POLITECNICO DI TORINO
Repository ISTITUZIONALE

Bridging the gap between Natural and Medical Images through Deep Colorization

Original
Bridging the gap between Natural and Medical Images through Deep Colorization / Morra, L., Piano, L., Lamberti, F., Tommasi, T.. - STAMPA. - (2021), pp. 835-842. (25th International Conference on Pattern Recognition, ICPR 2020 Milano (Italy) 10-15 gennaio 2021) [10.1109/ICPR48806.2021.9412444].

Availability:
This version is available at: 11583/2837062 since: 2022-04-06T08:57:42Z

Publisher:
IEEE

Published
DOI:10.1109/ICPR48806.2021.9412444

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

14 November 2026

Bridging the gap between Natural and Medical Images through Deep Colorization

Lia Morra[1], Luca Piano[1], Fabrizio Lamberti[1], Tatiana Tommasi[1,2]
[1] Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129, Torino, Italy
[2] Istituto Italiano di Tecnologia, Torino, Italy
email: {name.surname}@polito.it

arXiv:2005.10589v2 [cs.CV] 19 Oct 2020

***Abstract*—Deep learning has thrived by training on large-scale datasets. However, in many applications, as for medical image diagnosis, getting massive amount of data is still prohibitive due to privacy, lack of acquisition homogeneity and annotation cost. In this scenario, transfer learning from natural image collections is a standard practice that attempts to tackle shape, texture and color discrepancies all at once through pretrained model fine-tuning. In this work, we propose to design a dedicated network module that focuses on color adaptation, thus pre-processing the input into a form (RGB) that is closer to the domain the classification backbone was trained on. We combine learning from scratch of the color module with transfer learning of different classification backbones, obtaining an end-to-end, easy-to-train architecture for diagnostic image recognition on X-ray images. Extensive experiments showed how our approach is particularly efficient in case of data scarcity and provides a new path for further transferring the learned color information across multiple medical datasets.**

I. Introduction

In the last years, deep learning has led to major breakthroughs in computer vision, even touching superhuman performance in some task [1]. One of the crucial ingredients of this success is the availability of large curated image collections on which deep models can be trained. Thanks to the widespread diffusion of RGB cameras, collecting massive amount of images is relatively easy: the expensive part remains getting reliable labels for the image content. This becomes particularly challenging when moving beyond natural images towards specific applications dealing with multispectral information as remote sensing [2] or single channel data as in medical diagnosis [3]. In particular, the latter needs access to private hospital records and costly manual annotation from experienced doctors. Only very recently some efforts have been done to provide the community with large reference medical datasets [4], [5], which still cover only some specific acquisition modalities, body parts and pathological markers.

One way to compensate for the lack of extensive data collections is to transfer knowledge from domains where data are abundant. Although appealing, being able to leverage on photo-based source models for diagnostic tasks means developing approaches to bridge the gap between natural and medical images, and deal with distribution shifts due to texture, shape as well as color. The most standard strategy for this goal is to exploit existing network architectures (*e.g.* ResNet [6], DenseNet [7]) pretrained on large datasets like ImageNet [8], and simply fine-tune their parameters with a limited amount of medical images [9]. In this way the network maintains a fixed knowledge capacity and the original model's weights are progressively forgotten in favor of those useful to the new domain. A general trade off is adopted between the freedom provided to weight modification and the cardinality of the training data: only larger collections support a full learning of the model parameters and in those cases the pretrained model remains a relevant tool to reduce the training time with respect to learning from scratch.

In our work we propose a new transfer strategy that combines the benefit of learning from scratch with that of exploiting pretrained source models, by focusing on color adaptation. The information coded in the three standard color channels for natural images is extremely different from that saved in grayscale X-ray scans, and we show how recomposing this shift provides an essential support to learning with limited annotated data. Specifically, we introduce an efficient colorization module that can be combined with different backbones whose weights are pretrained on ImageNet. The result is an end-to-end, easy-to-train architecture in which the color module is learned from scratch while the backbone can be either kept fixed or finetuned for diagnostic multi-label recognition. Once learned, the colorization module can be further transferred and re-used on medical tasks similar to the primary one (see Fig. 1 for a schematic overview).

In the following we first provide a brief summary on previous work about transfer learning in medical imaging, also presenting an overview on existing image colorization approaches (see sec. II). Then in section III we describe the details of our multi-stage approach, discussing on how to design the colorization module and presenting three variants for it. Finally section IV introduces our experimental setup and section V showcases the performance of the proposed method on three state of the art X-ray datasets. Finally section VI summarizes our findings and provides directions for future research.

II. Related Work

A. Transfer learning in medical imaging

Convolutional Neural Networks (CNNs) trained on ImageNet have proven to transfer extremely well to a large variety of medical imaging tasks [9], [10], [11], [12], [13], [14], to the point that transfer learning from ImageNet has rapidly become a standard practice in the field [15]. However, medical images

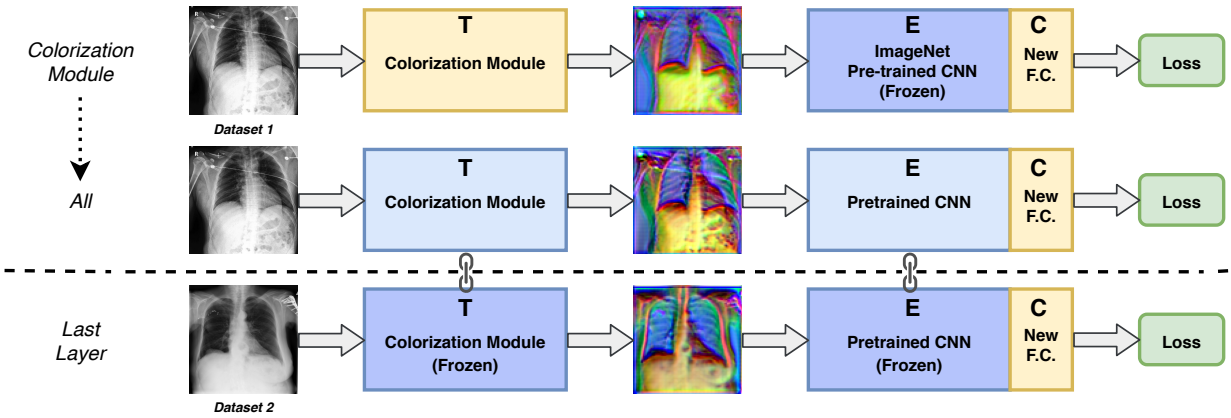

Fig. 1. A schematic representation of the proposed multi-stage transfer learning pipeline. First, the colorization module T is trained from scratch together with the classifier C, while the pre-trained CNN backbone E is kept frozen, to learn the mapping which maximizes classification accuracy. Then, the entire network is fine-tuned to learn useful features for the target task, while simultaneously adjusting the colorization mapping. Finally, the complete trained network is evaluated on a *different* dataset: here we freeze the $\{T, E\}$ modules, and train only the final classification layer C.

have distinctive features compared to RGB images: they are usually grayscale, often 12-bit, high resolution images, with a strong textural content. Natural image classification relies predominantly on local contrast, due to changes in illumination conditions, whereas in medical images the intensity values are often informative of the class of tissue class or pathology, and in selected cases may be even standardized. Thus, a reasonable question is whether transfer learning from medical images is more effective than transfer learning from RGB (non medical) images or, more in general, which strategies are available to facilitate the transfer from one domain to the other.

Few head-to-head comparisons are available on *medical vs. non-medical source transfer*, and their results are not always conclusive, also due to different data cardinality, transfer strategies and imaging modalities. A study by Cheplygina *et al*. [16] reviewed twelve papers comparing difference source-target combinations with contrasting results: roughly half of the works indicate that transferring from the non-medical domain outperforms transferring from another medical dataset, with the other half supporting the opposite thesis. Evidence suggests that larger data sets are not necessarily better for pre-training, and that diversity may play an important role, whereas medical datasets tend to be very homogeneous compared to their RGB counterparts [9]. Pre-training from medical datasets is preferable mostly across the same modality [17] and across the same body part [13], [12]. Previous experiments on Chest X-ray [13] showed that features learned on medical datasets may be quite specific to a given anatomical site or modality. In fact, while pre-training on another chest X-ray dataset yielded remarkable performance even when fine-tuning on very small datasets, pre-training on images from other anatomical sites (e.g., musculoskeletal X-ray) did not offer any advantage compared to pre-training on ImageNet.

Fewer studies have compared *transfer learning with training from scratch*. They suggested that fine-tuning pre-trained models is at least as good as training from scratch, and offers substantial advantages for small scale datasets [11], [16]. In recent works on large scale datasets, fine-tuning offered no advantage in terms of performance, but still allowed a remarkable speedup in training [18]. Nonetheless, collecting large scale datasets for all possible modalities and diseases is simply unfeasible, and transfer learning from ImageNet remains one of the prominent strategies for training deep learning models on medical tasks, especially for 2D images.

Alternative transfer solutions are those based on *pseudo-colorization*, but the related literature is rather limited. Pseudo-color images can be generated from CT scans by applying different handcrafted windows/level settings on each channel, in a way similar to how radiologists enhance the contrast of different tissues [11]. This technique, however, cannot be transferred to other modalities, such as X-ray, where the intensity values are not standardized. Teare *et al*. designed a genetic algorithm to discover a pseudo-color enhancement scheme for mammography [19]: each genome encodes three sequences of preprocessing functions along with their parameters (*e.g*., blurring, contrast enhancement, masking), one for each color channel, and the preprocessing that maximizes the accuracy of the CNN is selected. This approach, while resulting in a substantial increase in accuracy, is computationally expensive, since it requires training and testing many different networks. To reduce the computational requirements, it is necessary to use a shallower network or a smaller dataset for this step.

Our work relates to all transfer learning directions described above, but introduces a new colorization-based method for leveraging on non-medical sources. We propose to let the network learn its own optimal color transformation to close the domain gap while solving the medical recognition task of interest. In this way, our approach can take full advantage of all the data available for training at the low cost of updating only a dedicated part of the network, it does not require any strong assumption on which transformations are potentially useful, and can be seamlessly adapted to any type of input modality and any pre-trained CNN architecture.

B. Image colorization

Image colorization methods generate RGB images starting from a single channel or grayscale one. Unsurprisingly, in recent years, CNN-based methods have largely outperformed shallow learning models based on hand-crafted features, by leveraging massive amount of natural images from which the color is artificially removed. Indeed, the most common approaches learn a parametric mapping between pairs of corresponding grayscale and RGB images, either explicitly coding the probability distribution of each color [20], [21], or doing it implicitly through Generative Adversarial Networks [22], [23]. Of course, those strategies cannot be pursued in the medical domain, where the target color space is unknown.

Besides the works mainly related to style transfer, colorization techniques have been also investigated for several industrial applications on thermal images [24] as well as in robotics where single channel depth images are widely used [25], [26]. Of particular interest is the work in [27], in which the authors recently proposed to hallucinate colors from depth through a process guided directly by depth image classification. The adopted architecture can take full advantage of the pre-trained ImageNet-based models given that, although significantly different in texture, depth images contain similar objects shapes as their RGB counterparts. Inspired by this work, we propose to go beyond it, defining tailored colorization modules for medical images and a multi-stage transfer procedure involving end-to-end update of the main pretrained backbone.

III. Transfer learning through colorization

A. Learning Strategy

Let us assume to observe $\mathcal{D} = \{x_i, \boldsymbol{y}_i\}_{i=1}^N$ labeled data, where x_i is a grayscale medical image and $\boldsymbol{y}_i \subseteq \{-1, 1\}^C$ is its multi-label annotation. Specifically each annotation describes presence (1) or absence (-1) for a set of C pathological clues. Our final goal is to train a classification model able to predict the correct label for each new test sample. To simplify the explanation of the learning procedure, we assign different names to the main components of our deep architecture. We indicate with E the feature encoder, corresponding to the main convolutional backbone of the network. We use C to specify the final fully connected part of the architecture, whose role is to classify the sample given the features produced by E. Standard feature extractors take as input colorful images, thus are not suited to work with grayscale samples. We propose to learn the best colorization function $T : x \rightarrow \boldsymbol{x}$ such that the obtained colorful sample $\boldsymbol{x}$ can be processed by E and then produce the highest classification performance through C. All the parameters of the learning architecture are $\{\theta_E, \theta_C, \theta_T\}$, with the subscript indicating the respective component.

As a first step we leverage on an encoder model $E\big(\boldsymbol{x}, \theta_E^I\big)$ pre-trained on ImageNet, keeping it fixed and updating only T and C from scratch. Thus, the objective function is

$$\Big\{\theta_T^M, \theta_C^M\Big\} = \underset{\theta_T, \theta_C}{\arg\min} \sum_{i=1}^{N} \mathcal{L}\big(C\big(E\big(T(x_i, \theta_T), \theta_E^I\big), \theta_C\big), \boldsymbol{y}_i\big) \ , \tag{1}$$

where $\mathcal{L}$ is the binary cross-entropy loss. In the following we will indicate this learning procedure as *Color Module* (superscript M) since the focus is on learning to colorize the original grayscale images. Starting from the obtained model, we also define a second learning stage where T and E are fine-tuned, while C is learned from scratch to solve the following optimization problem

$$\Big\{\theta_T^A, \theta_E^A, \theta_C^A\Big\} = \underset{\theta_T, \theta_E, \theta_C}{\arg\min} \sum_{i=1}^{N} \mathcal{L}\big(C\big(E\big(T(x_i, \theta_T), \theta_E\big), \theta_C\big), \boldsymbol{y}_i\big) \ , \tag{2}$$

initialized with $\theta_T = \theta_T^M$, $\theta_E = \theta_E^I$. In the following we indicate this procedure as *All* (superscript A).

Till here we were dealing only with a colorization strategy for transfer learning on medical images from an RGB large scale collection as ImageNet. Still, the knowledge acquired in this way can also be shared across different medical image sets. Indeed, we can transfer the learned component T, when facing new labeled observations $\mathcal{B} = \{x_i, \boldsymbol{y}_i\}_{i=1}^{N_B}$, with $\mathcal{B} \neq \mathcal{D}$. Here the difference among the datasets can be of various nature: similar body parts and task but different annotation procedure, or even completely different body parts and labeling. By transferring T and keeping it fixed together with E, the problem reduces to solving

$$\Big\{\theta_C^L\Big\} = \underset{\theta_C}{\arg\min} \sum_{i=1}^{N} \mathcal{L}\big(C\big(E\big(T(x_i, \theta_T), \theta_E\big), \theta_C\big), \boldsymbol{y}_i\big) \ . \tag{3}$$

We use the tag *Last Layer* (superscript L) to indicate this strategy, which can have different variants, depending on which $\{T, E\}$ are used as starting point, with the two options $\{\theta_T^M, \theta_E^I\}$ or $\{\theta_T^A, \theta_E^A\}$. Alternatively, the $\{T, E\}$ components could be transferred and further fine-tuned while learning C from scratch, with the same optimization problem in eq. (2).

At test time, the prediction on a new image will be obtained as $\tilde{\boldsymbol{y}} = \big(C\big(E\big(T(x, \theta_T^*), \theta_E^*\big), \theta_C^*\big)$ with the superscript * indicating the parameters obtained at the end of the chosen learning strategy.

B. Colorization Module

The main challenge at the basis of the described learning approach concerns how to design the colorization module T. Building over [27], we consider two architectures that use a convolution and pooling layer to transform the $x = \{1 \times 320 \times 320\}$ grayscale image into a $\boldsymbol{x} = \{64 \times 80 \times 80\}$ tensor, which is then given as input to a sequence of 8 residual blocks, as shown in Fig. 2. The two architectures differ in the final up-sampling part: we indicate as *DECONV* the case using a Transpose Convolution as in [27], and with *Pixel Shuffle* our modification based on the homonym technique originally developed for super-resolution applications [28].

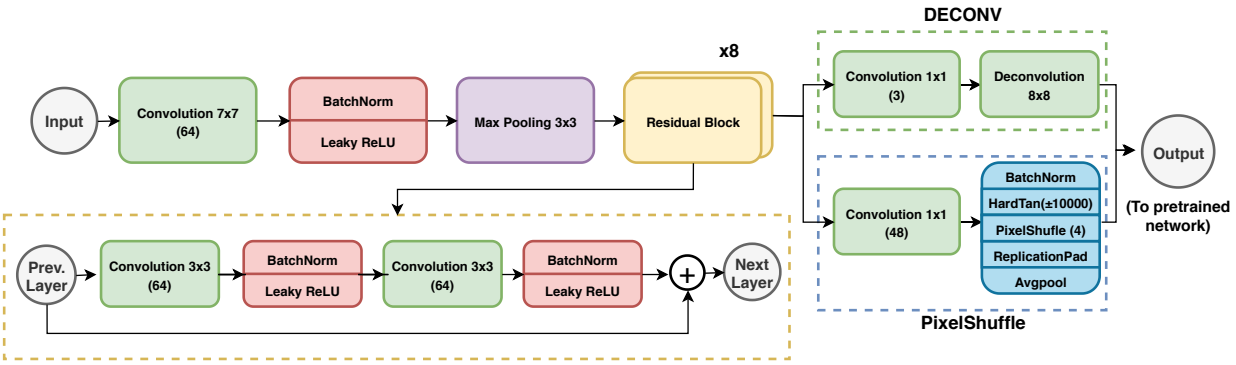

Fig. 2. Overview of the DECONV and Pixel Shuffle colorization modules: their share the main structure with a difference only in the final upsampling layer. Details of the residual block are shown in the dashed, bottom box.

This technique allows to reduce the checkerboard artifacts introduced by up-sampling.

We also defined a third variant of the colorization module[1], indicated in the following as *ColorU* (Fig.3). It takes inspiration from the U-Net architecture [29], [30] but presents different blocks. The input image passes through 3 down-sampling blocks (*ColorDown*), 2 up-sampling blocks (*ColorUp*) and another final up-sampling block (*ColorOut*) that generates the colorful output image. All these blocks use strided convolutions instead of spatial pooling functions, allowing the models to learn how to up/down-sample rather than using fixed methods, as proposed in [31]. Each ColorDown block is composed by two convolutional layers (conv with kernels 3×3 and 4×4) followed by a BatchNorm Layer. Each ColorUp block is composed by a Transpose Convolutional layer (4×4), followed by two conv layers and a BatchNorm layer. The ColorOut model has a similar structure to the ColorUp, but without the final BatchNorm layer. All conv layers have a Leaky ReLU activation, applied before BatchNorm, except for the final conv layer which outputs the colorized image. ColorU is much more efficient than DECONV and Pixel Shuffle, with a ratio of about $1/3$ in terms of needed learning parameters.

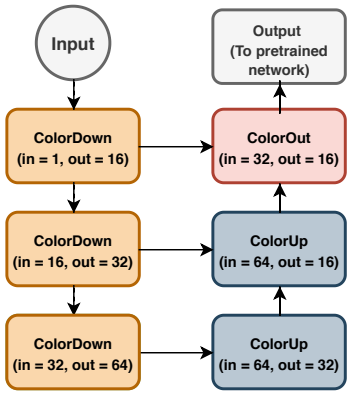

Fig. 3. ColorU architecture. The input image passes through 3 down-sampling blocks (*ColorDown*), 2 up-sampling blocks (*ColorUp*) and a final up-sampling and colorization block block (*ColorOut*). For each block, the number of input and output channels is reported.

IV. Experiments

A. Datasets

1) Chest X-ray: We experimented on two large chest X-ray datasets, CheXpert [4] and ChestX-ray14 [32].

The CheXpert training set contains 224,316 chest radiographs from 65,240 patients, whereas ChestX-ray14 contains 112,000 images from 32,717 patients. Both datasets are labeled for the presence or absence of 14 independent observations starting from radiological reports using two different semi-automatic labellers. We focus here on five pathological cues: Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion, which are common to both datasets and for which prior results were available in literature [4].

CheXpert ground truth takes into account uncertainty in the reporting and labeling process by assigning each observation one of three values: positive, negative and uncertain. Different policies were previously compared in [4] to handle uncertain labels during training: converting uncertain to positive (U-Ones) or negative values (U-Zeros), or ignoring uncertain samples (U-Ignore). For each observation, we applied the best policy based on previous results: Atelectasis (U-Ones), Cardiomegaly (U-Zeros), Consolidation (U-Zeros), Edema (U-Ones), and Pleural Effusion (U-Ones) [4]. We did not employ any specific training strategy to account for data imbalance during training.

2) MURA: MURA [5] comprises 40,561 multi-view musculoskeletal radiographs from about 12,000 patients, representing various body parts as either normal or abnormal. For our analysis we considered only one anatomical part, the shoulder, corresponding to 8,379 images (2,694 patients).

B. Experimental setup

We performed experiments on two baseline networks, with the E module of our architecture corresponding to either ResNet18 or DenseNet121.

When working on CheXpert and ChestX-ray14 the output layer is the combination of five binary heads ($C = 5$), one

[1]All our trained colorization modules are available at https://gitlab.com/grains2/DeepMedicalColorization

for each observation. For MURA the task reduces to a simple binary classification problem with $C = 1$.

Standard data augmentation is applied during training: non-square images are randomly cropped and rescaled to the network input size (320×320), then random rotation ($-10°$ – $+10°$) and zoom (0% – 10%) are applied with 75% probability. All the images were normalized with the mean and standard deviation calculated over CheXpert. All the networks were trained until convergence using the SGD optimizer with the One Cycle Policy [33]. The Learning Rate (LR) finder was used to determine the maximum learning rate for each network [34]. Weights were saved every 4800 iterations and we used the checkpoint with the highest validation performance.

C. Evaluation

Performance was assessed using the Area under the ROC curve (AUC). For multi-label datasets (CheXpert and ChestX-ray14), the AUC was calculated for each observation, and then the mean AUC was used to summarize the performance. For each configuration, the training was repeated three times with different random initialization. Paired t-test was used to compare different transfer learning procedures. We controlled for multiple hypothesis testing using the Benjamini–Hochberg procedure [35]; an adjusted p-value < 0.05 indicated statistical significance.

V. Results

A. Transfer learning through colorization

In this section we compare the different transfer learning procedures illustrated in section III with the standard approach, *i.e.* without the colorization module. Specifically, the baseline model takes as input a three channel image obtained by copying the single channel sample, and is constituted by the encoder E, followed by final fully connected component C. The baseline can be trained by freezing the weights θ_E^I (denoted in the following as *Baseline*) or fine-tuning all the layers (denoted in the following as *Baseline All*). For the colorization module T, we adopted the multi-stage transfer procedure described in Section III: first we train T from scratch (*Color Module*), and then both T and E are fine-tuned, while C is learned from scratch (*All*). We report the performance after each phase.

Average AUC values for the ResNet18 and DenseNet121 architectures are shown in Table I. Training the color module T with fixed E improves classification performance over the Baseline by roughly 7%: results are statistically significant for both ResNet18 (AUC=78.4 vs. 83.4, p=0.037) and DenseNet121 (AUC=78.6 vs. 84.3, p=0.043). These results confirm that color is indeed an important component of the domain shift from RGB to medical images. The obtained accuracy is still lower with respect to Baseline All, which however requests the update of 10^7 network parameters against the 10^5 needed with the Color Module solutions. A further fine-tuning of the whole network with the All strategy gets us back to the Baseline All results (ResNet18: AUC=89.6 vs. 88.9, p= 0.078; DenseNet121: AUC=89.8 vs. 89.6, p=0.142).

TABLE I
AVERAGE AUC FOR DIFFERENT TRANSFER LEARNING STRATEGIES ON CHEXPERT.

Colorization Module	Learning Strategy	Mean AUC
Backbone: ResNet18		
-	Baseline	78.4 ± 0.5
DECONV	Color Module	84.0 ± 0.3
PixelShuffle	Color Module	83.4 ± 0.4
ColorU	Color Module	83.9 ± 0.8
-	Baseline All	89.6 ± 0.2
DECONV	All	88.9 ± 0.3
PixelShuffle	All	88.9 ± 0.1
ColorU	All	89.3 ± 0.3
Backbone: DenseNet121		
-	Baseline	78.6 ± 0.2
DECONV	Color Module	83.5 ± 0.3
PixelShuffle	Color Module	84.3 ± 0.8
ColorU	Color Module	83.9 ± 0.5
-	Baseline All	89.8 ± 0.2
DECONV	All	89.2 ± 0.2
PixelShuffle	All	89.6 ± 0.1
ColorU	All	89.7 ± 0.3

TABLE II
RESULTS FOR INDIVIDUAL OBSERVATIONS ON THE CHEXPERT DATASET WITH DENSENET121. THE RESULTS FROM [4] ARE REPORTED FOR COMPLETENESS ALTHOUGH OBTAINED WITH AN ENSEMBLE OF 30 CHECKPOINTS.

Learning strategy	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
DenseNet121 [4]	85.8	84.0	93.2	94.1	93.4
Baseline All	**85.4 ± 0.5**	82.7 ± 0.6	**94.9± 1.0**	93.2 ± 0.3	89.8 ± 0.2
PixelShuffle All	84.7 ± 1.7	**85.0 ± 1.8**	93.8 ± 0.8	**93.8 ± 0.9**	**91.0 ± 0.6**

Regarding the specific choice of E, both DenseNet121 and ResNet18 achieve similar performance. DenseNet121 seems to offer a small advantage over ResNet18 in conjuction with PixelShuffle (AUC=89.6 vs. 88.9, p=0.10), but the baseline models are comparable (AUC=89.8 vs. 89.6, p=0.4). These findings are in agreement with previous studies discussing how networks characterized by different depth behave similarly on medical datasets [18]. This is likely due to the limited number of classes (compared to the thousand classes of ImageNet) and relatively homogeneous image content.

Regarding the specific choice of T, we can claim that all the considered variants offer similar advantages in terms of performance. We found a small, albeit not significant, advantage of ColorU over PixelShuffle for ResNet18 (AUC=83.9 vs. 83.4, p=0.132), but not for DenseNet121 (p=0.60); PixelShuffle is always equal or better than DECONV. Overall, differences are numerically small and suggest that colorization is stable across a wide range of architectures and encoders, including both very deep and more shallow networks.

A different point of view on the results is offered by the colorful images obtained as a byproduct of the learning process. As shown in Figure 4, they depend both on the specific used module T and on whether E is fixed or fine-tuned. The output of the colorization module is suggestive of the network learning process: anatomical features become more prominent

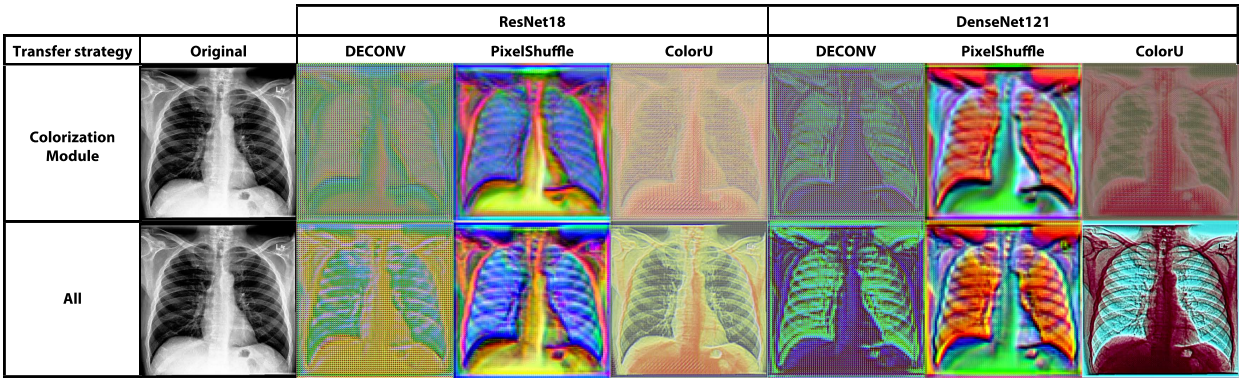

Fig. 4. Output of different colorization modules optimized for the ResNet18 and DenseNet121 architectures. In the top row, only the colorization module C is trained, freezing the encoder E. The bottom row shows results after fine-tuning all the layers end-to-end.

after the whole network is fine-tuned, suggesting that the pre-trained encoder is somewhat unable to identify important image features and more information on shape and texture need to be adapted besides the overall color. We underline that the goal of our work is not that of producing human-readable images, rather that of letting the network assign the most useful colors to get the maximum recognition accuracy. Whether and to what extent the colorization highlights areas that are relevant to the final diagnosis is an interesting avenue for future work.

Finally, *a per-class analysis* on the Pixel Shuffle All DenseNet121 results reveals that adding the colorization module increases the performance on Cardiomegaly (p=0.128) and Edema (p=0.024). On the other hand, Baseline All outperforms PixelShuffle All for Consolidation (p=0.07). Differences are less decisive for Atelectacsis (p=0.40) and Pleural Effusion (p=0.37) (see Table II). Only for this evaluation we are endorsing statistical significance for $p < 0.15$ to take into consideration the higher variability among different runs in every class. The obtained results suggest that different learning strategies may converge to different feature representations, with comparable overall performance, but complementary strengths and weaknesses.

B. Effect of training set size

The optimal transfer policy naturally depends on the size of the training set. Since the proposed colorization strategy allows to perform transfer learning on a budget of learning parameters with respect to whole network fine-tuning, we expect its advantage to become more visible in case of limited amounts of training data. CheXPert, while significantly smaller than ImageNet, is by far larger than most medical datasets. To study the effect of colorization in the small and very small data regimes, we conducted additional experiments on subsets ranging from 10% (22,431 images) to 0.1% (224) of the original CheXPert size. This range is representative of small dataset (5,000-10,000 images) to few-shot learning scenarios. Indeed, the distribution of positive, negative and uncertain labels varies widely among classes (see Figure 5) and, in the 0.1% scenario, the number of positive examples can

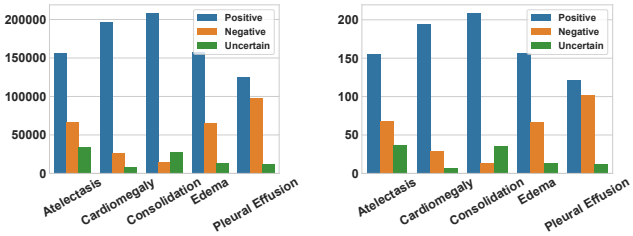

Fig. 5. Number of positive, negative and uncertain labels per observation in the entire CheXpert dataset (left) and in the 0.1% subset (right). The proportion of positive, negative and uncertain cases is the same across all dataset sizes.

be as low as 23. For each sampling size, three random datasets were selected; we report the average and standard deviation of the AUC across the three datasets. All experiments were conducted with the ResNet18 backbone and the PixelShuffle colorization module.

We compare the baseline and the colorization-enhanced network in two settings: when the encoder E is frozen (PixelShuffle Color Module vs Baseline), and when E is fine-tuned (Pixel Shuffle All vs Baseline All). As shown in Figure 6, regardless of the dataset size, the Baseline with frozen weights is outperformed by all the other configurations, indicating that even a very small dataset is sufficient to learn better feature representation for this task. Adding a colorization module consistently improves results in all data regimes, and the gap in performance is much higher in the small data regime (400 – 2,000 images).

PixelShuffle is equal (Color Module) or better (All) than Baseline All in the very small data regime (400 – 2,000 images). The proposed multi-stage transfer procedure (PixelShuffle All) outperforms training only the colorization module (PixelShuffle Color Module) across all dataset sizes, thus training T as a first step provides a more effective starting point for learning novel features useful for the new task. In the small data scenario, the network has limited ability to learn new features; instead, the cheap colorization module

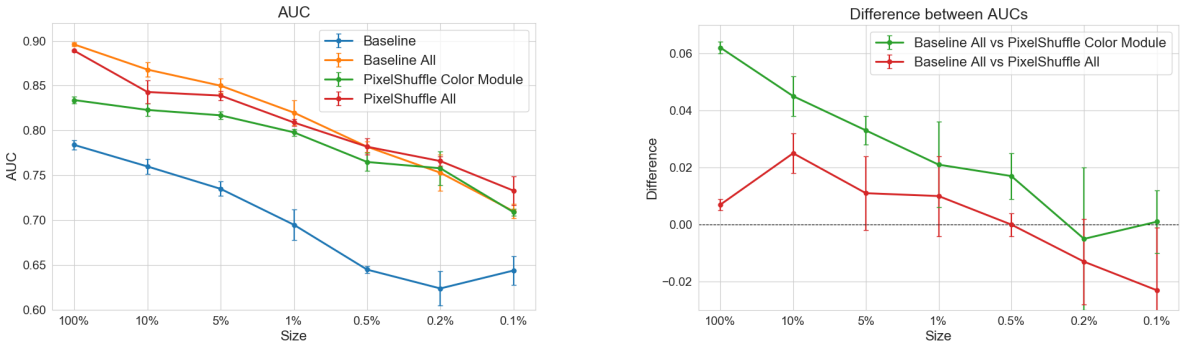

Fig. 6. Performance of different transfer strategies according to the dataset size. (a) Average AUC for four learning strategies: without colorization module (Baseline) and with PixelShuffle Color module. Results are reported both with the baseline is kept frozen (blue and green lines) as well as when fine-tuning the entire network (yellow and red lines). (b) Difference between the PixelShuffle architecture and the fine-tuned baseline network. When the all the architecture is fine-tuned (red line), PixelShuffle outperforms the baseline architecture in the limited data scenario. Training only the colorization module (green line) is less effective, but the difference progressively reduces for smaller datasets.

TABLE III
AVERAGE AUC FOR DIFFERENT TRANSFER LEARNING STRATEGIES ACROSS DATASETS. ALL THE EXPERIMENTS ARE PERFORMED WITH RESNET18 BACKBONE.

Backbone E	Colorization Module T	Learning strategy	Mean AUC
Mura			
(θ_E^I) ImageNet	-	Baseline	73.0 ± 0.11
(θ_E^I) ImageNet	PixelShuffle (θ_T^M CheXpert)	Last Layer	67.1 ± 1.2
ChestX-ray14			
(θ_E^I) ImageNet	-	Baseline	66.9 ± 0.1
(θ_E^I) ImageNet	PixelShuffle (θ_T^M CheXpert)	Last layer	68.3 ± 1.4
(θ_E^I) ImageNet	PixelShuffle (θ_T Scratch)	Color Module	72.9 ± 0.1
(θ_T^A) CheXpert	PixelShuffle (θ_T^A CheXpert)	Last layer	77.3 ± 0.5
(θ_E^I) ImageNet	-	Baseline All	79.9 ± 0.06
(θ_E^I) ImageNet	PixelShuffle (θ_T^M CheXpert)	All	79.4 ± 0.06
(θ_E^I) ImageNet	PixelShuffle (θ_T Scratch)	All	79.2 ± 0.3
(θ_T^A) CheXpert	PixelShuffle (θ_T^A CheXpert)	All	81.0 ± 0.1

can be effectively trained from scratch. On larger datasets, the advantage of the colorization module is reduced and the limited increase in parameter count may explain the lower performance. It should be noticed that all these experiments were conducted on a rather shallow backbone (ResNet18), and that the advantage of the proposed technique could be even higher on deeper models, where the chance of overfitting will be proportionally higher.

C. Transferability across datasets

We evaluated transferability of the colorization module T on two datasets, MURA and ChestX-ray14, representing two distinct scenarios: same modality/different body parts (MURA), and same modality/same body part (ChestX-ray14). Both MURA and ChestX-ray14 are smaller than CheXpert, which can therefore act as an intermediate step between RGB images and the medical domain. Multiple factors need to be taken into account in designing the transfer strategy: which encoder E is used as starting point ($\{\theta_T^M, \theta_E^I\}$ or $\{\theta_T^A, \theta_E^A\}$), whether T and E are frozen or further fine-tuned on the target dataset. As in previous experiments, the baseline model is defined by the encoder E and by the final layer C; the latter is trained from scratch in all configurations.

On MURA, when starting from the initial configuration $\{\theta_T^M, \theta_E^I\}$, the performance degrades with respect to the baseline (AUC=67.1 vs. 73.0, p=0.040). Differently, on ChestX-ray14, the performance largely increases when the colorization module is transferred ($\{\theta_T^M, \theta_E^I\}$, AUC=68.3 vs. 66.9, p=0.005), as well as when both colorization module and backbone are transferred ($\{\theta_T^A, \theta_E^A\}$, AUC=77.3 vs. 66.9, p=0.017). These results strongly suggest that the colorization module transfers well to similar datasets, but not across different body parts. Recent literature have also discussed how transferring from ImageNet is better than transferring from a medical source collection capturing a different body part than the target one [13]. This behavior likely arises from the lack of variety in the CheXpert dataset, which induces the network to learn very specific features and/or colorization patterns. Hence, we did not further insist on MURA, moving our focus towards more experiments on ChestX-ray14.

Transferring both the colorization module and backbone from CheXpert ($\{\theta_T^A, \theta_E^A\}$, Pixel Shuffle Last Layer) achieves slightly inferior results with respect to the Baseline All on ChestX-ray14 (AUC=77.3 vs. 79.9, p=0.016). The best results are obtained when the network is pre-trained on CheXpert and then fine-tuned on ChestX-ray14 ($\{\theta_T^A, \theta_E^A\}$, Pixel Shuffle All, AUC=81.0 vs. 79.9, p=0.008). We remark that, even though the task between CheXpert and ChestX-ray14 is the same, there is still a considerable domain shift between the two datasets, both due to the acquisition and pixel intensity distribution, as well as to the labelling process. Nonetheless, pre-training on CheXpert, also through the colorization module, can offer a significant boost in performance, which we expect would be even higher across more similar datasets or in case of extremely reduced data availability.

VI. CONCLUSIONS

This paper presented a novel strategy for transfer learning from the non-medical to the medical domain. It leverages a deep colorization module to learn an optimal color mapping for a given pre-trained convolutional neural network. We

propose a multi-stage transfer learning procedure where the lightweight colorization module is first trained from scratch, keeping the backbone frozen, followed by a refinement stage where both modules are fine-tuned towards the target task.

Our approach produces a significant gain in performance when the feature encoder is frozen with two different backbones and over two datasets. These results indicate that learning how to colorize medical images already helps to reduce the gap with RGB data from the non-medical domain with a minimal computational effort. Further fine-tuning the backbone is helpful to fully adapt to differences in shape and texture, besides color. This is confirmed by visual analysis of the hallucinated color images.

The experiments in the small and very small data regime highglighted how our colorization-based multi-stage transfer procedure is particularly effective when the target dataset is small, which is often the case in many biomedical applications. When a large dataset is available, as in the case of CheXpert, it opens a new path for transferring learned color information to similar medical datasets (*i.e*. containing the same body part).

Future work will investigate the effectiveness and generality of our approach on different medical imaging datasets, different pre-trained backbones, and different tasks beyond classification. The proposed colorization modules can be easily extended to manage multiple inputs, thus offer a promising solution for leveraging on pre-existing source models even when the target samples are 3D and multi-channel (*e.g*. magnetic resonance imaging).

REFERENCES

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, 2016.

[2] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, June 2013.

[3] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[4] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI*, 2019.

[5] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," in *MIDL*, 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[9] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[10] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[11] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[12] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, C. D. Richter, and K. H. Cha, "Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets," *IEEE Transactions on Medical Imaging*, vol. 38, no. 3, pp. 686–696, 2018.

[13] M. Romero, Y. Interian, T. Solberg, and G. Valdes, "Targeted transfer learning to improve performance in small medical physics datasets," *Medical Physics*, 2020.

[14] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," in *ISBI*, 2017.

[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[16] V. Cheplygina, "Cats or CAT scans: transfer learning from natural or medical image source datasets?" *Current Opinion in Biomedical Engineering*, 2019.

[17] H. G. Kim, Y. Choi, and Y. M. Ro, "Modality-bridge transfer learning for medical image classification," in *CISP-BMEI*, 2017.

[18] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *NeurIPS*, 2019.

[19] P. Teare, M. Fishman, O. Benzaquen, E. Toledano, and E. Elnekave, "Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement," *Journal of digital imaging*, vol. 30, no. 4, pp. 499–505, 2017.

[20] R. Zhang, P. Isola, and A. Efros, "Colorful image colorization," in *ECCV*, 2016.

[21] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *ECCV*, 2016.

[22] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[23] X. Yang, D. Xie, and X. Wang, "Crossing-domain generative adversarial networks for unsupervised multi-domain image-to-image translation," in *ACM International Conference on Multimedia*, 2018.

[24] D. Colucci, L. Morra, X. Zhang, D. Fissore, and F. Lamberti, "An automatic computer vision pipeline for the in-line monitoring of freeze-drying processes," *Computers in Industry*, vol. 115, p. 103184, 2020.

[25] U. Asif, M. Bennamoun, and F. A. Sohel, "RGB-D object recognition and grasp detection using hierarchical cascaded forests," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 547–564, June 2017.

[26] M. Gao, J. Jiang, G. Zou, V. John, and Z. Liu, "RGB-D-Based object recognition using multimodal convolutional neural networks: A survey," *IEEE Access*, vol. 7, pp. 43 110–43 136, 2019.

[27] F. M. Carlucci, P. Russo, and B. Caputo, "DE^{2}CO: Deep depth colorization," in *ICRA*, 2018.

[28] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015.

[30] V. Billaut, M. de Rochemonteix, and M. Thibault, "ColorUNet: A convolutional classification approach to colorization," *Preprint ArXiv:1811.03120*, 2018.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR workshop track*, 2015.

[32] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases," in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*. Springer, 2019, pp. 369–392.

[33] L. N. Smith and N. Topin, "Super-convergence: Very fast training of residual networks using large learning rates," *Preprint ArXiv:1708.07120*, 2017.

[34] L. N. Smith, "Cyclical learning rates for training neural networks," in *WACV*, 2017.

[35] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.