Weakly Supervised Domain-Specific Color Naming Based on Attention

Lu Yu 1,2, Yongmei Cheng 1, Joost van de Weijer 2
1.Key Laboratory of Information Fusion Technology, School of Automation, Northwestern Polytechnical University, Xi’an, China
2. Computer Vision Center, Universitat Autonoma de Barcelona, Barcelona, Spain
Email: luyu@cvc.uab.es, chengym@nwpu.edu.cn, joost@cvc.uab.es

Abstract—The majority of existing color naming methods focuses on the eleven basic color terms of the English language. However, in many applications, different sets of color names are used for the accurate description of objects. Labeling data to learn these domain-specific color names is an expensive and laborious task. Therefore, in this article we aim to learn color names from weakly labeled data. For this purpose, we add an attention branch to the color naming network. The attention branch is used to modulate the pixel-wise color naming predictions of the network. In experiments, we illustrate that the attention branch correctly identifies the relevant regions. Furthermore, we show that our method obtains state-of-the-art results for pixel-wise and image-wise classification on the EBAY dataset and is able to learn color names for various domains.

I. INTRODUCTION

Color is a basic characteristic of visual objects in the world. As one of the important features of visual data, colors are crucial in understanding the world, and they can be used to distinguish one object from another in our daily life. Humans use color names to refer to a specific color and to communicate color information with other humans. Examples of color names are ‘blue’, ‘crimson’ and ‘amber’. Computational color naming aims to identify color names in images; this is usually done by learning a mapping between color values and color names. Computational color naming is important for applications in human computer interaction, including online shopping, fashion analysis, image retrieval and person re-identification [1]–[3].

For the purpose of this article, we divide computational color naming models in methods which are trained in a supervised or semi-supervised manner. Supervised methods are based either on labeled color patches [4], [5] or on pixel ground-truth masks, providing the color names for all the relevant items in the image [1], [2]. The work of Van de Weijer et al. [6] proposed a method to learn color names from images retrieved from Google in a semi-supervised manner. We refer with semi-supervised to the fact that the provided label describes the color of the principal object in the image, but no information on the exact pixels which are described by the label is given. An advantage of semi-supervised methods is that they reduce the label effort significantly. However, the existing unsupervised methods [1], [2], [6]–[8] still require pixel masks at the testing phase. The methods are therefore semi-supervised at training, but supervised at test time.

The vast majority of the existing color naming approaches use the eleven basic color terms [1], [2], [3] which were defined in the seminal study of Berlin and Kay [9]. Even though these color names are widely used, many applications apply different domain-specific color names. In Fig. 1 several examples of color names within various applications are provided: a ‘champagne’ colored horse, ‘almond’ colored hair and ‘coral red’ lips. Because different application domains use different sets of color names, the laborious labeling process would need to be performed repeatedly. Therefore, in this paper we aim for a method which can learn from weakly labeled data, and which does not require any supervision at testing time. Learning from weakly labeled data has been studied before for image classification [10], [11], image segmentation [12], [13], saliency detection [14], object detection [15], [16], and object recognition [17]–[20].

To address the drawback of the deep learning approaches for color naming, we propose a weakly-supervised deep learning framework based on attention. The main contribution of our paper is a new two-branch network design for color naming based on attention, which is capable of automatically discovering relevant regions related to weak image labels, and simultaneously learn a mapping between color values and color names. In addition, we collect a large-scale dataset using a Web image search engine, which contains 11 basic color naming images for 4 categories, and a dataset for domain-specific color naming which includes color names for horses, eyes colors, lips colors, and the tomato growing stages. Experiments show that our attention network correctly identifies the relevant image regions, and at the same time
II. ATTENTION MODULATION FOR COLOR NAMING

The aim of this article is to predict the color name that best describes the principal object in the image. The method is to be trained from weakly-labeled data, which means a color name label is provided for the image, but that no segmentation mask or bounding box is provided to identify the principal object. We assume that images contain a single principal object which can be described by a single color name.

To train from the weakly-supervised data the method has to perform two tasks: identify the principal object in the scene and predict the color name which best describes its color. In the design of the network, which is provided in Fig. 2, we have two parallel branches, one for each task. The first branch is a shallow convolutional neural network which aims to predict a pixelwise color name map. The second branch computes an attention map which identifies the regions which contain the relevant color name information. The two branches are combined with a modulation part which combines the automatically learned attention map with each channel of the predicted color naming map.

A. Color Naming Network (CN-CNN)

The color naming network takes a color image $I \in \mathbb{R}^{H \times W \times 3}$ as an input and produces an estimate of the color name distribution $Y \in \mathbb{R}^{H \times W \times C}$ where $C$ is the number of color names. The structure of the CN-CNN is illustrated in Fig. 3 (see also top row of Table I). Specifically, first it passes through several convolutional and pooling layers after which we apply deconvolution to arrive back at the original image size. Then the features after one convolutional layer from the original input are concatenated to the part after the deconvolution layer with a skip layer [21]. One soft-max layer is then added to normalize the distribution of all dimensions.

Before training the CN-CNN in an end-to-end fashion, we initialize the network by using the weak-labels of the images. We train the CN-CNN by minimizing a weighted cross entropy loss:

$$L = \sum_i \sum_y \sum_x m_i(x, y) \log Y_i(x, y, l_i)$$

where the summations are over the spatial coordinates $x$ and $y$ and image indexes $i$, and $l_i$ is the ground truth label of image $i$ and $m_i$ is a mask. Since not all the pixels in the image are correctly described by the weak-label of the image, we use a mask which is computed with a standard saliency algorithm [22] that has a value of one for the salient part of the image and zero otherwise. This mask provides a very rough estimate of the important parts of the image, but we found this to be sufficient to provide an initialization of the network. Note that this loss is not used when training end-to-end with the whole two-branch network.

B. Visual Attention Network (VA-CNN)

Direct training on images with only the weak-labels is expected to lead to unsatisfactory performance. To further improve the visual attention network (VA-CNN), which should identify the relevant parts of the principal object in the image. To obtain this, we propose to use an attention network branch (in purple in Fig. 2). This branch has a color image as input
and aims to compute an attention map $A \in \mathbb{R}^{H \times W \times 1}$ as output. The architecture of the attention network is based on the popular fully convolutional semantic segmentation network FCN-8s [21] followed by one ReLu layer (see for details Table I). The final output of the network provides the importance of each pixel for the task of color naming.

One drawback of FCN is that it cannot learn the spatial prior. However, the principal salient object in the image is most likely in the center of the image [23]. We therefore add a spatial prior layer into the visual attention network. This layer exists of a single pixel input with value equal to one, followed by a deconvolution layer which outputs the spatial prior. This spatial prior is then used to modulate the downsampled features of the FCN network, in the same way as the modulation layer which is explained in the following section. By backpropagating, the weights of this deconvolutional layer learns the spatial prior of the dataset.

Attention model have been applied in various network architectures. They are used to attend to the relevant region in the image related to text in captioning [24] or visual question answer networks [25]. Also they have been studied in various computer vision tasks, including image recognition [20], and saliency detection [26].

### C. Modulation Layer

In this weakly supervised learning system, neither the ground-truth color names of each pixel nor the ground-truth of the confidence map is provided for directly training the CN-CNN or VA-CNN. We therefore propose an indirect method to jointly train both branches with only weak-labels. For this purpose the pixel wise color name predictions $Y$ and the visual attention map $A$ with a modulation layer to output the final color name prediction for the image $Z \in \mathbb{R}^{C}$. The modulation layer does a channel wise multiplication of the feature maps of $Y$ with the attention map $A$ according to

$$\hat{Y}_k(x, y, l) = A(x, y) Y_k(x, y, l)$$  

(2)

where $Y_k$ denotes the $k$-th channel with $k = \{1, \cdots, C\}$; $A$ is the attention map; Score aggregation is then performed on $\hat{Y}$ using average pooling to predict image-level score $\hat{y}$ for the $k$-th category.

The back propagation for the modulation layer is as follows:

$$\frac{\partial (\hat{Y})}{\partial (A)} = \sum_{k=1}^{C} Y_k$$  

(4)

### D. Network Training

Both CN-CNN and VA-CNN can be trained by minimizing the cross entropy loss $L(l, \hat{y})$, where $l$ is the ground truth label. We found that it was difficult to train the network jointly, and therefore propose an alternating training scheme. Specifically, after the CN-CNN is trained, we fix this part and fine-tuning the VA-CNN to learn attention map. After several epochs we stop training the VA-CNN branch and freeze it, and change to train the CN-CNN part again, and we repeat this process till the loss converges.

### III. COLOR NAMING DATA COLLECTION

For the purpose of this paper we collect two datasets: one of domain-specific color names, and one class-specific basic color term dataset. Both datasets are weakly-labeled.

**Domain-specific color name dataset:** we collect several domain-specific datasets from Google search engine by using the query of ‘color name + object’: 5 colors of eyes, 7 colors of lips, 9 colors of horses and tomatoes in 6 growing stages. Then, we manually removed the noisy images. Each class has 40 images for training, 10 for validation and 20 for testing. In total, 50 images for each class of each group for domain-specific color naming learning. The dataset is available at [https://github.com/yulu0724/AttentionColorName](https://github.com/yulu0724/AttentionColorName). Examples are shown in Fig. 5.

**Class-specific basic color term dataset:** Since existing methods report on the eleven based color terms we also collect a class-specific dataset for these color names. We collected

| CN-CNN | Type | Filters | Stride or Upsample | Output |
|--------|------|---------|--------------------|--------|
|       |      | 72      | 1*1               | 227*227 |
|       | Maxpool | 233 / 2 | 113*113           | 227*227 |
|       | Deconv+BN+Relu | 3*3 / 2 | 227*227           | 72 / 1 |
|       | Conv+BN+Relu | 72      | 1*1               | 227*227 |
|       | Concat | 72      | 1*1               | 227*227 |
| VA-CNN | Type | FCN-8S (1-31) | FC+Relu | FCN-8S (36-43) |
|       | Filters | 512 | 1*1 | 1*1 |
|       | Stride or Upsample | 7*7 | 1*1 | 3*3 / 4 |
|       | Output | 8*8 | 8*8 | 512 |
|       | Deconv | 3*3 / 2 | 227*227 |
|       | Modulation | 1 |
|       | Average pool | 1*1 |
|       | Softmax | 1*1 |

**TABLE I**

Details of our network
2200 images from Google Image on-line by using the query of 'color name'+object'. We choose the 11 basic color names as the indicated in [27], the difference is that four specific categories 'car','dress','pottery' and 'shoes' are selected as our 'object' class (the same categories as the EBAY color name dataset in Section IV-B) to decrease the probability of false positives, and adapt to our method. Hence for red, the query is 'red+car', 'red+dress', 'red+pottery' and 'red+shoes'. Then, we manually removed the noisy images. We retrieve 50 images for each color and object, so 200 images in total for each color name. Four special categories examples for the 11 color names are given in Fig.4.

IV. EXPERIMENTS

A. Implementation Details

We implemented our method with Matconvnet framework. The CN-CNN part is first pre-trained using the saliency method [22] to get a rough mask of the principal object as explained in Section II-A. Next we perform alternating training of the two branches using the weakly labeled data. The newly added layers in our network are initialized with Xavier method. All the training images are resized to $227 \times 227$ in our experiments for both of the CN-CNN and VA-CNN. Both of the models are optimized using Stochastic Gradient Descent (SGD) method with a batch size of 32 and 6 respectively, and a momentum of 0.9. The learning rate is set to 0.01 initially and divided by 10 after 20 epochs.

B. Color Naming from Weakly Labeled Data

Most existing methods on color naming are trained with the eleven basic color terms. We start with an ablation study to evaluate our method, and next compare it to other methods.

We compare results on the EBAY dataset which contains a total of 440 images, consisting of ten images for the eleven color names for four different categories (cars, shoes, dresses, and pottery). All images come with a mask image which identifies the pixels which belong to the named object. This Evaluation is only performed for the pixels in the mask, see example in Fig. 6.

Ablation Study: We perform an ablative study to analyze the contribution of the critical components of our proposed model. The results are on our class-specific dataset and summarized in Table III. They show a drop of about 2% without applying alternating learning, 2.5% drop without further adding centric information, and a significant drop when removing all of these, which demonstrates the relevance of the components we propose.

Comparison with the State-of-the-art: In Table IV we compare our results testing on the EBAY dataset with previous...
TABLE II
Comparison of state-of-the-art methods, testing on the EBAY dataset, training with class-agnostic dataset and new class-specific dataset. We indicate with test type which methods are supervised (S) or unsupervised (U).

| dataset              | Method     | test type | car    | dress  | shoes  | pottery | overall | car    | dress  | shoes  | pottery | overall |
|----------------------|------------|-----------|--------|--------|--------|---------|---------|--------|--------|--------|---------|---------|
| class-agnostic       | PLSA       | S         | 56.00  | 80.00  | 77.00  | 70.00   | 74.00   | 74.00  | 83.00  | 94.00  | 82.00   | 83.40   |
|                      | SS_net     | S         |        |        |        |         |         | 73.18  | 91.82  | 91.18  | 83.36   | 84.89   |
|                      | Ours       | S         | 51.38  | 80.27  | 77.64  | 71.03   | 71.83   | 71.32  | 86.36  | 88.18  | 80.91   | 81.30   |
|                      | Ours       | U         |        |        |        |         |         | 63.64  | 79.07  | 81.82  | 74.55   | 74.77   |
|                      | Classification | U  |        |        |        |         |         | 66.36  | 78.18  | 70.91  | 72.73   | 72.05   |
|                      | Ours       | S         | 57.88  | 85.35  | 78.32  | 75.54   | 74.27   | 73.64  | 94.55  | 94.55  | 86.36   | 87.27   |
|                      | Ours       | U         |        |        |        |         |         | 72.72  | 94.54  | 84.55  | 87.27   | 86.59   |
|                      | Human      | -         | -      | -      | -      | -       | -       | -      | -      | -      | -       | 88.98   |

TABLE III
Comparison of our model learned using different components on the EBAY dataset. We abbreviate attention, centric information and alternating learning as AM, C, AL.

|                      | Accuracy   |
|----------------------|------------|
| Ours                 | 55.45      |
| Ours+AM              | 84.09      |
| Ours+AM+C            | 84.77      |
| Ours+AM+C+AL         | 86.59      |

Table IV gives the results of color naming for the Eye, Lip and Horse color and Tomato growing stage. Our method outperforms the classification network on all groups. The attention network manages to identify the relevant objects as can be seen from the attention maps of some testing images in Fig. 7 where highlighted yellow regions indicate high-interest parts, and blue means low-interest parts. The smaller gains on the Horse and Tomato groups can be explained by the fact that the main object occupies most of the image and in that case the classification network also manages to extract the relevant color name.

C. Domain-Specific Color Naming

The main objective of our paper is to provide a method which can be applied to new sets of domain-specific color names with only weakly labeled data. Here we evaluate our method on the four groups from our domain-specific dataset. We compare to the previously discussed classification network; the other methods cannot be applied in this setting.

Table IV gives the results of color naming for the Eye, Lip and Horse color and Tomato growing stage. Our method outperforms the classification network on all groups. The attention network manages to identify the relevant objects as can be seen from the attention maps of some testing images in Fig. 7 where highlighted yellow regions indicate high-interest parts, and blue means low-interest parts. The smaller gains on the Horse and Tomato groups can be explained by the fact that the main object occupies most of the image and in that case the classification network also manages to extract the relevant color name.
TABLE IV

| Dataset      | Ours | Classification |
|--------------|------|----------------|
| Eye          | blue | overall        |
| Accuracy     | 65.00| 59.00          |
|              | brown| overall        |
|              | 85.00| 49.00          |
|              | gray | overall        |
|              | 65.00| 60.00          |
|              | green| overall        |
|              | 70.00| 55.00          |
|              | hazel| overall        |
|              | 10.00| 50.83          |
| Lip          | classic_red | overall |
| Accuracy     | 65.00| 65.00          |
|              | sheer_peach | overall   |
|              | 40.00| 85.00          |
|              | coral_red | overall      |
|              | 55.00| 59.00          |
|              | mandarin | overall     |
|              | 70.00| 45.00          |
|              | nude | overall        |
|              | 60.00| 35.00          |
|              | plum | overall        |
|              | 65.00| 50.83          |
|              | wine | overall        |
| Horse        | black | overall      |
| Accuracy     | 80.00| 60.00          |
|              | dark_brown | overall   |
|              | 45.00| 55.00          |
|              | bright_reddish | overall |
|              | 15.00| 65.00          |
|              | dark_gray | overall    |
|              | 85.00| 45.00          |
|              | champagne | overall   |
|              | 80.00| 55.00          |
|              | chestnut | overall    |
|              | 70.00| 55.00          |
|              | dun | overall        |
|              | 30.00| 60.00          |
|              | white | overall      |
|              | 90.00| 58.89          |
|              | brown | overall      |
|              | 45.00| 50.00          |
| tomato       | green | overall      |
| Accuracy     | 55.00| 55.72          |
|              | breakers | overall   |
|              | 25.00| 55.00          |
|              | tuning | overall      |
|              | 25.00| 55.00          |
|              | pink | overall        |
|              | 35.00| 55.00          |
|              | light_red | overall    |
|              | 65.00| 45.00          |
|              | red | overall        |
|              | 60.00| 53.33          |

V. CONCLUSIONS

In this paper we have proposed a new network for the learning of domain-specific color names from weakly labeled data. This two-branch network learns, in an end-to-end fashion, a color name probability map for each pixel and an attention map. When joined, these maps result in a color name prediction for the image. Our method is the first color name method which does not require hand-labeled masks at testing time. Results show that the attention maps correctly identify the relevant image regions and that the network successfully learns domain-specific color names. In addition, we show that the pixel-wise and image-wise predictions of the network obtain state-of-the-art results on the EBAY dataset.

ACKNOWLEDGEMENT

This work was supported by TIN2016-79717-R of the Spanish Ministry and the CERCA Programme / Generalitat de Catalunya, the EU Project CybSpeed MSCA-RISE-2017-777720, and Chinese National Natural Science Foundation under Grant 61603364. We also acknowledge the generous GPU support from Nvidia.

REFERENCES

[1] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, “Fashion parsing with weak color-category labels,” IEEE Transactions on Multimedia, vol. 16, no. 1, pp. 253–265, 2014.
[2] Z. Cheng, X. Li, and C. C. Loy, “Pedestrian color naming via convolutional neural network,” in Asian Conference on Computer Vision. Springer, 2016, pp. 35–51.
[3] L. Yu, L. Zhang, J. van de Weijer, F. S. Khan, Y. Cheng, and C. A. Parraga, “Beyond eleven color names for image understanding,” Machine Vision and Applications, vol. 29, no. 2, pp. 361–373, 2018.
[4] R. Benavente, J. Van de Weijer, M. Vanrell, C. Schmid, R. Baldric, J. Verbeek, and D. Larlus, “Color names,” in Color in Computer Vision. Wiley, 2012.
[5] A. Mopsilovic, “A computational model for color naming and describing color composition of images,” IEEE Transactions on Image Processing, vol. 14, no. 5, pp. 690–699, 2005.
[6] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, “Learning color names for real-world applications,” IEEE Transactions on Image Processing, vol. 18, no. 7, pp. 1512–1523, 2009.
[7] Y. Wang, J. Liu, J. Wang, Y. Li, and H. Lu, “Color names learning using convolutional neural networks,” in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015, pp. 217–221.
[8] Z. Yuan, B. Chen, J. Xue, N. Zheng et al., “Illumination robust color naming via label propagation,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 621–629.
[9] C. L. Hardin and L. Maffi, Color categories in thought and language. Cambridge University Press, 1997.
[10] V. Mnih and G. E. Hinton, “Learning to label aerial images from noisy data,” in Proceedings of the 29th International conference on machine learning (ICML’12), 2012, pp. 567–574.
[11] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, “Learning from massive noisy labeled data for image classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2691–2699.
[12] D. Pathak, P. Krahenbuhl, and T. Darrell, “Constrained convolutional neural networks for weakly supervised segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1796–1804.
[13] S. Joon Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, “Exploiting saliency for object segmentation from image level labels,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4410–4419.
[14] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, “Learning to detect salient objects with image-level supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR), 2017, pp. 136–145.
[15] X. Chen and A. Gupta, “Weby supervised learning of convolutional networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1431–1439.
[16] H. Bilen, M. Pedersoli, and T. Tuytelaars, “Weakly supervised object detection with convex clustering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1081–1089.
[17] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, “Learning object categories from internet image searches,” in Proceedings of the IEEE, vol. 98, no. 8, pp. 1453–1466, 2010.
[18] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, “The unreasonable effectiveness of noisy data for fine-grained recognition,” in European Conference on Computer Vision. Springer, 2016, pp. 301–320.
[19] L. Niu, W. Li, and D. Xu, “Visual recognition by learning from web data: A weakly supervised domain generalization approach,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2774–2783.
[20] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, “Multi-label image recognition by recurrently discovering attentional regions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 464–472.
[21] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 4, pp. 640–651, 2017.
[22] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in Advances in neural information processing systems, 2007, pp. 545–552.
[23] T. Judd, K. Euchner, F. Durand, and A. Torralba, “Learning to predict where humans look,” in Computer Vision, 2009 IEEE 12th international conference on. IEEE, 2009, pp. 2106–2113.
[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudin, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in International Conference on Machine Learning, 2015, pp. 2048–2057.
[25] H. Xu and K. Saenko, “Ask, attend and answer: Exploring question-guided spatial attention for visual question answering,” in European Conference on Computer Vision. Springer, 2016, pp. 451–466.
[26] J. Kuen, Z. Wang, and G. Wang, “Recurrent attentional networks for saliency detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3668–3677.
[27] B. Berlin and P. Kay, Basic color terms: Their universality and evolution. Springer, 2016, pp. 301–320.