Is It a Plausible Colour?
UCapsNet for Image Colourisation

Rita Pucci
Department of Computer Science
University of Udine
rita.pucci@uniud.it

Christian Micheloni
Department of Computer Science
University of Udine
Christian.Micheloni@uniud.it

Gian Luca Foresti
Department of Computer Science
University of Udine
gianluca.foresti@uniud.it

Niki Martinel
Department of Computer Science
University of Udine
niki.martinel@uniud.it

Abstract

Human beings can imagine the colours of a grayscale image with no particular effort thanks to their ability of semantic feature extraction. Can an autonomous system achieve that? Can it hallucinate plausible and vibrant colours? This is the colourisation problem. Different from existing works relying on convolutional neural network models pre-trained with supervision, we cast such colourisation problem as a self-supervised learning task. We tackle the problem with the introduction of a novel architecture based on Capsules trained following the adversarial learning paradigm. Capsule networks are able to extract a semantic representation of the entities in the image but loose details about their spatial information, which is important for colourising a grayscale image. Thus our UCapsNet structure comes with an encoding phase that extracts entities through capsules and spatial details through convolutional neural networks. A decoding phase merges the entity features with the spatial features to hallucinate a plausible colour version of the input datum. Results on the ImageNet benchmark show that our approach is able to generate more vibrant and plausible colours than exiting solutions and achieves superior performance than models pre-trained with supervision. **Keywords:**Colourisation, Vision for Graphics, CapsNet, Capsules, UNet, Encoder, Decoder, Self supervised learning

1 Introduction

In colourisation, an observer is supposed to know which colours to add to a grayscale image to make it accurate and realistic. Willing to deal with colourisation with an autonomous model, we aim to obtain a realistic/plausible colourisation for a grayscale image by the means of models based on statistical dependencies between the semantic of the objects in the image and its grayscale texture. Existing works mainly rely on Convolutional Neural Networks (CNNs) to obtain de-saturated image colours [2, 1, 3], with a few exceptions [13, 10, 22, 11] achieving

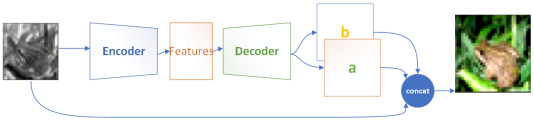

Figure 1: Exemplar model structure: the grayscale input image is goes into an encoded that extracts features of the entities present in the image. The encoded representation is then decoded to predict the channels responsible of colourisation.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

plausible colourisation for a limited set of samples through auxiliary losses. We believe that such unsatisfactory results are due to the limited abilities of CNNs in understanding the semantic information of an entity in relation to its parts [18]. Indeed, it is a matter of fact that colours are closely related to the semantics of a scene. This motivates us to introduce an approach that grasps semantic information and generates a plausible colourisation that will potentially fool an observer. We cast the colourisation problem as a self-supervised learning task (Fig.1) under the adversarial learning framework. Specifically, we introduce the **UCapsNet** architecture that leverages the convolutional operators to get spatial features and combines these with capsules [8, 18] for the semantic features extraction process. A capsule consists of a group of neurons that collaborate to generate anactivity vector. The activity vector represents the probability of an entity existence in the image (length of the vector) and its instantiated parameters/features (direction of the vector). We entangle a strong interaction between the spatial features and the entities information through an encoder-decoder solution that exploits skip connections. Such a colour generator shows a spatial and semantic comprehension of the entities present in the images and a consistent ability of colourisation that is optimised through the adversarial min-max game with a discriminator. Results on ImageNet show that UCapsNet outperforms existing solutions (some of which exploits pre-trained models).

2 Related Work

Colourisation algorithms propose to extract data useful to map the grayscale channel onto the coloured ones. The literature can be clustered into non-parametric and parametric methods. The former exploits image analogies based on a multiscale autoregression between two images, where colours are transferred onto the input image taking into consideration analogous regions of the other images [9, 21, 7, 14, 23]. The latter extracts knowledge from training datasets, it can be later applied for colourisation, in this paper our approach embrace the same idea. In [3], authors apply the LEARCH framework to balance pixelwise accuracy and spatial error and scene labelling to pick the appropriate scene specific regressor. In [13], a fully connected pre-trained VGG16 model incorporates semantic features and a colour histogram framework to predict colourisation. Taking up the idea of collaboration between pre-trained layers, [6] introduces a model of Dense Blocks structured following the UNet [16] architecture. With the idea of considerign semantic and spatial information, in [10], four component collaborate to extract classification information along side chrominance features by the means of CNNs layers that merge local patch information with the global one obtained from the entire image. In [22], a model based on CNNs layers is trained to learn a quantised representation of colours per each pixel. In [11], a Generative Adversarial Model (GAN) is trained to obtained a complete autonomous colourisation system. In [15], a pre-trained VGG19 [19] model is added to the CapsNet [18] architecture and fine-tuned for colourisation. Differently from all such methods, UCapsNet (i) explores the collaboration between CNNs and capsules layers to merge spatial and semantic features (ii) with the aim of generating a plausible colorization (iii) by means of a self-supervised learning solution that does not hinge on supervised pre-training.

3 Approach

Image preprocessing: we followed the previous literature [22, 10, 11, 13, 15] and considered images in the CIELab colourspace. This colourspace considers three channels: *L* (luminance), *a* and *b* (chrominance). It is a perceptually linear colourspace as it establishes a mapping between the colours in the Euclidean space and the colours as perceived by humans (i.e., unlike the RGB colourspace, CIELab is designed to approximate human vision).
Colour quantization: we take the idea of colours quantization from [22]. The *ab* channels 2D-space is quantized into $Q = 313$ values in gamut using a grid size of 10. The aim is to learn a mapping $G : \mathbf{I} \mapsto \mathbf{Z}$ with $\mathbf{I}$ being the grayscale input image and $\mathbf{Z} \in [0, 1]^{H \times W \times Q}$ corresponding to the per-pixel probability distribution over the Q colours.
Overall Architecture: The proposed network architecture is shown in Fig. 2. The main computational blocks are the *double block down* (green box) and the *double block up* (blue box). We followed a "U"-shaped architecture presenting a downsample (encoding) phase that applies the *double block down* to reduce the spatial definition of the input by extracting multiple features at different levels. After each block, the output is sent to the next block and to the corresponding *double block up* in the deconding phase (i.e., through a skip-connection). With this we aim to extract and maintain the spatial information of the images which is crucial for colourisation.

Details: The *downsample phase* starts with a preprocess block (orange box in Fig. 2) that consists of a $2D$ convolution having a 7×7 kernel and stride of 2×2 with Batch Normalization (BN) and Rectified Linear Unit (ReLU) activation function. A max pooling layer is used in the skip-connection to ensure that the output spatial dimension is compliant with the upsampling correspondent layer. All the subsequent operators are *double block down* layers. Each consists of two $2D$ convolutional layers with 3×3 kernels and 1×1 strides. BN and ReLU layers follow each convolutional one. Each *double block down* doubles the number of input features maps while halving their spatial resolution. At the end of downsampling, we obtain 512 features maps of size 24×24. These are fed to the capsules (Primary Caps layers) to extract information about the entities present in the input image. The 16 blue capsules in *Primary Caps Down* consist of convolutional layers applied to extract features used during the Routing by agreement, as described in [18], to extract high level entities information. The entities information extracted are fed to the *Primary Caps Up* that performs an UpSampling operation representing the beginning of the *upsample phase*. These *Primary Caps up* consists of 16 capsules. The output of the*Primary caps up* is the input of the first *double block up*. This block receives the entities features from the Primary Caps layers and the spatial features from the skip-connected *double block down*. The upsampling phase proceeds with three double blocks up that reconstruct the predicted quantization of colours of the input image. Each *double block up* consists of two UpSampling layers activated with BN and ReLU layers. Given a 224×224 input, the network outputs a matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{56 \times 56 \times 313}$.

Colourisation: starting from $\hat{\mathbf{Z}}$, an RGB image is reconstructed by first applying an inverse mapping from the $Q = 313$ values in gamut to the ab coordinates, then by the concatenation of the results with the input grayscale datum.

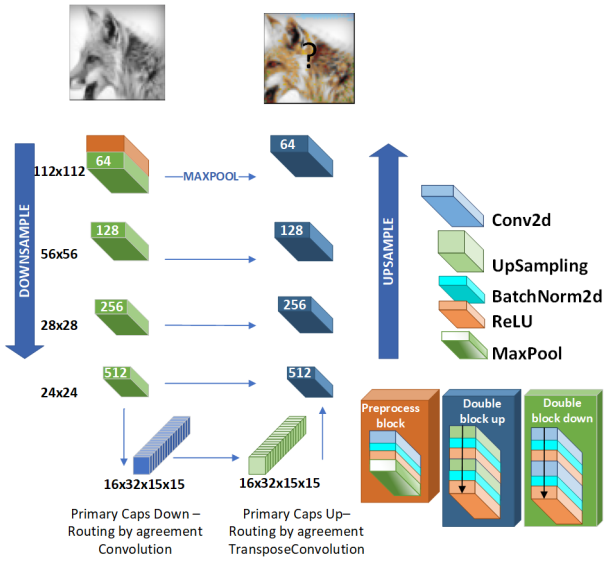

Figure 2: UCapsNet architecture. Numbers within each box denote the number of feature maps, while values on the sides indicate the corresponding spatial dimensions. Best viewed in colours.

Quantization loss: to enforce generation of vibrant colours, we adopted the multinomial cross entropy loss proposed in [22]:

$$\mathcal{L}_{cl}(\hat{\mathbf{Z}}, \mathbf{Z}) = \sum_{q} \mathbf{Z}_{h,w,q} log(\hat{Z}_{h,w,q}) \tag{1}$$

that compares the predicted $\hat{\mathbf{Z}}$ quantization against the ground truth $\mathbf{Z}$.

Training procedure: to train our UCapsNet model, we considered a self-supervised learning procedure where, given an image, the input to the model is the corresponding L channel and the ground-truth are the remaining ab ones. We followed a Generative Adversarial Network (GAN) [5, 4] where the Markovian discriminator (PatchGAN) [11] has been considered as the discriminator $D(\cdot)$ and and our proposed UCapsNet defined the generator $G(\cdot)$. The optimisation problem considers a combination of $\mathcal{L}_{GAN}(G, D) = E_y[logD(y)] + E_{x,z}[log(1 - D(G(x, z))]$, the traditional loss $\mathcal{L}_{L1}(G) = E_{x,y,z}[\|y - G(x, z)\|_1]$, and the $\mathcal{L}_{cl}$ defined in equation (1):

$$G^* = \arg \min_{\text{G}} \max_{\text{D}} \mathcal{L}_{GAN}(G, D) + \mathcal{L}_{L1}(G) + \mathcal{L}_{cl} \tag{2}$$

4 Experimental Results[1]

Dataset: to validate the performance of our approach we have considered the ImageNet dataset [17]. The training set has been used for model fitting. For a fair comparison with other methods, from the

[1] https://github.com/Riretta/Colourisation_w_Capsules

ImageNet evaluation set, we considered the same 1000 samples adopted in [20] for colourisation.
Optimisation: the Adam optimiser [12] with a learning rate of 2×10^{-5} has been used to minimise the objective losses for both the generator and the discriminator. We trained our model for 100 epochs with a batch size of 32 using the PyTorch framework and an NVIDIA Titan Xp.
Model variants: we present results obtained with UCapsNet trained with and without the GAN framework (in the latter case, only the quantization loss is considered). We also report on the performance obtained by learning to directly predict the ab channels (UCapsNet AB(GAN)), thus predicting $\hat{\mathbf{Z}_{ab}} \in \mathcal{R}^{56\times56\times2}$. In such a case the model is optimised with the mean squared error loss in place of $\mathcal{L}_{cl}$.

Performance: To visually assess the performance of our approach we computed the results in Fig. 3. These show that UCapsNet has promising colourisation capabilities. Comparable colourisation performance with more complex structures are obtained even if UCapsNet we do not apply mechanisms to deepen by the means of pre-trained model (e.g., [20]). Results obtained by Larsson [13] and by UCapsNet AB(GAN) tend to output muted colours that tend to be more "brownish" than the vibrant colours obtained by UCapsNet Q(GAN) and Zhang [22]. UCapsNet is also able to reconstruct plausible information by adding vivid colours that better match the entity in the image. In fact, the second, third, and sixth rows demonstrate that the colours generated by our solution are more plausible than the ones predicted by other methods. The inconsistent splotches obtained by UCapsNet Q are generally well addressed by means of the adversarial approach (second and fifth rows). Also, colour boundaries are not perfectly detected by UCapsNet Q and Larsson [13]. To have an overall analysis of our performance, we followed a common approach [13, 20] and computed the peak signal-to-noise ratio (PSNR) of the predicted *ab* images with respect to the ground truth and compared to those obtained for other fully automatic methods. The results shown in Tab. 1 demonstrate that our model has the best overall performance. This results indicate a better colourisation performance throughout the test dataset.

Figure 3: Qualitative colourisation ImageNet performance. Comparison between the results obtained with different UCapsNet variants and the existing solutions.

5 Conclusion

In this paper, the UCapsNet architecture based on Capsule Network (CapsNet) is designed for the image colourisation problem under a self-supervised learning setup. The proposed method is a fully automatic, end-to-end, deep model that explores the collaboration between spatial features extracted by convolutional layers and entity features extracted by Capsule layers. This model exploits the generative adversarial network framework to learn a plausible colourisation model. Experiments with the ImageNet dataset show that UCapsNet has superior performance than exiting works considering pre-trained models. Future works will focus on designing a deeper UCapsNet architecture and on investigating the application of our colourisation approach as a pretext task. **Acknowledgement** we thank Harry S. Rugg for proofreading.

Table 1: PSNR (dB) results on 1000 ImageNet validation samples.

Model	PSNR (dB)
Isola et al [11]	21.57
Larsson et al [13]	24.93
Zhang et al [22]	22.04
Vitoria et al [20]	25.57
Ours (Q)	28.43
Ours (Q GAN)	28.42
Ours (AB GAN)	28.57

References

[1] G. Charpiat, M. Hofmann, and B. Schölkopf. Automatic image colorization via multimodal predictions. In *European conference on computer vision*, pages 126–139. Springer, 2008.

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

[3] A. Deshpande, J. Rock, and D. Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 567–575, 2015.

[4] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] S. Gu, R. Timofte, R. Zhang, M. Suin, K. Purohit, A. N. Rajagopalan, A. N. S., J. B. Pinjari, Z. Xiong, Z. Shi, C. Chen, D. Liu, M. Sharma, M. Makwana, A. Badhwar, A. P. Singh, A. Upadhyay, A. Trivedi, A. Saini, S. Chaudhury, P. K. Sharma, P. Jain, A. Sur, and G. Özbulak. Ntire 2019 challenge on image colorization: Report. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2233–2240, 2019.

[7] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong. Image colorization using similar images. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 369–378, 2012.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

[10] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016.

[14] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng. Intrinsic colorization. In *ACM SIGGRAPH Asia 2008 papers*, pages 1–9, 2008.

[15] G. Ozbulak. Image colorization by capsule networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[18] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *NIPS*, pages 3856–3866, 2017.

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

[20] P. Vitoria, L. Raad, and C. Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2445–2454, 2020.

[21] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 277–280, 2002.

[22] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[23] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.