GAN-based Image Colourisation with Feature Reconstruction Loss

Laia Tarrés[1] Marc Gorriz[3] Xavier Giro-i-Nieto[1,2] Marta Mrak[3]
[1]*Universitat Politècnica de Catalunya*
[2]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC*
[3]*BBC Research & Development*
laia.tarres@upc.edu

Figure 1: Example of coloured images using our GAN model compared to the state-of-the-art approaches in *Pix2Pix* [5] and Górriz et al[3]

.

1. Introduction

Image colourisation is the task of adding plausible colour to grayscale images. This transformation requires obtaining a three dimensional colour-valued mapping from a real-valued grayscale image, which leads to an undetermined problem because the gray-scale semantics and texture provide cues for multiple possible colour mappings. The goal of image colourisation in not to recover the ground truth colour in a manner that it is perceived as natural by a human observer.

Our work takes as a baseline a scheme based on an end-to-end trainable convolutional neural network (CNN) trained with a smooth L1 loss to predict the ab channels of a colour image given the L channel. We introduce an extra perceptual reconstruction loss during training to improve the capabilities of a adversarial adversarial model, that we adopt as a baseline. Figure 1 presents some examples of the results achieved by our method.

2. Related Work

Image colourization networks are typically trained in a self-supervised set up in which colour images are converted to grayscale [14]. This allows quickly gathering data suitable for training deep neural networks. A first approach of image colourization with deep learning was proposed by Cheng et al. [13] by formulating a least square minimization problem solved with deep neural networks.

The capabilities of Generative Adversarial Networks (GANs) [2] for producing realistic samples was firstly applied for image colourization in *Pix2Pix* Isola et al. [5]. Some training improvements to his set up were proposed by multiple authors [9, 12, 3]. In particular, Gorriz et al. [3] increased the colour saturation obtained by an off-the-shelf *pix2pix* model by adding batch and instance normalization to the training, as well as multiple discriminators.

3. Methodology

In our work we add the feature (or perceptual) reconstruction loss and include it in the objective function used in *Pix2Pix* [5]. This loss was proposed by Johnson et al. [6] for image translation tasks, defined as the squared and normalized Euclidean distance between activations produced in the early layers of the network for the output image and the target image. Mahendran et al. [8] showed that using a feature reconstruction loss for training image transformation networks encourages the output image to be perceptually similar to the target image, but does not force them to match exactly.

In our implementation, rather than using only per-pixel loss functions depending only on low-level pixel information, we train our networks using added perceptual loss functions that depend on high-level features from a pretrained loss network. During training, perceptual losses measure image similarities more robustly than per-pixel losses. This way, when we feed an image to a pretrained network for image classification, the model has already learned the perceptual and semantic information that we

would like to measure. So comparing the network's activations from the ground truth and the generated image provides perceptual information.

The computation of the feature reconstruction loss corresponds to the squared and normalized Euclidean distance between the activations of a selected layer produced by the real image and generated image, when forwarded through the perceptual loss network.

For our experiments, we have tried different types of pretrained neural networks to extract the features to be compared. The first group of experiments used either a VGG16 network [11] or ResNet50 network [4] classification networks, were both were pretrained for image classification on the ImageNet dataset [1]. The second group of experiments is based on either U-Net [10] or FPN [7] segmentation networks, both composed by classification networks pretrained on ImageNet and COCO dataset, respectively.

4. Experimental Results

4.1. Implementation Details

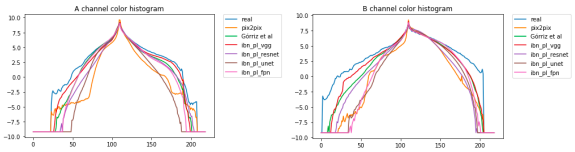
We define as baselines a U-Net [10] architecture as generator, and PatchGAN as the discriminator, the same approach as is *pix2pix* [5]. Training data are extracted from the ImageNet dataset. We select 50,000 RGB images that represent 50 images per class for training, and 10,000 test images selected as 10 images per each class. All classes are converted to CIE Lab colour space. The best results were obtained with the feature reconstruction loss on *block3 conv3* layer from the VGG16 network, with a loss weight of 0.00001. This configuration was trained for 23 epochs, during 36 hours.

4.2. Quantitative results

Evaluating the quality of a colourised image in a quantitative way is a challenging task, and still remains to be solved. Therefore, quantitative measures reflecting how close the outputs are to the ground truth data may not characterise the human perception of the problem. Nevertheless, we have used quantitative measures in order to quantitatively compare the results of the proposed methods to others in the literature.

The plots in Figure 2 represent the colour histogram of the real coloured images, *pix2pix* baseline [5], Górriz et al. [3] and our model with different backbones. As quantitative metrics, we have chosen the Kullback Liebler Divergence. Furthermore, as state-of-the-art methods on colourisation, we have also included peak signal to noise ratio (PSNR). They are represented in Table 1.

Our model with the VGG16 backbone has the most similar histogram to the real histogram, denoting more vivid colours in the image. PSNR is a measure that does not penalize desaturated results, so pix2pix performs better.



(a) A channel (b) B channel

Figure 2: Comparison of logarithmic colour histograms for both AB channels for CIE lab colourspace. The wider the histograms, the more colours they are representing, and more vivid the resulting images are.

Table 1: Quantitative metrics for the models

Models	Backbone	JS divergence		PSNR
		a	b	
pix2pix	-	0.13	0.13	**26.70**
Górriz et al.	-	**0.06**	0.06	25.14
Our model	VGG16	0.009	**0.05**	25.13
	ResNet	0.12	0.13	25.23
	Unet	0.23	0.19	25.19
	FPN	0.15	0.19	25.24

Table 2: Qualitative metrics for the models

Model	Naturalness
Ground Truth	0.87
pix2pix [5]	0.53
Górriz et al. [3]	0.26
Ours with VGG16	0.38

4.3. Qualitative results

A perceptual realism study was performed, similarly to the one presented in ChromaGAN [12]. Images were shown to non-expert participants, where some are ground-truth colourisation and others the results of a colourisation method. The colourisation methods included were: our method with VGG16 backbone, pix2pix [5] and Górriz [3] implementation. For each image shown, the participant indicates if the image has real or generated colours.

The qualitative study was run for 150 ground truth images and 150 images for each model. Each participant had 50 images to label, and the study was performed 35 times. Perceptual realism corresponds to the % of pictures noted as real from each model.

The results presented in Table 2 show how *pix2pix* achieves better perceptual realism, as participants tend to classify desaturated images as real. Our model produces more colourful results that are perceptually coherent, so it is suitable when aiming at equally vibrant results.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[3] M. Górriz, M. Mrak, A. Smeaton, and N. O'Connor. End-to-end conditional gan-based architectures for image colourisation. 2019. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2

[6] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, 2016. 1

[7] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 2

[8] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. *In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[9] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. pages 85–94, 2018. 1

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *LNCS*, 9351:234–241, 2015. 2

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014. 2

[12] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: An adversarial approach for picture colorization. In *WACV*, 2019. 1, 2

[13] Q. Yang Z. Cheng and B. Sheng. Deep colorization. *In Proceedings of the IEEE International Conference on Computer Vision*, page 415–423, 2015. 1

[14] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 1