https://doi.org/10.2352/EI.2023.35.8.IQSP-306

What Are We Looking At? An Investigation on the Use of Deep Learning Models for Image Quality Assessment

Ha Thu Nguyen, Seyed Ali Amirshahi
Colourlab, Norwegian University of Science and Technology, Gjøvik, Norway
hatn@stud.ntnu.no, s.ali.amirshahi@ntnu.no

Abstract

In recent years different Image Quality Metrics (IQMs) that are focused on comparing the feature maps extracted from different pre-trained deep learning models have been introduced. While such objective IQMs have shown a high correlation with the subjective scores little attention has been paid on how they could be used to better understand the Human Visual System (HVS) or how observers evaluate the quality of images. In this study, by using different pre-trained Convolutional Neural Networks (CNNs) we identify the most relevant features for image quality assessment. By this our goal is to have a better understanding of which features play a dominant role when evaluating the quality of images. Experimental results on four benchmark datasets show that the most important feature maps represent repeated textures such as stripes or checkers, and feature maps linked to colors blue, or orange also play a crucial role. Additionally, when it comes to calculating the quality of an image based on a comparison of different feature maps, a higher accuracy can be reached when only the most relevant feature maps are used in calculating the image quality instead of using all the extracted feature maps from a CNN model.

Introduction

Over the last few decades different studies have focused on evaluating the quality of images and videos resulting in different Image and Video Quality Metrics (IQMs and VQMs respectively). In the case of images, IQMs try to predict the subjective evaluation done by observers and provide a consistent measure which can also be used for quality optimization. Until recently the introduced IQMs were based on traditional image processing techniques and took advantage of handcrafted features. With the introduction of different CNN based IQMs, we now have access to a huge amount of data extracted from images which can help us better understand the Human Visual System (HVS) and the process observers take to evaluate the quality of images.

While the CNN based IQMs show a higher performance compared to traditional IQMs [1, 2, 3], little attention has been paid on how such metrics work and the information we can acquire by analyzing their performance. A deep dive in how such metrics work will not only allow us to have a better understanding of the HVS but also how we can improve the performance of different CNN based IQMs. With this goal in mind, in this study we aim to detect how different feature maps extracted from the image would affect the overall performance of an IQM. Our focus will be on IQMs which are based on comparing different feature maps between the test and reference image.

Our results show that independent of the dataset, image, and the type of distortion that affects the image, specific features play a prominent role in the quality assessment of the image. Furthermore, while CNN based methods outperform traditional metrics, by using a limited number of such deep features, we would not only be able to reduce the computational costs but also improve the performance of the IQMs.

In the rest of the paper we first provide a short overview of the previous works done on CNN based IQMs. The methodology used in our work is then introduced in the next section followed by experimental results and conclusion of the work.

Related Works

Traditional IQMs simply measure the difference or the similarity between the reference and the test image (in the case of the full-reference and reduced-reference metrics) or how the quality image is with regards to an idea case scenario (no-reference metrics). The first approach, which is referred to as error visibility, calculates the error at the pixel level. At each location, a value corresponding to the difference between the pixels in the test (distorted) and the reference image is computed. A pooling operation, which is normally an average of all the results is then applied to the set of error values at all pixel positions to get a single value representing the quality of the image. Three representatives for this category of IQMs are Mean Square Error (MSE), PNSR, and ΔE_{ab}. Another group of IQMs take into account the tendency in the HVS to create mathematical models of the metrics [4]. Structural Similarity Index (SSIM) [5] is a perceptual quality metric, which is based on the principle that the HVS is adapted to extract structural information from images. The local structure similarity, which is constructed from three components: luminance, contrast, and structural comparison, is leveraged to evaluate the quality of images. Over years, different variants of SSIM have been proposed such as MS-SSIM [6], which apply SSIM at multiple scales of the image, F-SSIM [7] calculating similarity at low-level features, or IW-SSIM [8] extending a content-based weighted on measuring local similarity. The aforementioned SSIM methods show a better performance than their error visibility counterparts in image quality assessment. Some other metrics use the information-theoretic to measure the fidelity between the reference and distorted images such as VIF [9], or combine the above approaches as in VSNR [10].

With the development of deep neural networks, CNNs are considered as the efficient model for many computer vision tasks. Due to the lack of enough data to train and test a CBB model, initial IQMs which took advantage of deep learning techniques [2, 1, 3] used a pre-trained CNN model to extract different features from the image which then they used in their propsoed IQM.

Amirshahi et al .[2] proposed a full-reference IQM measuring the similarity between the extracted features from the test and the reference images at multiple layers. Following this approach, they extract the feature maps at different convolutional layers of a pre-trained CNN model and compared the feature maps using traditional IQMs such as PSNR and SSIM [1]. Around the same time, Gao et al. [3] used a deeper pre-trained CNN, the VGG model [11] for feature extraction, and calculate the local similarity of the feature maps similar to the main idea behind the SSIM IQM. The experimental results show that the metric that compare the extracted feature from the pre-trained CNN model achieve a higher accuracy in evaluating image quality than traditional IQMs.

Methodology

In this study we focus on IQMs which use a comparison of the feature maps extracted from a pre-trained CNN model. In the previous section we introduced few methods which take such an approach [1, 2, 3]. To find the features which play the most important role in evaluating the quality of an image we use a wrapper-based approach (forward selection) [12] which has previously been used in similar fields of research [13, 14]. In such an approach, we start with an empty set of feature maps and iteratively add the feature maps which result in the highest increase in the correlation between the subjective scores and the proposed IQM. Using this approach, a combination of feature maps which provide the highest correlation with the subjective scores is detected.

The detailed step-by-step approach that iteratively builds a set of feature maps that provide the best performance of a given IQM can be found in the following.

1. An empty set of feature maps is created.
2. Feature maps are extracted from the test ($\mathscr{I}_T$) and the reference ($\mathscr{I}_R$) images at all convolutional layers.
3. The quality score for all the test images in our dataset at each feature map is then calculated by

$$q((\mathscr{F}_{n,m}(\mathscr{I}_{T_1})) = IQM((\mathscr{F}_{n,m}(\mathscr{I}_{T_1}),(\mathscr{F}_{n,m}(\mathscr{I}_{R_1})) \quad (1)$$

   where $\mathscr{F}_{n,m}(\mathscr{I}_{T_1})$ corresponds to the m^{th} feature map in the n^{th} convolutional layer for the test image $\mathscr{I}_{T_1}$, IQM indicates the IQM used in our calculations, and $q((\mathscr{F}_{n,m}(\mathscr{I}_{T_1}))$ corresponds to the quality score of image $\mathscr{I}_{T_1}$ using feature map $\mathscr{F}_{n,m}(\mathscr{I}_{T_1})$.
4. The correlation between the quality score of each feature map and the subjective score is calculated.
5. The feature map with the highest correlation score is then added to the feature set.
6. The remaining feature maps are then individually combined with the feature set and the quality of the test images in our dataset are then calculated.
7. The feature map which in combination with the feature set we have provides the highest correlation with the subjective scores is then added to the feature set.
8. This process (steps six and seven) continues until adding any new feature map to the feature set does not result in an increase in the correlation scores.

The selected feature set is then visualized to better understand how observers evaluate the quality of an image.

Table 1: Number of features in each feature set which reach the highest correlation rate along with the corresponding non-linear Pearson correlation achieved using the feature set compared to when all features are calculated in the AlexNet model.

Dataset	Num. of selected feature maps	Non-linear Pearson correlation	
		Using selected feature maps	Using all feature maps
CID:IQ 100cm [15]	2	0.89	0.80
CID:IQ 50 cm [15]	12	0.87	0.68
CSIQ [16]	16	0.90	0.87
TID2013 [17]	9	0.95	0.91
CIDGD [18]	13	0.75	0.61

Experiment and Results

To create each feature set we tested our approach on four benchmark subjective datasets namely, CID:IQ [15], CSIQ [16], TID2013 [17], and CID:GD [18]. The feature maps were extracted from a pre-trained AlexNet [19] model and compared to each other using the SSIM IQM [5] similar to what was proposed in [1]. Depending on the dataset, different number of feature maps were selected in our feature set (Table 1). From the results, we can observe that compared to the original IQM approach which is based on more than 1000 feature maps in the AlexNet model the proposed approach uses relatively a small number of features. This is while the correlation values obtained using the collected feature set show an increase compared to when all the feature maps in the AlexNet model is used. This could be a good evidence that the selected feature maps play an important role in assessing the quality of images.

What Features Play an Important Role in Image Quality Assessment?

To have a better understanding of the feature maps which play an important role in the quality assessment of images we tried to visualize and interpret them using different approaches. In the first approach, we use the DeepVis [20] method which provides detail on how the pixels in an image affect the response at a feature map in a hidden layer of a CNN model. From the visualized maps (Figure 1) the common pattern which appears in the visualization maps is texture where the activated regions are not smooth (see feature maps 8, 9, and 27). Meanwhile, feature maps 15 and 40 seem to indicate the sky/color blue, and feature map 52 corresponds to the color orange in the first image (Figure 1(a)).

As the DeepVis technique significantly depends on the content of the input image, misinterpretation can happen. Thus, we used the Activation Maximization approach [21] to synthesize the input image and indicate the property that the model learns at each feature map. The generated image for each feature map (Figure 2) provides us with a better sense of the feature maps that play an important role in image quality evaluation, such as lined patterns (feature maps 27 and 9), and bluish color (feature maps 15).

Finally, the semantic representation of the NetDissect approach [22] was integrated into the analysis. This method of latent space visualization for deep networks aims at interpreting the semantic meaning of the feature extracted from the intermediate convolutional layers. In the method, a dataset containing several visual concepts such as object, texture, or color was collected and

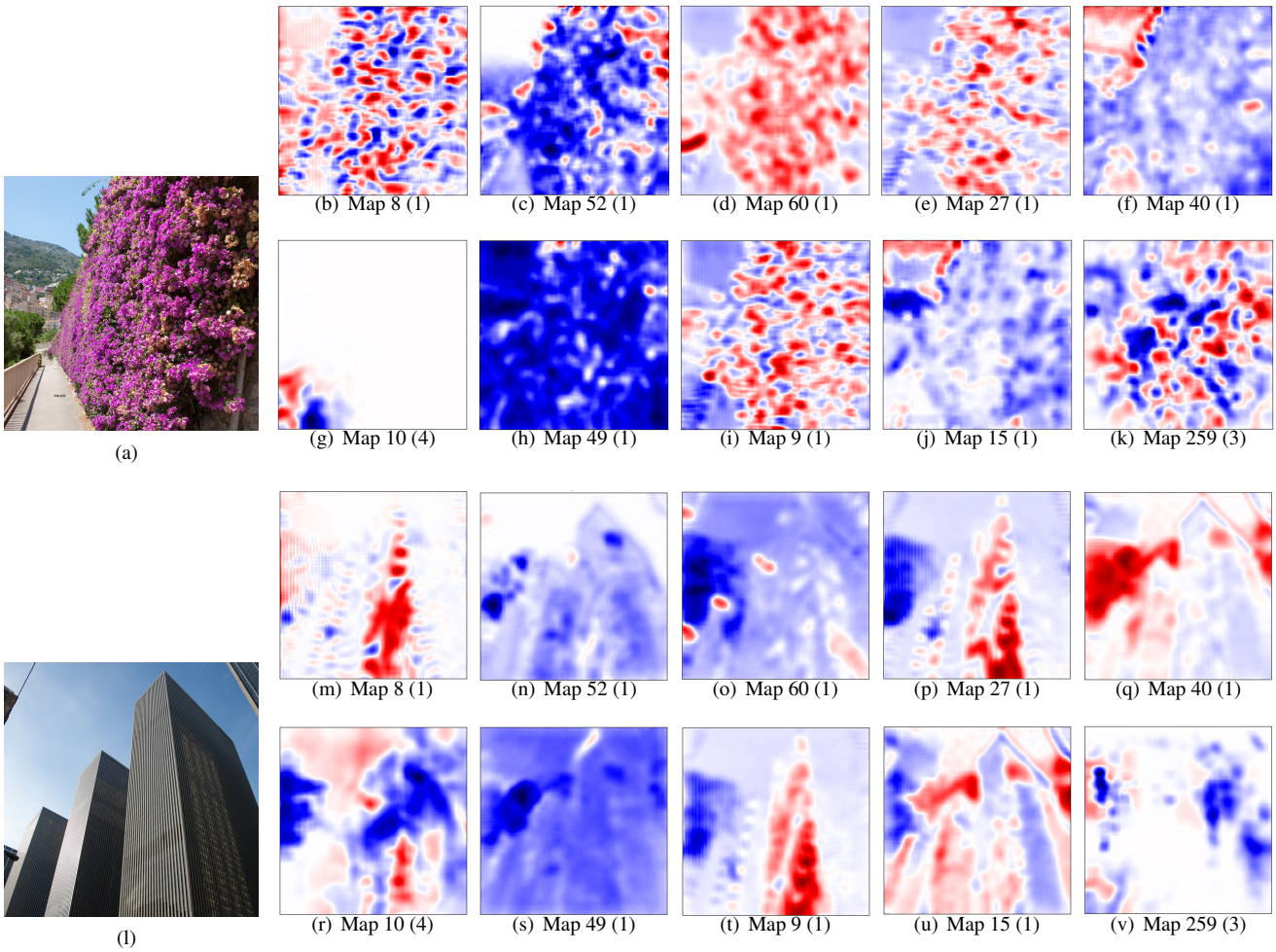


Figure 1. *Sample images from the CID:IQ [15] dataset ((a) and (l)) along with the 10 most important features ordered based on importance and visualized using the DeepVis [20] approach ((b)-(k) and (m)-(v) respectively) using the AlexNet model. The numbers inside the parenthesis indicate the convolutional layer that the feature map was extracted from. Red pixels represent what the model learns in each feature map.*

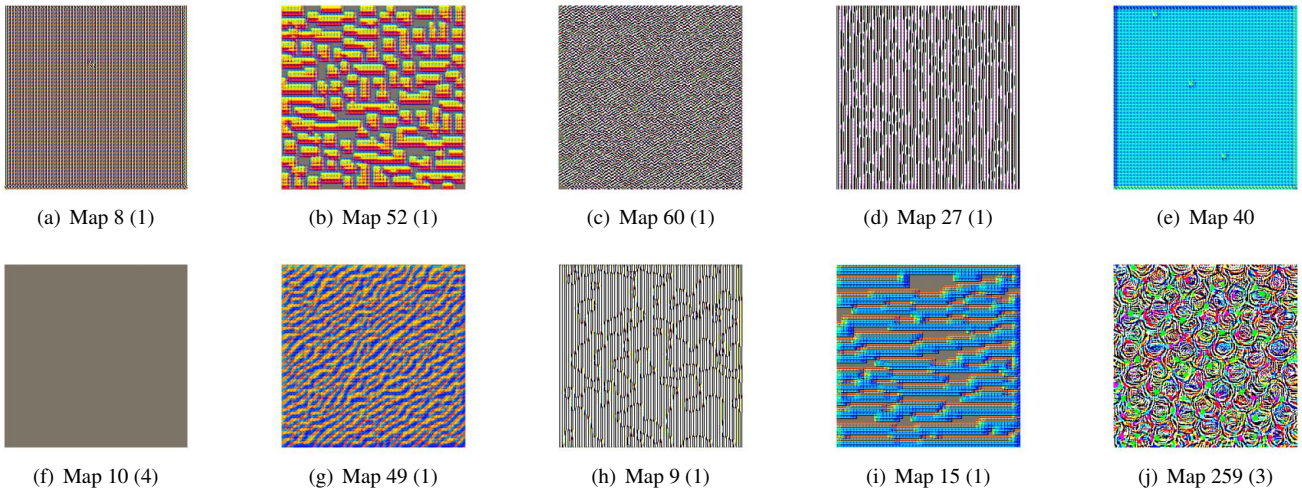
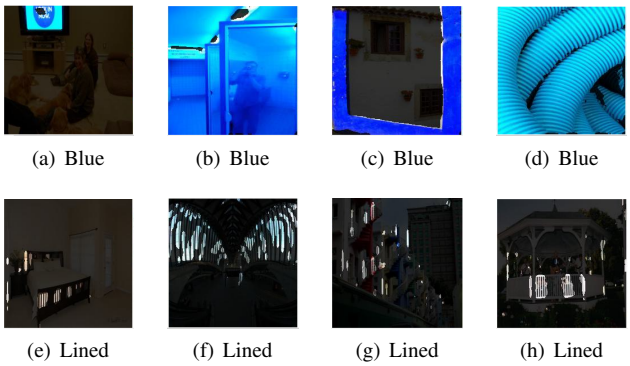


Figure 2. *The selected feature maps for the CID:IQ [15] dataset using the AlexNet model visualized by the activation maximization approach [21]. The numbers inside the parenthesis indicate the convolutional layer that the feature map was extracted from.*



(a) Blue (b) Blue (c) Blue (d) Blue

(e) Lined (f) Lined (g) Lined (h) Lined

Figure 3. *Visual concept detector by NetDissect [22] for feature map 15 ((a)-(d)) and feature map 27 ((e)-(h)) at convolutional layer 1 of the AlexNet model. (a)-(d) shows the segmentation generated by feature map 15 on the images of the highest activation. (e)-(h) are the result for feature map 27. The captions below each image indicate the detected visual category that the segmented part represents.*

densely labeled at pixel level by human annotators. The feature activated at each feature map from the CNN model is considered as a segmentation mask for each concept. A feature map is interpreted by a concept if its corresponding segmentation mask and the ground-truth annotation of images in that category match together (Figure 3). Although there are many objects in the images, the highlighted response region share the same visual meaning. For example, in the case of the 15^{th} feature map in the first convolutional layer (Figure 3(a)-(d)) which is the interpretation result for feature map 15 of layer 1, highlight the color blue. It suggests that the feature that is extracted from this feature map represents the color blue. Similarly, feature map 27 (Figure 3(e)-(h)) seems to correspond to a lined pattern in the image. The interpretation for the remaining eight selected maps is: lined (feature map 8 - layer one), color orange (feature map 52 - layer one), lacelike (feature map 60 - layer one), blue sky (feature map 40 - layer one), wheel (feature map 10 - layer four), ball pit (feature map 49 - layer one), lined (feature map 9 - layer one), honeycombed (feature map 259 - layer three).

From running the three different approaches, similar interpretations can be made. Although the selected feature set in each dataset are different (Figure 4), they represent a similar visual concept. That is, in general features related to repeated patterns such as lines, checkers, or stripes, and features that have an emphasis on the colors blue or orange, tend to play important roles in determining the quality of an image.

To investigate if the depth of the CNN model plays a role in the features selected in our feature set, we also used the same approach on a pre-trained VGG16 [11] model (Table 2). The VGG network was first introduced in 2014, with the purpose of image recognition. This model was also trained on the ImageNet benchmark, but compared to AlexNet it contains a higher number of convolutional layers. There are three main variants of VGG: VGG11, VGG16, and VGG19. The VGG16 model which consists of 13 convolutional layers was used in our experiments. We can see that compared to the AlexNet model, in the case of most datasets a higher number of feature maps were chosen for our feature set while at the same time the accuracy has slightly improved. As most of the feature maps are extracted from the deep layers (Figure 5), their representations are not easy to explain.

Table 2: Number of features in each feature set which reach the highest correlation rate along with the corresponding non-linear Pearson correlation achieved using the feature set compared to when all features are calculated in the VGG16 model.

Dataset	Num. of selected feature maps	Non-linear Pearson correlation	
		Using selected feature maps	Using all feature maps
CID:IQ 100cm [15]	35	0.93	0.85
CID:IQ 50 cm [15]	28	0.92	080
CSIQ [16]	25	0.91	0.83
TID2013 [17]	24	0.90	0.88
CIDGD [18]	5	0.71	0.62

Conclusion and Future work

In this study we tried to investigate what features play an important role in evaluating the quality of images. For this goal, we used an IQM which takes advantage of a pre-trained CNN model and compares the feature maps between the reference and test images. Our results show that by using a limited number of features we are able to not only reduce the computational costs but also increase the accuracy of the IQM. When it comes to the selected feature maps, it is clear that in general features related to repeated patterns such as lines, checkers, or strips and the colors blue and orange play a dominant role in evaluating the quality of the images.

While the findings of the study could be seen as a promising first step there is still a huge room for improvement and investigation on how the quality of the image is evaluated by observers. Also, with the introduction of new datasets that provide the individual scores of observers [23, 24] it would be interesting to see if the feature maps selected in the feature set are different for different observers.

References

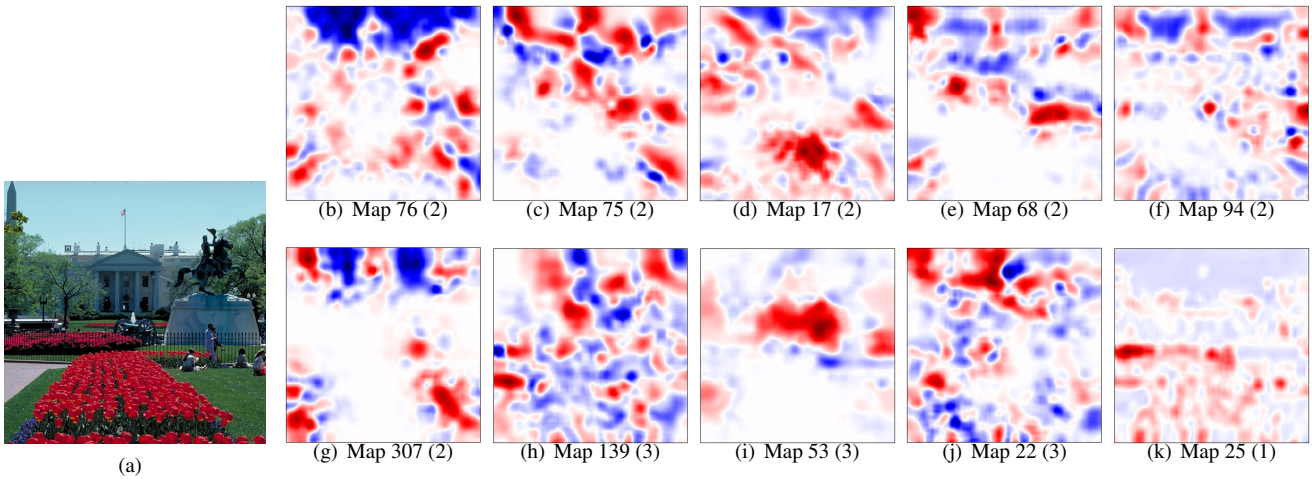
[1] Seyed Ali Amirshahi, Marius Pedersen, and Azeddine Beghdadi. Reviving traditional image quality metrics using cnns. In *Color and imaging conference*, volume 2018, pages 241–246. Society for Imaging Science and Technology, 2018.

[2] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu. Image quality assessment by comparing cnn features between images. *Journal of Imaging Science and Technology*, 60(6):60410–1, 2016.

[3] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017.

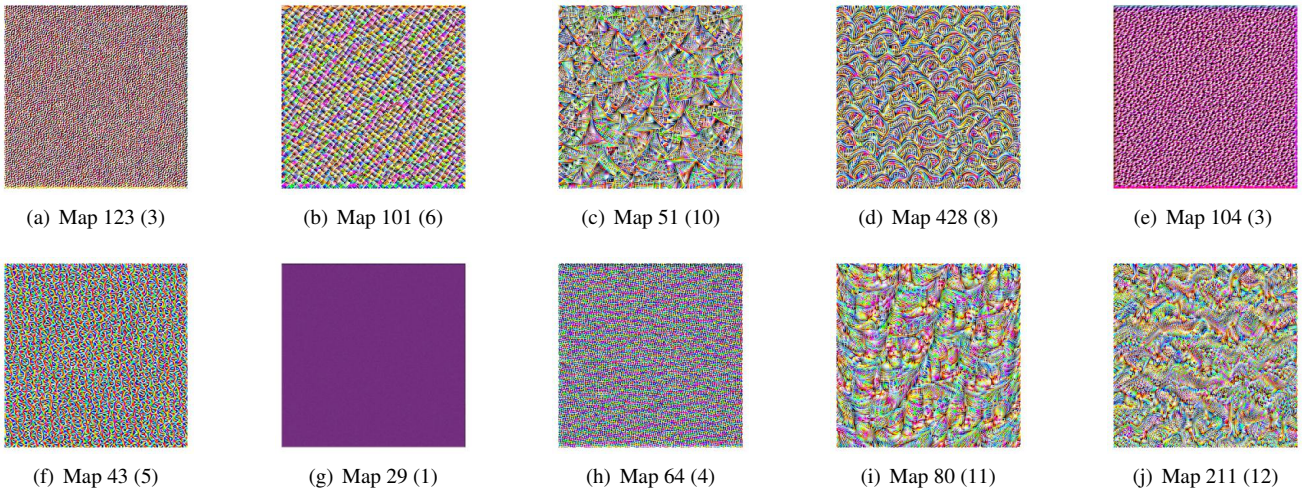
[4] Farah Torkamani-Azar and Seyed Ali Amirshahi. A new approach for image quality assessment using svd. In *2007 9th International Symposium on Signal Processing and Its Applications*, pages 1–4. IEEE, 2007.

[5] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.



(a) (b) Map 76 (2) (c) Map 75 (2) (d) Map 17 (2) (e) Map 68 (2) (f) Map 94 (2) (g) Map 307 (2) (h) Map 139 (3) (i) Map 53 (3) (j) Map 22 (3) (k) Map 25 (1)

Figure 4. *Sample image from the CSIQ [16] dataset (a) and selected feature maps visualized using the DeepVis [20] approach. The numbers inside the parenthesis indicate the convolutional layer that the feature map was extracted from. Red pixels represent what the model learns in each feature map.*



(a) Map 123 (3) (b) Map 101 (6) (c) Map 51 (10) (d) Map 428 (8) (e) Map 104 (3) (f) Map 43 (5) (g) Map 29 (1) (h) Map 64 (4) (i) Map 80 (11) (j) Map 211 (12)

Figure 5. *The selected feature maps using the VGG16 for the CID:IQ [15] dataset visualized by the activation maximization approach [21]. The numbers inside the parenthesis indicate the convolutional layer that the feature map was extracted from.*

[6] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[7] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[8] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on image processing*, 20(5):1185–1198, 2010.

[9] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[10] Damon M Chandler and Sheila S Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE transactions on image processing*, 16(9):2284–2298, 2007.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[13] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006.

[14] Seyed Ali Amirshahi and Joachim Denzler. Judging aesthetic quality in paintings based on artistic inspired color features. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017.

[15] Xinwei Liu, Marius Pedersen, and Jon Yngve Hardeberg. Cid: Iq–a new image quality database. In *International Conference on Image and Signal Processing*, pages 193–202. Springer, 2014.

[16] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010.

[17] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

[18] Marius Pedersen and Seyed Ali Amirshahi. Colourlab image database: Geometric distortions. In *Color and Imaging Conference*, volume 2021, pages 258–263. Society for Imaging Science and Technology, 2021.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[20] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

[21] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[22] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[23] Olga Cherepkova, Seyed Ali Amirshahi, and Marius Pedersen. Analyzing the variability of subjective image quality ratings for different distortions. In *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2022.

[24] Olga Cherepkova, Seyed Ali Amirshahi, and Marius Pedersen. Analysis of individual quality scores of different image distortions. In *Color and Imaging Conference*, volume 2022. Society for Imaging Science and Technology, 2022.

Author Biography

Ha Thu Nguyen is currently a master's student at the Erasmus+ Joint Master program in Computational Color and Spectral Imaging at Norwegian University of Science and Technology, Norway. She obtained her BSc in Electronics and Telecommunications in 2019 from the Hanoi University of Science and Technology, Vietnam. Her research interests include image processing, image quality assessment, and deep learning.

Seyed Ali Amirshahi is an Associate Professor at the Norwegian University of Science and Technology (NTNU). His work is focused on image/video quality assessment and computational aesthetics. He received his PhD from the Friedrich Schiller University of Jena in Germany (2015). Prior to his current position he was a Marie Curie post-doctoral Fellow at NTNU and a visiting researcher at University Sorbonne Paris Nord. Prior to that he was a post-doctoral Fellow at the International Computer Science Institute in Berkeley, CA.