Deep Learning for Automatic Colorization of Legacy Grayscale Aerial Photographs

Quentin Poterek, Pierre-Alexis Herrault, Grzegorz Skupinski, and David Sheeren

***Abstract*—Legacy grayscale aerial photographs represent one of the main available sources for studying the past state of the environment and its relationship to the present. However, these photographs lack spectral information thereby hindering their use in current remote sensing approaches that rely on spectral data for characterizing surfaces. This article proposes a conditional generative adversarial network, a deep learning model, to enrich legacy photographs by predicting color channels for an input grayscale image. The technique was used to colorize two orthophotographs (taken in 1956 and 1978) covering the entire Eurométropole de Strasbourg. To assess the model's performances, two strategies were proposed: first, colorized photographs were evaluated with metrics such as peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM); second, random forest classifications were performed to extract land cover classes from grayscale and colorized photographs, respectively. The results revealed strong performances, with PSNR $= 25.56 \pm 2.20$ and SSIM $= 0.93 \pm 0.06$ indicating that the model successfully learned the mapping between grayscale and color photographs over a large territory. Moreover, land cover classifications performed on colorized data showed significant improvements over grayscale photographs, respectively, +6% and +17% for 1956 and 1978. Finally, the plausibility of outputs images was evaluated visually. We conclude that deep learning models are powerful tools for improving radiometric properties of old aerial grayscale photographs and land cover mapping. We also argue that the proposed approach could serve as a basis for further developments aiming to promote the use of aerial photographs archives for landscapes reconstruction.**

***Index Terms*—Colorization, deep learning, generative adversarial network (GAN), grayscale imagery, historical aerial photograph, remote sensing.**

I. Introduction

HISTORICAL aerial photographs provide crucial data for efficient long-term environmental monitoring and change detection [1]–[4]. They offer unique background information at a very high spatial resolution [2], [5]. However, reproducible works based on these data can be challenging due to their inherent and heterogeneous properties, such as their spatial and radiometric resolution [6], [7]. Until the late 1930s, aerial photography was solely monochromatic, with sensors limited to the visible and near-infrared ranges [6], [7]. Thereafter, color photography rapidly superseded the use of grayscale shots [6]. Despite the current prominence of color aerial photographs, there remain many aerial photograph libraries that consist mainly of grayscale products. For instance, the French National Mapping Agency (IGN) distributes aerial photographs collected since 1919. Of 20 345 flights, 18 054 were completed using panchromatic sensors. The first IGN color photographs were only made available in 1959 [8].

Although grayscale photographs contain valuable historical information, they remain poorly exploited due to their heterogeneous specifications and quality, e.g., scale, lens properties, spectral sensitivity, and film development [6]. However, these products can still serve diverse purposes, including cartography, landscape dynamics analysis, and photogrammetry [6], [9]. Such applications usually rely on digital processing, such as feature extraction techniques, to supplement the available information and analysis potential provided by grayscale imagery and photography. The contribution of color, texture, and lightness features for land monitoring using remote sensing is well known, for example, Heller *et al.* [10] highlight the importance of color for identifying tree species from aerial photographs by showing that identification accuracy improved by up to 27% when foresters used photographs obtained with RGB films compared to panchromatic stills. The integration of infrared film further enhances interpretation from aerial photographs and highlights the importance of providing the widest possible range of spectral information for analyzing spatial data. Palsson *et al.* [11] and Cavallaro *et al.* [12] pointed out that without ancillary data, RGB channels alone improved classification results in urban areas by more than 40% in some cases when compared to the use of panchromatic stills. Attempts to provide the models with other discriminating features, such as texture, do not necessarily improve results. Feng *et al.* [13] compared the performances of two random forest classifiers trained on UAV imagery for urban vegetation mapping. The first was trained using RGB data only, whereas the second benefited from Haralick textures, computed from panchromatic pictures and provided as ancillary data. The authors reported an improvement in overall accuracy due to additional texture features, with a 23% increase for their first test site, and a 13% increase for the second one. Hence, insufficient spectral information in legacy aerial photographs would prove to be a disincentive to their use in the current remote

Manuscript received November 19, 2019; revised January 28, 2020 and April 3, 2020; accepted April 20, 2020. Date of publication May 28, 2020; date of current version June 16, 2020. This work was supported by the *Zone Atelier Environnementale Urbaine* that is part of the French Long Term Ecosystem Research Network. *(Corresponding author: Quentin Poterek.)*

Quentin Poterek, Pierre-Alexis Herrault, and Grzegorz Skupinski are with the Laboratoire Image Ville Environnement, CNRS UMR 7362, University of Strasbourg 67083, Strasbourg, France (e-mail: quentin.poterek@live-cnrs.unistra.fr; pierre-alexis.herrault@live-cnrs.unistra.fr; grzegorz.skupinski@live-cnrs.unistra.fr).

David Sheeren is with the DYNAFOR, INRA UMR 1201, University of Toulouse, 31320 Castanet-Tolosan, France (e-mail: david.sheeren@ensat.fr).

Digital Object Identifier 10.1109/JSTARS.2020.2992082

sensing landscape, where fully automated workflows prevail and profit from rich datasets. Therefore, standard remote sensing techniques may benefit from colorized legacy aerial photographs, which would help improve their spectral resolution.

Photography and movie industries, facing similar issues, were the first to develop colorization techniques, which allow superimposing color to grayscale stills [14]. While the process highly benefited from computer science, it is still mainly restricted to conventional photography, rather than remote sensing products, despite being of fundamental heritage and scientific interests.

A. Traditional Colorization Techniques

In the context of this article, techniques are considered traditional if they are unrelated to deep learning. Hand coloring was not investigated, as the process is time-consuming, and most legacy aerial photographs are digitally distributed.

Until the mid-2010s, various approaches tackled computer-based colorization mainly in a semiautomatic manner. These methods were either based on color transfer or color scribbles. On the one hand, transfer-based techniques—first proposed by Reinhard *et al.* [15] and developed further by various authors—require a human operator to select a source color image. Chromatic information is then transferred to a grayscale picture by matching various attributes, such as lightness and texture [16]–[21]. On the other hand, scribble-based techniques rely on the placement of color scribbles on a grayscale image. Chromatic information is then passed to nearby and similar pixels, using various optimization algorithms and attributes [22]–[25]. In both cases, input from an external operator is not a concern in the case of a single simple task. However, these techniques become inappropriate when dealing with vast spatial data, such as city-scale orthoimagery, due to the quantity and spatial extent of the images. The continued dependence on human operators encouraged the use of traditional learning-based techniques in the early-2010s [26], [27]. Despite tailored solutions, these methods require upstream consideration. A set of features must be selected, computed from a panchromatic picture and fed into a model during training. In turn, this produces a mapping from grayscale to color. Therefore, expertise is still required and remains a crucial step in this process.

B. Deep Learning Based Colorization Techniques

The advent of deep learning has enabled new opportunities for automated colorization. The operating process of the corresponding algorithms allows automatic learning of an entire set of high-level features before producing the desired output [28]. Thus, it helps reducing the emphasis on data engineering and preparation. Architectures developed in artificial vision are diverse and increasingly applied to remote sensing applications. Included among these applications are photogrammetry [29], classification and segmentation of images [30], [31], and super-resolution [32].

Recent years have witnessed an increase in automated colorization techniques that are deployed using various implementations and deep learning backends (see Table I).

TABLE I
OVERVIEW OF DEEP LEARNING TECHNIQUES DEVELOPED FOR AUTOMATING GRAYSCALE IMAGE COLORIZATION

Author(s)	Year	Model(s)	Modeling	Data
[33]	2015	DNN	Regression	Sift Flow
[34]	2016	CNN	Classification, regression	CIFAR-10
[35]	2016	VAE	Generation	Wild FLW, LSUN-Church, ILSVRC
[36]	2016	CNN	Classification, regression	Places
[37]	2016	GAN	Generation	Cityscapes, CMP Facades, Google Maps, etc.
[38]	2016	CNN	Classification	ImageNet, SUN-A, SUN-B
[39]	2016	CNN	Regression	Custom
[40]	2016	CNN	Classification	ILSVRC, SUN
[41]	2016	CNN	Classification	ImageNet
[42]	2017	GAN, WGAN	Generation	LSUN
[43]	2017	GAN	Generation	Safebooru
[44]	2017	CNN, GAN, WGAN	Regression, generation	Safebooru
[45]	2017	WGAN	Generation	ImageNet
[46]	2017	VAE	Generation	CIFAR-10, ILSVRC
[47]	2017	CNN	Classification	Custom
[48]	2017	GAN	Generation	Custom
[49]	2017	CNN	Classification	SUN
[50]	2018	CNN	Classification	AID, RSSCN7

The first-ever category of models being used for colorizing grayscale images corresponds to simple deep neural networks (DNNs), organized around fully connected layers of artificial neurons. Cheng *et al.* [33] deployed a DNN, trained with color images, their grayscale counterparts, and a set of features computed at different scales to learn the intended mapping. DNNs were not pursued further on account of their lack of flexibility and the computational resources required for working on raster datasets.

Convolutional neural networks (CNNs) were proposed as an alternative for image processing. Convolutions allow retrieval of information from an image, using a sliding window defined by a kernel size and weights associated with each of its cells. More than half of the currently proposed colorization techniques use CNNs for 1) classification [34], [36], [38], [40], [41], [47], [49], [50] and 2) regression [34], [36], [39], [44]. However, regression loss functions, mostly based on Euclidean distance, produce blurry and unsaturated outputs as they tend to minimize the prediction error. While the classification setting offers crisp colors, authors usually must tweak the corresponding losses to conform to colorization requirements [37].

Generative adversarial networks (GAN) are a more recent category of deep learning models proposed by Goodfellow *et al.* [51]. They rely on a pair of entities, a generator G and a discriminator D, both trained in an adversarial framework. G learns how to generate new samples by capturing a reference distribution, whereas D learns to differentiate between real and generated samples. It outputs the probability of belonging to a reference distribution, which is used as a feedback mechanism to help G improve its performances during training. In their original setting, GANs take a noise vector z as an input, and generate data close to ground truth samples [51]. In the framework of colorization, grayscale-to-color mapping must be learned. Thus,

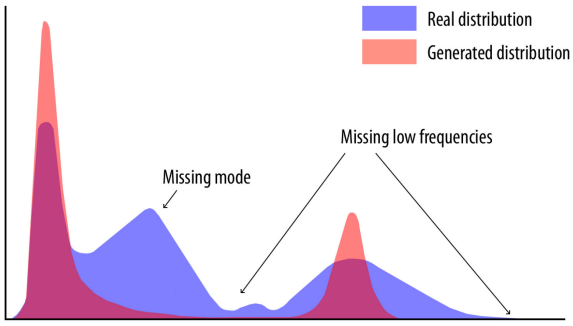

Fig. 1. Example of a generated distribution, compared to the reference used for training. Note the mode collapse effect, represented by the prominence of a pair of modes and a lack of low frequencies within the generated distribution.

the noise vector can be replaced by a panchromatic sample to condition the input, preserve spatial information, and learn the grayscale-to-color mapping only. Such workflow can be carried out by conditional generative adversarial networks (cGAN), derived from GANs and designed to handle mapping operations [52]. Authors who used generative models as a base framework for colorization all relied on a conditional setting [37], [42]–[45], [48]. However, generative models are known for being unstable during training. They also sometimes learn only a part of the target distribution, a phenomenon called mode collapse (see Fig. 1). In this setting, G learns only part of the modal values of the reference data, and thus outputs very similar samples. Various algorithm variations were created to tackle these problems, including BEGANs [53], DRAGANs [54], and WGANs [55]. They provide the user with regularization techniques to allow learning more diverse distributions and low-frequency samples. Hence, regularization techniques help generating crisp and plausible colors without the hassle of developing a loss function suited for colorization, such as for classification-based CNN.

Furthermore, only a very limited number of works have tackled the problem of colorizing spatial products [47], [50], let alone historical ones. Song *et al.* [47] used a pretrained VGG-16 network to extract high-level features from single-polarization SAR satellite images and reconstruct their full-polarization counterparts. Liu *et al.* [50] trained a multitask autoencoder to perform colorization and superresolution on VHR optical satellite imagery. Although both approaches present novel techniques for processing spatial products, they focused on current satellite imagery only, and they are not ideally suited for legacy aerial photographs due to the various domain gaps and the inherent characteristics of these data. Some research teams freely distribute their models and learned parameters. However, their training datasets differ greatly from those required for colorizing legacy aerial photographs. Spatial semantics learned from horizontal and oblique pictures do not match those of vertical imagery, as shown in Fig. 2.

Taking all of these points into account, we can therefore see that most shallow and deep techniques do not quite fit the requirements for an easy and automatic processing of legacy aerial photographs. As shallow techniques remain semiautomatic for the most part, they are unfit for managing photographs

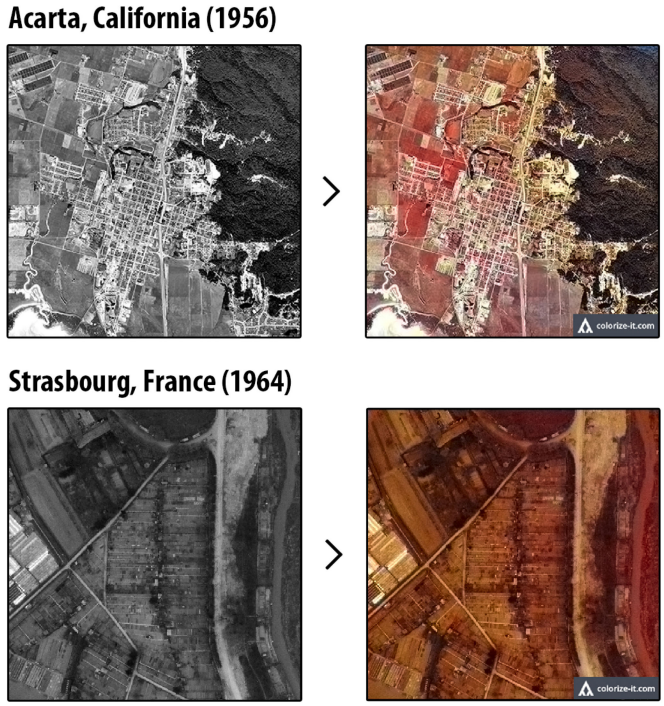

Fig. 2. Colorization examples, using the model trained by Zhang *et al.* [41] with ImageNet data. Note that the model did not manage to properly predict chrominance over grayscale shots.

taken at different dates, with ever-evolving sensors and platform parameters. Indeed, they would require adaptations based on photographs and mosaics specifications. Regarding deep techniques, GANs appeared to be interesting models as they allow the use of fully convolutional networks, and do not require to develop tailored loss functions, unlike standard CNN-based classification methods. Moreover, deep learning allows to better take into account variability within a series, thanks to data augmentation techniques that are better integrated than with traditional techniques.

Given the lack of literature addressing the colorization of legacy aerial photographs, we propose a novel methodology to provide new spatial products, suitable for better integration within current remote sensing workflows (e.g., classification). We deploy deep learning techniques to learn grayscale-to-color mapping based on current photographs using 1) a custom photo library and 2) a conditional DRAGAN architecture that requires virtually little to no prior knowledge. To assess the quality of generated samples, both at train- and test-time, we propose 3) various metrics at the pixel and land cover class levels. Moreover, even though the specifications of current-day images used for learning this mapping function do not necessarily match those of legacy photographs, especially due to pixel size, noise, illumination and capture date, we explore various solutions for taking into account such heterogeneity.

II. Methodology

A simplified version of the proposed methodology appears in Fig. 3. These points are covered in the following sections.

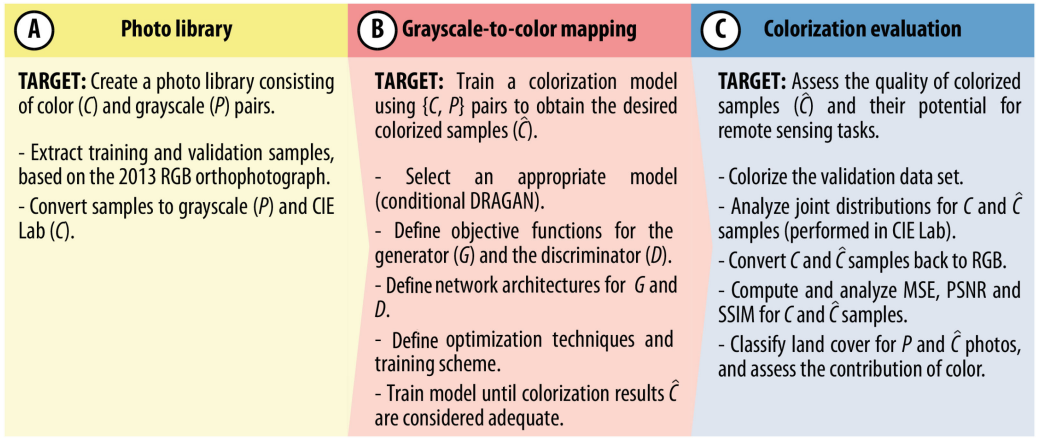

Fig. 3. Graphical abstract of the proposed methodology.

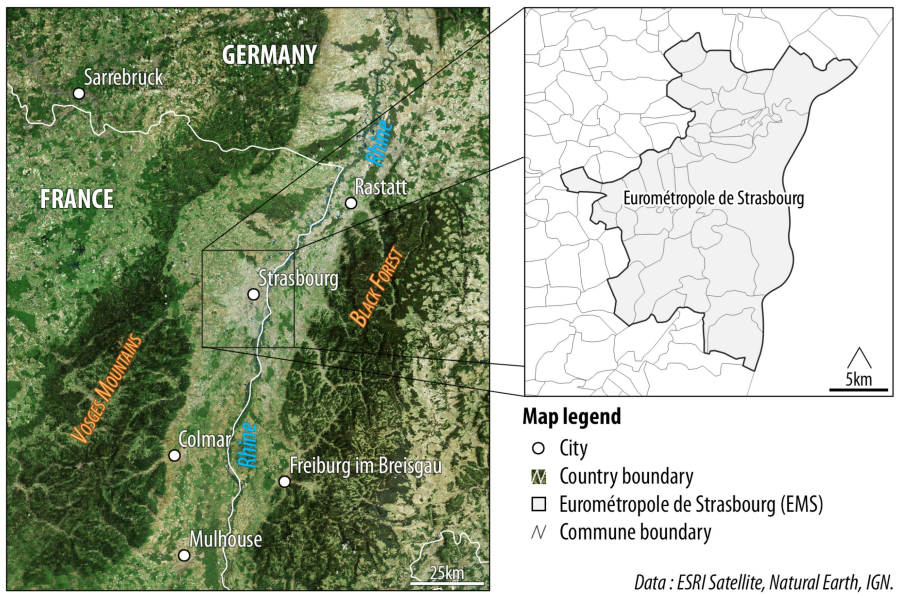

Fig. 4. Study site location and surroundings.

A. Study Area

The proposed method was tested on aerial photographs of the city of Strasbourg (France) and its surroundings (see Fig. 4). With an estimated population of 550 000 in 2015, the *Eurométropole de Strasbourg* (EMS) is the sixth most populated city in France. Located in the Upper Rhine Graben, it covers a surface of 340 km^2 and is bordered by the Rhine River and the Vosges Mountains. Due to its history, topography, and hydrography, Strasbourg and nearby cities have a dense urban landscape and an open-field agricultural system.

B. Available Data and Preprocessing

The colorization model developed in this article relied on an extensive historical spatial database designed by the *Zone Atelier Environnementale Urbaine* that is part of the French Long-Term Ecosystem Research network. The database consists of numerous orthophotos, which cover the entire EMS, from 1932 to 2013 (see Table II). Four dates were candidates for the colorization process as they were the only ones with available grayscale photographs: 1932, 1956, 1964, and 1978. Due to deterioration (1932) and spectral incompatibility with the other stills (1932 and 1964), we focused this research only on the 1956 and 1978 aerial photographs.

To learn the semantics suitable for aerial photography, we designed a custom dataset based on the available photographs

TABLE II
OVERVIEW OF PHOTOGRAPHS AVAILABLE IN THE DATABASE

Year	Season	Spectral content	Spatial resolution (cm)	Sources
1932	Summer	PAN	20	EMS
1956	Winter	PAN	50	EMS
1964	Spring	PAN	20	EMS, LIVE
1978	Summer	PAN	30	EMS, LIVE
1998	Summer	R, G, B	50	IGN
2007	Summer	R, G, B	50	IGN
2013	Summer	R, G, B	16	EMS

TABLE III
SOURCES OF SAMPLE IMAGES CONTAINED IN THE TRAINING PHOTOGRAPHIC LIBRARY

Name	Samples	Spatial resolution (cm)	Source
2013 Orthophoto	8550	30	EMS
2013 Orthophoto	8550	50	EMS
Historical supplement	900	Mixed	IGN
INRIA data set	2900	30	[57]

series. Learning the mapping from grayscale to color required the following photographic library: $\Delta = \{C, P\}$, C and P corresponding to pairs of images taken from the chromatic and panchromatic sets, respectively.

To create the chromatic set C, we used a color orthoimage taken in 2013 (see Table I) as a baseline. It was downsampled to 30 and 50 cm to learn a multiscale model that was capable of processing both the 1956 and 1978 photographs. We applied a stratified sampling scheme to extract 17 100 128×128 spatially independent color images at random. Stratification was based on the following four major land cover classes defined by the level-1 CIGAL dataset:

1) artificial surfaces;
2) agricultural areas;
3) forest and seminatural areas;
4) water bodies.

These ancillary data were digitized on top of an aerial photograph acquired in 2012 over Alsace, a historical French region. It is meant for use at scales inferior or equal to 1/250 000 for the first level of the nomenclature [56]. We also added 3800 images from a variety of older aerial sources to the photo library after a manual selection. This step was required to learn rare spatial semantics, such as historical buildings and geometric distortions induced by the photographic process, e.g., camera tilt and topographic displacement in urban areas [4]. All sources are described in Table III and amount to a total of 20 900 samples.

The panchromatic set P was created by converting all color images to grayscale, leading to the desired $\{C, P\}$ pairs. Conversion was conducted using the following formula from Scikit-Image [58]:

$$P = 0.2125 \times R + 0.7154 \times G + 0.0721 \times B. \quad (1)$$

Hence, panchromatic aerial photographs were not used in the process, but they were replaced by pseudopanchromatic counterparts. Such a choice was fundamental, as perfectly similar and coregistered pairs are required to learn the desired spectral mapping. Otherwise, a slight change in elevation, illumination or position would disturb the model, due to geometric or radiometric discrepancies.

We converted color images to the CIE Lab color space, as recommended by numerous authors [34], [36], [37], [41], [44]–[46], [50], using Scikit-Image. Chroma is denoted by both a and b. These channels are uncorrelated, unlike those of RGB-based color spaces. This strategy allowed the learning of only two channels rather than three, as L refers to lightness and corresponds actually to the (pseudo) panchromatic image [59]. Separating lightness and chrominance could also allow to fix any potential deterioration that may affect legacy photographs, such as vignetting, noise, and scuffs.

We finally created an independent validation dataset $\Delta_{\text{val}} = \{C_{\text{val}}, P_{\text{val}}\}$, following the same methodology. In addition to the 20 900 training samples, 512 color images of 128×128 pixels each were extracted randomly from the 2013 orthophoto, 128 for each of the four CIGAL land cover classes. We also applied conversion to grayscale and the CIE Lab color space to this dataset.

C. Grayscale-to-Color Mapping Using Deep Learning

Due to the unstable nature of GANs in their vanilla setting, a conditional DRAGAN was trained to learn the panchromatic-to-color mapping. Initially proposed by Kodali *et al.* [54], DRAGANs, a subcategory of GANs, were developed from the assumption that mode collapse and instability can be explained by the model converging toward a nonoptimal local equilibrium. Therefore, tweaking the objective functions by penalizing the discriminator's gradients can help avoid such a situation [54]. In the case of colorization, this technique helps learn a proper color distribution, including low-frequency samples, and avoids similar color generation over inherently different land covers and spatial semantics.

1) Objective: In the conditional setting of a DRAGAN for image colorization, the objective functions can be expressed as

$$L_G = E\left[\log(D(G(p), p))\right] \quad (2)$$

$$L_D = E\left[\log(D(c, p))\right] + E\left[\log(1 - D(G(p), p))\right] + \alpha E\left[\left(|\nabla D| - 1\right)^2\right]. \quad (3)$$

Here, both p and c represent panchromatic P and color C samples from the training dataset $\Delta = \{C, P\}$. Colorized samples $\hat{C}$ are described by $G(p)$, the generated images obtained after passing p to the generator G. We followed the recommendations provided by Kodali *et al.* [54] by having the discriminator's gradients ∇ weighted by $\alpha = 10$ in order to stabilize the model. On the one hand, G attempts to minimize its objective function L_G. This is done by fooling the discriminator D into believing $G(p)$ is truly part of the reference distribution c. On the other hand, D tries to minimize its objective function L_D, to better differentiate between the actual and generated samples, c and $G(p)$, respectively. We also conditioned both networks via the panchromatic still p that contains both lightness and spatial information, and is represented by the second parameter in $D(c, p)$ and $D(G(p), p)$.

As suggested by Isola *et al.* [37], we also supplemented the generator objective with a Euclidean-based $L1$ distance, weighted by a factor $\lambda = 100$. The end function used to optimize G can be noted as

$$L_{G*} = L_G + \lambda \times \sum_{i=1}^{n} |c_i - G(p_i)|. \quad (4)$$

This regularization technique helped to model low-frequency colors, whereas the usual adversarial objective was used to

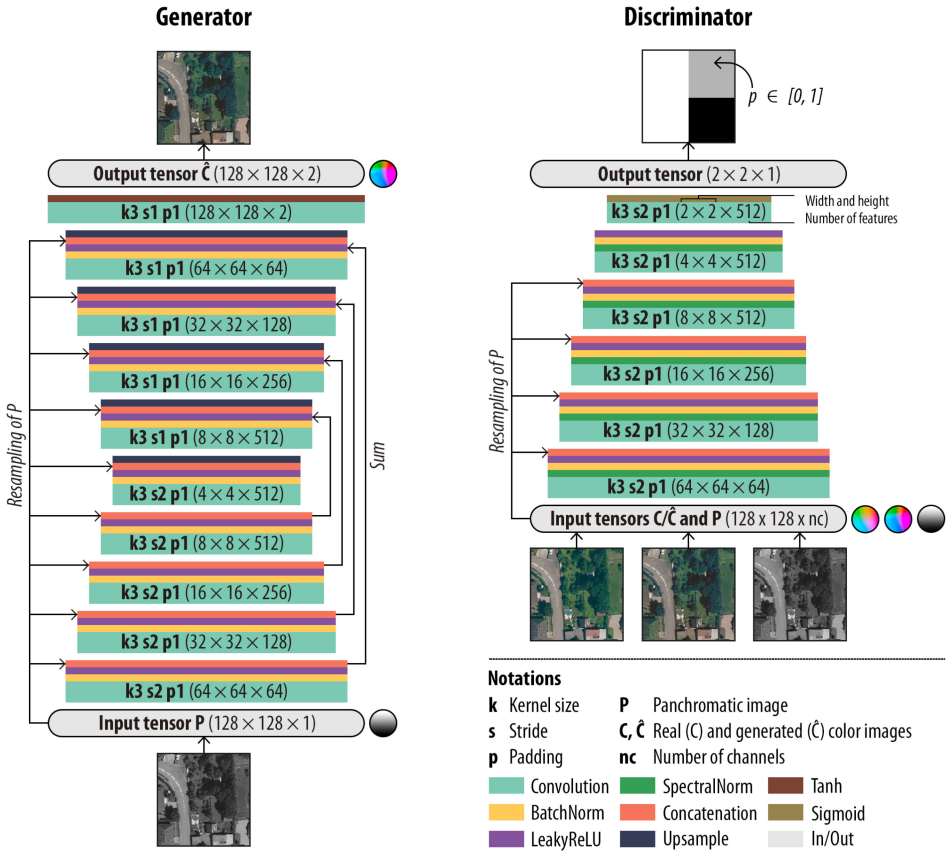

Fig. 5. Architectures of the generator and discriminator networks, both part of the conditional DRAGAN used for learning the grayscale-to-color mapping.

capture the overall distribution, represented by high frequencies [37].

2) Network Architecture: The architecture of the proposed model (see Fig. 5) followed numerous recommendations from [37] and [42].

The generator G of our DRAGAN was based on a UNet-like architecture [60] that helped capture spatial semantics. Composed solely of convolutional layers, G consisted of the following:

1) a downsampling section in which sets of feature maps were learned for different scales;
2) a bottleneck;
3) an upsampling section in which spatial information was restored and combined with previously learned features.

Unlike [37], we did not use skip-connections when combining features from the downsampling and upsampling sections. Due to technical limitations, we simply summed features from both portions in a symmetric fashion.

Our discriminator D was based on a PatchGAN architecture [37]. A series of convolutions evaluated whether or not each colorized patch belonged to a real distribution. We applied spectral normalization to each layer of the discriminator to stabilize training and generate higher quality samples [61], [62].

In addition, both G and D were fully conditioned by P, from top to bottom. Inspired by Cao *et al.* [42], panchromatic samples were automatically resampled and stacked to each layer in G and D during each pass. This design was proposed to enforce spatial consistency at each scale, and in generated images.

3) Optimization and Training: In addition to the DRAGAN objective functions and spectral normalization, we implemented other regularization techniques, such as label smoothing [63]. We also initialized both networks' parameters with values sampled randomly from a Gaussian distribution [51].

The generator and discriminator were trained alternately, each at one step at a time, using a batch size of 256. We used different optimizers to update both networks. For the discriminator, we used SGD as it works best with spectral normalization [61]. Learning rate lr and momentum were set at 2×10^{-4} and 0.9, respectively. For the generator, Adam was preferred, having a learning rate of 2×10^{-4}, and momentum parameters of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Different values for the hyperparameters have been empirically tested. We did not observe any particular instability, improvement or deterioration of the results in these tests. Thus, we ended up using the hyperparameters proposed by Isola *et al.* [37] for the generator, and by Miyato *et al.* [61] and Zhang *et al.* [62] for the discriminator.

The overall training scheme is presented in Algorithm 1. We also implemented fully random data augmentation techniques to obtain more samples artificially, such as rotations, as well as vertical and horizontal flips. A random Gaussian blur was also proposed, with $\sigma \in [1.0, 1.25]$, to mimic the appearance of some of the legacy photographs.

Algorithm 1: Training Algorithm.

input : Batch of C and P images pairs
output: Batch of colorized images $\hat{C}$
for $i \leftarrow 0$ **to** $n(numberofepochs)$ **do**
  Sample batch of color C and pseudo-panchromatic pairs P;
  # Training step for D;
  Augment data and pass P to generator G to generate colorized images $\hat{C}$;
  Update discriminator D according to its objective function, by comparing P, C and $\hat{C}$;
  # Training step for G;
  Again, randomly augment data and pass P to generator G to generate colorized images $\hat{C}$;
  Pass generated samples $\hat{C}$ to discriminator D;
  Update generator G according to its objective function, after getting feedback from discriminator D;
end

Once the model was trained, the discriminator served no other further purpose. From then on, it was possible to pass new panchromatic images to the generator, which in turn produced colorized samples. The entire colorization process was implemented as part of a sliding window, convenient when processing large mosaic datasets. The fully convolutional architecture also avoided being forced into using a single size of image at test time.

D. Evaluating the Colorization

Numerous metrics are already available for evaluating outputs produced by regular machine learning techniques, e.g., classification and regression. Despite their increased use over the last few years, generative models still lack a proper means of assessing the quality of their generated outputs [64]. Therefore, we proposed more traditional metrics to evaluate the performances of the colorization system.

It also must be noted that the generator's outputs are in CIE Lab color space. However, visualization and part of the metrics used for evaluating the colorization process require working in the RGB color space. Thus, among the evaluation techniques proposed below, only the analysis of bivariate distributions remained in CIE Lab. All other steps were fulfilled after converting the colorized samples back to RGB.

It is important to note that most articles on colorization include a comparison with the results obtained through the works of other authors, particularly in deep learning. However, most research teams work with public datasets (see Table I), such as ImageNet [38], [41], [45], CIFAR-10 [34], [46], and SUN [38], [40], [42], [49], which allow for comparison. At the time of writing this article, all models available online were based on parameters learned from nonaerial images. Comparison with outputs generated through such networks would not yield meaningful results, as shown in Fig. 2. Another solution would be to train the models proposed by authors from scratch, by reusing the suggested hyperparameters. However, hyperparameters can be specific to certain datasets, and would require further fine-tuning in order to properly account for spatial semantics extracted from aerial photographs. The same issue can be raised for traditional techniques, that would require tinkering with hyperparameters for each mosaic, or even possibly each picture. Thus, we do not propose a comparison with other colorization techniques, as we considered it would not yield meaningful information regarding aerial photographs colorization.

1) Statistical Evaluation: As noted previously, generative models require consideration of samples' plausibility rather than raw accuracy. Although we suspected a generic evaluation would not be able to capture such an abstract concept, it was conducted nonetheless for exploratory purposes.

Based on the validation dataset $\Delta_{\text{val}} = \{C_{\text{val}}, P_{\text{val}}\}$, P_{val} was passed to the generator G, which in turn produced colorized samples $\hat{C}_{\text{val}}$. First, we compared the bivariate distributions of the a and b color channels, by randomly sampling joint pixels from C_{val} and $\hat{C}_{\text{val}}$. This specific step is the only part of the evaluation process where colorized and reference samples remained in the CIE Lab color space.

The second part of this evaluation compared C_{val} and $\hat{C}_{\text{val}}$ using mean squared error (MSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), averaged over all color channels. For a simpler analysis of all three metrics, we converted C_{val} and $\hat{C}_{\text{val}}$ samples back to RGB beforehand.

MSE measures the dissimilarity between the real and predicted samples. Its values range from 0 (similar) to $+\infty$ (not similar). MSE can be defined as

$$\text{MSE}(c, \hat{c}) = \frac{1}{n} \times \sum_{i=1}^{n} (c_i - \hat{c}_i)^2. \tag{5}$$

PSNR measures the similarity between real and predicted samples, and is usually used to assess the quality of a reconstructed signal. PSNR values range from 0 (not similar) to $+\infty$ (similar) and can be defined as

$$\text{PSNR}(c, \hat{c}) = 10\log_{10} \times (255^2/\text{MSE}(c, \hat{c})). \tag{6}$$

SSIM measures the similarity between real and predicted samples, based on structure, lightness and contrast [65]. Images are processed at kernel-level, using convolution. SSIM values range from 0 (not similar) to 1 (similar). SSIM—having three constants c_1, c_2, and c_3 that depend on bit depth—can be defined as

$$\text{SSIM}(c, \hat{c}) = \frac{(2\mu_c\mu_{\hat{c}} + c_1)(2\sigma_c\sigma_{\hat{c}} + c_2)(2cov_{c\hat{c}} + c_3)}{(\mu_c^2 + \mu_{\hat{c}}^2 + c_1)(\sigma_c^2 + \sigma_{\hat{c}}^2 + c_2)(\sigma_c\sigma_{\hat{c}} + c_3)}. \tag{7}$$

2) Evaluating Land Cover Classification: To assess the contribution of predicted colors, we complemented the quantitative analysis of the results by comparing land cover classifications between the grayscale P and colorized photographs $\hat{C}$. To meet present-day standards in remote sensing, we developed a simple but standard workflow, with a traditional sampling scheme and the computation of ancillary features. This step was not meant to indicate whether or not colorization produced plausible results, but rather to shed light on the contribution of generated chrominance to a very common task in remote sensing. Reference data were obtained by digitizing six generic land use and land cover (LULC) classes using the panchromatic photographs P from 1956 and 1978 to serve as references. The selected classes were: artificial surfaces, roadway, agricultural

TABLE IV
PERCENTAGE OF PIXELS FOR THE DIGITIZED LAND COVER CLASSES FOR THE 1956 AND 1978 PHOTOGRAPHS

% of pixels/Class/Year	1956	1978
Agricultural surface	48.9	14
Tree vegetation	25.6	28.7
Herbaceous vegetation	6.8	22.8
Built-up surface	11.1	17
Roadway	3.2	6.6
Other surfaces	4.4	10.9

areas, herbaceous vegetation, tree vegetation, and other surfaces (see Table IV).

We produced a straightforward and readily reproducible classification scheme for assessing both the 1956 and 1978 photographs. First, we computed two sets of textures using P: 1) default and uniform local binary patterns [66] and 2) simple Haralick textures [67] computed with a kernel size of 5. These texture features were then stacked to both P and $\hat{C}$. Two random forest classifiers [68] were trained, one for each of the two stacks—grayscale and colorized—after a fine-tuning of the hyperparameters using a random grid search. We then validated the model using stratified fivefold cross-validation (80%/20% training/validation split) for both years. Finally, we compared $F1$ scores obtained for the classified P and $\hat{C}$ pictures, at the scene and land cover class levels.

III. RESULTS

Before investigating the colorization results provided by the proposed model, an analysis of the effects of various parameters was carried out with a 20% sample of the training set. Among the available parameters, three were tested: batch size, learning rate, and lambda (see Sections II-C1 and II-C3 for more information). Visual results and quality metrics (SSIM and PSNR) obtained with different values for each parameter are presented in Fig. 6 for the full validation set.

Changes in batch size (bs) did not result in major variations, be it qualitatively or quantitatively. Larger batch sizes resulted in slightly better quality metrics, going from $\overline{\text{PSNR}} = 25.11 \pm 4.41$ and $\overline{\text{SSIM}} = 0.92 \pm 0.05$ for $bs = 64$, to $\overline{\text{PSNR}} = 25.44 \pm 4.61$ and $\overline{\text{SSIM}} = 0.93 \pm 0.04$ for $bs = 256$. However, with lower batch sizes, one can note more plausible colors for bare soils, while colorization mistakes appeared to be reinforced over green surfaces, which, for some, turned navy blue. Overall, the best results were obtained for $bs = 256$. Such observations do not necessarily match those of the available literature. Indeed, larger batch sizes are usually responsible for a larger generalization gap during test time, which was not the case here [69], [70]. However, it could be argued that the differences in bs were not large enough to expose such phenomenon.

Changes in learning rate (lr) for both the generator and the discriminator had a strong effect on the predicted colors. With $lr = 2e^{-7}$, the model gave a colorization with magenta artifacts for bare soils and artificial surfaces. Vegetated surfaces showed better results, despite hue being not quite right. With $lr = 2e^{-1}$, colorization results were not plausible and turned various shades of red, which resulted in the lowest values for the quality metrics with $\overline{\text{PSNR}} = 5.04 \pm 1.18$ and $\overline{\text{SSIM}} = 0.083 \pm 0.06$. Overall,

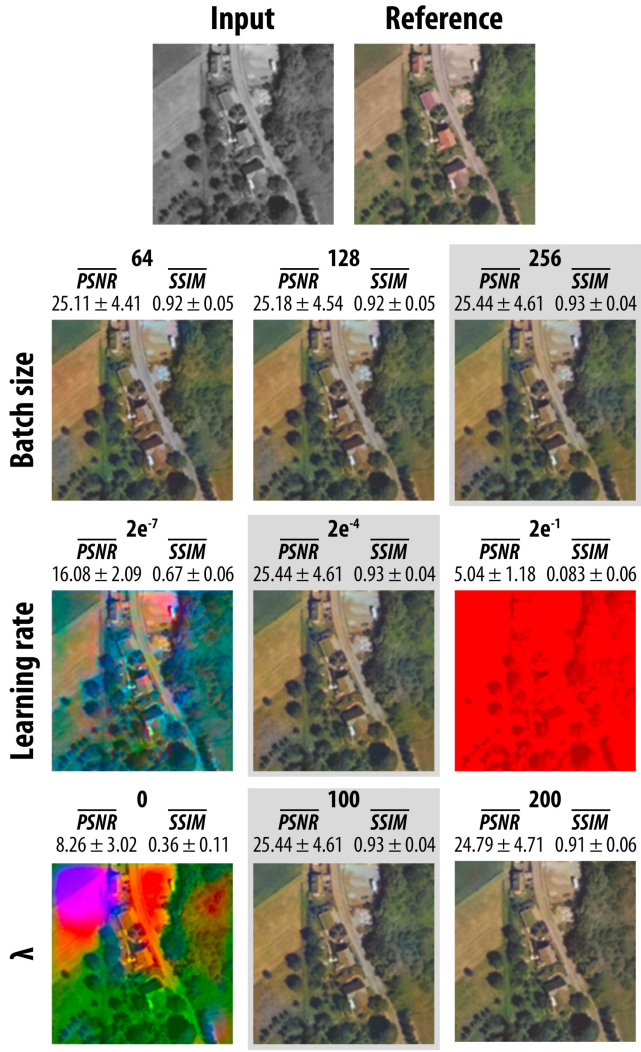

Fig. 6. Qualitative and quantitative results obtained for three parameters set with different values. For each test, the model was trained for 500 epochs with only 20% of the full training set, due to hardware limitations. SSIM and PSNR were computed for the validation set. Gray panels correspond to the baseline values proposed in this article for each parameter. For each test, one parameter was tested, whereas the other two were set to their baseline value.

the best results were obtained for $lr = 2e^{-4}$. The learning rate controls the speed at which the model moves toward a hypothetical global optima [28]. A low learning rate may result in slow convergence, as seen here with $lr = 2e^{-7}$, or sometimes even in an improper solution. A high learning rate may result in higher training error and in the model passing the global optima. Too much of an update in the weights of the model may have led to an improper solution, which could explain the result obtained with $lr = 2e^{-1}$.

Changes in the generator's loss function through the λ parameter had consequences on the distribution and saturation of the predicted colors. With $\lambda = 0$, colors were too saturated, even though the greens were properly placed over vegetated surfaces. It resulted in low values for the quality metrics, with $\overline{\text{PSNR}} = 8.26 \pm 3.02$ and $\overline{\text{SSIM}} = 0.36 \pm 0.11$. When setting $\lambda = 0$, Isola *et al.* [37] obtained results of much better quality with crisp colors, which may be due to the low number of epochs

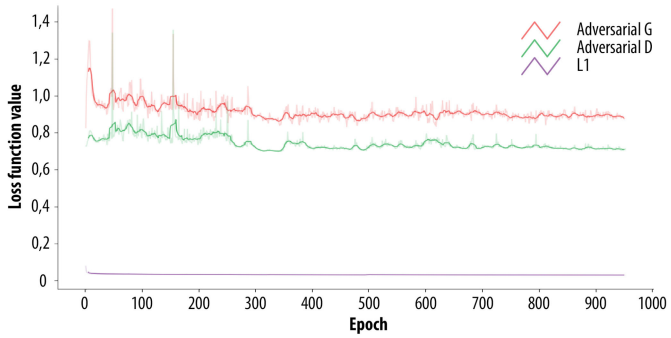

Fig. 7. Adversarial and $L1$ losses during training.

Fig. 8. Example of colorization results for the validation dataset (2013). Generated samples were randomly selected.

that we used for testing the effects of parameters. With $\lambda = 200$, the colorization results were less saturated than with $\lambda = 100$, but there were no artifacts such as the navy blue spots that were previously observed. This effect was illustrated by Isola *et al.* [37], who showed how weighting the generator's loss with λ helped reduce the number of artifacts, at the cost of color saturation. Overall, the best results were obtained for $\lambda = 100$, even though $\lambda = 200$ was close in terms of color and quality metrics, with $\overline{\text{PSNR}} = 24.79 \pm 4.71$ and $\overline{\text{SSIM}} = 0.91 \pm 0.06$.

The proposed model was trained for 950 iterations with the parameters for which the best results were obtained, as described in Section II-C. Each epoch required about an hour, the full training step being over after 40 days. It was not possible to train more due to technical limitations. Indeed, this article was carried out on a paid Google Cloud virtual machine instance, with a half-K80 GPU (12 GB video memory). However, the outputs generated using this number of epochs were good and encouraging. Values for the adversarial and $L1$ losses are shown in Fig. 7.

Results for the validation set Δ_{val} are presented in Fig. 8. As a reminder, they correspond to the colorizations obtained for the year 2013, for which we have a color reference. Colorization results for legacy photographs, without any reference, are presented later in this section.

A. Visual Evaluation

For the validation dataset and legacy photographs, the model produced a crisp color delineation and a consistent spatial distribution of chrominance (see Figs. 8 and 9). This indicated a good performance of the proposed model. We observed a diverse spectrum of colors for vegetation in urban areas. Colors ranged from light green for the most exposed areas, to dark green for the most shaded ones. Results for asphalt surfaces were also good as gray intensities differed between footpaths and roadways. However, rooftops lacked color diversity, being mostly gray and brown. This lack of diversity reflected how instability or mode collapse can still affect generative adversarial models despite use of the DRAGAN framework. In rural areas, colorized photographs again highlighted the strong performance of our approach, especially for agricultural parcels. The variation in the beige and brown tones for soils distinguished between bare and cropped areas. In addition, the roughness and overall texture of fields were more visible, for example, we could identify furrowed fields traced by plows or other agricultural machinery. The different intensities of green also helped estimate vegetation height and the heterogeneity of intraparcel coverage. Finally, color was well generated for water bodies, and shadows cast by adjacent trees were properly reconstructed. However, we did observe some failures cases around areas with specific semantics. The result was blobs of abnormal color, mostly deep blues, yellows, and magenta. This is especially true for legacy photographs, for which entire areas are poorly colorized. This effect is for the most part observed for tall buildings with visible facades, or in dense urban areas and narrow crops.

B. Statistical Evaluation

Comparing the joint distributions of the a and b color channels for the validation dataset Δ_{val} (see Fig. 10) provided insight into how the model generates colors.

For all classes, very low frequency values were not captured by the model, as shown by the white trims around joint plots of $\hat{C}$. Although the overall shapes of all distributions were quite similar between the real and generated samples, many modes in the generated outputs were missing, displaced, or possessed higher or lower frequencies. Such phenomena were markedly noticeable for all land cover classes; the exception was artificial surfaces that presented very similar distributions for the colorized and reference samples. These observations matched the usual drawbacks related to the use of generative networks.

All three metrics—MSE, PSNR, and SSIM—testified to the strong performance of the developed model (Fig. 11), with $\overline{\text{MSE}} = 210.29 \pm 166.99$, $\overline{\text{PSNR}} = 25.56 \pm 2.20$, and $\overline{\text{SSIM}} = 0.93 \pm 0.06$. At a global level, these scores indicated both a high similarity between the colorized and reference products, and a high quality of image reconstruction relative to the original samples. Indeed, the SSIM measure indicates a close structure between reference and generated samples, with $Q1_{\text{SSIM}} = 0.907$, $Q3_{\text{SSIM}} = 0.963$, and $IQR_{\text{SSIM}} = 0.055$. Nevertheless, when focusing on LULC classes, a deeper analysis of the same outputs revealed a more mixed outcome.

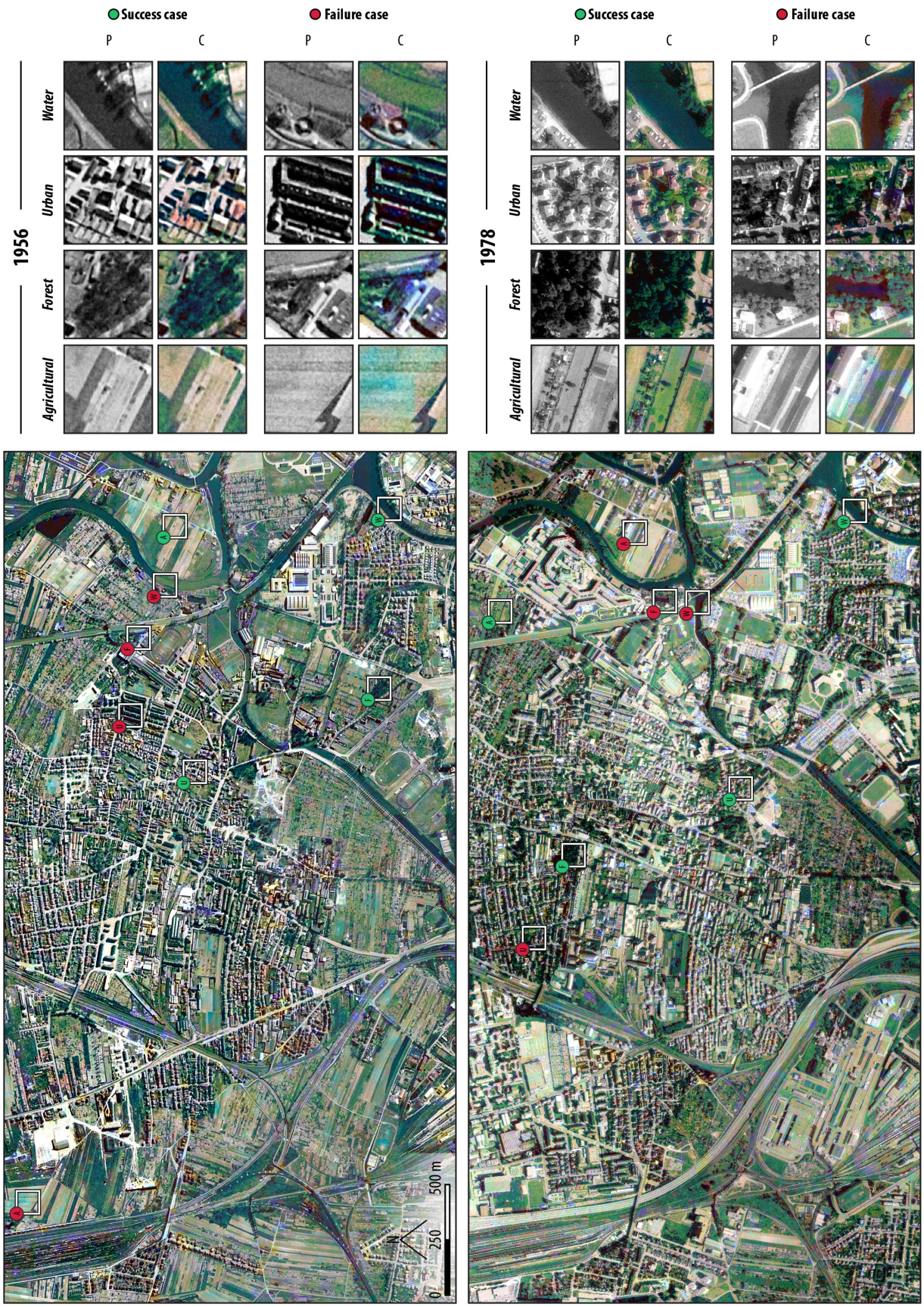

Fig. 9. Colorization results for the 1956 and 1978 panchromatic mosaics covering the western side of Strasbourg. Cases of success and failure are shown for four different LULC classes.

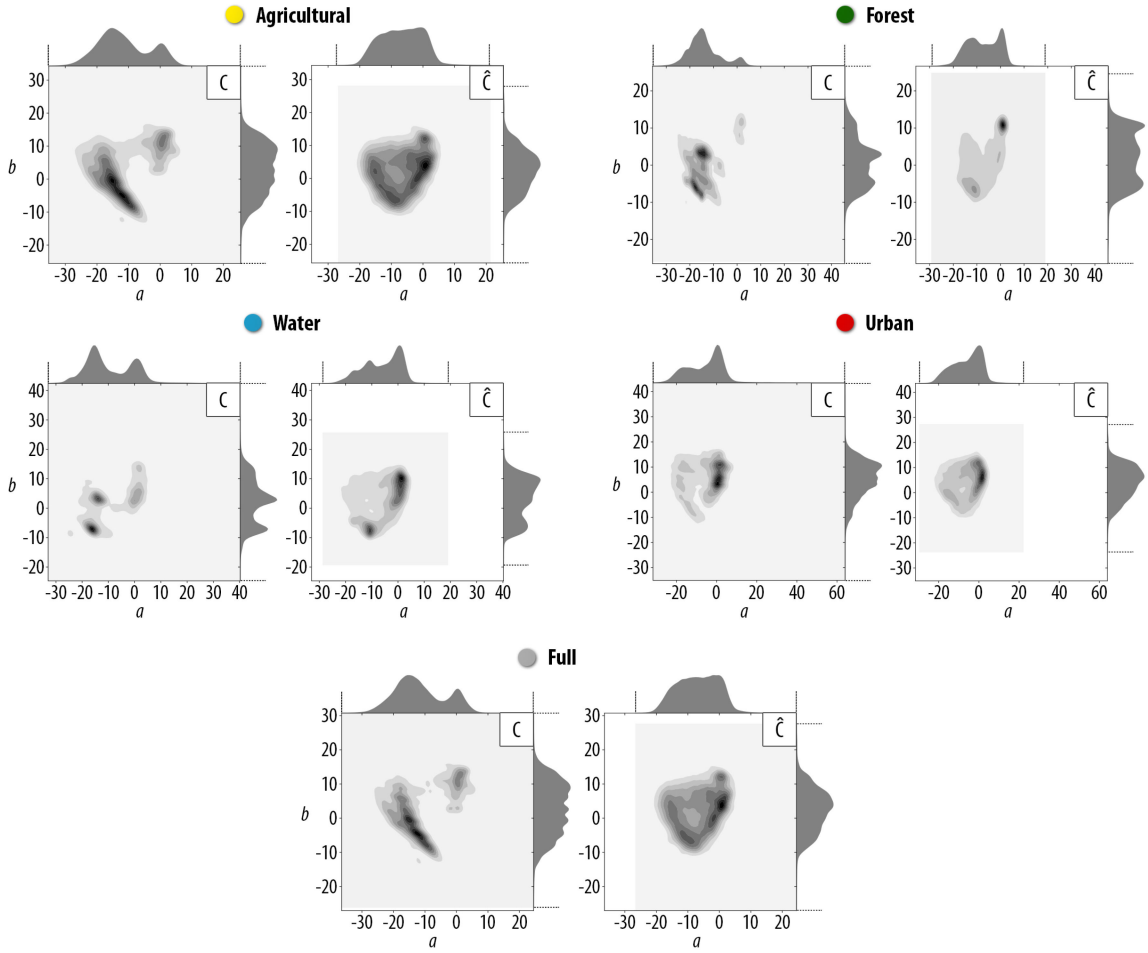

Fig. 10. Joint distributions of a and b channels computed for the reference (C) and colorized samples ($\hat{C}$) taken from the validation dataset after training. These metrics were computed at the LULC class and scene levels.

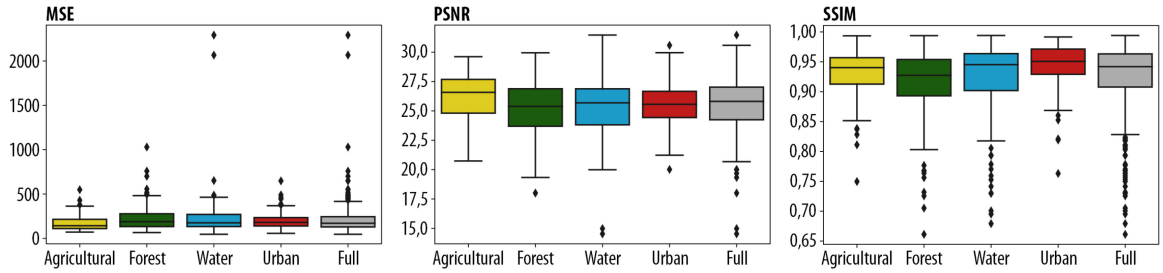

Fig. 11. Metrics computed for the colorized samples ($\hat{C}$) taken from the validation dataset after training. These metrics were computed at the LULC class and scene levels.

The best performances were obtained for urban areas, which exhibited low MSE ($\overline{\text{MSE}} = 202.08 \pm 90.31$) and high PSNR and SSIM ($\overline{\text{PSNR}} = 25.44 \pm 1.76$ and $\overline{\text{SSIM}} = 0.94 \pm 0.04$). This is consistent with the joint plots, that showed very similar chrominance distributions between the colorized and reference images for built-up areas. Performances then decreased for other LULC classes: water ($\overline{\text{MSE}} = 263.14 \pm 266.34$, $\overline{\text{PNSR}} = 25.33 \pm 2.42$, $\overline{\text{SSIM}} = 0.91 \pm 0.07$), tree vegetation ($\overline{\text{MSE}} = 232.42 \pm 149.34$, $\overline{\text{PNSR}} = 25.18 \pm 2.43$, $\overline{\text{SSIM}} = 0.91 \pm 0.06$), and agricultural surfaces ($\overline{\text{MSE}} = 170 \pm 85.58$, $\overline{\text{PNSR}} = 26.27 \pm 1.93$, $\overline{\text{SSIM}} = 0.93 \pm 0.04$). Based on these metrics, higher potential colorization failures could be expected for these classes.

C. Evaluating Land Cover Classification

The classification results are presented in Table V and Fig. 12. For the 1978 photograph, the mean $F1$ score at the scene level

Fig. 12. Classification results for panchromatic photographs taken in 1956 and 1978, and their colorized counterparts.

TABLE V
$F1$ SCORES OBTAINED AFTER CLASSIFYING PANCHROMATIC PHOTOGRAPHS (P) TAKEN IN 1956 AND 1978, AND THEIR COLORIZED COUNTERPARTS ($\hat{C}$)

F1 scores	**1978**			**1956**		
LULC classes	P	$\hat{C}$	**Gain (%)**	P	$\hat{C}$	**Gain**
Agricultural surface	0.59	0.72	22	0.80	0.83	4
Tree vegetation	0.75	0.78	4	0.64	0.66	3
Herbaceous vegetation	0.59	0.66	12	0.14	0.21	50
Built-up surface	0.48	0.66	38	0.43	0.50	16
Roadway	0.37	0.56	51	0.29	0.40	38
Other surfaces	0.42	0.52	24	0.40	0.45	13
(Photo) Mean F1 score	**0.58**	**0.68**	**17**	**0.64**	**0.68**	**6**

(including all classes) improved from 0.58 (P) to 0.68 ($\hat{C}$) (17% higher), indicating an improvement due to the colorization process. Each land cover class investigated in this classification benefited from chrominance, as $F1$ scores were higher than with monochromatic-based features. Artificial areas—roadway, built-up surfaces, and other surfaces—take full advantage of generated chromatic information, with gains of 51%, 38%, and 24%, respectively. Agricultural areas, herbaceous vegetation, and tree vegetation benefited the least from colorization, with gains of 22%, 12%, and 4%, respectively. Despite a less marked impact for the 1956 set, the colorization workflow also helped refine classification, raising the mean $F1$ score from 0.64 (P) to 0.68 ($\hat{C}$), an increase of 6.25%. Again, we observed the best results for artificial surfaces, mainly roads (38% increase), followed by built-up areas (16%) and other surfaces (12%). Among seminatural land cover classes, herbaceous surfaces benefited the most from colorization, with an increase of 50% for the $F1$ score, followed far behind by agricultural and tree surfaces, having, respectively, increases of only 4% and 3%.

Visual inspection of the output LULC maps (see Fig. 12) also confirmed that the colorization process helped improve classification. Overall, a higher uniformity was observed in the color-based results. This was particularly true for rooftops and agricultural areas, where it was possible to note an important decrease in isolated and misclassified pixels. Thus, the colorization process clearly enhanced the distinction between classes from the classifier, such as for roads and built-up areas. All these improvements tended to simplify the interpretation of the image due to the improved spatial consistency.

IV. DISCUSSION

The proposed model successfully learned the mapping between grayscale and color aerial photographs. Our approach produced plausible colorization for legacy stills, thereby providing a novel solution to access underused data. This important advance stemmed from a series of choices that were specific to the initial challenge. First, the conditional DRAGAN framework that we propose helps to learn an extensive set of modes within the reference distribution of chrominance. This learning would likely not have been possible using a standard GAN due to the prevalence of mode collapse in its standard setting [71]. Second, the development of a fully convolutional architecture proved to be efficient when processing the dataset. Indeed, convolutions are usually faster and more intuitive than their fully connected counterparts when dealing with multiple dimensions [31]. We

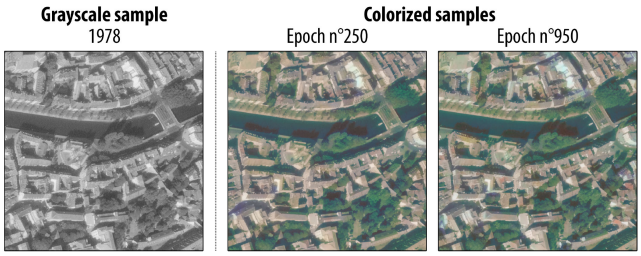

Fig. 13. Examples of colorized samples after 250 and 950 epochs. Note the emergence of new rooftop colors after 950 epochs, especially the red and orange tiles on the right of the image.

also considered the input data, given that aerial photographs are very heterogeneous in their specifications. The design of our multiscale photo library allows colorizing stills at spatial resolutions that range from 30 to 50 cm. In addition, data preparation techniques provide the model with proper pseudopanchromatic photograph, after conversion of the colored original. This conversion facilitates the learning of a full spectral mapping, irrespective of radiometric and geometric changes that would have resulted from the use of a nonhomologous image. The success of the colorization of two legacy aerial photographs, taken with actual panchromatic sensors, also implies that the learned grayscale-to-color mapping depends more on spatial semantics than pixel values.

However, visual analysis, quantitative metrics, and the evaluation of classification highlight the uneven quality of generated results, at both global and LULC classes scales. Global performances first should be viewed in relation to the number of iterations used for training the model. Although the adversarial objective functions remained stable during the second half of the training of both networks (see Fig. 7), the model still may have to find a global optimal solution for the grayscale-to-color mapping function. Such a phenomena could explain colorization mistakes, along with the late learning of new rooftop colors, red and orange in particular (see Fig. 13). Our results are encouraging, as they suggest that a longer training time would further improve the results. Performance at the scale of LULC classes was also unequal; the unbalanced distribution of spatial features and objects in the training set may explain the phenomenon. Indeed, despite a stratified sampling scheme, the mapping function was still learned using information contained within the images. However, territories are anisotropic by nature, leaving us with rare, sometimes even unavailable, semantics that are poorly perceived and generalized by the model. Furthermore, various colorization scenarios may be appropriate for the same object, representing different modes within the distribution of chrominance, for example, roofing tiles can be gray, black, red, etc. Despite the use of various regularization techniques, mode collapse remains a frequent issue with GANs. It explains why color may be harder to predict for classes or semantics having multiple possibilities, especially when training is not complete.

Given these previous points, the photo library that was designed for training is another factor that conditions model performance. Indeed, the remote sensing products that we selected for this article are heterogeneous, as the variability between the ever-improving sensors involves changes in spatial resolution, noise, and other degradation. Moreover, the training set was built around a recent, very high resolution and artifacts-free orthophoto, acquired during summer. Data preparation and augmentation techniques were proposed to make reference samples and legacy products compatible. However, not all discrepancies (e.g., seasonality and surface phenology) could be accounted for. This could in part explain why the colorized 1956 winter photograph is of a lesser quality than the one taken during the summer of 1978. However, from a visual point of view, surfaces that exhibit periodic life cycle events (natural and seminatural lands) display the most plausible chrominance values at both dates, compared to artificial surfaces. As noted above, for both sets of aerial photographs, spatial semantics surpass radiometric content (and therefore seasonality) when providing the generator with meaningful information. This could explain why artificial surfaces show lower quality results, given that these surfaces are more complex and require a wider set of spatial semantics to predict a proper color. Although seasonality could play a major role, other properties specific to photographs may better explain the quality of generated samples. For example, the photographic processing of analog pictures leads to various forms of deterioration, rarely or not at all present in digital photographs (e.g., vignetting, noise, blur, smudges) [6]. As the 1956 photograph was moderately blurred, a portion of the spatial semantics may differ from what was learned by the model, thus leading to lower quality results. Further research is required, however, to identify those factors that influence the mapping process and determine the impact of these identified factors.

Despite some errors in colorization, predicted chrominance produced an information gain, as evidenced by LULC classifications. The performance of the random forest classifier was improved when compared to the results obtained for the panchromatic pictures. Rather than requiring perfect colorization, our developed classification techniques benefit from a simple distinction between the desired LULC classes. In this case, color helps distinguish between pixels, that are otherwise too similar based on panchromatic and texture features. The ambiguity between herbaceous vegetation and smooth artificial surfaces (rooftops, roads) is an obvious example. Also, based on the $F1$ scores and visual inspection, the information gain for the 1956 photographs was less than for the 1978 set. This observation reinforces our idea of the model being unable or less able to predict proper chrominance for degraded photographs or spatial semantics. Classification may be an efficient means of measuring the quality of generated products, and the information gain provided by the model. However, the multimodal nature of colorization and generative networks properties must be considered, as better classification results do not necessarily imply an appropriate prediction of color.

We also demonstrated existing inconsistencies between the human visual system and quantitative metrics usually suggested for comparing images. MSE and PSNR only provide pixel-wise information; they do not take context into account. The SSIM index—first proposed as an indicator highly correlated to human vision [65]—was unable to provide more accurate results in relation to visual evaluation. A possible explanation may be the

strong relationship between MSE and SSIM, as demonstrated by Dosselmann and Yang [72]. The phenomenon is particularly strong for spatial features and objects that possess numerous plausible colors; urban areas where rooftops often appear brown or gray, instead of orange, for example, are considered to be the best outputs according to MSE, PNSR, and SSIM. These indicators are also unable to assess correctly the quality of other colorizations, such as for agricultural areas that obtain visually plausible results with crisp colors (green, brown, and yellow mostly). Indeed, unsaturated results are likely to be favored by these metrics, due to a minimization of the global error, despite visually unsettling colorization. Varga and Sziŕanyi [40] have drawn similar conclusions when evaluating their results with QSSIM, another structural index based on SSIM [73]. Despite obtaining an average QSSIM of 0.93, the authors point out the nonlinearity of the relationship between this metric and the quality of their outputs, along with inconsistencies of evaluation. Thus, despite being time-consuming and sensitive to subjectivity, visual analysis of generated samples currently remains the most appropriate evaluation technique for colorized aerial photographs.

It must be stressed that no other work has tackled the problem of colorizing grayscale aerial photographs in this manner, let alone legacy stills. However, Song *et al.* [47] and Liu *et al.* [50], who, respectively, colorized radar and optical satellite imagery using CNN-based classifiers, experienced similar issues and results. Liu *et al.* [50] observed the same colorization inconsistencies using present-day satellite images. This may indicate that the grayscale-to-color mapping does not depend necessarily on the quality of the grayscale-based products, but rather requires comprehensible spatial semantics, regardless of the medium. Song *et al.* [47] showed that their reconstructed full-polarization SAR images could be used in various PolSAR applications (e.g., Freeman–Durden decomposition) and provide plausible results. These points thus highlight the potential of reconstructed data in remote sensing workflows.

Solutions for the colorization of aerial photographs: Different solutions can be proposed to improve colorization results and evaluation techniques. Although we demonstrated that the DRAGAN framework was efficient enough to produce plausible outputs, new algorithms are released regularly and could help reduce undesirable effects, such as mode collapse. Thus, novel models based on alternative regularization techniques must continue to be explored.

An operational system must also be able to generate chrominance over much larger territories, beyond the scale of our test regions (i.e., Strasbourg and Alsace). Due to the specific agricultural landscape (open-field) and urban morphology (dense habitat) within our study photographs, semantics learned by the generator may not generalize well when applied to other landscapes. A solution would be to add many more samples to the photo library to meet the baseline image number of current colorization techniques (see Fig. 14). A greater number of reference photographs would also help reduce colorization artifacts by providing rare or currently unavailable semantics.

Providing hints to the model, as proposed by Zhu *et al.* [74], may further improve the performances. Indeed, various authors have already managed to successfully train models and obtain better results with ancillary data. Some existing works used additional features maps derived from the grayscale image, such as multistage features [40], multiscale normalized pyramids [39], combined patch-descriptors [33] or radiometric and height features based on three-dimensional point clouds and orthoimagery [75]. Other authors have also used classification and semantic segmentation models for feature extraction. A multipart model, based on four networks, was trained by Iizuka *et al.* [36]. In this model, three out of four networks were dedicated to the extraction of low-, mid-, and high-level features. All features were then concatenated and passed to the colorization network. A two-part model based on an extractor and a translator was also proposed in [47]. The extractor was tasked with the prediction of a hypercolumn deep of 1153 descriptors for each pixel. It was then passed to the translator whose role was similar to that of a colorization network. These ancillary data give further information regarding the nature of each pixel and help learn a more accurate mapping function [76], [77]. Thus, such techniques may produce better colorization results by feeding additional features to the generator. However, due to their unique channel, grayscale aerial photographs are unfit for the computation of radiometric indices, such as the vegetation index. Moreover, a colorization model trained for processing a time series needs to take data heterogeneity into account, which is particularly challenging for legacy photographs. Indeed, each air mission and photograph displays specific characteristics, such as phenology, scale, noise, vignetting, and sun spots. Features that depend on spatial resolution, such as texture indices [78], thus vary inside the time series and may not be compatible with that of the data used for training. These factors explain why radiometric and texture indices were not used in this article. Features obtained by the means of classification or segmentation models could provide meaningful information. However, the remote sensing community still lacks large annotated databases such as ImageNet [30]. Moreover, there is currently no baseline implementation for processing long-term remotely sensed products with high heterogeneity, analogous to AlexNet [30], Inception-v4 [79], or VGG-19 [80]. Finally, directly providing a semantic map to the colorization network might be useful, so as to create a corpus of LULC and

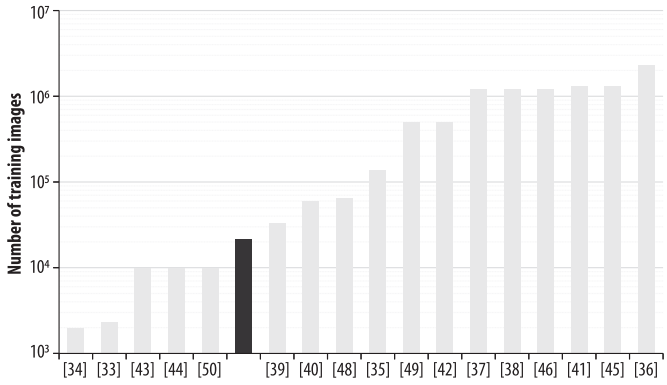

Fig. 14. Number of images used to train colorization models in various deep learning papers. The dark bar corresponds to the number of training samples used in this article.

colors associations. Nevertheless, it would require to have LULC maps for the entire time series at hand, which we did not.

Apart from the computation of ancillary features, new data augmentation algorithms could be developed in order to better account for the various specifications of legacy photographs, particularly radiometric (e.g., noise, illumination, and holes) and geometric properties (e.g., scale, camera height, and camera angle). Finally, other machine learning and image processing techniques could be developed to restore legacy photographs prior to colorization.

At the same time, the quantitative evaluation of colorization remains a crucial step toward establishing an operational system. However, classic metrics fail to match human evaluation. Comparing various quantitative and qualitative measures, Borji [64] demonstrated the difficulty in evaluating generative models. There is currently no technique that allows for the simultaneous evaluation of diversity and fidelity. Other metrics may be more appropriate, however, they are based on specific deep learning models. For example, the Inception Score could be used, yet it relies on an Inception model trained using the ImageNet dataset that contains images taken at an eye-level camera angle [63], [81]. Thus, geographic data increases the complexity of evaluating generative models as the vertical or high-angle shot domain it belongs to is very specific, revealing an underexplored problem. Consequently, new indicators must be developed to assess the quality of generated products, based on remotely sensed images and photographs.

V. Conclusion

This article presents a novel methodology for colorizing legacy grayscale aerial photographs. The proposed DRAGAN is fully conditioned by grayscale samples that are derived from current color stills. The model achieves a proper grayscale-to-color mapping, demonstrated by its rather convincing colorization outputs. Quantitative analysis of the results illustrated that high-frequency colors are learned accurately, whereas difficulties arose with lower frequency values. Model performance also varied among the four studied land categories; this variability may be related to the complexity of spatial structures and the semantics that depend on the processed area. Our colorized samples can then be fed into traditional remote sensing pipelines, as demonstrated by our pixel-level image classification, confirming the utility of color for characterizing land cover. We also report the lack of studies involving legacy spatial products in both remote sensing and deep learning; this is particularly evident for image restoration and quality assessment. Thus, this preliminary article underscores the importance of research aimed at maximizing the value of legacy photographs in modern remote sensing applications.

References

[1] L. Bowden and W. Brooner, "Aerial photography: A diversified tool," *Geoforum*, vol. 1, no. 2, pp. 19–32, 1970.

[2] R. Welch, "Spatial resolution requirements for urban studies," *Int. J. Remote Sens.*, vol. 3, no. 2, pp. 139–146, 1982.

[3] F. Goldsmith, *Monitoring for Conservation and Ecology* (Conservation Biology). London, U.K.: Chapman and Hall, 1991.

[4] J. L. Morgan, S. E. Gergel, and N. C. Coops, "Aerial photography: A rapidly evolving tool for ecological management," *BioScience*, vol. 60, no. 1, pp. 47–59, 2010.

[5] B. C. Forster, "An examination of some problems and solutions in monitoring urban areas from satellite platforms," *Int. J. Remote Sens.*, vol. 6, no. 1, pp. 139–151, 1985.

[6] J. S. Aber, I. Marzolff, and J. Ries, *Small-Format Aerial Photography: Principles, Techniques and Geoscience Applications*, 1st ed. New York, NY, USA: Elsevier, 2016.

[7] D. P. Paine and J. D. Kiser, *Aerial Photography and Image Interpretation*. Hoboken, NJ, USA: Wiley, 2012.

[8] Institut Géographique National, "Remonter le temps," 2016. [Online]. Available: https://remonterletemps.ign.fr/. Accessed on: Mar. 11, 2019.

[9] P.-A. Herrault *et al.*, "Combined effects of area, connectivity, history and structural heterogeneity of woodlands on the species richness of hoverflies (Diptera: Syrphidae)," *Landscape Ecol.*, vol. 31, no. 4, pp. 877–893, 2015.

[10] R. C. Heller *et al.*, "Identification of tree species on large-scale panchromatic and color aerial photographs," U.S. Department of Agriculture, Forest Service, Washington, DC, USA, 1964.

[11] F. Palsson, J. R. Sveinsson, J. A. Benediktsson, and H. Aanaes, "Classification of pansharpened urban satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 281–297, Feb. 2012.

[12] G. Cavallaro, M. D. Mura, J. A. Benediktsson, and A. Plaza, "Remote sensing image classification using attribute filters defined over the tree of shapes," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3899–3911, Jul. 2016.

[13] Q. Feng, J. Liu, and J. Gong, "UAV remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote Sens.*, vol. 7, no. 1, pp. 1074–1094, 2015.

[14] J. Yumibe, *Moving Color: Early Film, Mass Culture, Modernism* (Techniques of the Moving Image). New Brunswick, NJ, USA: Rutgers Univ. Press, 2012.

[15] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 34–41, Jul./Aug. 2001.

[16] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, 2002.

[17] R. Irony, D. Cohen-Or, and D. Lischinski, "Colorization by example," in *Proc. 16th Eurograph. Conf. Rendering Techn.*, 2005, pp. 201–210.

[18] G. Charpiat, M. Hofmann, and B. Schölkopf, *Automatic Image Colorization via Multimodal Predictions* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5304. New York, NY, USA: Springer, 2008, part 3, pp. 126–139.

[19] X. Liu *et al.*, "Intrinsic colorization," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 152:1–152:9, 2008.

[20] A. Y.-S. Chia *et al.*, "Semantic colorization with internet images," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 156:1–156:8, 2011.

[21] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, p. 369.

[22] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1214–1220, 2006.

[23] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1120–1129, May 2006.

[24] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Proc. 18th Eurograph. Conf. Rendering Techn.*, 2007, pp. 309–320.

[25] D. Sýkora, J. Dingliana, and S. Collins, "LazyBrush: Flexible painting tool for hand-drawn cartoons," *Comput. Graph. Forum*, vol. 28, no. 2, pp. 599–608, 2009.

[26] A. Bugeau and V. Ta, "Patch-based image colorization," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 3058–3061.

[27] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 567–575.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[29] J. Žbontar and Y. Le Cun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2015, vol. 1, pp. 1592–1599.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, vol. 1, pp. 1097–1105.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.

[32] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 105–114.

[33] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 415–423.

[34] M. Agrawal and K. Sawhney, "Exploring convolutional neural networks for automatic image colorization," Stanford Univ., Stanford, CA, USA, Tech. Rep. n°409, 2016.

[35] A. Deshpande, J. Lu, M.-C. Yeh, M. Jin Chong, and D. Forsyth, "Learning diverse image colorization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6837–6845.

[36] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!" *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.

[37] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 5967–5976.

[38] G. Larsson, M. Maire, and G. Shakhnarovich, *Learning Representations for Automatic Colorization* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9908. New York, NY, USA: Springer, 2016, pp. 577–593.

[39] M. Limmer and H. P. Lensch, "Infrared colorization using deep convolutional neural networks," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl.*, 2016, pp. 61–68.

[40] D. Varga and T. Szirányi, "Fully automatic image colorization based on convolutional neural network," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 3691–3696.

[41] R. Zhang, P. Isola, and A. A. Efros, *Colorful Image Colorization* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9907. New York, NY, USA: Springer, 2016, pp. 649–666.

[42] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, *Unsupervised Diverse Colorization via Generative Adversarial Networks* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10534. New York, NY, USA: Springer, 2017, pp. 151–166.

[43] K. Frans, "Outline colorization through tandem adversarial networks," 2017, *arXiv:1704.08834*.

[44] K. Fu, Y. Wang, and B. Liu, "CS229 final project automatic colorization for line arts," Stanford University, Stanford, CA, USA, 2017.

[45] S. Lal, V. Garg, and O. P. Verma, "Automatic image colorization using adversarial training," in *Proc. 9th Int. Conf. Signal Process. Syst.*, 2017, pp. 84–88.

[46] A. Royer, A. Kolesnikov, and C. H. Lampert, "Probabilistic image colorization," 2017, *arXiv:1705.04258*.

[47] Q. Song, F. Xu, and Y. Jin, "Radar image colorization: Converting single-polarization to fully polarimetric using deep neural networks," *IEEE Access*, vol. 6, pp. 1647–1661, 2018.

[48] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Learning to colorize infrared images," in *Proc. Int. Conf. Practical Appl. Agents Multi-Agent Syst.*, 2017, vol. 619, pp. 164–172.

[49] D. Varga and T. Szirányi, "Twin deep convolutional neural network for example-based image colorization," in *Computer Analysis of Images and Patterns*, M. Felsberg, A. Heyden, and N. Krüger, Eds. Cham, Switzerland: Springer, 2017, pp. 184–195.

[50] H. Liu, Z. Fu, J. Han, L. Shao, and H. Liu, "Single satellite imagery simultaneous super-resolution and colorization using multi-task deep neural networks," *J. Vis. Commun. Image Represent.*, vol. 53, pp. 20–30, 2018.

[51] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[52] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[53] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*.

[54] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*.

[55] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*.

[56] A. Crepin, "Land cover interpretation in 2011-2012 and measurement of urban sprawl in Alsace," SIRS, Villeneuve d'Ascq, France, Rep. v130614, 2013.

[57] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[58] S. van der Walt *et al.*, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, 2014, Art. no. e453.

[59] D. Margulis, *Photoshop LAB Color: The Canyon Conundrum and Other Adventures in the Most Powerful Colorspace*. Berkeley, CA, USA: Peachpit Press, 2005.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[61] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[62] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*.

[63] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[64] A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vision Image Understanding*, vol. 179, pp. 41–65, 2019.

[65] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[66] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[67] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[68] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[69] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: Closing the generalization gap in large batch training of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1731–1741.

[70] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," 2017, *arXiv:1711.00489*.

[71] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[72] R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image Video Process.*, vol. 5, no. 1, pp. 81–91, 2009.

[73] A. Kolaman and O. Yadid-Pecht, "Quaternion structural similarity: A new quality index for color images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1526–1536, Apr. 2012.

[74] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2223–2232.

[75] E. Maltezos, N. Doulamis, A. Doulamis, and C. Ioannidis, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042620.

[76] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," 2015, *arXiv:1511.06390*.

[77] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4676–4689, Sep. 2018.

[78] C. Fern and T. A. Warner, "Scale and texture in digital image classification," *Photogramm. Eng. Remote Sens.*, vol. 68, no. 1, pp. 51–63, 2002.

[79] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[81] S. Barratt and R. Sharma, "A note on the inception score," 2018, *arXiv:1801.01973*.

Quentin Poterek received the B.Sc. degree in geography and M.Sc. degree in geomatics and earth observation, in 2016 and 2018, respectively, from the University of Strasbourg, Strasbourg, France, where he has been working toward the Ph.D. degree with the Laboratoire Image Ville Environnement Laboratory, since 2019.

His research interests include remote sensing, archive spatial products processing, machine learning, and landscape ecology.

Pierre-Alexis Herrault received the M.Sc. degree in geomatics from the University of Rennes, Rennes, France, in 2011, and the Ph.D. degree in geography from the University of Toulouse, Toulouse, France, in 2015.

His thesis focused on the reconstruction of forest changes to evaluate the effects of history on current biodiversity. From 2015 to 2017, he was a Postdoctoral Research Associate with the Centre d'Études Spatiales de la Biosphère Laboratory. Since 2017, he has been an Associate Professor with the Laboratoire Image Ville Environnement, University of Strasbourg, Strasbourg, France. His research interests include remote sensing for landscape ecology and estimating biodiversity, old spatial data processing and uncertainty assessment in land cover changes from spatial data.

Grzegorz Skupinski received the engineering degree in cartography and land surveying from the Wroclaw Academy of Agriculture, Wroclaw, Poland, in 2005, and the M.Sc. degree in environmental geography from the University of Strasbourg, Strasbourg, France, in 2007.

Since 2010, he has been an Engineer with the Laboratoire Image Ville Environnement Laboratory, University of Strasbourg. His research interests include field surveying, processing dense time series of satellite images, historical analysis of impervious surfaces, LiDAR processing for landslide monitoring, and web GIS technologies.

David Sheeren received the B.Sc. degree in geomatics and land surveying from the University of Liège, Liège, Belgium, in 1999, the M.Sc. degree in GIS from the French National School of Geographic Sciences (ENSG), Paris, France, in 2001, and the Ph.D. degree in computer science (AI) from the University of Paris 6, Paris, France, in 2005.

His Ph.D. thesis was prepared at the ex-COGIT Laboratory, Institut Géographique National, the French National Mapping Agency and focused on spatial database integration. In 2005, he was a Postdoctoral Research Fellow in Remote Sensing with the ICube Laboratory and LIVE Laboratory, CNRS/University of Strasbourg, Strasbourg, France. Since 2006, he has been an Associate Professor in Geo-Information Sciences with Toulouse INP-ENSAT (Engineering Faculty of Life Sciences), Castanet-Tolosan, France, and joined the interdisciplinary research unit DYNAFOR (INRAE). His research interests include remote sensing for landsacpe ecology, image analysis for mapping forest ecosystems and agro-ecological infrastructures, and the modeling of species-habitat relationships with earth observation data. He is also motivated by advanced methods of machine learning and their application in habitat mapping and biodiversity monitoring.