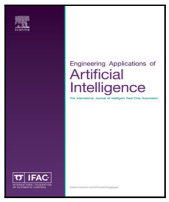
Engineering Applications of Artificial Intelligence 100 (2021) 104204

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



A better way of extracting dominant colors using salient objects with semantic segmentation

Ayse Bilge Gunduz *, Berk Taskin, Ali Gokhan Yavuz, Mine Elif Karsligil

Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey

ARTICLE INFO

Keywords:
Deep neural networks
Dominant colors
k-means clustering
SALGAN
Salient object detection
Semantic segmentation

ABSTRACT

One of the most prominent parts of professional design consists of combining the right colors. This combination can affect emotions, psychology, and user experience since each color in the combination has a unique effect on each other. It is a very challenging to determine the combination of colors since there are no universally accepted rules for it. Yet finding the right color combination is crucial when it comes to designing a new product or decorating the interiors of a room. The main motivation of this study is to extract the dominant colors of a salient object from an image even if the objects overlap each other. In this way, it is possible to find frequent and popular color combinations of a specific object. So, first of all, a modified Inception-ResNet architecture was designed semantically segmentate objects in the image. Then, *SALGAN* was applied to find the salient object in the image since the aim here is to find the dominant colors of the salient object in a given image. After that, the outputs consisted of the *SALGAN* applied image and segmented image were combined to obtain the corresponding segment for the purpose of finding the salient object on the image. Finally, since we aimed to quantize the pixels of the corresponding segment in the image, we applied k-means clustering which partitions samples into K clusters. The algorithm works iteratively to assign each data point to one of the K groups based on their features. Data points were clustered according to feature similarity. As a result the clustering, the most relevant dominant colors were extracted. Our comprehensive experimental survey has demonstrated the effectiveness of the proposed method.

1. Introduction

Playing a vital role in designs, color has been used by artists and designers for decades. Color is a key element for professional design. Color combinations can affect emotions, psychology and user experience since each color in the combination has a unique effect on each other and the balance of the combination has significant importance (Machajdik and Hanbury, 2010; Terwogt and Hoeksma, 1995). Furthermore, the majority of people match colors with an intuitions developed at a young age. As there is no universally accepted combination of colors, it is a very challenging task to determine popular combinations of colors to be preferred.

Finding the right color combination is crucial for each and every aspect of our lives from designing a new product or combining clothes to create the perfect combination to decorating a space either at home or in the office. However, there is no single right color combination but rather preferred or popular color combinations. In this paper, we present a novel approach that finds the most significant dominant colors of the salient object in an image containing an extensive number of different objects either separate from or overlapped the salient object in question. Thus, our approach enables people to perceive the most suitable combination of colors for a given occasion. Designing a new product or dressing for a special event could be considered as an example for this occasion. Furthermore, when our approach is applied to images of different categories, including fashion, fragrance, interior design, etc., the spatial, temporal, and cultural variations of popular color combinations in these categories could be discerned. This data could also be used to predict the future color combination preferences for a given category e.g. determining a country level color preference for a certain type of object, or yearly most popular colors, etc.

It is common to have more than one object in an image, unless there is a special shooting setup for photographs. Moreover, whether professional or casual, in a there is always a background in a photograph. First of all we segmentate the image via semantic segmentation to obtain different objects in order to determine the salient object. Then, the original image is submitted to salient object detection and the outputs of this stage are intersected with the segmented image from the previous step. This process gives us the salient object as a whole. Finally, the colors of the salient object are quantized and then clustered to determine the dominant colors. Our contributions are as follows:

* Corresponding author.
E-mail address: aysebilgegunduz@gmail.com (A.B. Gunduz).

https://doi.org/10.1016/j.engappai.2021.104204
Received 17 August 2020; Received in revised form 10 February 2021; Accepted 12 February 2021
Available online 23 February 2021

- we have extracted the most accurate dominant colors of the focused object in a given image when compared to existing research focusing primarily on extracting the dominant colors of either the entire image or the foreground only.
- we successfully applied deep learning architectures in order to semantically understand and detect the focused object in a given image.

The rest of the manuscript is organized as follows: Related Work is given in Section 2, both of the models developed for our novel approach is introduced in Section 3, Section 4 is used to represent Experimental Results with the Conclusion given in Section 5.

2. Related work

A review of related work on dominant color extraction reveals the fact that this method has been mainly used for color theme extraction and content-based image retrieval. Salient regions and salient features are only mention on Liu et al. (2018) and Xing et al. (2018) respectively. However, they do not perform an independent salient object detection as both research focuses on images containing a single object. Therefore, the aforementioned saliency only refers to the most notable regions and/or features of a single object within the image. All the other studies focuses on the image as a whole. The following paragraphs provide a comprehensive review of the related work in categories as mentioned above.

CBIR (Content Based Image Retrieval) can simply be defined as an approach to facilitate image searching in large databases using image contents rather than metadata (Kato, 1992). Since the goal of CBIR is to find images, by using helpful characteristic features such as color, texture, shape, etc. In this context, the exact composition of dominant colors can supply this characteristic contextual information. Even when the color information provided is insufficient for CBIR, it could still help reducing the amount of data to be searched in the database. In Yan et al. (2015), researchers studied images to detect misused or fake logos and trademarks. In this research, for the purpose of color image retrieval Linear Block Algorithm (LBA), k-means clustering, and spatial feature extraction were combined. In order to extract dominant colors, LBA and K-Means algorithms were combined by using color quantization in RGB color space. The dominant colors were extracted from adjacent partitions of the color clusters in an agglomerative fashion (Yang et al., 2008). The clusters continued to be merged until their Euclidean distance was smaller than a given threshold value. The resulting clusters represented dominant colors. Instead of using previous data-sets Yan et al. prepared their own data-set consisting of 266 images. The images were tagged with MPEG-7, blurred with Gaussian mixture, and noised with Salt-and-Pepper and the aforementioned CBIR approach was applied. However, performance measurement is not clear and comparisons are only conducted between the methods they proposed. Chauhan et al. proposed a method for dominant color extraction and content-based image retrieval (CBIR) in Chauhan et al. (2018). K-means clustering was used to extract a dominant color in HSV color space. The dominant color was represented as the largest color bin. It was then used as a similarity measure for CBIR. Wang database, containing 1000 images with 10 classes, was used for CBIR experiments and overall accuracy rate was stated as 85%. In Zhou et al. (2019a), researchers proposed a fashion recommendation approach. The model they proposed consists of 2 distinctive parts; one aims to conduct CBIR and the other aims to recommend stylish combinations for full-body images with clothes. The dominant color histogram was extracted in RGB color space by using k-means clustering for the purpose of content based image retrieval. The colors were then matched with the most similar Pantone colors to be used then as features. The dataset used in this study consisted of crawled images obtained from the websites of popular fashion stores. Only expert opinions were used as a performance metric.

A color theme is a collection of conspicuous color swatches that represent or define color choices in a design or an object (Jahanian et al., 2015). Therefore, color theme extraction can be defined as finding those conspicuous color swatches. A color theme extraction study was specifically conducted on fabric images by Liu et al. in Liu et al. (2018). K-means clustering was used to extract dominant colors from both the background and foreground. Cheng et al.'s salient region detection method (Cheng et al., 2013) was used to differentiate the foreground from background. Since finding the respective dominant colors is important for fabrics, Liu et al. held a user survey to evaluate the efficiency of their proposed method. According to the survey results, their proposed method was found successful. Xing et al. proposed a method for dominant color extraction for Chinese traditional costume images (Xing et al., 2018). The photographs in the dataset were particularly taken in front of the same solid background with a professional camera in a studio. Using RGB color space, researchers extracted three sub-images for each color bin from the original image. Then, the median filter was applied to each sub-image. The filtered images were converted into Lab color space and afterward, they were segmented based on the background color. Finally, mean shift clustering (Liu et al., 2012) was used to extract the dominant colors. CMC(2:1) method was used to do a performance comparison between the proposed approach and manual extraction results (Teel et al., 1992). As a result of the comparison, the results of the proposed approach which extracted the dominant colors from costumes were similar to manual results. In Liu and Luo (2016), hierarchical emotional color themes approach is proposed. These colors are obtained from dominant colors extracted via k-means clustering from an image. Then, users are asked to select a color theme from a set of hierarchical color themes, thus a set of candidate emotion colors is determined for further use of the study. The quality of the user-selected color themes was controlled with Emotional Entropy. An emotion value is assigned to each pixel by using the Pearson Correlation Coefficient to calculate the emotional entropy (Ou et al., 2004). The model represents eight basic emotions: warm, cool, heavy, light, soft, hard, passive, and active. The ground truth data for this research was determined by user opinions used to measure the performance of the proposed model. The obtained emotional color themes were then optimized using Least Absolute Shrinker and Selection Operator (LASSO) to build an emotional relationship between the selected theme and the candidate theme. After that, the optimized color theme was then applied to the image, and the color transfer was accomplished (Tibshirani, 1996). The main purpose of the researchers is that users can freely select corresponding emotional color themes based on their feelings. As a result, the subjective feelings of users can be transferred into an image. In Feng et al. (2018) Feng et al. also proposed a color theme extraction approach utilizing the comparability principle of human visual perception by taking into consideration the relationship between the pixels and its color information. A community finding algorithm is used by considering certain factors including number, accuracy, and span. The superpixels of the image are segmented uniformly to determine pixels that represent candidate color themes. Then, a set of candidate color theme was extracted by using improved Simple Linear Iterative Clustering (SLIC) in a CIELAB color space. They run a survey with participants to create ground truth color themes then a methodology that best suits human visual perception is constructed to select the best color theme among the candidate color themes. Wu and Han assessed the color compatibility of Web Pages and proposed a method for color theme extraction of a Web page in Wu and Han (2018). The outlier-aware clustering method (Forero et al., 2012) was used for color theme extraction. The aim of this study is to propose colors for future web designs. However, the limitation here is that only the colors of the static parts of a web page were under study but the elements like fonts, images, etc. were left out. Also, the proposed method was evaluated by surveying the opinions of participants consisted of unqualified users. Takahashi et al. conducted a study on the analysis of color characteristics in TAKAHASHI et al. (2018). They collected

images of restaurant websites from Tokyo and proposed a method to express the relationship between the prices in these restaurants and the their representative colors. K-means clustering was used to extract the colors, the number of which was then reduced using color bins. In order to define ground truth values, a test group was selected to determine prices for each restaurant by solely looking at the colors of restaurant web-pages. Kendall rand correlation coefficients (Abdi, 2007) were used as a performance metric of the proposed method.

3. System design

The main motivation of this study is to extract the dominant colors of a salient object from an image even if the objects overlap each other. It is obvious that the dominant colors of an image do not always represent the dominant colors of the salient object within the image. Therefore, methods proposed in the related work section is mostly limited to dominant colors extraction of the entire image. It is evident that objects in an image must be segmentated before extracting the dominant colors. But, features like color and/or texture will not be sufficient enough to distinguish different objects within the image. So, a simple object segmentation using only colors and/or textures will provide unsatisfactory results. Therefore, we have chosen to semantically segmentate objects in a given image. Then, the results of the segmentated objects and salient object detection were combined. Finally, dominant color extraction was applied to the object filtered out by the salient object detection (Fig. 1).

3.1. Semantic segmentation

Segmentation, in the context of computer vision, is the idea of separating each object in a scene according to their distinguishing features such as color, texture, or motion. However, objects exhibiting more than one feature or consisting of variations of the same feature makes it almost impossible to segmentate the object as a whole. For example, in order to segmentate a full-body image with a sweater and jeans, both the colors and textures of the clothes must be taken into consideration. This is a common obstacle in the way of successful object segmentation. Therefore, it is crucial to understand a scene in its entirety in order to create a productive segmentation model.

Traditional methods for segmentation rely either on clustering techniques or on region/edge-based approaches. Clustering techniques aim at grouping similar pixels together. On the other hand, region-based approaches group neighboring pixels with similar features. Edge-based approaches try to separate objects by using the transitions between them. The transitions are detected using edge finding filters (Yuheng and Hao, 2017). One may also consider using image descriptors as the basis of image segmentation. However, the literature suggests that this approach is especially efficient for image matching and reconstruction rather than segmentation (Wang et al., 2019, 2018). Consequently, none of the traditional methods are successful at completely separating an object from an image consisted of several different objects. Therefore, in our study, we used semantic segmentation. Semantic segmentation is an object-oriented deep learning-based approach that provides semantic labeling for objects. As a deep learning approach, semantic segmentation utilizes features extracted via convolutional layers. In the first few layers of such systems, local features of varying scale are learned. The deeper the architecture goes, the more abstract these features become. Differentiation between these layer-wise outputs is the reason why deep learning architectures are successful at making sense of a given input. By understanding the textures and patterns as a whole, the deep learning architecture is able to bridge this ‘semantic gap’ and thus provide a better representation of features extracted from the image.

One of the most crucial aspects of semantic segmentation is the ground truth data. There are a number of ways to generate the ground truth data, but as we used the Tensorflow/Keras framework we chose a dataset matching the requirements of the framework (Chollet et al., 2015). The ground truth data in our dataset was obtained by converting each pixel of manually labeled images into a one-hot vector whose dimension matches the class number defined in the dataset. Our framework employs an encoder–decoder structure to do semantic segmentation. The encoder part of the semantic segmentation network takes an input image, extracts features from it, and then the decoder part tries to recreate an image of the same size with object labels. The difference between the ground truth data and the network output is fed back into the encoder–decoder structure as a loss function. As a result, at the end of the semantic segmentation, we obtain both an understanding of the scene presented in the image and objects with dense labeling. In order to achieve semantic segmentation a number of different architectures such as FCN (Long et al., 2015), SegNet (Badrinarayanan et al., 2017), Unet (Ronneberger et al., 2015), ResNet (He et al., 2016), Xception (Chollet, 2017), Inception-ResNet (Szegedy et al., 2017), DeepLab (Chen et al., 2017), etc. could be used.

3.1.1. Approach and methodology

Semantic segmentation is an intrinsically challenging task since it requires fine-grained labeling of each and every pixel in an image for each object. This current backbone used for our semantic segmentation was trained and tested with MIT Scene Parsing dataset (Zhou et al., 2017, 2019b). Training and fine-tuning of the network was made via different hyper-parameters using categorical cross-entropy loss as an approximation metric, until no further approximation between validation epochs for the loss was possible. MIT Scene Parsing dataset has 151 object labels including background, therefore the output of our network is dimensioned as HxW. Extracting this much information while staying robust against color and lighting changes requires deep feature extractors, which constitute the encoder part of the network. The output of the encoder should be fed into the decoder part of the network in order to achieve the recreation of the original images. In the next sub-sections, parts of the designed network which are given in Fig. 2 will be explained in their forward pass order.

3.1.2. Feature extractor — encoder

The encoder part of our network mainly extracts features. The feature extractor works as a series of convolutional filters, and these filters distill object information in a given image to acquire the corresponding compressed form. The current state of the art approach for semantic segmentation on Pascal VOC dataset (Everingham et al., 2007) is DeepLabv3+ JFT. However, this approach could not be used directly in our case due to the fact that MIT Scene Parsing dataset with its large object count does not contain a sufficient number of images for each object type to train the network efficiently. On the other hand, FCNs (Long et al., 2015) and SegNet (Badrinarayanan et al., 2017) structures are not strong enough to extract necessary features. After some initial testing with available models, we have seen that using a backbone of modified Inception-ResNet with depth-wise separable convolutions seems to work best with 80% categorical cross-entropy accuracy. Depthwise separable convolution, a concept that MobileNet (Howard et al., 2017) pioneered, consists of two parts. The first part performs a spatial convolution over each input channel, and the second part is a point-wise (1x1) convolution (Chollet, 2017). As a result, the number of parameters in the convolution steps is reduced since point-wise convolution projects its results to a new channel. Such strong feature extractors are prone to over-fitting. After each block of convolution, dropout layers with varying weights in a certain range were used to overcome over-fitting. Any increase on dropout rate above a certain threshold results in reduced accuracy on the training of the neural network. Additionally, the original feature extractor was shortened to allow for a skip connection block to be placed. The purpose of this block is to provide the latest features to the next part of the architecture.

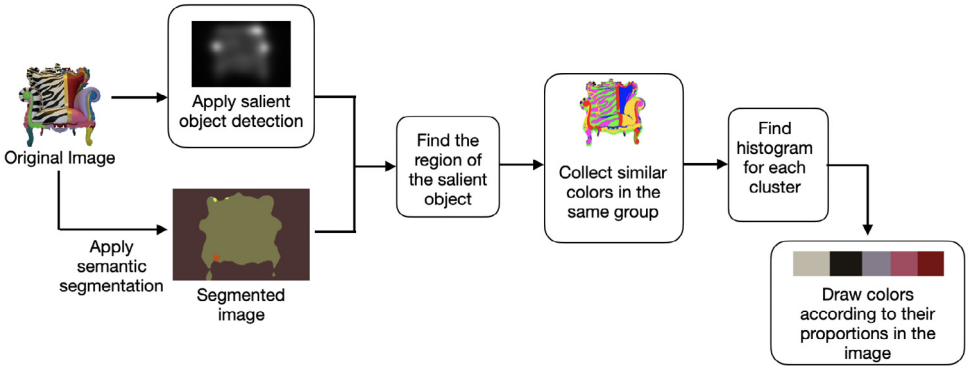


Fig. 1. Block diagram of the proposed dominant color extraction approach is shown in this figure. Original image is processed to create segmented image and salient image separately. Then those two separate images are intersected to obtain the salient object in the original image. Finally, dominant colors are extracted by applying k-means clustering. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

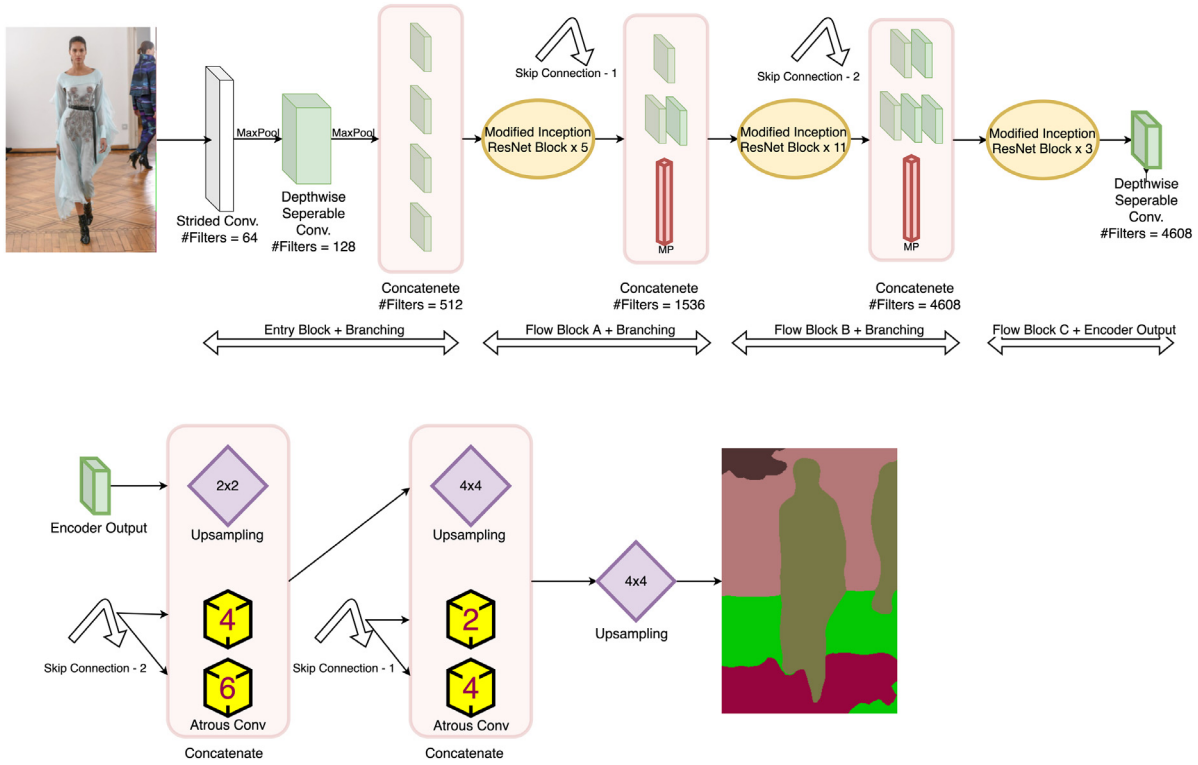


Fig. 2. Our custom designed semantic segmentation architecture for the proposed dominant color extraction approach is depicted in this figure. Upper part of this figure represents the encoder (feature extraction) side of our network while the lower part represents the decoder (image recreation). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1.3. *Image recreation — decoder*

The decoder should be able to recreate the original image using the features extracted by the encoder part of the network. In addition to the original image, the decoder should also produce fine-grained pixel labels for the objects within the image. Usage of FCNs (Long et al., 2015) and SegNet (Badrinarayanan et al., 2017) relied solely on extracted features for both object identification and localization. It has been shown (Zhao et al., 2017) that using intermediate layers from earlier convolutions and different scaled atrous convolutions after feature extraction (Chen et al., 2017) are helpful towards achieving a better result. This is due to the fact that most of the original input images can be retained at least as carryover information. Furthermore, much like DeepLab, we utilized skip connections starting from the entry block of the encoder to the pooling layers of the decoder as a way of retaining as much information as possible from the original image. The decoder itself is comprised of three separate parts. The first part takes the encoded features and upsamples them by 4x4, then the second part takes the skip connections from the middle of the encoder and passes them through a series of layers, each with different atrous convolution rates, and concatenates them together. The last part of the decoder simply reshapes and upsamples everything to match the input image size of the network.

3.2. *Extraction of the salient object*

In order to obtain the objects in a given image, we apply semantic segmentation. However, as dominant color extraction must be applied to the dominant object within the image, straight-forward object segmentation is not satisfactory. Therefore, it is necessary to determine the dominant object within the image first. So, we run salient object detection on the whole image and match the results of this detection with the objects extracted previously.

Fig. 3. (a) shows the original image, (b) shows the output of the salient object detection, and (c) shows the segmented image. There are a total of 8 segments numbered from 1 to 8. Segment 3 corresponds to the salient object.

Table 1
Comparison of SALGAN with other state-of-the-art solutions on the MIT300 benchmarks (Bylinskii et al., 0000). Values in brackets correspond to performances worse than SALGAN (Pan et al., 2017; Kummerer et al., 2017; Liu and Han, 2018; Huang et al., 2015; Pan et al., 2017; Jetley et al., 2016; Cornia et al., 2016; Kümmerer et al., 2014; Pan et al., 2016; Zhang and Sclaroff, 2013).

Detection model	AUC-J ↑	Sim ↑	AUC-B ↑	sAUC ↑
Humans	0.92	1.00	0.88	0.81
Deep Gaze II (Kummerer et al., 2017)	(0.84)	(0.43)	(0.83)	0.77
DSCLRCN (Liu and Han, 2018)	0.87	0.68	(0.79)	0.72
SALICON (Huang et al., 2015)	0.87	(0.60)	0.85	0.74
SALGAN (Pan et al., 2017)	0.86	0.63	0.81	0.72
PDP (Jetley et al., 2016)	(0.85)	(0.60)	(0.80)	0.73
ML-NET (Cornia et al., 2016)	(0.85)	(0.59)	(0.75)	(0.70)
Deep Gaze I (Kümmerer et al., 2014)	(0.84)	(0.39)	0.83	(0.66)
SalNet (Pan et al., 2016)	(0.83)	(0.52)	0.82	(0.69)
BMS (Zhang and Sclaroff, 2013)	(0.83)	(0.51)	0.82	(0.65)

For example, in Fig. 3, there are a total of 8 segments, but only segment 3 corresponds to the salient object. To determine the dominant colors of the salient object, we map the salient objects area to the corresponding segment within the image. Then we run our dominant extraction algorithm. This is one of the main points where our approach distinguishes itself from the studies given in the related work section. Applying dominant color extraction directly to the whole image without segmentation and salient object detection yields poor results. Without our approach, it is possible that various non-salient objects having the same color distribution could contribute to the dominant colors more than the color distribution of the salient object.

As it was not our intention to produce a better salient object detection approach for the purposes of this study, we investigated current salient object detection approaches to determine the best-suited one to incorporate it into our dominant colors detection model. Table 1 shows the performance comparison of the state-of-the-art salient object detection models on the MIT300 benchmark. As part of this benchmark, ground truth was established with a survey involving human judgment, and an AUC-J of 0.92 was obtained (Pan et al., 2017). Therefore, all the other model performances were evaluated against this value. Table 1 shows that only two other models outperform the SALGAN architecture by just 0.01 percent point. Thus, we decided to use SALGAN architecture in our proposed model. There is always a possibility of improving our proposed dominant colors detection model as better salient object detection models are introduced into the literature.

3.3. Dominant color extraction

The selection of color spaces may vary in image processing, and the most well-known color spaces are Red, Green, Blue (RGB), CIELab, and Hue, Saturation, Value(HSV) color spaces. Although the RGB color space is common, it is non-linear and is device-dependent with redundant values. It also has perceptual non-uniformity (Baldevbhai and Anand, 2012; Schwarz et al., 1987). The CIELab color space consists of three parameters; the first one is L, which represents the luminosity of the color, and the other ones are a and b components that represent the chromatic information, respectively. The CIELab color space defines colors independently of how they were created or displayed. The HSV color space is the alternative representation of the RGB color space. In this color space, hue color is arranged in a radial slice, around a central axis of neutral colors, ranging from black at the bottom to white at the top. Therefore, this provides a better alignment with the human eye perception of colors (Schwarz et al., 1987). The human perception of colors heavily relies on the luminance information. In order to exploit this information, several color spaces have been developed other than the previously mentioned ones. These color spaces separate the luminance and chrominance information in a given image. YIQ, YUV, and YCbCr are the main color spaces based on this approach (Cattin, 2016). In this representation, Y stands for the luminance information, whereas the chrominance is given with the rest of the components. As part of this research, we tried our approach in different color spaces, and we did not observe any significant improvement in terms of dominant color extraction performance. Therefore, we chose to operate in the RGB color space.

Even if a region in a real world image has the same color, the pixels representing this region in the digitized image could end up as different colors because of external factors like light, contrast, or imperfections of the camera system. Therefore, the colors of the pixels must be quantized before the extraction of the dominant colors. To this end, we applied k-means clustering to the pixels. However, since we cannot foresee the number of main colors in a given image, the value of k must be determined dynamically. So we applied the elbow method to all of the images representing salient objects to determine the optimal k value. As depicted in Algorithm 1, for each image, we started with k=2 and kept it increasing by 1 at each round and observed the rate of decrease in the within-cluster sum of squares (WCSS) value (Ng, 2012; Arthur and Vassilvitskii, 2006; Ketchen and Shook, 1996). Our tests showed a huge drop in the WCSS value for k values 2 through 5. After 5 there was only a minimal drop; hence we chose 5 as the optimal value for k. Fig. 4 shows the elbow curve for a sample image depicted in Fig. 5. As mentioned before, the drop rate in the elbow curve for this image becomes more and more minimal for k values greater than 5. In Fig. 5, the sub-figure(a) shows the original image whereas the sub-figure(b) represents the k-clustered image with each color representing a different cluster.

Fig. 4 shows the relation of the number of clusters to the WCSS value for the object given in Fig. 5a. For this instance, the optimal value of k was determined as 5. The value of k was used to quantize the colors of the salient object for each instance. The obtained clusters (Fig. 5b) were then sorted in descending order according to the sizes of the clusters. Thus, dominant colors were obtained (Algorithm 2). Table 2 gives detailed comparison of our proposed detection model and existing models.

Table 2
Comparison of the studies given in the related work section and our approach.

Study	Main purpose	Method	Salient object detection	Evaluation metric
Yan et al. (2015)	Content based image retrieval	LBA, k-means clustering & spatial feature ext. for color quantization	No	Similarity
Yang et al. (2008)	Content-based image retrieval	Custom dominant color ext. method in agglomerative based clustering	No	None
Chauhan et al. (2018)	Content-based image retrieval	k-means clustering for general color ext. + dominant color is selected as largest color bin	No	None
Zhou et al. (2019a)	Content-based image retrieval	k-means clustering for dominant color histograms + the colors are mapped to the Pantone chart to be used as a feature for recommendation	No	None
Liu et al. (2018)	Color theme extraction	k-means clustering on fabric images for background and foreground dominant color extraction separately + Global Contrast Based Salient Region Detection[24]	No	User Survey
Xing et al. (2018)	Color theme extraction	All photographs in dataset were taken in studio and has same solid background. Segmentation was done according to this background. Mean Shift Clustering for dominant color ext.	No	Color-difference comparison
Liu and Luo (2016)	Color theme extraction	k-means clustering for dominant color extraction + Least Absolute Shrinker for optimization	No	User Survey
Feng et al. (2018)	Color theme extraction	Simple Linear Iterative Clustering for dominant color ext.	No	Similarity
Wu and Han (2018)	Color theme extraction	Outlier aware-clustering [25] for dominant color ext.	No	User Survey
TAKAHASHI et al. (2018)	Color theme extraction	k-means clustering for general color ext. + clusters are reduced by using color bins	No	Kendall rand correlation coeff.
Our Approach	Color theme extraction	Modified Inc.V4 + ASPP are used for semantic segmentation & SALGAN is used for salient object det. Results are combined to find the segment of salient object. k-means clustering is applied for dominant color extract.	Yes	User Survey + Color-difference (Sharma et al., 2005)

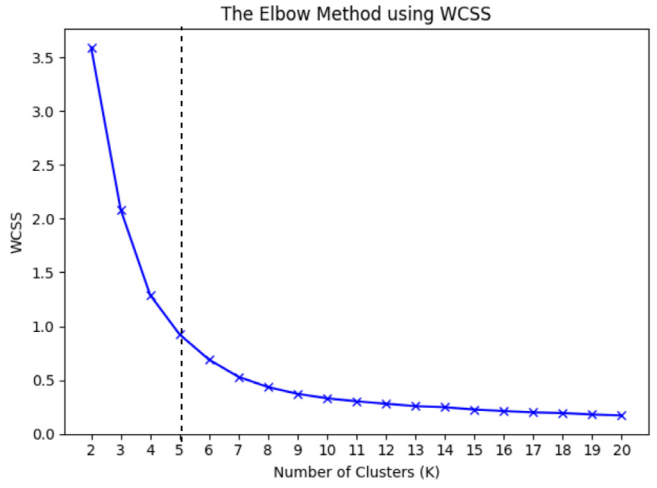


Fig. 4. Determination of the k value using the elbow method. This figure shows the elbow curve for the image given in Fig. 5(a). It is evident that the drop rate of the WCSS value becomes minimal after 5.

Algorithm 1: Determination of the k value using the elbow method.

Input: *candidate_image*
Output: *k_value*
$i = 2$
Load *limit_value* with the highest possible clustering value
while $i <$ *limit_value* **do**
  k=i
  Apply *k-means clustering* in each round on *candidate_image*
  $WCCS[i]$ = sum of clusters' *inertia*
  i = i+1
end
Draw a linear line between lowest and highest $WCCS$ index
Calculate *distance* from every $WCCS$ value to the line
k_value = $WCCS$ index corresponding to maximum *distance*

Algorithm 2: Extraction of the Dominant Color

Input: *segmented_Image*, *original_Image*, *salient_Image*
Output: *dominant_colors*
salient_object = *bitwise_and*(*segmented_Image*, *salient_Image*)
Binarize *salient_object*
Get the region of the salient object from *salient_object*
Use region to get the corresponding *segment* of the salient object
Set k with elbow method result
Apply k-means clustering to the *segment* by using k
Sort clusters in descending order according to the number of instances
Output *dominant_colors*

4. Experimental results

In this section, we demonstrated and put forward the experiments carried out to ascertain the performance of our proposed approach. We used two different datasets to obtain the experimental results. The first dataset is the MIT Scene Parsing dataset, which was used for semantic segmentation (Zhou et al., 2017). The second dataset is a custom dataset provided by HueData (Lechner and Harrington, 0000). It was used to evaluate the performance of our dominant colors detection approach. The MIT Scene Parsing dataset consists of 22K

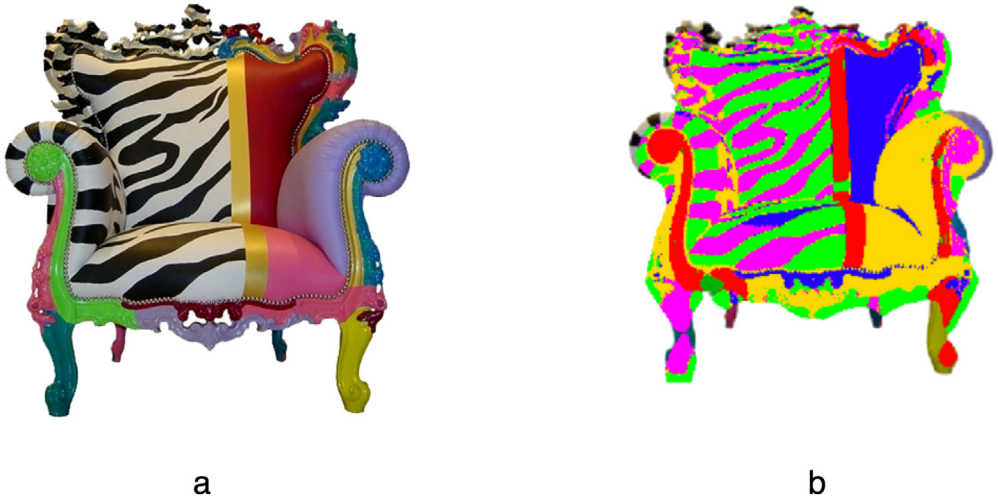


Fig. 5. (a) shows the original image, (b) shows the clustered image where each color represents a different cluster. The colors of the clusters were chosen randomly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

images with 150 semantic categories, whereas the HueData dataset consists of 3K images with six different categories. Furthermore, the MIT Scene Parsing dataset consisting of 22K samples was divided into two subsets of 20K and 2K. The subset with 20K samples was used to train our proposed model, and the remaining 2K samples were used for validation. The 3K instances of the HueData dataset were solely used to test the performance.

We used two metrics, pixel accuracy (Long et al., 2015) and Intersection over Union (IoU) (Rezatofighi et al., 2019) for semantic segmentation network and model accuracy evaluation. We did the training phase on an NVIDIA RTX 2080 TI with 11GB of VRAM. Our training methodology was carried out in three distinct steps. In the first step, we used SGD to optimize our network's weights with a learning rate of 0.001. In addition to that, we reduced this value by a fraction every 4000 iterations, roughly 5 times per epoch. In the second step of training, 3 dropout layers were added to the network in the encoder, right after repeating layers. The learning rate was reduced to 0.0001, and we switched the optimizer from SGD to Adam. During the last phase, we froze weights on the encoder and decoder in turn for 5 epochs each. We also added 3 dropout layers to the network on the decoder part and changed our learning rate again to 5e-5. Afterward, we left the system to train until no further approximation was possible. The learning rate reduction was limited to once per epoch in this step. The results of our trials with different models can be seen in Table 3. While other architectures such as VGG19 and ResNet50 were also tested, none of them gave results that would be considered comparable. Thus, they were not included in the table. SALGAN architecture was used to find the salient object as part of the dominant color extraction process with standard parameters.

We have successfully demonstrated that an effective dominant color extraction can be obtained using our proposed approach, as depicted in Fig. 6. In this figure, column (d) shows the dominant colors obtained merely by processing the image as a whole without regard to the salient object as in common to all previous studies. On the other hand, Fig. 6e contains the output of our proposed dominant colors detection model in which segmented image (Fig. 6b) and salient object within the image (Fig. 6c) were used together.

A closer look into the images in Fig. 6a reveals that some of them have a different number of objects of various sizes, either separate or overlapped in the background as opposed to a solid background. Consequently, dominant color extraction methods utilizing the image as a whole do also include these objects within the extraction process. Thus, the dominant colors obtained in that manner do not necessarily

Table 3
Performance comparison of semantic segmentation approaches including our architecture (highlighted row) on MIT scene parsing dataset.

Architecture name	Time per Epoch	Pixel accuracy	IoU
ResNet-101 + PsP	50 mins	74%	0.38
DeepLabV3	150 mins	75%	0.40
InceptionV4 + ASPP	120 mins	69%	0.31
Modified Inc.V4 + ASPP	50 mins	81%	0.42

belong to the salient object. On the other hand, a solid background manifests itself as the primary or secondary dominant color depending on the background's size, when the dominant colors are extracted over the whole image. These two facts are illustrated in rows 1, 2, 3, 4 and 8, 11, 12 in Fig. 6 respectively. Column (e) in Fig. 6 shows the results of the dominant color extraction of our proposed method for the same images. The results clearly show that more accurate results were obtained consequent to using semantic segmentation and salient object detection as per our proposed method. Furthermore, we also manifested the success of our results by conducting a survey with a randomly selected set of participants.

In order to quantify the success of our proposed method, we asked the same participants to manually extract dominant colors for 20 images from our dataset. For each image, manual dominant colors were determined based on the majority of the participants' choices. Then, we calculated the color-difference values for the first dominant color pointed out by the participants against the first dominant color determined using the whole image and our proposed method, respectively. The color difference values were calculated using the CIEDE2000 color-difference formula (Sharma et al., 2005). The results are given in Table 4. In this table, the first column gives the image label with respect to the labels in Fig. 8, whereas the first ΔE value represents the color-difference between manual extraction and whole image extraction, and the second ΔE value gives the color-difference between manual extraction and our proposed method. The results clearly show the improvement in dominant color extraction using our proposed method. It is also an indication that participants were paying much more attention to the salient object than the background when determining the dominant colors, which furthermore supports the validity of our proposed approach.

On the other hand, we also determined that our model's performance is prone to three factors, namely *lighting conditions*, *performance*

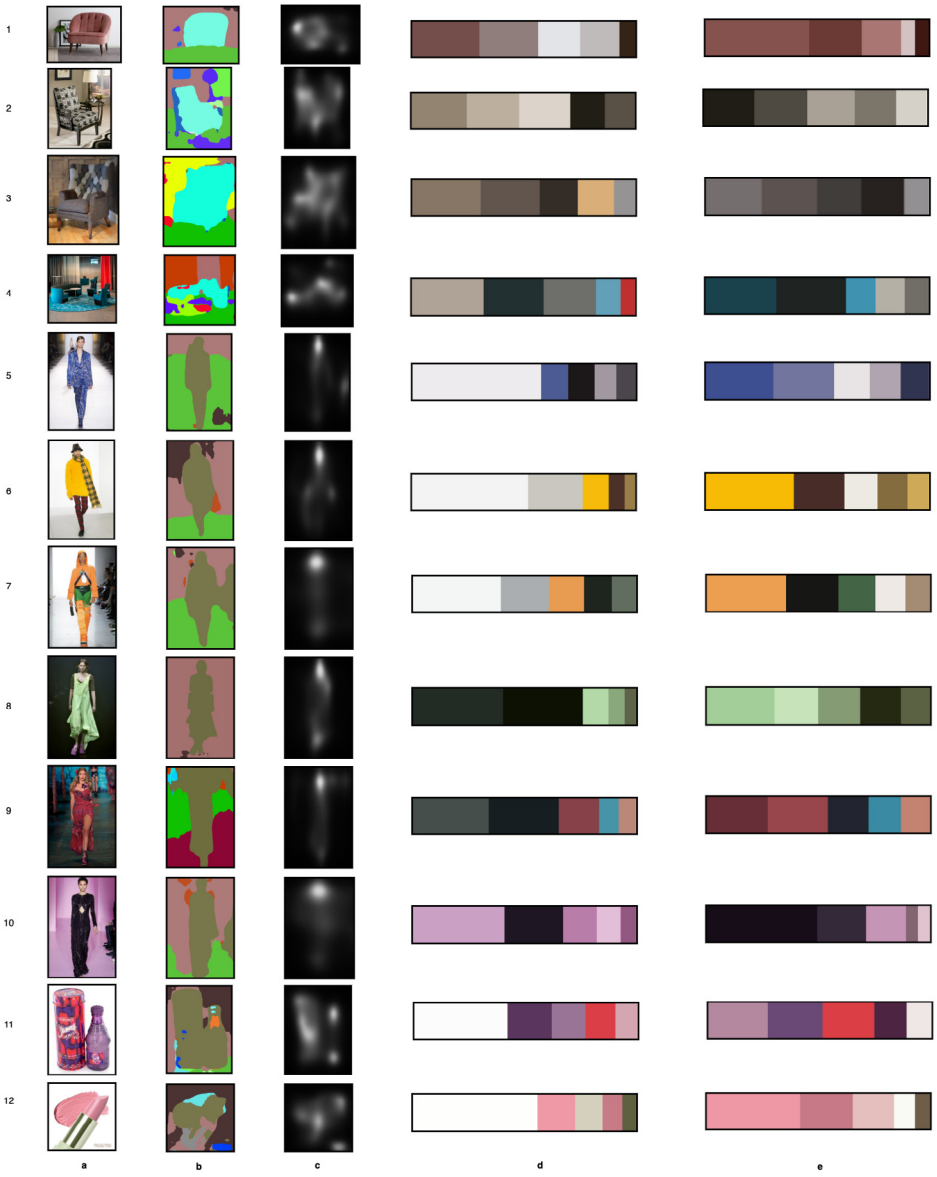


Fig. 6. Successful samples (a) the original image, (b) the segmented image, (c) the salient object image, and (d) dominant colors extracted from the whole image in proportionally descending order, and (e) dominant colors (extracted via the approach proposed by us) in proportionally descending order. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the semantic segmentation, and *performance of the salient object detection*. The lighting conditions still impose a major problem in colored image processing applications, and it also causes human misjudgments. However, the effect of illumination in color determination is a study on its own, and it is not focused on in this study (Barnard, 1998). As presented in the related work section, we followed a similar approach and left out the illumination problem. As to the other two factors, their theoretical performance limitations were discussed in the respective sections previously. Fig. 7 gives a summary of the unsuccessful outputs of our proposed model due to the aforementioned factors. In Fig. 7 image #1, semantic segmentation puts the foreground model and audiences within the same segment. Furthermore, the salient object detection detects both front model and model in the back as the salient objects. Thus, the dominant colors detection tries to extract the dominant colors from the union of those two objects. In image #2, semantic segmentation segments the front model and the audience as one. As a result, the dominant colors detection extracts the dominant colors from the union of the model and the audiences even if the salient object was detected correctly. In image #3, the couch was successfully extracted and detected by the semantic segmentation and salient object detection models. Unfortunately, as a result of shadows caused by poor lighting conditions and the same color manifesting itself in different shades, dominant color extraction was not able to produce the correct result. Finally, image #4 is an example of one of the rare occasions when the salient object detection produces confusing results. Consequently, the dominant color detection runs on unrelated segments within the same image and produces incorrect results.

5. Conclusion

In this paper, we have investigated the problem of extracting the dominant colors of the focused object in a given image and successfully presented a novel approach for the dominant color extraction. Our approach is primarily based on semantic segmentation complemented by salient object detection. Our key contribution is the extraction of the most representative dominant colors of the salient object in an

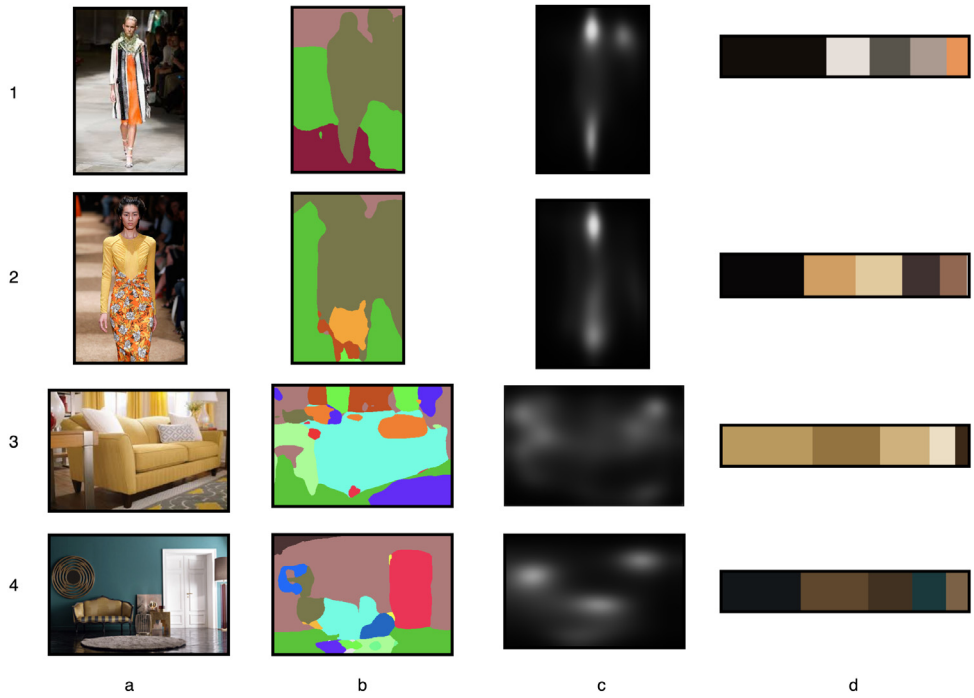


Fig. 7. Unsuccessful samples (a) the original image, (b) the segmented image, (c) the salient object image, and (d) dominant colors (extracted via the approach proposed by us) in proportionally descending order. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The color-difference values according to Sharma et al. (2005) for the images given in Fig. 8. The first column corresponds to the image label in the figure and the second and the third columns give the respective ΔE values.

Image label	ΔE Whole image	ΔE Proposed method
1	44.04	4.22
2	70.54	19.56
3	12.51	5.76
4	37.91	1.51
5	32.70	21.55
6	69.37	2.49
7	57.31	4.84
8	30.69	0.54
9	32.74	1.53
10	53.91	3.82
11	33.37	14.42
12	53.97	1.75
13	32.01	3.14
14	56.60	2.20
15	33.45	4.54
16	57.88	6.02
17	99.17	1.55
18	30.20	3.57
19	29.66	0.88
20	22.65	2.05

image in contrast to the current research, which produces the dominant colors treating the image as a whole without considering the individual objects within the image. As a result of our proposed method, the performance was typically increased by 14% on average than the approaches using the images as a whole (Fig. 8). Although the ΔE value exhibits some variation for different types of images, it appears that it is consistently better than the "whole image" approach. In addition, our proposed model could be used as a framework on a number of different occasions, both in academic research and commercial applications. For instance, our proposed model could be a part of the cycle of a new product in interior design, fashion, fragrance, etc. Furthermore, it can open up new opportunities to filter objects based on colors in e-commerce sites or search engines. This data could also be used to predict the future color combination preference for a given category. To the best of our knowledge, this work is an initial attempt at combining the dominant color extraction method with deep learning approaches to demonstrate the importance of color compatibility to possibly drive business value.

CRediT authorship contribution statement

Ayse Bilge Gunduz: Conceptualization, Methodology, Software, Visualization, Writing - review & editing. **Berk Taskin:** Methodology, Software, Visualization. **Ali Gokhan Yavuz:** Writing - review & editing, Supervision. **Mine Elif Karsligil:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Name: A. Bilge GUNDUZ Address: Davutpasa Mah. Yildiz Technical University, Department of Computer Engineering 34220 Esenler/Istanbul/Turkey Phone: +90 554 666 18 90 E-mail: aysebilgegunduz@gmail.com

Acknowledgments

We would like to thank the *HUEDATA* for providing us with necessary datasets. Special thanks to VRAY GUILLAUME MARC GEORGES and M. YASIN SAGLAM for their contribution in running the experimental tests.

Fig. 8. The results of dominant color extraction manually pointed out by the participants, determined from the whole image and using our proposed method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Abdi, H., 2007. The Kendall rank correlation coefficient. In: Encyclopedia of Measurement and Statistics. Citeseer, Sage, Thousand Oaks, CA, pp. 508–510.

Arthur, D., Vassilvitskii, S., 2006. k-means++: The Advantages of Careful Seeding. Tech. Rep., Stanford.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Baldevbhai, P.J., Anand, R., 2012. Color image segmentation for medical images using L* a* b* color space. IOSR J. Electron. Commun. Eng. 1 (2), 24–45.

Barnard, K., 1998. Modeling scene illumination colour for computer vision and image reproduction: A survey of computational approaches. Comput. Sci. Simon Fraser Univ. 39.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A., 0000. MIT saliency benchmark. http://saliency.mit.edu/.

Cattin, P., 2016. Digital Image Fundamentals: Introduction to Signal and Image Processing. University of Basel.

Chauhan, S., Prasad, R., Saurabh, P., Mewada, P., 2018. Dominant and LBP-based content image retrieval using combination of color, shape and texture features. In: Progress in Computing, Analytics and Networking. Springer, pp. 235–243.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Cheng, M.-M., Warrell, J., Lin, W.-Y., Zheng, S., Vineet, V., Crook, N., 2013. Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1529–1536.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Chollet, F., et al., 2015. keras.

Cornia, M., Baraldi, L., Serra, G., Cucchiara, R., 2016. A deep multi-level network for saliency prediction. In: 2016 23rd International Conference on Pattern Recognition. ICPR, IEEE, pp. 3488–3493.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2007. The pascal visual object classes challenge 2007 results.

Feng, Z., Yuan, W., Fu, C., Lei, J., Song, M., 2018. Finding intrinsic color themes in images with human visual perception. Neurocomputing 273, 395–402.

Forero, P.A., Kekatos, V., Giannakis, G.B., 2012. Robust clustering using outlier-sparsity regularization. IEEE Trans. Signal Process. 60 (8), 4163–4177.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, X., Shen, C., Boix, X., Zhao, Q., 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 262–270.

Jahanian, A., Vishwanathan, S., Allebach, J.P., 2015. Autonomous color theme extraction from images using saliency. In: Imaging and Multimedia Analytics in a Web and Mobile World 2015, Vol. 9408. International Society for Optics and Photonics, 940807.

Jetley, S., Murray, N., Vig, E., 2016. End-to-end saliency mapping via probability distribution prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5753–5761.

Kato, T., 1992. Database architecture for content-based image retrieval. In: Image Storage and Retrieval Systems, Vol. 1662. International Society for Optics and Photonics, pp. 112–123.

Ketchen, D.J., Shook, C.L., 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strateg. Manag. J. 17 (6), 441–458.

Kümmerer, M., Theis, L., Bethge, M., 2014. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045.

Kummerer, M., Wallis, T.S., Gatys, L.A., Bethge, M., 2017. Understanding low-and high-level contributions to fixation prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4789–4798.

Lechner, A., Harrington, L., 0000. Hue dataset. URL (Online). Available: http://www.hue-data.com.

Liu, N., Han, J., 2018. A deep spatial contextual long-term recurrent convolutional network for saliency detection. IEEE Trans. Image Process. 27 (7), 3264–3274.

Liu, S., Jiang, Y., Luo, H., 2018. Attention-aware color theme extraction for fabric images. Text. Res. J. 88 (5), 552–565.

Liu, S., Luo, H., 2016. Hierarchical emotional color theme extraction. Color Res. Appl. 41 (5), 513–522.

Liu, L.-x., Tan, G.-z., Soliman, M., 2012. Color image segmentation using mean shift and improved ant clustering. J. Central South Univ. 19 (4), 1040–1048.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Machajdik, J., Hanbury, A., 2010. Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM International Conference on Multimedia. pp. 83–92.

Ng, A., 2012. Clustering with the k-means algorithm. Mach. Learn.

Ou, L.-C., Luo, M.R., Woodcock, A., Wright, A., 2004. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color Res. Appl. 29 (3), 232–240.

Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X., 2017. Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081.

Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E., 2016. Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 598–606.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 658–666.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Schwarz, M.W., Cowan, W.B., Beatty, J.C., 1987. An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models. ACM Trans. Graph. 6 (2), 123–158.

Sharma, G., Wu, W., Dalal, E.N., 2005. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. In: Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, Vol. 30. (1), Wiley Online Library, pp. 21–30.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence.

TAKAHASHI, N., SHOJI, H., SAKAMOTO, T., Kato, T., 2018. An analysis on color characteristics of website images of restaurants according to price range. In: International Symposium on Affective Science and Engineering ISASE2018. Japan Society of Kansei Engineering, pp. 1–4.

Teel, D.R., Aspland, J.R., Jarvis, J.P., Dunlap, K.L., 1992. Improved methods for colour inventory management in the apparel industry. Int. J. Cloth. Sci. Technol. 4 (2–3), 66–70.

Terwogt, M.M., Hoeksma, J.B., 1995. Colors and emotions: Preferences and combinations. J. Gen. Psychol. 122 (1), 5–17.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Wang, C., Wang, X., Li, Y., Xia, Z., Zhang, C., 2018. Quaternion polar harmonic Fourier moments for color images. Inform. Sci. 450, 141–156.

Wang, C., Wang, X., Xia, Z., Ma, B., Shi, Y.-Q., 2019. Image description with polar harmonic fourier moments. IEEE Trans. Circuits Syst. Video Technol. 30 (12), 4440–4452, URL https://doi.org/10.1109/TCSVT.2019.2960507.

Wu, O., Han, M., 2018. Screenshot-based color compatibility assessment and transfer for Web pages. Multimedia Tools Appl. 77 (6), 6671–6698.

Xing, L., Zhang, J., Liang, H., Li, Z., 2018. Intelligent recognition of dominant colors for Chinese traditional costumes based on a mean shift clustering method. J. Text. Inst. 109 (10), 1304–1314.

Yan, Y., Ren, J., Li, Y., Windmill, J., Ijomah, W., 2015. Fusion of dominant colour and spatial layout features for effective image retrieval of coloured logos and trademarks. In: 2015 IEEE International Conference on Multimedia Big Data. IEEE, pp. 306–311.

Yang, N.-C., Chang, W.-H., Kuo, C.-M., Li, T.-H., 2008. A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval. J. Vis. Commun. Image Represent. 19 (2), 92–105.

Yuheng, S., Hao, Y., 2017. Image segmentation algorithms overview. arXiv preprint arXiv:1707.02051.

Zhang, J., Sclaroff, S., 2013. Saliency detection: A boolean map approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 153–160.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.

Zhou, W., Mok, P., Zhou, Y., Zhou, Y., Shen, J., Qu, Q., Chau, K., 2019a. Fashion recommendations through cross-media information retrieval. J. Vis. Commun. Image Represent. 61, 112–120.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 633–641.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A., 2019b. Semantic understanding of scenes through the ade20k dataset. Int. J. Comput. Vis. 127 (3), 302–321.