

ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution

Patricia Vitoria, Lara Raad and Coloma Ballester
 Department of Information and Communication Technologies
 University Pompeu Fabra, Barcelona, Spain
 {patricia.vitoria, lara.raad, coloma.ballester}@upf.edu

Abstract

The colorization of grayscale images is an ill-posed problem, with multiple correct solutions. In this paper, we propose an adversarial learning colorization approach coupled with semantic information. A generative network is used to infer the chromaticity of a given grayscale image conditioned to semantic clues. This network is framed in an adversarial model that learns to colorize by incorporating perceptual and semantic understanding of color and class distributions. The model is trained via a fully self-supervised strategy. Qualitative and quantitative results show the capacity of the proposed method to colorize images in a realistic way achieving state-of-the-art results.

1. Introduction

Colorization is the process of adding plausible color information to monochrome photographs or videos (we refer to [37] for an interesting historical review). Currently, digital colorization of black and white visual data is a crucial task in areas so diverse as advertising and film industries, photography technologies or artist assistance. Although important progress has been achieved in this field, automatic image colorization still remains a challenge.

Colorization is a highly undetermined problem, requiring mapping a real-valued grayscale image to a three-dimensional color-valued one, that has not a unique solution. Before the emergence of deep learning techniques, the most effective methods relied on human intervention, usually through either user-provided color scribbles or a color reference image. Recently, convolutional neural network strategies have benefit from the huge amount of publicly available color images in order to automatically learn what colors naturally correspond to the real objects and its parts. In this paper, we propose a fully automatic end-to-end adversarial approach called ChromaGAN. It combines the strength of generative adversarial networks (GANs) with semantic class distribution learning. As a result, ChromaGAN is able to perceptually colorize a grayscale image from the



Figure 1. ChromaGAN is able to colorize a grayscale image from the semantic understanding of the captured scene.

semantic understanding of the captured scene. To illustrate this, Fig. 1 shows vibrant and diverse colorizations frequently achieved. Moreover, ChromaGAN shows variability by colorizing differently some objects belonging to the same category that may have several real colors, as for example, the birds in Fig. 1. The user-based perceptual ablation study show that the effect of the generative adversarial learning is key to obtain those vivid colorizations.

The contributions of this work include:

- An adversarial learning approach coupled with semantic information leading to a three term loss combining color and perceptual information with semantic class distribution.
- An unsupervised semantic class distribution learning.
- A perceptual study showing that semantic clues coupled to an adversarial approach yields high quality results.

The outline of this paper is as follows. Section 2 reviews the related work. In Section 3 the proposed model, architecture and algorithm are detailed. Section 4

presents qualitative and quantitative results. Finally, the paper is concluded in Section 5. The code is available at <https://github.com/pvitoria/ChromaGAN>.

2. Related Work

In the past two decades, several colorization techniques have been proposed. They can be classified in three classes: scribble-based, exemplar-based and deep learning-based methods. The first two classes depend on human intervention. The third one is based on learning leveraging the possibility of easily creating training data from any color image.

Scribble-based methods. The user provides local hints, as for instance color scribbles, which are then propagated to the whole image. These methods were initiated with the work of Levin *et al.* [23]. They assume that spatial neighboring pixels having similar intensities should have similar colors. They formalize this premise optimizing a quadratic cost function constrained to the values given by the scribbles. Several improvements were proposed in the literature. Huang *et al.* [16] improve the bleeding artifact using the grayscale image edge information. Yatziv *et al.* [37] propose a luminance-weighted chrominance blending to relax the dependency of the position of the scribbles. Then, Luan *et al.* [25] use the scribbles to segment the grayscale image and thus better propagate the colors. These methods suffer from requiring large amounts of user inputs in particular when dealing with complex textures. Moreover, choosing the correct color palette is not a trivial task.

Exemplar-based methods. They transfer the colors of a reference image to a grayscale one. Welsh *et al.* [36], propose to match the luminance values and texture information between images. This approach lacks of spatial coherency which yields unsatisfactory results. To overcome this, several improvements have been proposed. Ironi *et al.* [18] transfer some color values from a segmented source image which are then used as scribbles in [23]. Tai *et al.* [33] construct a probabilistic segmentation of both images to transfer color between any two regions having similar statistics. Charpiat *et al.* [7] deal with the multimodality of the colorization problem estimating for each pixel the conditional probability of colors. Chia *et al.* [9] use the semantic information of the grayscale image. Gupta *et al.* [14] transfer colors based on the features of the superpixel representation of both images. Bugeau *et al.* [4] colorize an image by solving a variational model which allows to select the best color candidate, from a previous selection of color values, while adding some regularization in the colorization. Although this type of methods reduce significantly the user inputs, they are still highly dependent on the reference image which must be similar to the grayscale image.

Deep learning methods. Recently, different approaches have been proposed to leverage the huge amount of grayscale/color image pairs. Cheng *et al.* [8] first proposed

a fully-automatic colorization method formulated as a least square minimization problem solved with deep neural networks. A semantic feature descriptor is proposed and given as an input to the network. In [11], a supervised learning method is proposed through a linear parametric model and a variational autoencoder which is computed by quadratic regression on a dataset of color images. These approaches are improved by the use of CNNs and large-scale datasets. For instance, Iizuka *et al.* [17] extract local and global features to predict the colorization. The network is trained jointly for classification and colorization in a labeled dataset. Zhang *et al.* [39] learn the color distribution of every pixel. The network is trained with a multinomial cross entropy loss with rebalanced rare classes allowing unusual colors to appear. Mouzon *et al.* [26] couple the resulting distribution in Zhang *et al.* [39] with the variational approach proposed in [28]. This allows to select for each pixel a color candidate from the pixel color distributions while regularizing the result. Also, it avoids the halo artifacts noticed in [39]. Larsson *et al.* [22] train a deep CNN to learn per-pixel color histograms. They use a VGG network to interpret the semantic composition of the scene and the localization of objects to then predict the color histograms of every pixel. The network is trained with the KL divergence.

Other CNN based approaches are combined with user interactions. For instance, Zhang *et al.* [40] propose to train a deep network given the grayscale version and a set of sparse user inputs. This allows the user to have more than one plausible solution. Also, He *et al.* [15] propose an exemplar-based colorization method using a deep learning approach. The colorization network jointly learns faithful local colorization to a meaningful reference and plausible color prediction when a reliable reference is unavailable. These hybrid methods yield results containing rare colors. Recently, Yoo *et al.* in [38] capture rare color instances without human interaction using a memory network together with a triplet loss without the need of labels.

Some methods use GANs [12] to colorize grayscale images. Their ability in learning probability distributions over high-dimensional spaces of data such as color images has found widespread use for many tasks (*e.g.*, [6, 19, 21, 24, 29, 35, 42]). Isola *et al.* [19] propose to use conditional GANs to map an input image to an output image using a U-Net based generator. They train their network by combining the L_1 -loss with an adapted GAN loss. An extension is proposed in [27] where the authors claim to generalize [19] to high resolution images, speed up and stabilize the training. Cao *et al.* [5] also use conditional GANs and obtain diverse possible colorizations by sampling several times the input noise. It is incorporated in multiple layers in their architecture, which consists of a fully convolutional non-stride network. Notice that none of these GANs based methods use additional information such as classification while our

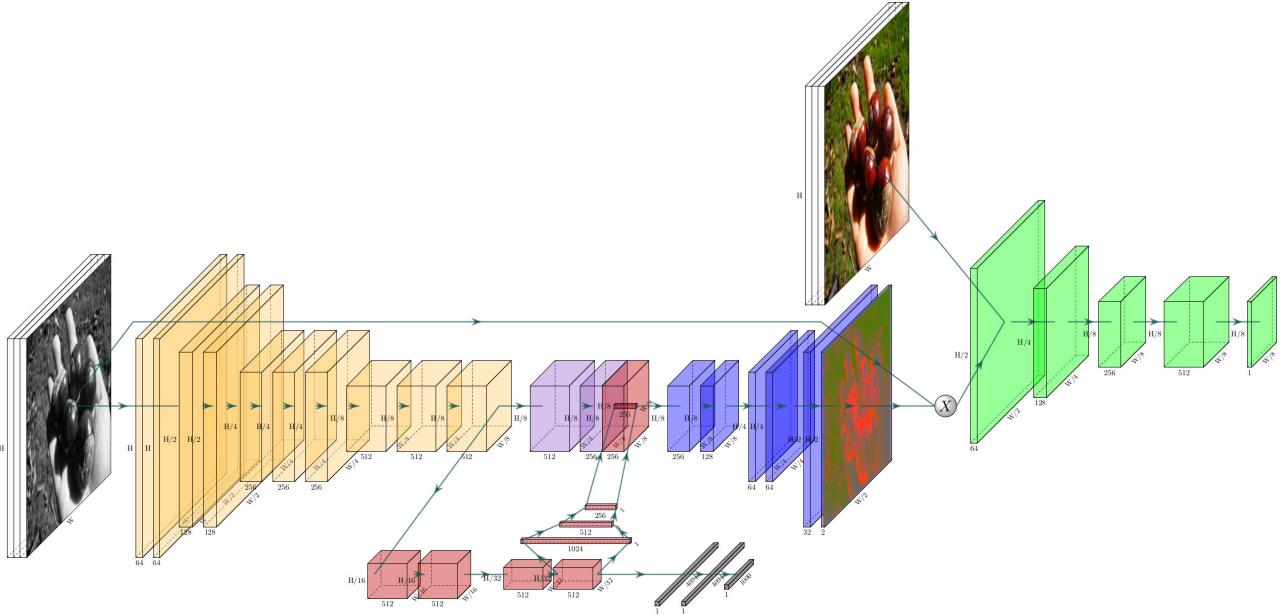


Figure 2. Overview of our model, ChromaGAN, able to automatically colorize grayscale images. It combines a Discriminator network, D_w (in green), and a Generator network, G_θ . G_θ consists of two subnetworks: $\mathcal{G}_{\theta_1}^1$ (yellow, purple, red and blue layers) that outputs the chrominance information $(a, b) = \mathcal{G}_{\theta_1}^1(L)$, and $\mathcal{G}_{\theta_2}^2$ (yellow, red and gray layers) which outputs the class distribution vector, $y = \mathcal{G}_{\theta_2}^2(L)$.

KL and the Jensen-Shannon divergences which produce the vanishing gradients problem. Also, the WGAN alleviates the mode collapse problem which is interesting when aiming to be able to capture multiple possible colorizations.

The proposed \mathcal{L}_g loss favours perceptually real results. As the experiments in Section 4 show and has been also noticed by some authors [19], the adversarial GAN model produces sharp and colorful images favouring the emergence of a perceptually real palette of colors instead of ochreish outputs produced by colorization using only terms such as the L_2 or L_1 color error loss. It is in agreement to the analysis in [3]. The authors define the perceptual quality of an image as the degree to which it looks like a natural image and argue that it can be defined as the best possible probability of success in real-vs-fake user studies. They show that perceptual quality is proportional to the distance between the distribution of the generated images and the distribution of natural images, which, at the optimum generator and discriminator, corresponds to the \mathcal{L}_g value.

Color Error Loss. In some colorization methods [22, 39] the authors propose to learn a per-pixel color probability distribution allowing them to use different classification losses. Instead, we chose to learn two chrominance values (a, b) per-pixel using the L_2 norm. As mentioned, only using this type of loss yields ochreish outputs. However, in our case the use of the perceptual GAN-based loss relaxes this effect making it sufficient to obtain notable results.

Class Distribution Loss. The KL-based loss $\mathcal{L}_s(\mathcal{G}_{\theta_2}^2)$ (3) compares the generated density distribution vector $y = \mathcal{G}_{\theta_2}^2(L)$ to the ground truth distribution $y_v \in \mathbb{R}^m$. The latter is computed using the VGG-16 [32] pre-trained on ImageNet dataset [10]. The VGG-16 model was trained on color images; in order to use it without any further training, we re-shape the grayscale image as (L, L, L) . The class distribution loss adds semantic interpretation of the scene. The effect of this term is analyzed in Section 4.

3.2. Detailed Model Architecture

The proposed GAN architecture is conditioned by the grayscale image L through the proposed loss (1). It contains three distinct parts. The first two, belonging to the generator, focus on geometrically and semantically generating a color image information (a, b) and classifying its semantic content. The third one belongs to the discriminator network learning to distinguish between real and fake data.

An overview of the model is shown in Fig. 2. In the remaining of the section we will describe the full architecture. More details are available in the supplementary material.

Generator Architecture. The generator G_θ is made of two subnetworks (denoted by $\mathcal{G}_{\theta_1}^1$ and $\mathcal{G}_{\theta_2}^2$) divided in three stages with some shared modules between them. Both of them will take as input a grayscale image of size $H \times W$. The subnetwork $\mathcal{G}_{\theta_1}^1$ outputs the chrominance information, $(a, b) = \mathcal{G}_{\theta_1}^1(L)$, and the subnetwork $\mathcal{G}_{\theta_2}^2$ outputs the com-

puted class distribution vector, $y = \mathcal{G}_{\theta_2}^2(L)$. Both subnetworks are jointly trained with a single step backpropagation.

The first stage (displayed in yellow in Fig. 2) is shared between both subnetworks. It has the same structure as the VGG-16 without the three last fully-connected layers at the top of the network. We initialize them with pre-trained VGG-16 weights which are not frozen during training.

From this first stage on, both subnetworks, $\mathcal{G}_{\theta_1}^1$ and $\mathcal{G}_{\theta_2}^1$, split into two tracks. The first one (in purple in Fig. 2) process the data by using two modules of the form Conv-BatchNorm-ReLu. The second track (in red), first processes the data by using four modules of the form Conv-BatchNorm-ReLu, followed by three fully connected layers (in red). This second path (in gray) outputs $\mathcal{G}_{\theta_2}^2$ providing the class distribution vector. To generate the probability distribution y of the m semantic classes, we apply a softmax function. Notice that $\mathcal{G}_{\theta_2}^2(L)$ is a classification network and is initialized with pre-trained weights. But, as part of this subnetwork is shared with $\mathcal{G}_{\theta_1}^1$, once the network is trained, it has learnt to give a class distribution close to the output of the VGG-16 and to generate useful information to help the colorization process. This could be understood as fine tuning the network to learn to perform two tasks at once.

In the third stage both branches are fused (in red and purple in Fig. 2) by concatenating the output features. Later on, (a, b) will be predicted by processing the information through six modules of the form Convolution-ReLu with two up-sampling layers in between. Note that while performing back propagation with respect to the class distribution loss, only the second subnetwork $\mathcal{G}_{\theta_2}^2$ will be affected. For the color error loss, the entire network will be affected.

Discriminator Architecture. The discriminator network D_w is based on the Markovian discriminator architecture (PatchGAN [19]). The PatchGAN discriminator keeps track of the high-frequency structures of the generated image compensating the fact that the L_2 loss $\mathcal{L}_e(\mathcal{G}_{\theta_1}^1)$ fails in capturing high-frequency structures but succeeds in capturing low-level ones. In order to model the high-frequencies, the PatchGAN discriminator focuses on local patches. Thus, rather than giving a single output for the full image, it classifies each patch as real or fake. We follow the architecture defined in [19] where the input and output are of size $H \times W$ and $H/8 \times W/8$, respectively.

4. Experimental Results and Discussion

In this section the proposed method is evaluated both quantitatively and qualitatively. Notice that evaluating the quality of a colorized image in a quantitative way is a challenging task. For some objects several colors could perfectly fit. Therefore, quantitative measures reflecting how close the outputs are to the ground truth data are not the best measures for this type of problem. For that reason, qualitative comparisons are provided as well as a user-based per-

ceptual study quantifying the realism of colorized images by the proposed method. Nevertheless, in order to quantitatively compare the results of the proposed methods to others in the literature a quantitative measure will also be used.

To assess the effect of each term of our loss function in the entire network, we perform an ablation study by evaluating three variants of our method: *ChromaGAN*, using the adversarial and classification approach, *ChromaGAN w/o Class* avoiding the classification approach ($\lambda_s = 0$) and *ChromaNet* avoiding the adversarial approach ($\lambda_p = 0$).

4.1. Dataset

We train each variant of the network end-to-end on 1.3M images from the subset of images [30] taken from ImageNet [10]. It contains objects from 1000 different categories and of different color conditions, including grayscale images. Due to the presence of fully connected layers in our network, the input size to the class distribution branch has to be fixed. We chose to work with input images of 224×224 pixels as is done when training the VGG-16 [32] on ImageNet. Thus, we have resized each image in the training set and converted it to a three channels grayscale image by triplicating the luminance channel L .

4.2. Implementation Details

We train the network for a total of five epochs and set the batch size to 10, on the 1.3M images from the ImageNet training dataset resized to 224×224 . A single epoch takes approximately 23 hours on a NVIDIA Quadro P6000 GPU. The prediction of the colorization of a single image takes an average of 4.4 milliseconds. We minimize our objective loss using Adam optimizer with learning rate equal to $2e-5$ and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We alternate the optimization of the generator \mathcal{G}_{θ} and discriminator D_w . The first stage of the network (displayed in yellow in Fig. 2), takes as input a grayscale image of size 224×224 , and is initialized using the pre-trained weights of the VGG-16 [32] trained on ImageNet.

4.3. Quantitative Evaluation

A perceptual realism study is performed to show the strength of coupling the adversarial approach with semantic information. Furthermore, as state-of-the-art methods on colorization, a quantitative assessment of our method is included in terms of *peak signal to noise ratio* (PSNR).

The following perceptual realism study was performed. Images are shown one by one to non-expert participants, where some are natural color images and others the result of a colorization method such as ChromaGAN, ChromaGAN w/o classification, ChromaNet and the four state-of-the-art methods [17, 22, 39, 19]. For each image shown, the participant indicates if the colorization is realistic or not.

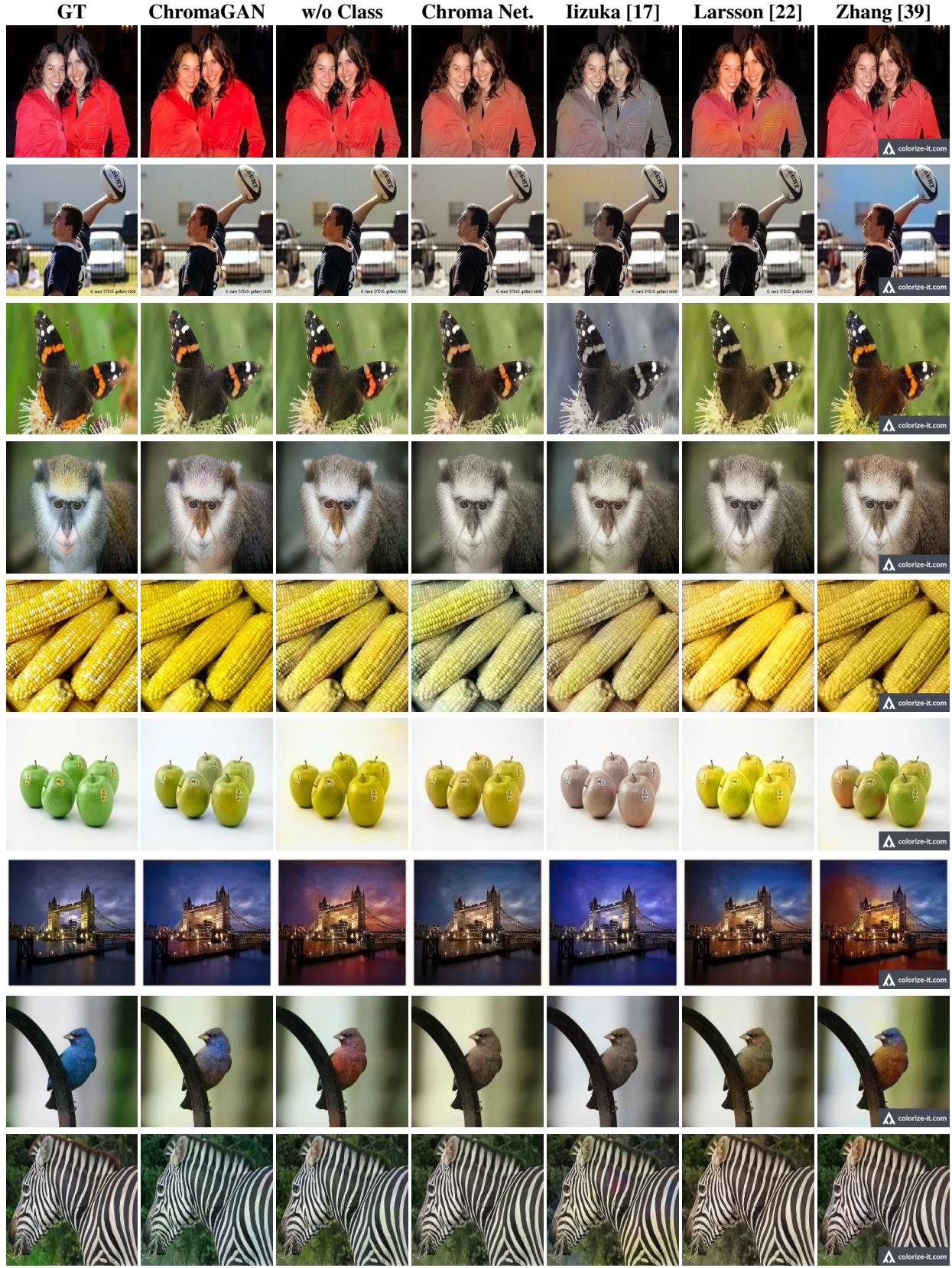


Figure 3. Some qualitative results using, from right to left: Gray scale, ChromaGAN, ChromaGAN w/o Classification, ChromaNet, Iizuka *et al.* [17], Larsson *et al.* [22] and Zhang *et al.* [39]

Method	Naturalness
Real images	87.1%
ChromaGAN	76.9%
ChromaGAN w/o Class	70.9%
ChromaNet	61.4%
Iizuka <i>et al.</i> [17]	53.9%
Larsson <i>et al.</i> [22]	53.6%
Zhang <i>et al.</i> [39]	52.2%
Isola <i>et al.</i> [19]	27.6%

Table 1. Numerical detail of the perceptual test. The values shows the mean naturalness over all the experiments of each method.

Method	PSNR (dB)
ChromaGAN	24.98
ChromaGAN w/o Class	25.04
ChromaNet	25.57
Iizuka <i>et al.</i> [17]	23.69
Larsson <i>et al.</i> [22]	24.93
Zhang <i>et al.</i> [39]	22.04
Isola <i>et al.</i> [19]	21.57

Table 2. Comparison of the average PSNR values for automatic methods, some extracted from the table in [40]. The experiment is performed on 1000 images of the ILSVRC2012 challenge set [31].

Fifty images are taken randomly from a set of 1600 images composed of 200 ground truth images (from ImageNet and Places datasets [41]) and 200 results from each method (ChromaGAN, ChromaNet, ChromaGAN w/o Class and [17, 22, 39, 19]). The study was performed 62 times. Table 1 shows the results of perceptual realism for each method. One can observe that ChromaGAN who couples the adversarial approach with semantic information yields perceptually more realistic results. Moreover, by comparing the results of ChromaNet, ChromaGAN w/o Class and [17, 22, 39, 19], one can conclude that the adversarial approach plays a more important role than the semantic information for the generation of natural color images.

Additionally, the PSNR of the obtained (a, b) images is computed with respect to the ground truth and compared to those obtained for other fully automatic methods as shown in Table 2. The table shows the average of this measure over all the test images. One can observe that, in general, our PSNR values are higher than those obtained by [17, 22, 39]. Moreover, comparing the PSNR of our three variants the highest one is achieved by ChromaNet. This is not surprising since the training loss of this method gives more importance to the quadratic color error term compared to the losses of ChromaGAN and ChromaGAN w/o Class.

4.4. Qualitative Evaluation

Our results are compared to those obtained in [17, 22, 39] by using the publicly available online demos. The methods are trained on ImageNet dataset in the case of [39, 22] and on Places dataset in the case of [17]. Notice that while ImageNet contains 1.3M training images, Places contains 2.4M. Several colorization results are shown on the validation set of ImageNet dataset in Fig. 3 and on Places in Fig. 5. As can be observed, the method in [17] and ChromaNet tend to output muted colours in comparison to the lively colors obtained with ChromaGAN, ChromaGAN w/o class and [22, 39]. Also, ChromaGAN is able to reconstruct color information by adding natural and vivid colors in almost all the examples (specially the first, fifth, sixth, eighth and ninth rows). Notice that the deep CNN in [39] is trained with a multinomial cross entropy loss with rebalanced rare classes allowing in particular for unusual colors to appear in the colorized image. Desaturated results are mainly obtained by [17] and with our method without using the adversarial approach (specially in the first, second, third, fourth, fifth and eighth rows), in some cases also by [22] (second, fourth and eighth rows). Also, color boundaries are not clearly separated generally in the case of [17] and sometimes by GAN w/o Class (sixth row) and [22] (third, fourth and eighth rows). Inconsistent chromaticities can be found in the second and sixth row by [39] where the wall is blue and the apples green and red at the same time. Third and seventh rows display some failure cases of our method: the bottom-right butterfly wing is colored in green. In fact, the case of the seventh row shows a difficult case for all the methods. For the sake of comparison, some results on Places dataset are shown in Fig. 5 for ChromaGAN trained on ImageNet, together with the results of [17] trained on Places dataset. In Fig. 6 the results of ChromaGAN are compared to those of the adversarial approach in [19]. The results in the second column of Fig. 6 were taken directly from [19]. Overall, their results provide sometimes more vivid colors as in the second row and sometimes uncolored results as in the first row. Furthermore, the results of ChromaGAN yield better results in terms of consistent hue. In the second column unnatural color stains are observed as for instance the green marks under the bird (second row) and the yellow and red stains on the coffee pot (third row).

Legacy Black and White Photographs. ChromaGAN is trained using color images where the chrominance information is removed. Due to the progress in the field of photography, there is a great difference in quality between old black and white images and modern color images. Thus, generating color information in original black and white images is a challenging task. Fig. 4 shows some results. Additional examples can be found in the supplementary material, where we also include results applied on paintings.



Figure 4. Colorization results of historical black and white photographs using the proposed ChromaGAN. Note that old black and white photographs are statistically different than actual ones, thus, making the process of colorize more difficult.

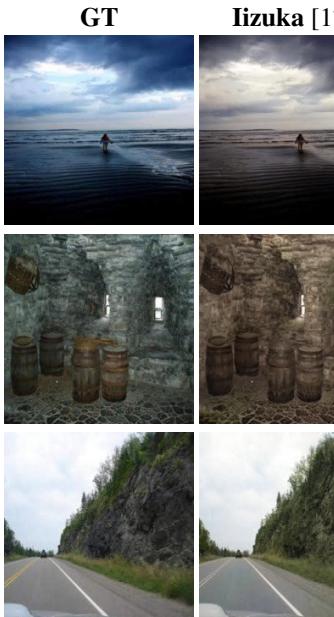


Figure 5. Results on images from the validation set of the Places Dataset. Left: Ground truth, middle: Iizuka *et al.* [17], right: ChromaGAN. Notice that the model of [17] is trained on Places Dataset having the double of training images. Our model is trained on the ImageNet dataset. The results are comparable.

5. Conclusions

A novel colorization method is presented. The proposed ChromaGAN model is based on an adversarial strategy that captures geometric, perceptual and semantic information. A variant of ChromaGAN which avoids using the distribution of semantic classes also shows satisfying results. Both cases prove that our adversarial technique provides photo-realistic colorful images. The quantitative comparison to state-of-the-art methods show that our method outperforms

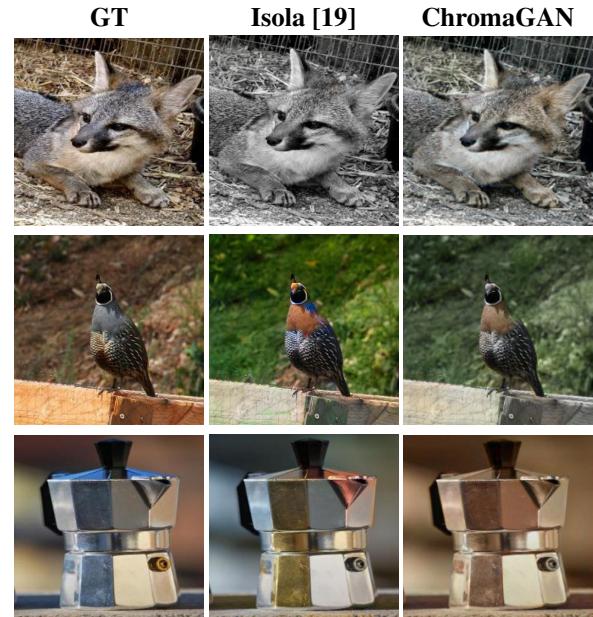


Figure 6. Results on GAN-based approaches. From left to right: Ground Truth, results from Isola [19] and results from ChromaGAN. The results from Isola *et al.* were extracted from [19].

them in terms of PSNR and perceptual quality while qualitative comparisons show their visual high quality.

Acknowledgements

This work was partially funded by MICINN/FEDER UE project, reference PGC2018-098625-B-I00 VAGS, and by H2020-MSCA-RISE-2017 project, reference 777826 No-MADS. We thank the support of NVIDIA Corporation for the donation of GPUs used in this work, Rafael Grompone von Gioi for valuable suggestions and José Lezama for his help with the user study.

- [33] Y.-W. Tai, J. Jia, and C.-K. Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 747–754. IEEE, 2005.
- [34] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
- [35] P. Vitoria., J. Sintes., and C. Ballester. Semantic image inpainting through improved wasserstein generative adversarial networks. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP*, pages 249–260. INSTICC, SciTePress, 2019.
- [36] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 277–280. ACM, 2002.
- [37] L. Yatziv and G. Sapiro. Fast image and video colorization using chrominance blending. *IEEE transactions on image processing*, 15(5):1120–1129, 2006.
- [38] S. Yoo, H. Bahng, S. Chung, J. Lee, J. Chang, and J. Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11283–11292, 2019.
- [39] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.
- [40] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.
- [41] 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*, 40(6):1452–1464, 2018.
- [42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.