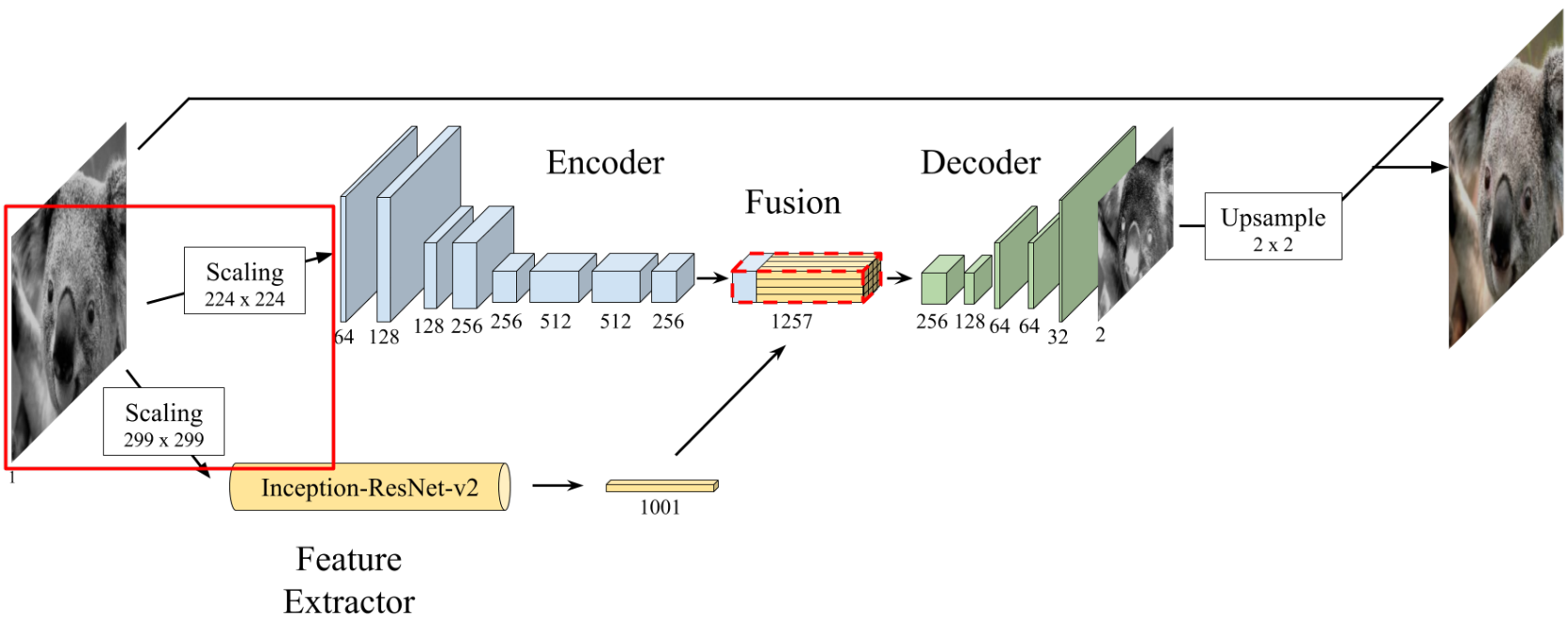
Architecture Overview

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



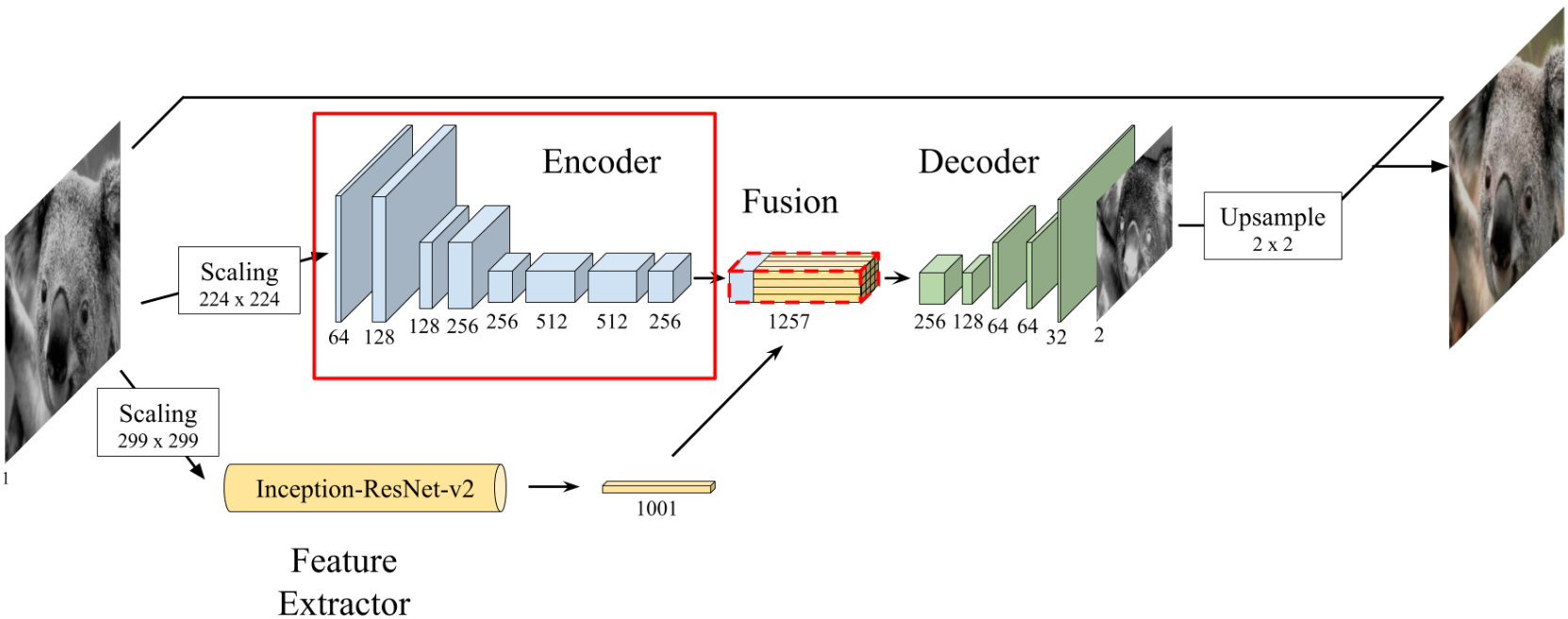
Preprocessing

- Pad and scale the images to allow for batch training:
  - 299x299 for Inception-ResNet-v2
  - 244x244 for our Model
- Center the pixel values in the range $[-1, 1]$



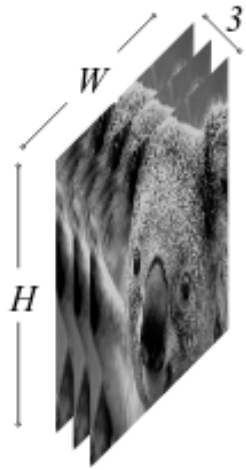
Encoder

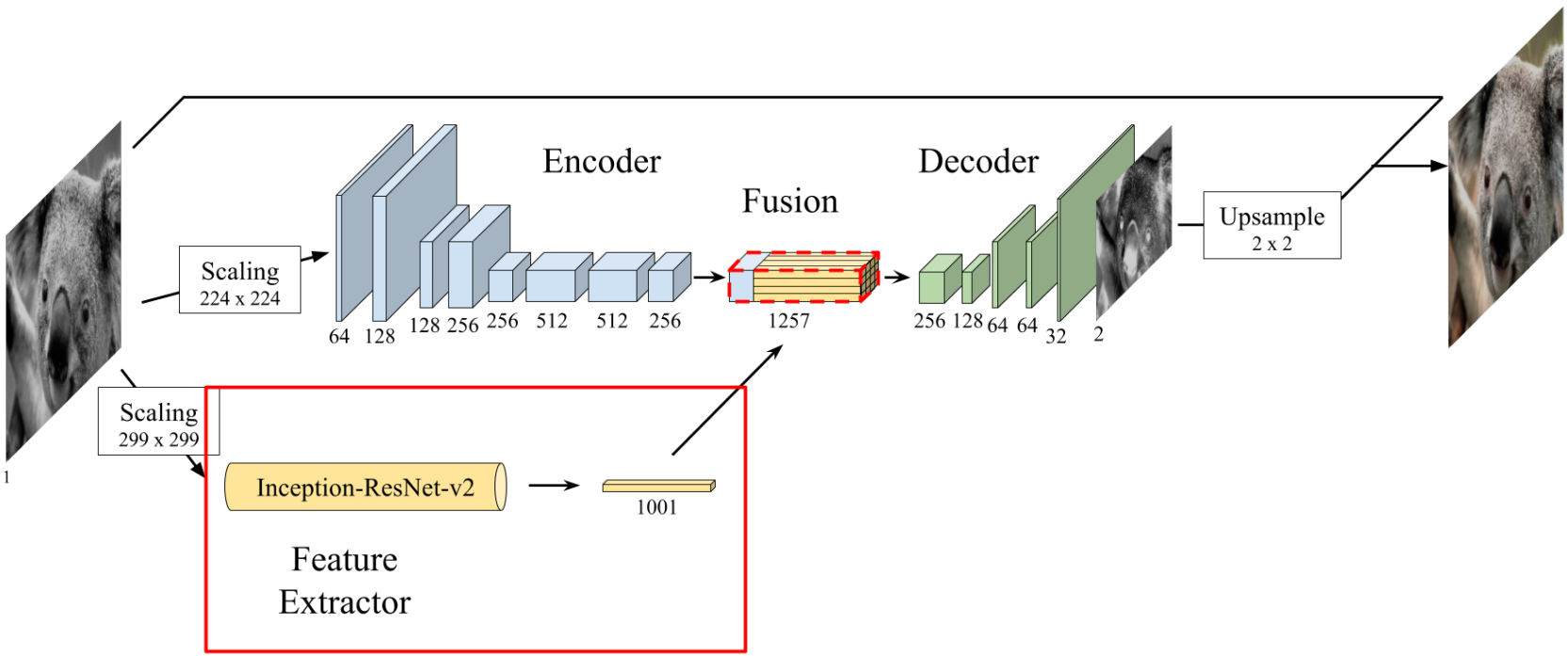
- Input: $H \times W \times 1$ (L component)
- Output: $H/8 \times W/8 \times 512$ feature representation
- 8 convolutional layers with 3×3 kernels that alternate:
  - Stride 1 and padding to preserve the input size
  - Stride 2 to halve the input size



Feature Extraction

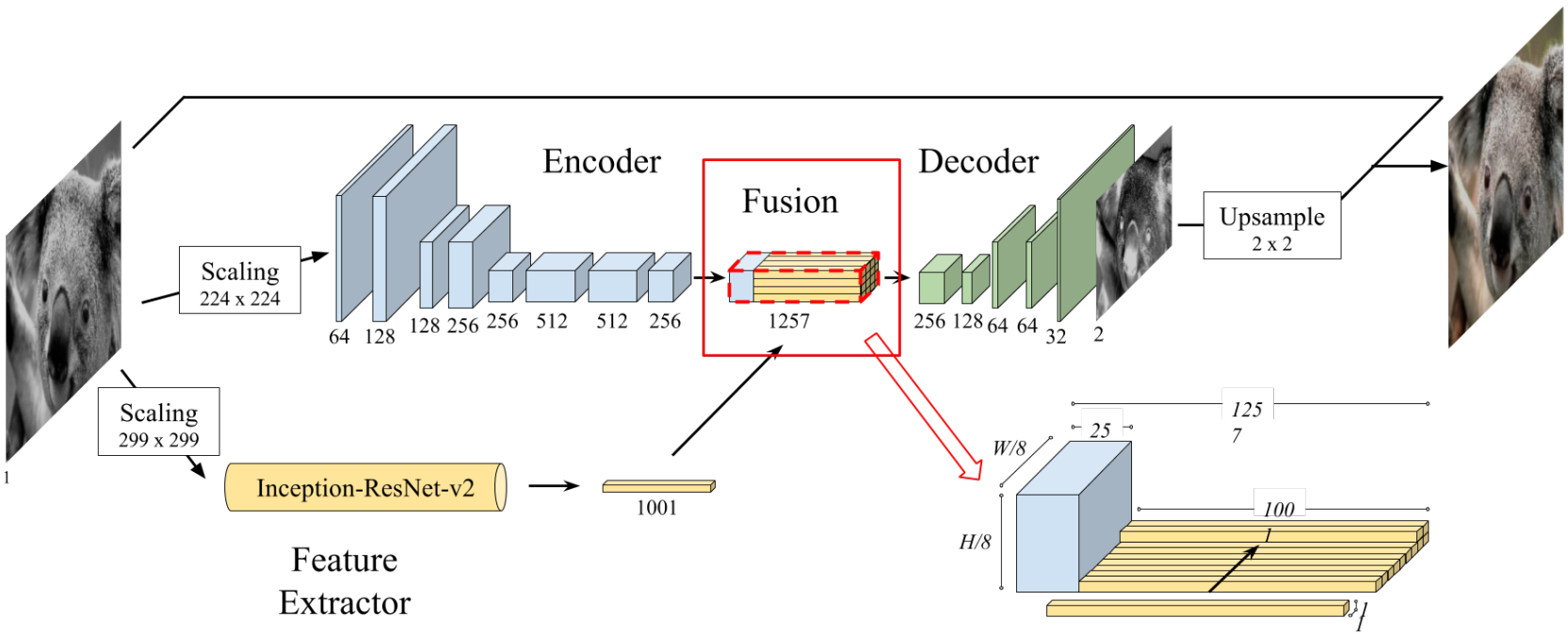
- Using a pre-trained Inception model
- Stack the luminance 3 times with itself (dimension requirements)
- Extract the features from the last layer before the softmax function





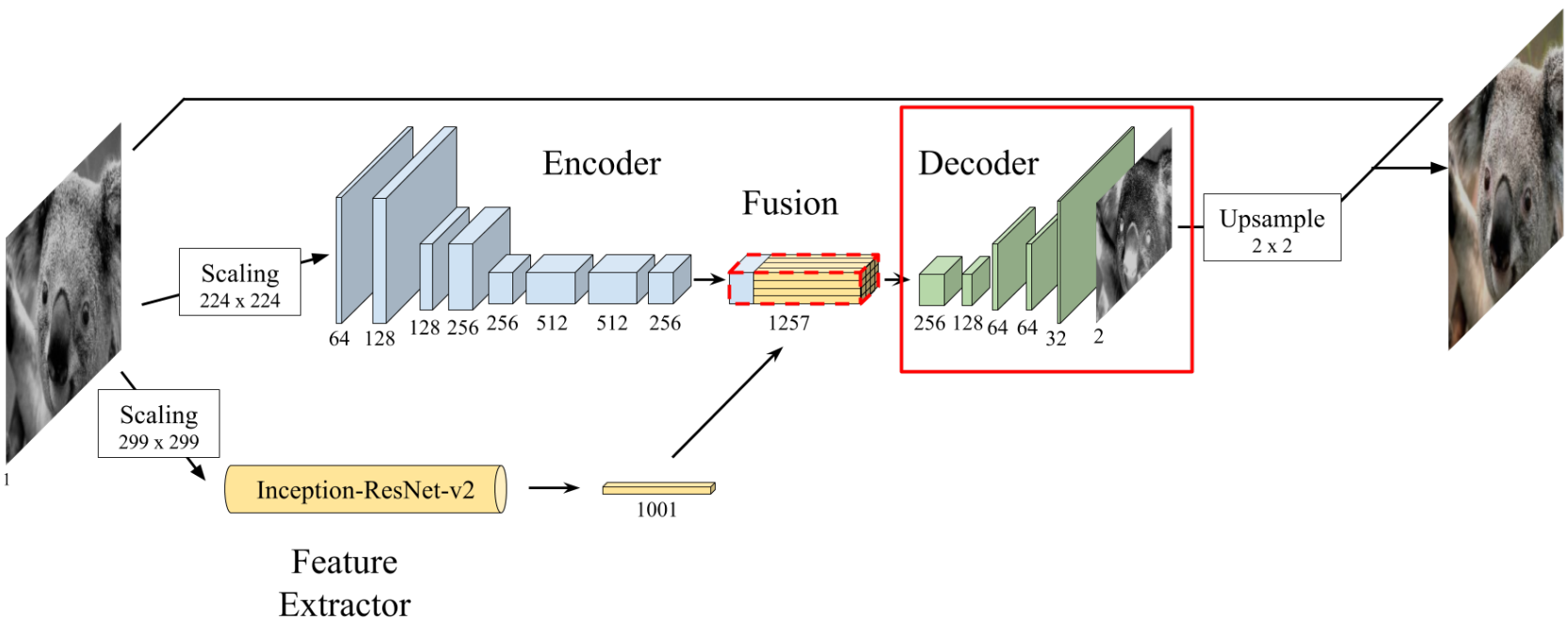
Fusion Layer

- Replicate the embedding $HW/8^2$ times
- Concatenate it to the encoder's feature volume along the depth axis
- Independent to the initial image size



Decoder

- Alternating
  - convolutional layers
  - up-sampling (basic nearest neighbor approach)
- Final output: $H \times W \times 2$ (a*b* components)



Training

- Mean Squared Error between predicted pixels colors and their real value:

$$C(\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{2HW} \sum_{k \in \{a,b\}} \sum_{i=1}^{H} \sum_{j=1}^{W} (X_{k_{i,j}} - \tilde{X}_{k_{i,j}})^2$$

- Stochastic Gradiend Descent with Adam Optimizer ($\eta = 0.001$)
- Fixed input image size to allow batch processing

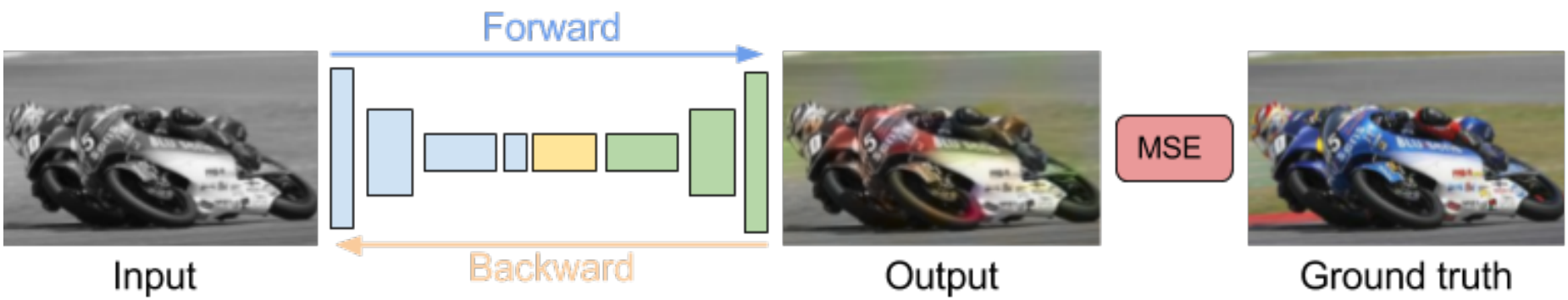
ImageNet Dataset

- More than 1.400.000 images
- Free access
- Only $\sim$ 60.000 images used for training

Training

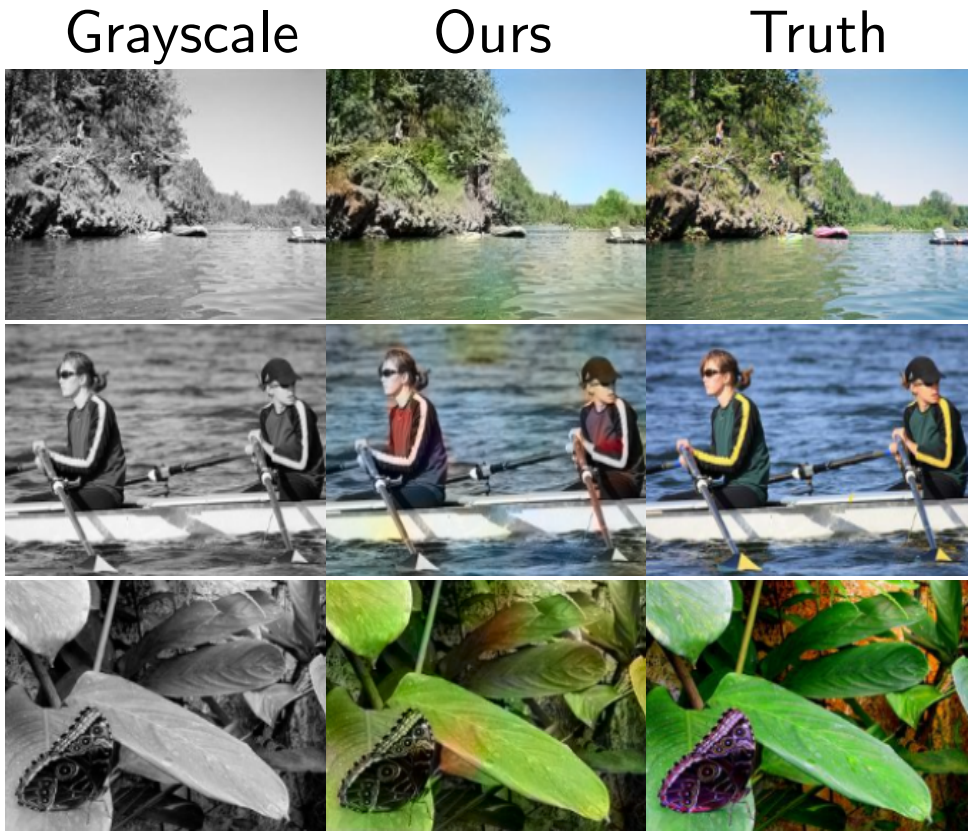
Training:

- 90% training - 10% validation/test
- Batch size of 100 images
- $\sim$23 hours training the network
- Keras with TensorFlow backend
- PDC-KTH: NVIDIA Tesla K80 GPU

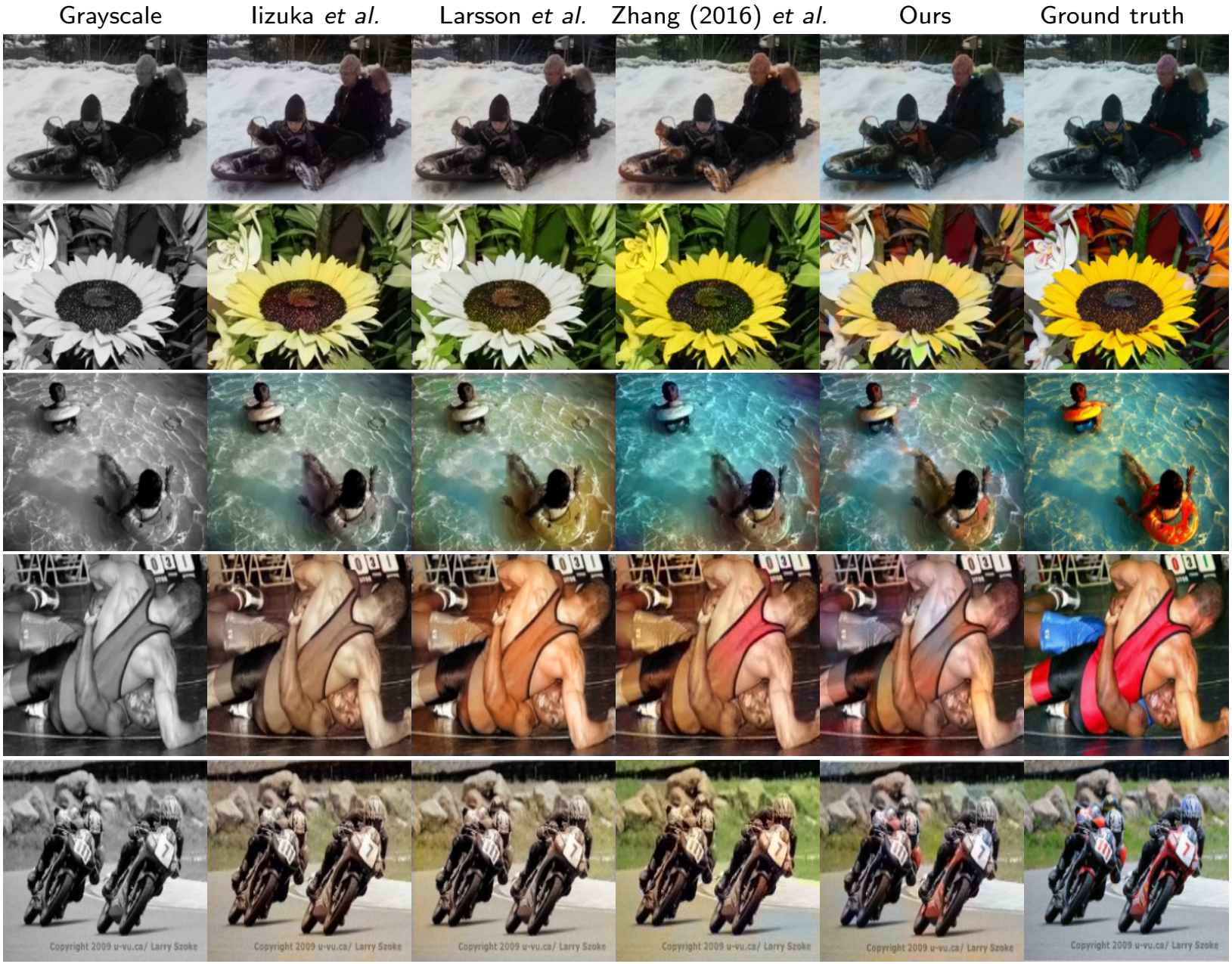


Results

- Near-photorealistic colorization for some images
- Alternative but realistic colored estimates
- Reduced dataset: the network performances depend on which global features are present in the test images



Comparison with others



User acceptance Study

- 12 images in the test
  - 3 original
  - 9 colorized samples chosen from our very best ones
- 41 users were asked the question:

Is this image Fake or Real?

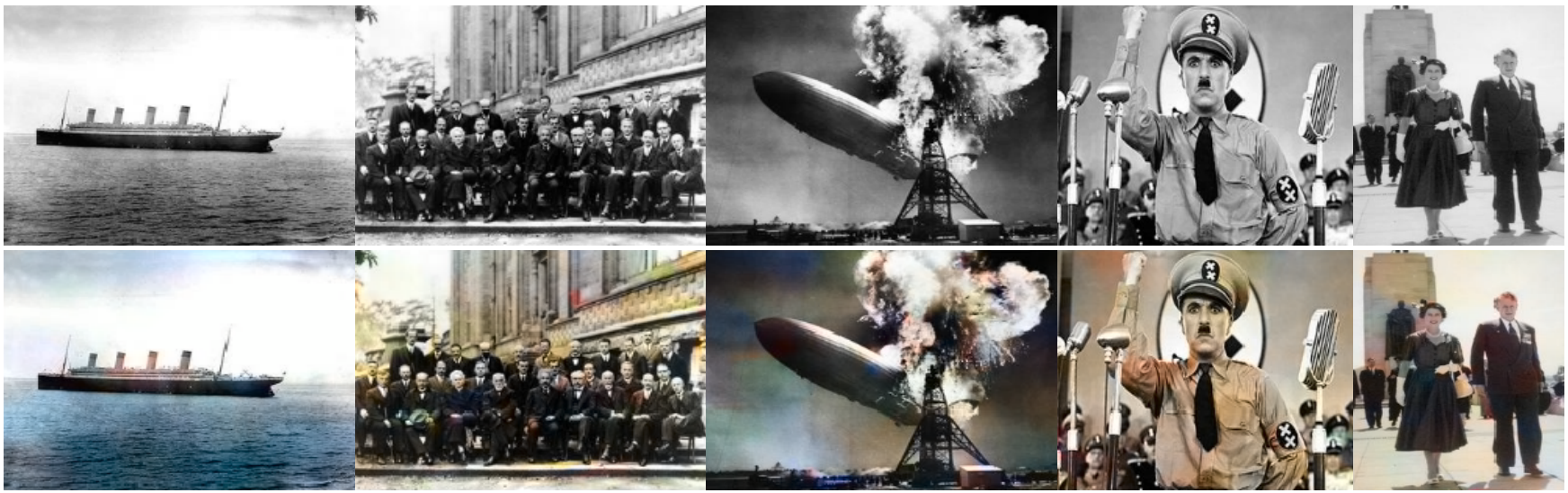
User acceptance Study

- 45.87% of the users miss-classified recolored images as originals
- Recolored images for the user study were selected from our best results.

79.0% 69.1% 64.1% 54.3% 42.4%

35.9% 35.1% 17.9% 15.0%

Historical pictures



Conclusion

Achievements:

- Successful colorization of high-level image components such as the sky, the sea or the forest
- Lower accuracy for smaller details on certain type of images

Future work:

- Training over the entire dataset as in the current trained network only a small portion of the spectrum of possible subjects is represented, therefore, the performance on unseen images highly depends on their specific contents
- Experiments on per pixel probabilistic approach as it has shown to perform better
- Video colorization by adding temporal coherence

Questions?