

A HYPOTHESIS FOR THE COGNITIVE DIFFICULTY OF IMAGES

Anonymous authors

Paper under double-blind review

ABSTRACT

This paper proposes a hypothesis to analyze the underlying reason for the cognitive difficulty of an image from two perspectives, *i.e.* a cognitive image usually makes a DNN strongly activated by cognitive concepts; discarding massive non-cognitive concepts may also help the DNN focus on cognitive concepts. Based on this hypothesis, we use multi-variate interactions to represent cognitive concepts and non-cognitive concepts contained in an image, and further design a set of image revision operations to decrease the cognitive difficulty of the image. In experiments, we found that the revised image was usually more cognitive than the original one. Besides, we also discovered that strengthening cognitive concepts and weakening non-cognitive concepts could improve the aesthetic level of an image.

1 INTRODUCTION

Nowadays, there are plenty of studies devoted to evaluating and improving the image quality, such as image super-resolution (Yang et al., 2010; Dong et al., 2015) and image aesthetic assessment (Jin et al., 2019b; Deng et al., 2018). In contrast, this paper focuses on the cognitive difficulty of an image, *i.e.* whether this image is easy for people to recognize.

Unfortunately, there does not exist a proper method to define and model the cognitive difficulty of an image, because the cognitive difficulty of the image has many aspects. For example, an object with bright colors, clear contours, and simple contents is usually believed to be more cognitive (Farah, 1992). Although such cognitive information can be learned by a model in a supervised manner, this paper focuses on the mechanism for the cognitive difficulty. *I.e., we do not directly teach the model which object is cognitive, but we expect this model automatically to push the grass greener and flowers redder in Fig. 4.*

To this end, there remain two essential challenges to model the cognitive difficulty of an image. (1) The cognitive process is subjective, and there is no proper method to directly model the information-processing mechanism of the human brain. (2) There lacks the theoretical foundation to explain the mechanism of distinguishing cognitive visual concepts and non-cognitive visual concepts in the human brain. Here, the visual concept is referred to as features for certain shapes or textures, instead of representing semantics. Note that the cognitive concept has the dramatic difference from the visual saliency, which will be discussed later.

Fortunately, the fast development of deep learning provides us new ways to deal with the aforementioned two challenges. First, we try to use an artificial deep neural network (DNN) learned for object classification on a huge dataset (*e.g.* the ImageNet dataset (Krizhevsky et al., 2012)) as a substitute to mimic the human brain.

Second, we discover that the interaction defined in game theory can be used to explain cognitive concepts and non-cognitive concepts. The basic idea is that both the human brain and the DNN do not use a single pixel/patch for inference, but they use groups of highly interacted pixels/patches to form interaction patterns, instead. Thus, these interaction patterns can be considered to represent visual concepts.

Then, let us introduce why and how to use the game-theoretic interaction between multiple patches to mimic human cognition. Given a pre-trained DNN f and an image with n patches $N = \{1, 2, \dots, n\}$, we use the Harsanyi dividend $I(S)$ (Harsanyi, 1963) to measure the multi-variate interaction between

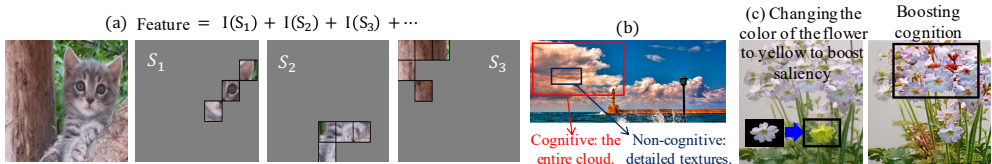


Figure 1: (a) Sketch for the multi-variate interaction. The feature representation of a DNN can be decomposed into interaction utilities of different visual concepts. (b) The cloud, as an entire object, is regarded as a cognitive concept, but the detailed textures inside the cloud are non-cognitive and forgettable. (c) Difference between boosting saliency and boosting cognition. Boosting saliency changes the color of the flower to attract more attention, while boosting cognition makes flowers redder to be recognized more easily. Please see Appendix A for the detailed difference between saliency and cognitive concepts.

a set of input patches $S \subseteq N$. The Harsanyi dividend is a well-known and standard metric to estimate the marginal benefit when patches in S collaborate with each other *w.r.t.* the case when they work independently. For example, let us focus on Fig. 1 (a), $S_1 = \{\text{mouth, eye, ear}\} \subseteq N$. The mouth, ear, and eye patches in S_1 collaborate with each other to trigger a feature, which represents the cat head concept for inference. This triggered feature is referred to as the interaction utility $I(S_1) \in \mathbb{R}^d$ (here we extend the Harsanyi dividend from a scalar to a vector). The absence of any patch will deactivate the feature of the head concept. More crucially, the Harsanyi dividend ensures that the feature can be decomposed into utilities of different concepts, $\text{feature} = \sum_{S \subseteq N} I(S)$. For example, in Fig. 1 (a), the feature of the image can be represented as the sum of the interaction utility of the head concept $I(S_1)$, the utility of the paw concept $I(S_2)$, the utility of the wood concept $I(S_3)$, etc.

Furthermore, the Harsanyi dividend enables us to distinguish cognitive concepts and non-cognitive concepts. There are 2^n concepts theoretically, but only very sparse concepts have significant effects $\|I(S)\|_2$ on the feature, namely *cognitive concepts*. Whereas, a large ratio of concepts have negligible influences $\|I(S)\|_2$ on the feature, namely *non-cognitive concepts*. Because the DNN is trained using massive samples, we consider such cognitive concepts and non-cognitive concepts can mimic human cognition, to some extent.

Therefore, we propose a hypothesis for the cognitive difficulty of an image that **a cognitive image usually contains many cognitive concepts which are strongly activated by the human brain; discarding massive non-cognitive concepts may also help the human focus on cognitive concepts**, *i.e.* triggering clean and strong signals of cognitive concepts to the brain and letting people ignore non-cognitive concepts. Based on the hypothesis, we can quantify and decrease the cognitive difficulty of an image by simply changing the hue, saturation, brightness, and the sharpness of different regions, without changing the layout or the content of the image. Besides, in experiments, we discover that enhancing cognitive concepts and discarding non-cognitive concepts of an image may improve its aesthetic level. It may be because such operation can reduce cognitive costs of the human.

The essential difference between cognitive concepts and salient pixels is as follows. Saliency is usually defined at the pixel level. Specifically, saliency methods (Mechrez et al., 2019) simply classify all pixels into salient pixels and non-salient pixels as a black and white problem or assess the saliency numerically from 0 to 1. In short, the goal of saliency methods is to attract the human attention based on the contrast or brightness of each pixel, without concerning the cognition of each pixel. In comparison, the cognitive difficulty mainly depends on the conceptual representation. In other words, a pixel/patch is cognitive, if and only if it collaborates with other pixels/patches to form a cognitive concept. For example, although the detailed textures within the cloud in Fig. 1 (b) are non-cognitive and forgettable, the cloud, as an entire object, is cognitive. Moreover, Fig. 1 (c) shows that boosting saliency changes the color of the flower to make it more distinctive and attract more attention. Whereas, boosting cognition makes flowers redder to be recognized more easily. Please see Appendix A for discussions about the difference between attribution methods and cognitive concepts.

Contributions of this paper are summarized as follows. (1) We propose a hypothesis for the cognitive difficulty of an image. (2) We use the multi-variate interaction to mimic the concept of human cognition. (3) We propose a set of operations to revise an image to decrease its cognitive difficulty.

2 RELATED WORK

Cognitive science mainly focuses on how the human brain represents, processes and transforms information (Thagard, 2008). Some studies used Bayesian models to mimic the mechanism of human cognition (Xu & Tenenbaum, 2007; L Griffiths et al., 2008; Tenenbaum, 1999), and some researches evaluated human cognition through behavioral studies (Laina et al., 2020; Jia et al., 2013). However, previous studies did not propose a direct hypothesis to explain the mechanism of cognitive concepts. In contrast, this paper uses the game-theoretic interaction to model cognitive and non-cognitive visual concepts.

Cognitive difficulty of an image is an important issue, but it does not receive much attention. Existing studies related to this topic mainly focused on the memorability of an image, *i.e.* whether an image was memorable or forgettable (Baveye et al., 2016; Fajtl et al., 2018; Khosla et al., 2015; Goetschalckx et al., 2019; Sidorov, 2019). However, these studies were mainly conducted by training a model to fit the manually annotated memorability in a supervised manner, without proposing a hypothesis for memorability. Meanwhile, these studies also did not provide theoretical principles to ensure that the model could successfully reflect the memorability of humans. In comparison, this paper proposes a hypothesis for the cognitive difficulty of the image, which explores the underlying reason for cognitive concepts and non-cognitive concepts without the help of human annotations.

Image aesthetics can be roughly classified into image aesthetic assessment (Talebi & Milanfar, 2018; Sheng et al., 2018; Hosu et al., 2019; Jin et al., 2019b) and image aesthetic manipulation (Deng et al., 2018; Sheng et al., 2018; Wang et al., 2019; Moran et al., 2020). However, previous studies were mainly conducted in an end-to-end manner without modeling the image aesthetics. In this paper, we extend the hypothesis for the cognitive difficulty to partially explain the image aesthetics, *i.e.* discarding non-cognitive concepts and strengthening cognitive concepts can improve the aesthetic level of an image, which is successfully verified by experimental results.

3 HYPOTHESIS FOR THE COGNITIVE DIFFICULTY OF AN IMAGE

In this paper, we propose a hypothesis for the cognitive difficulty of an image that *a cognitive image usually makes a DNN strongly activated by cognitive concepts; discarding massive non-cognitive concepts may also help the DNN focus on cognitive concepts.*

In other words, if a DNN is triggered by massive visual concepts of an image, but none of them are strong enough to dominate the cognition (inference), then we may consider this image is difficult to recognize. To verify this hypothesis, we use the multi-variate interaction as a new tool to decompose visual concepts from features encoded in the neural network, and further distinguish cognitive concepts and non-cognitive concepts. Based on the hypothesis, we revise the image to reduce its cognitive difficulty, and experimental results verify the hypothesis.

3.1 MULTI-VARIATE INTERACTION

Definition of the multi-variate interaction. We use the multi-variate interaction between multiple pixels/patches to represent cognitive concepts and non-cognitive concepts of an image in game theory. Given a pre-trained DNN f and an image consisting of n patches, $N = \{1, 2, \dots, n\}$, $f(N) \in \mathbb{R}^d$ denotes the feature of an intermediate layer. The DNN usually does not use a single patch for feature representation, but it uses groups of highly interacted patches to form interaction patterns. In this way, we quantify the interaction between multiple patches as the additional utility to the feature representation when these patches *collaborate with each other*, in comparison with the case of *they working individually*.

To this end, we regard a visual concept encoded in the DNN as a salient interaction pattern. Let us consider a set of patches {ear, eye} in Fig. 1 (a) for the representation of a cat as a toy example, in order to understand the multi-variate interaction. If we add the *mouth* patch into this set and obtain the concept $S_1 = \{\text{ear, eye, mouth}\}$, then the collaboration between the eye, the nose, and the mouth will add a complementary interaction utility $I(S_1)$ to the feature of the cat. The absence of any patch in the concept S_1 will destroy this concept, thereby removing the interaction utility $I(S_1)$.

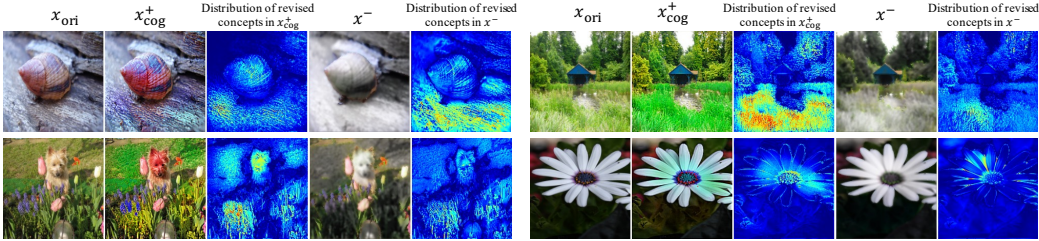


Figure 2: Visualization of the original image x_{ori} , the image x_{cog}^+ revised by strengthening cognitive concepts ($\min_{\theta} \text{Loss}_2(\theta)$, defined in Eq. (6)), the image x^- revised by weakening concepts ($\min_{\theta} \text{Loss}_1(\theta) - \text{Loss}_2(\theta)$). Note that cognitive concepts are not just foreground objects, but are determined by their interaction utilities for the feature representation. Hence, we decrease the cognitive difficulty of the image by pushing collaborations inside concepts towards more cognitive ones (e.g. making the grass greener), rather than just changing the contrast of some pixels like saliency methods. Here, we revised images from the ImageNet dataset based on the VGG-16 model. Please see Appendix G for more results.

Specifically, we use the Harsanyi dividend (Harsanyi, 1963), a well-known and standard metric, to measure the interaction utility $I(S) \in \mathbb{R}^d$. We discover that the feature $f(N)$ can be decomposed into interaction utilities of different visual concepts S , when we extend the Harsanyi dividend from a scalar to a vector. Please see Appendix B for the proof of this decomposition.

$$f(N) = \sum_{S \subseteq N} I(S), \quad s.t. \quad I(S) \stackrel{\text{def}}{=} \sum_{L \subseteq S} (-1)^{|S|-|L|} f(L). \quad (1)$$

$f(S)$ denotes the feature of an intermediate layer in the DNN when only patches in the concept S are given. Particularly, $I(\emptyset) = f(\emptyset)$, and $|\cdot|$ denotes the cardinality of the set. In the computation of $f(S)$, we mask patches in $N \setminus S$ to baseline values and keep patches in S unchanged, so as to approximate the removal of patches in $N \setminus S$. Specifically, we follow settings in (Dabkowski & Gal, 2017) to set the baseline value as the mean value of all patches over all images. Please see (Dabkowski & Gal, 2017) or Appendix C for details.

In this way, the Harsanyi dividend measures the marginal utility of the concept S , when interaction utilities of all smaller subsets of patches in S are removed, i.e. $\{I(L) | L \subsetneq S\}$. For a better understanding, let us consider the computation of the Harsanyi dividend *w.r.t* the cat head in Fig. 1 (a). The interaction utility of this cat head concept S_1 is defined as $I(S_1 = \{\text{ear, eye, mouth}\}) = f(\{\text{ear, eye, mouth}\}) - I(\emptyset) - I(\{\text{eye}\}) - I(\{\text{mouth}\}) - I(\{\text{ear}\}) - I(\{\text{eye, mouth}\}) - I(\{\text{eye, ear}\}) - I(\{\text{mouth, ear}\})$.

Understanding cognitive concepts and non-cognitive concepts encoded in a DNN. There are totally 2^n concepts in Eq. (1), but we discover that only a few concepts make significant effects on the feature $f(N)$, i.e. generating **large interaction utilities** $\|I(S)\|_2$. These concepts are regarded as **cognitive concepts**, because the DNN mainly uses these concepts to represent the feature. In comparison, most other concepts composed of random patches make negligible impacts on the feature, i.e. $\|I(S)\|_2$ **is small**. We consider such concepts as **non-cognitive concepts**, because they are usually ignored by the DNN. In this way, the overall feature $f(N)$ can be rewritten as the sum of interaction utilities of massive non-cognitive concepts $\Omega_{\text{non-cognitive}}$ and interaction utilities of sparse cognitive concepts $\Omega_{\text{cognitive}}$. Here, $\Omega_{\text{cognitive}} \cup \Omega_{\text{non-cognitive}} = \{S | S \subseteq N\}$.

$$f(N) = \sum_{S \subseteq N} I(S) = \sum_{S \in \Omega_{\text{cognitive}}} I(S) + \sum_{S \in \Omega_{\text{non-cognitive}}} I(S). \quad (2)$$

Note that non-cognitive concepts are NOT equivalent to random noises on images. Instead, these non-cognitive concepts are regarded as sets of patches, interactions inside which are not useful for the feature representation. For example, the detailed texture of the water in the top right of Fig. 2 may not provide significant features for the DNN, thereby being considered as a non-cognitive concept. In comparison, cognitive concepts represent sets of patches, whose interactions are used for the feature representation, NOT just representing foreground objects in classification. For example, the dog on the foreground and the grass on the background are both cognitive concepts in Fig. 2.

3.2 APPROXIMATE APPROACH TO REVISING VISUAL CONCEPTS

Based on the hypothesis and above analysis, we can reduce the cognitive difficulty of images by strengthening cognitive concepts with large $\|I(S)\|_2$ values. Moreover, we can also discard non-cognitive concepts with small $\|I(S)\|_2$ values to make the DNN ignore non-cognitive concepts and move attention to cognitive concepts. However, according to Eq. (2), the enumeration of all cognitive concepts and non-cognitive concepts is NP-hard. Fortunately, we find a close relationship between interaction utilities of visual concepts and Shapley values (Shapley, 1953). In this way, we can use the intermediate term $\Delta f(i, L)$ (defined in Eq. (3)) in the Shapley value to design two loss functions, which force the DNN to strengthen cognitive concepts and to discard non-cognitive concepts, respectively.

Definition of Shapley values and $\Delta f(i, L)$. Before the optimization of visual concepts, we first introduce Shapley values *w.r.t.* the feature representation of an intermediate layer in the DNN. The Shapley value is widely considered as an unbiased estimation of the numerical utility *w.r.t.* each patch in game theory (Weber, 1988). Without loss of generality, we extend the scalar Shapley value to a vectorized one. It is because this paper focuses on the utility of each patch to represent a certain feature, rather than the utility to a scalar classification result. Specifically, given a trained DNN f and an image x with n patches $N = \{1, 2, \dots, n\}$, let $f(x) \in \mathbb{R}^d$ denote the feature of an intermediate layer. The Shapley value is proposed to fairly divide and assign the overall effects on the feature representation to each patch, as follows.

$$\phi(i|N) = \sum_{L \subseteq N \setminus \{i\}} \frac{(n - |L| - 1)!|L|!}{n!} [\Delta f(i, L)]. \quad (3)$$

Here, $\phi(i|N) \in \mathbb{R}^d$ represents the utility of the patch i to the feature representation. $\Delta f(i, L) = f(L \cup \{i\}) - f(L) \in \mathbb{R}^d$ measures the *marginal utility* of the patch i to the feature representation, given a set of contextual patches L . Please see Appendix D for details.

Using $\Delta f(i, L)$ to strengthen/discard visual concepts. In order to overcome the aforementioned problem with the computational cost, we have proven that the term $\Delta f(i, L)$ in the Shapley value can be re-written as the sum of interaction utilities of some visual concepts (proved in Appendix E).

$$\Delta f(i, L) = \sum_{L' \subseteq L} I(S' = L' \cup \{i\}). \quad (4)$$

According to empirical observations, $\Delta f(i, L)$ is sparse among all contexts L . In other words, only a small ratio of contexts L have significant impacts $\|\Delta f(i, L)\|_2$ on the feature. In this way, we can roughly consider cognitive concepts are usually contained by the term $\Delta f(i, L)$ with a large L_2 norm, while the term $\Delta f(i, L)$ with a small L_2 norm only contains non-cognitive concepts.

Therefore, we define the following two loss functions. We use Loss_1 to weaken non-cognitive concepts included by $\Delta f(i, L)$ with a small L_2 norm, which forces the DNN to ignore non-cognitive concepts and move attention to cognitive concepts. Moreover, we propose Loss_2 to strengthen cognitive concepts contained in $\Delta f(i, L)$ with a large L_2 norm.

$$\text{Loss}_1 = \mathbb{E}_r \left[\mathbb{E}_{(i_1, L_1) \in \Omega_{\text{trivial}}, |L_1|=r} \left[\|\Delta f(i_1, L_1)\|_2 \right] \right], \quad (5)$$

$$\text{Loss}_2 = -\mathbb{E}_r \left[\mathbb{E}_{(i_2, L_2) \in \Omega_{\text{significant}}, |L_2|=r} \left[\|\Delta f(i_2, L_2)\|_2 \right] \right]. \quad (6)$$

Here, the set $\Omega_{\text{trivial}} \subseteq \{(i, L) | L \subseteq N, i \in N \setminus L\}$ contains m_1 pairs of (i, L) , which generate the m_1 smallest $\|\Delta f(i, L)\|_2$ values. In comparison, the set $\Omega_{\text{significant}} \subseteq \{(i, L) | L \subseteq N, i \in N \setminus L\}$ includes m_2 pairs of (i, L) , which generate the m_2 largest $\|\Delta f(i, L)\|_2$ values. In this way, we consider the set Ω_{trivial} is supposed to include non-cognitive concepts, and the set $\Omega_{\text{significant}}$ is supposed to include cognitive concepts. Considering the computation cost of Loss_1 and Loss_2 is huge, we optimize these loss functions by sampling pairs of (i, L) in implementation.

Note that in Eq. (4), (5), and (6), we only enumerate contexts L made up of $1 \leq r \leq 0.5n$ patches. There are two reasons. First, when r is larger, $\Delta f(i, L)$ contains more contexts L , and thus it is more difficult to maintain $\Delta f(i, L)$ sparse among all coalitions $\{(i, L)\}$. This hurts the sparsity assumption for Eq. (4). Second, collaborations of massive patches (a large r) usually represent interactions between mid-level patterns, instead of representing patch-wise collaborations (Cheng et al., 2021), which boost the difficulty of the image revision. Please see Appendix F for discussions.

3.3 OPERATIONS TO REVISE IMAGES

In the above section, we propose two loss functions to decrease the cognitive difficulty of an image from two perspectives. In this section, we further introduce a set of operations to revise images to decrease the loss functions. **A key principle for image revision is that the revision should not bring in additional concepts, but just revises existing concepts in images.** Specifically, this principle can be implemented as the following three requirements.

1. The image revision is supposed to simply change the hue, saturation, brightness, and sharpness of different regions in the image.
2. The revision should ensure the spatial smoothness of images. In other words, each regional revision of the hue, saturation, brightness, and sharpness is supposed to be conducted smoothly over different regions, because dramatic changes of the hue, saturation, brightness, and sharpness may bring in new edges between two regions.
3. Besides, the revision should also preserve the existing edges. *I.e.* the image revision is supposed to avoid the color of one object passing through the contour and affecting its neighboring objects.

In this way, we propose four operations to adjust the hue, saturation, brightness, and sharpness of images smoothly, without damaging existing edges or bringing in new edges. Thus, the image revision will not generate new concepts or introduce out-of-distribution features.

Operation to adjust the hue. Without loss of generality, we first consider the operation to revise the hue of the image x in a small region (with a center c). As Fig. 3 (c) shows, the revision of all pixels in the region with the center c is given as

$$Z^{(c, \text{hue})} = k^{(c, \text{hue})} \cdot (G \circ M^{(c)}), \quad (7)$$

where \circ denotes the element-wise production. $k^{(c, \text{hue})} \in \mathbb{R}$ indicates the significance of the hue revision in this region, which is learned via loss functions. $G \in \mathbb{R}^{d' \times d'}$ is a cone-shaped template¹, which is used to smoothly decrease the significance of the revision from the center to the border, according to Requirement 2. Specifically, each element in the template $G_{hw} = \max(1 - \lambda \cdot \|p_{hw} - c\|_2, 0)$ represents the significance of the revision for each pixel in the receptive field. Here, $\|p_{hw} - c\|_2$ indicates the Euclidean distance between the pixel (h, w) within the region and the center c . λ is a positive scalar to control the range of the receptive field. We set $\lambda = 1$ in experiments.

In order to protect the existing edges (Requirement 3), we also design a binary mask $M^{(c)} \in \{0, 1\}^{d' \times d'}$, which masks image regions that are not supposed to be revised. Specifically, all pixels blocked by the edge (*e.g.* the object contour) are not allowed to be revised. For example, in Fig. 3 (c), the effects of the sky are not supposed to pass through the contour of the bird and revise the bird head, when the center c is located in the sky.

Then, as Fig. 3 (b) shows, for each pixel (h, w) , the overall effect of its hue revision $\Delta_{hw}^{(\text{hue})}$ is a mixture of effects from surrounding pixels N_{hw} , as follows.

$$x_{hw}^{(\text{new, hue})} = \beta^{(\text{hue})} \cdot \Delta_{hw}^{(\text{hue})} + x_{hw}^{(\text{hue})}, \quad \Delta_{hw}^{(\text{hue})} = \tanh\left(\sum_{c \in N_{hw}} Z_{hw}^{(c, \text{hue})}\right), \quad (8)$$

where $0 \leq x_{hw}^{(\text{hue})} \leq 1$ denotes the original hue value of the pixel (h, w) in the image x . $\tanh(\cdot)$ is used to control the value range of the image revision, and the maximum magnitude of hue changes is limited by $\beta^{(\text{hue})}$. Note that the revised hue value $x_{hw}^{(\text{new, hue})}$ may exceed the range $[0, 1]$. Considering the hue value has a loop effect, the value out of range can be directly modified to $[0, 1]$ without being truncated. For example, the hue value of 1.2 is modified to 0.2.

Operations to adjust the saturation and brightness. Without loss of generality, let us take the saturation revision for example. Considering the value range for saturation is $[0, 1]$ without a loop

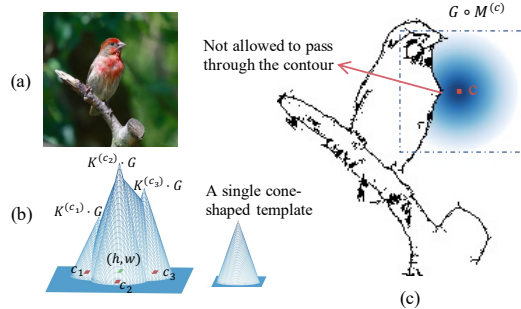


Figure 3: Cone-shaped templates for image revision.

¹We use a cone-shaped template, instead of a Gaussian template, in order to speed up the computation.

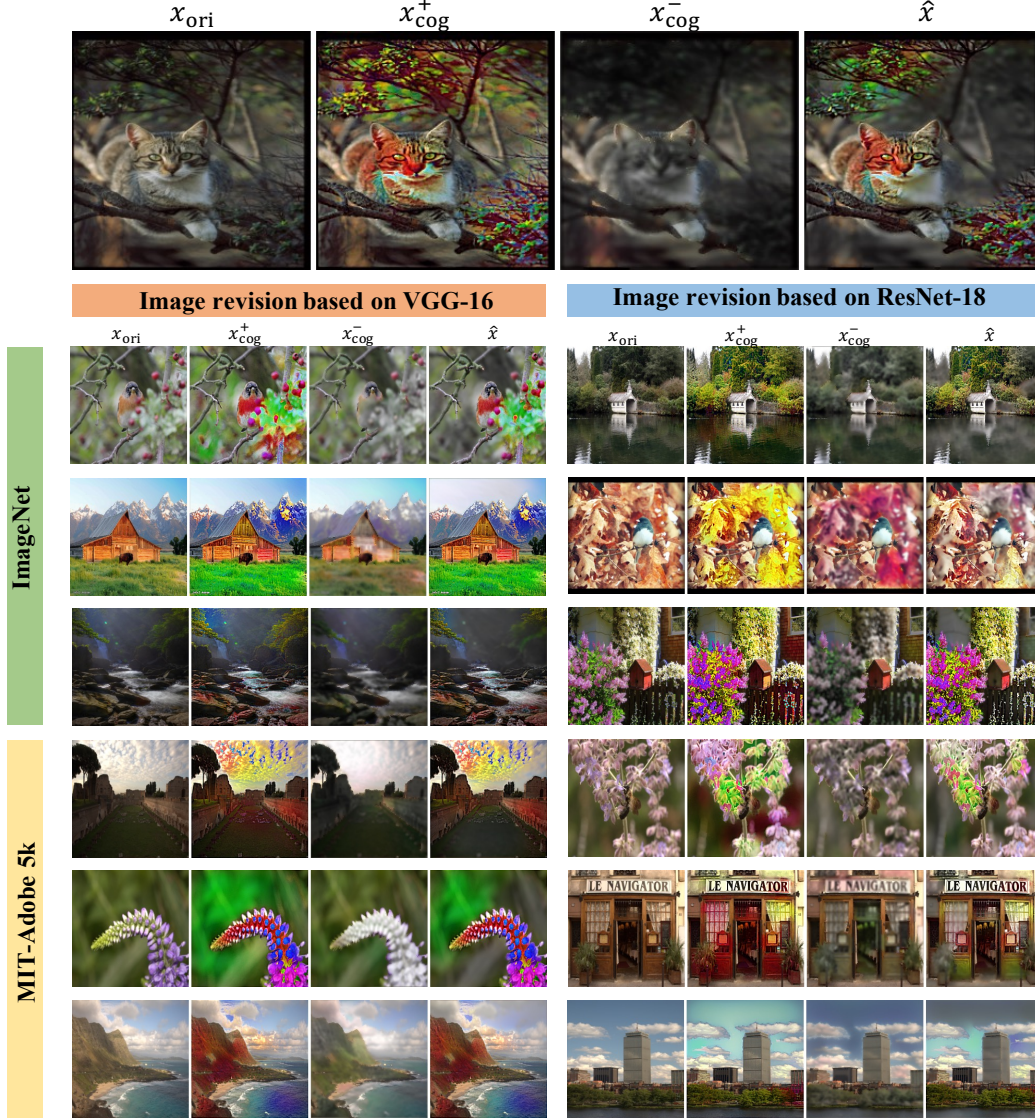


Figure 4: Comparisons between original images x_{ori} , images x_{cog}^+ revised by strengthening cognitive concepts ($\min_{\theta} \text{Loss}_2(\theta)$), images x_{cog}^- revised by discarding cognitive concepts ($\max_{\theta} \text{Loss}_2(\theta)$), and images \hat{x} revised by strengthening cognitive concepts and weakening non-cognitive concepts ($\min_{\theta} \text{Loss}_1(\theta) + \text{Loss}_2(\theta)$). We revised images from the ImageNet dataset and the MIT-Adobe 5K dataset based on the VGG-16 model and the ResNet-18 model, respectively. Please see Appendix G for more results.

effect, we use the sigmoid function to control the revised saturation value $x_{hw}^{(\text{new}, \text{sat})}$ within the range.

$$x_{hw}^{(\text{new}, \text{sat})} = \text{sigmoid}(\beta^{(\text{sat})} \cdot \Delta_{hw}^{(\text{sat})} + \text{sigmoid}^{-1}(x_{hw}^{(\text{sat})})), \quad \Delta_{hw}^{(\text{sat})} = \tanh\left(\sum_{c \in N_{hw}} Z_{hw}^{(c, \text{sat})}\right), \quad (9)$$

where $Z^{(c, \text{sat})} = k^{(c, \text{sat})} \cdot (G \circ M^{(c)})$, just like Eq. (7). Besides, the operation to adjust the brightness is similar to the operation to adjust the saturation.

Operation to sharpen or blur images. We propose another operation to sharpen or blur the image in the RGB space. Just like the hue revision, the sharpening/blurring operation to revise the image can be represented as follows.

$$x_{hw}^{(\text{new}, \text{blur/sharp})} = \Delta_{hw}^{(\text{blur/sharp})} \cdot \Delta x_{hw} + x_{hw}, \quad \Delta_{hw}^{(\text{blur/sharp})} = \tanh\left(\sum_{c \in N_{hw}} Z_{hw}^{(c, \text{blur/sharp})}\right), \quad (10)$$

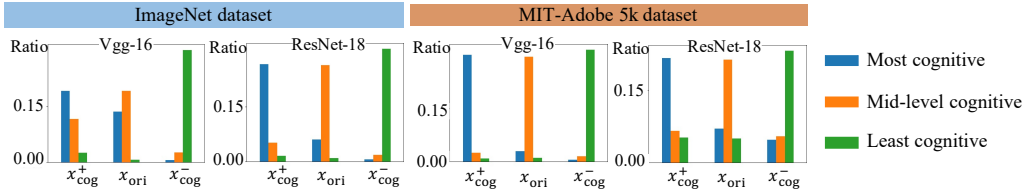


Figure 5: Human measure for the cognitive difficulty. The image x_{cog}^+ revised from the original image x_{ori} by strengthening cognitive concept was likely to be most cognitive. The image x_{cog}^- revised by discarding cognitive concept tended to be least cognitive.

where $\Delta x_{hw} = x_{hw}^{(\text{blur})} - x_{hw}$ indicates the pixel-wise change towards blurring the image, and $x^{(\text{blur})}$ is obtained by blurring the image using the Gaussian blur operation. Accordingly, $-\Delta x_{hw}$ describes the pixel-wise change towards sharpening the image. $\Delta_{hw}^{(\text{blur/sharp})} \in [-1, 1]$. If $\Delta_{hw}^{(\text{blur/sharp})} > 0$, then this pixel is blurred; otherwise, being sharpened. $Z^{(c, \text{blur/sharp})} = k^{(c, \text{blur/sharp})} \cdot (G \circ M^{(c)})$, just like Eq. (7).

In this way, we revise the hue, saturation, brightness, and sharpness of images to strengthen cognitive concepts by minimizing $\text{Loss}_2(\theta)$ in Eq. (6), and to discard non-cognitive concepts by minimizing $\text{Loss}_1(\theta)$ in Eq. (5). Here, $\theta = \{k^{(c, \text{hue})}, k^{(c, \text{sat})}, k^{(c, \text{bright})}, k^{(c, \text{blur/sharp})} | c \in N_{\text{center}}\}$, and N_{center} denotes a set of central pixels *w.r.t.* different regions in the image x . Note that our method can also be applied to other human-defined operations for image revision (*e.g.* changing the image style), besides the revision of the hue, saturation, brightness, and sharpness of images.

4 EXPERIMENTS

- Visualization of image revision.** We conducted experiments to test the cognitive difficulty, considering the following two settings: (1) strengthening cognitive concepts and (2) discarding non-cognitive concepts. Specifically, we computed loss functions based on the VGG-16 model (Simonyan & Zisserman, 2015) and the ResNet-18 model (Kaiming He & Sun, 2016) to revise images from the ImageNet dataset (Krizhevsky et al., 2012) and the MIT-Adobe 5K dataset (Bychkovsky et al., 2011), respectively. These DNNs were trained on the ImageNet dataset for object classification. For the ResNet-18 model, we used the feature after the first block as f . For the VGG-16 model, we used the feature of the `conv3_3` layer as f . In order to decrease the cognitive difficulty of the image, we set $m_1 = 0.1M$, and $m_2 = 0.6M$, where $M = |\{(i, L) | L \subseteq N, i \in N \setminus L\}|$. We divided an image into 28×28 patches, and set $\beta^{(\text{hue})} = 0.35$, $\beta^{(\text{sat})} = 3$, $\beta^{(\text{bright})} = 1.5$. We utilized the off-the-shelf edge detection method (Xie & Tu, 2015) to extract the object contour and to calculate the binary mask $M^{(p)}$. Besides, we used the Gaussian filter with the kernel size 5 to blur images.

Let us focus on columns of x_{ori} , x_{cog}^+ , and x_{cog}^- in Fig. 4. We found that the image x_{cog}^+ revised by strengthening cognitive concepts was more cognitive, compared with both the original image x_{ori} and the image x_{cog}^- revised by weakening cognitive concepts. Note that we allowed the color in the image to be significantly revised, instead of maintaining the original color, because the purpose of the image revision was to decrease the cognitive difficulty, *e.g.* making the tree greener in Fig. 4.

- Human measure for cognitive difficulty.** In order to examine whether strengthening cognitive concepts could decrease the cognitive difficulty of an image, we conducted an experiment with 384 human participants. In this experiment, we showed each participant a group of images, including one original image x_{ori} , the image x_{cog}^+ revised from x_{ori} by strengthening cognitive concepts, and the image x_{cog}^- revised from x_{ori} by weakening cognitive concepts. Then, we asked each participant to sort the cognitive difficulty of these three images. Please see Appendix H for an example.

To this end, we calculated the ratio that the revised image x_{cog}^+ was correctly recognized as the most cognitive one in each group. We also computed the ratio that the revised image x_{cog}^- was correctly considered as the least cognitive one in each group. Fig. 5 reports above ratios, which shows that the image revised by strengthening cognitive concepts tended to be most cognitive, and the image revised by weakening cognitive concepts tended to be least cognitive. Thus, our hypothesis was verified.

- Using the existing classifier to evaluate the memorability (which was related to the cognitive difficulty).** Besides the human measure, we also used an existing DNN, which was trained to evaluate the memorability of images in a supervised manner, to test the effectiveness of our method. Fajtl et al.

Table 1: Using the existing classifier to evaluate the memorability (which was related to the cognitive difficulty). The image x_{cog}^+ revised from the original image x_{ori} by strengthening cognitive concept tended to be most memorable.

Dataset	ImageNet						MIT-Adobe 5K					
Model	VGG-16			ResNet-18			VGG-16			ResNet-18		
Images	x_{cog}^+	x_{ori}	x_{cog}^-	x_{cog}^+	x_{ori}	x_{cog}^-	x_{cog}^+	x_{ori}	x_{cog}^-	x_{cog}^+	x_{ori}	x_{cog}^-
Memorability	0.759	0.734	0.710	0.751	0.734	0.720	0.694	0.661	0.659	0.679	0.661	0.658

Table 2: Using existing classifiers to evaluate the aesthetic level of images. The image \hat{x} revised by strengthening cognitive concepts and discarding non-cognitive concepts tended to be the most aesthetic one.

Dataset	ImageNet						MIT-Adobe 5K					
Model	VGG-16			ResNet-18			VGG-16			ResNet-18		
Images	\hat{x}	x_{cog}^+	x_{ori}	\hat{x}	x_{cog}^+	x_{ori}	\hat{x}	x_{cog}^+	x_{ori}	\hat{x}	x_{cog}^+	x_{ori}
$\gamma_{\text{aes}}^{\text{NIMA}}$	4.803	4.740	4.653	4.670	4.621	4.653	4.705	4.661	4.684	4.732	4.592	4.684
$\gamma_{\text{aes}}^{\text{ILGNet}}$	93.4%	83.5%	72.5%	81.3%	69.2%	72.5%	95.0%	86.0%	54.0%	81.7%	55.0%	54.0%

(2018) proposed the AMNet model to measure the memorability of the image, whose value range was $[0, 1]$, and a large value indicated the image was more memorable. Here, we used 120 images from the ImageNet dataset and 60 images from the MIT-Adobe 5K dataset for evaluation. Table 1 shows that the image x_{cog}^+ revised by strengthening cognitive concepts was most memorable.

- **Using the cognitive difficulty to explain the image aesthetics.** Beyond the cognitive difficulty, we found that strengthening cognitive concepts and discarding non-cognitive concepts could improve the aesthetic level of an image in experiments. It is because such operation could reduce the cognitive cost of humans and remove noises, which made the image more aesthetic, to some extent.

To examine the correctness of this finding, we used existing classifiers to evaluate the aesthetic level of images. These existing classifiers had been trained as off-the-shelf models to predict aesthetic qualities of images in a supervised manner. Specifically, we used the NIMA model (Talebi & Milanfar, 2018), and the ILGNet (Jin et al., 2019a) for evaluation, whose outputs were denoted by $\gamma_{\text{aes}}^{\text{NIMA}}$ and $\gamma_{\text{aes}}^{\text{ILGNet}}$, respectively. A large value of $\gamma_{\text{aes}}^{\text{NIMA}}$ or $\gamma_{\text{aes}}^{\text{ILGNet}}$ indicated the image tended to be more aesthetic. Table 2 shows that an image \hat{x} revised by strengthening cognitive concepts and discarding non-cognitive concepts tended to be more aesthetic than both the original image and the image revised by only strengthening cognitive concepts.

Besides, Fig. 4 also shows that the image revised \hat{x} by strengthening cognitive concepts and weakening non-cognitive concepts tended to be more aesthetic than x_{ori} , in terms of the *vivid color*, *object emphasis*, and *light* (Mavridaki & Mezaris, 2015). Moreover, compared to x_{cog}^+ , the combination of discarding non-cognitive concepts and strengthening cognitive concepts made the revised image more aesthetic. It may be because simply strengthening cognitive concepts just made the attention uniformly separated, while discarding non-cognitive concepts would reduce noises and avoid attracting attention. For example, on the top left of Fig. 4, leaves in \hat{x} were blurred and turned grey to emphasize the bird and the flower, which satisfied the rule *object emphasis* in the photography. In comparison, those leaves in x_{cog}^+ were all sharpened and turned greener to make the attention more uniformly separated.

5 CONCLUSION

In this paper, we propose a hypothesis for the cognitive difficulty of an image from two perspectives, *i.e.* a cognitive image usually makes a DNN strongly activated by cognitive concepts; discarding massive non-cognitive concepts may also help the DNN focus on cognitive concepts. To verify this hypothesis, we use multi-variate interactions to represent cognitive concepts and non-cognitive concepts encoded in the DNN. Furthermore, we define two loss functions to revise the hue, saturation, brightness, and sharpness of images, where one loss forces the DNN to strengthen cognitive concepts, and the other forces the DNN to discard non-cognitive concepts. We find that the revised images are more cognitive than the original ones. We also discover that strengthening cognitive concepts and discarding non-cognitive concepts can improve the aesthetic level of an image.

Reproducibility Statement. In terms of the algorithm of this paper, Appendix B and Appendix E provide complete proofs and discuss all assumptions for theoretical results in this paper. In terms of experiments, the first paragraph of Section 4, Appendix G, and Appendix H provide complete descriptions of all experimental details, which ensure the reproducibility. Nevertheless, we will release the code when the paper is accepted.

REFERENCES

- Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction: The emotional bias. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 491–495, 2016.
- Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pp. 97–104. IEEE, 2011.
- Xu Cheng, Chuntung Chu, Yi Zheng, Jie Ren, and Quanshi Zhang. A game-theoretic taxonomy of visual concepts in dnns. *arXiv preprint arXiv:2106.10938*, 2021.
- Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, 2017.
- Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 870–878, 2018.
- Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.
- Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6363–6372, 2018.
- Martha J Farah. Is an object an object an object? cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science*, 1(5): 164–169, 1992.
- Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.
- John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.
- Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9375–9383, 2019.
- Yangqing Jia, Joshua T Abbott, Joseph L Austerweil, Tom Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems*, pp. 1842–1850, 2013.
- Xin Jin, Le Wu, Xiaodong Li, Xiaokun Zhang, Jingying Chi, Siwei Peng, Shiming Ge, Geng Zhao, and Shuying Li. Ilnet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation. *IET Computer Vision*, 13(2): 206–212, 2019a.
- Xin Jin, Le Wu, Geng Zhao, Xiaodong Li, Xiaokun Zhang, Shiming Ge, Dongqing Zou, Bin Zhou, and Xinghui Zhou. Aesthetic attributes assessment of images. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 311–319, 2019b.

- Shaoqing Ren, Kaiming He, Xiangyu Zhang and Jian Sun. Deep residual learning for image recognition. *In CVPR*, 2016.
- Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pp. 2390–2398, 2015.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.
- Iro Laina, Ruth C Fong, and Andrea Vedaldi. Quantifying learnability and descriptibility of visual concepts emerging in representation learning. *arXiv preprint arXiv:2010.14551*, 2020.
- Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *2011 International Conference on Computer Vision*, pp. 2206–2213. IEEE, 2011.
- Eftichia Mavridaki and Vasileios Mezaris. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 887–891. IEEE, 2015.
- Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Saliency driven image manipulation. *Machine Vision and Applications*, 30(2):189–202, 2019.
- Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12826–12835, 2020.
- Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415. IEEE, 2012.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- L. Shapley. A value for n-person games. 1953.
- Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 879–886, 2018.
- Oleksii Sidorov. Changing the image memorability: From basic photo editing to gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In ICLR*, 2015.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.
- Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.
- Joshua B Tenenbaum. Bayesian modeling of human concept learning. *Advances in neural information processing systems*, pp. 59–68, 1999.
- Paul Thagard. *Cognitive science*. Routledge, 2008.
- Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6849–6857, 2019.

Robert J Weber. Probabilistic values for games, the shapley value. essays in honor of lloyd s. shapley (ae roth, ed.), 1988.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Fei Xu and Joshua B Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

A ESSENTIAL DIFFERENCE BETWEEN COGNITIVE CONCEPTS AND SALIENCY/ATTRIBUTION METHODS.

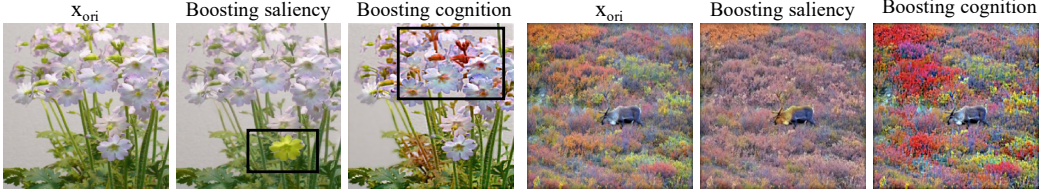


Figure 6: Difference between boosting saliency and boosting cognition. Boosting saliency changes the color of the flower to make it more distinctive and attract more attention. In comparison, boosting the cognition makes flowers redder to be recognized much more easily. x_{ori} represents the original image.

The essential difference between cognitive concepts and salient pixels is as follows. Saliency methods (Mechrez et al., 2019) simply classify all pixels into salient pixels and non-salient pixels based on their abilities to attract human attention. The goal of these saliency methods is to make each pixel more attractive by changing its contrast or color, instead of concerning whether each pixel is cognitive or not. In comparison, our method models the cognitive difficulty at the level of conceptual representation, instead of at the pixel-wise manner. *I.e.* the cognition difficulty of a pixel depends on how this pixel collaborates with surrounding pixels. In this way, the multi-variate interaction can better model how to push a pixel to collaborate with other pixels to form a cognitive concept, rather than simply change the contrast or the color of this pixel in a local manner. For example, in Fig. 6, boosting saliency changes the color of the flower to make it distinctive and attract more attention. In comparison, boosting cognition made the flower redder to be much more easily recognized.

Moreover, the difference between attribution methods (Sundararajan et al., 2017; Ribeiro et al., 2016) and cognitive concepts is as follows. The attribution methods usually reflect the importance of the target object to the result of object detection, while our method focuses on how the collaboration inside the concept contributes to the feature representation.

B PROOF OF THE DECOMPOSITION IN EQ. (1)

Mathematically, we have proven that the feature $f(N)$ of an intermediate layer in the DNN can be decomposed into interaction utilities of different visual concepts S , when we use the Harsanyi dividend (Harsanyi, 1963) to measure the interaction utility $I(S)$.

$$f(N) = \sum_{S \subseteq N} I(S) \tag{11}$$

• *Proof:*

$$\begin{aligned} \sum_{S \subseteq N} I(S) &= \sum_{S \subseteq N} \sum_{L \subseteq S} (-1)^{|S|-|L|} f(L) \\ &= \sum_{L \subseteq N} \sum_{K \subseteq N \setminus L} (-1)^{|K|} f(L) \quad \% \text{ Let } K = S \setminus L \\ &= \sum_{L \subseteq N} \left[\sum_{|K|=0}^{n-|L|} C_{n-|L|}^{|K|} (-1)^{|K|} \right] f(L) = f(N) \end{aligned}$$

C DETAILED INTRODUCTION OF SETTING BASELINE VALUES

Given a pre-trained DNN f and the image x consisting of n patches, $N = \{1, 2, \dots, n\}$, $f(S) \in \mathbb{R}^d$ denotes the feature of an intermediate layer in the DNN when only patches in the concept $S \subseteq N$ are given. In the computation of $f(S)$, we mask patches in $N \setminus S$ to baseline values by following settings

in (Dabkowski & Gal, 2017) and keep patches in S unchanged, so as to approximate the removal of patches in $N \setminus S$. In this way, $f(S)$ can be represented as follows.

$$f(S) = f(\text{mask}(x, S)), \quad \text{mask}(x, S) = x_S \sqcup b_{\bar{S}}, \quad (x_S \sqcup b_{\bar{S}})_i = \begin{cases} x_i, & i \in S \\ b_i, & i \in \bar{S} = N \setminus S \end{cases} \quad (12)$$

where $\text{mask}(x, S)$ denotes the masked image, and b_i denotes the baseline value of the i -th patch. \sqcup indicates the concatenation of x 's dimensions in S and b 's dimensions in $\bar{S} = N \setminus S$. Following settings in (Dabkowski & Gal, 2017), the baseline value of each patch is set to the mean value of this patch over all images, *i.e.* $b_i = \mathbb{E}_x[x_i]$.

D DETAILED INTRODUCTION OF SHAPLEY VALUES

The Shapley value (Shapley, 1953) defined in game theory is widely considered as an unbiased estimation of the numerical utility *w.r.t.* each input patch. Without loss of generality, we extend the scalar Shapley value to a vectorized one. It is because this paper focuses on the utility of each patch to the feature representation, rather than the utility to the scalar classification result. Given a trained DNN f and an image x with n patches $N = \{1, 2, \dots, n\}$, some patches may cooperate to form a context $L \subseteq N$ to influence a feature of an intermediate layer $y = f(x) \in \mathbb{R}^d$. To this end, the Shapley value is proposed to fairly divide and assign the overall effects on the feature to each patch, as follows.

$$\phi(i|N) = \sum_{L \subseteq N \setminus \{i\}} \frac{(n - |L| - 1)!|L|!}{n!} [\Delta f(i, L)], \quad (13)$$

Here, the Shapley value $\phi(i|N) \in \mathbb{R}^d$ represents the utility of the patch i to the feature. $\Delta f(i, L) \stackrel{\text{def}}{=} f(L \cup \{i\}) - f(L) \in \mathbb{R}^d$ measures the marginal utility of the patch i to the feature, given a set of contextual patches L .

Weber (1988) has proven that the Shapley value is a unique method to fairly allocate the overall utility to each patch that satisfies following properties.

(1) Linearity property: If two independent DNNs can be merged into one DNN, then the Shapley value of the new DNN also can be merged, *i.e.* $\forall i \in N, \phi_u(i|N) = \phi_v(i|N) + \phi_w(i|N); \forall c \in \mathbb{R}, \phi_{c \cdot u}(i|N) = c \cdot \phi_u(i|N)$.

(2) Nullity property: The dummy patch i is defined as a patch satisfying $\forall S \subseteq N \setminus \{i\}, v(S \cup \{i\}) = v(S) + v(\{i\})$, which indicates that the patch i has no interactions with other patches in N , $\phi(i|N) = v(\{i\}) - v(\emptyset)$.

(3) Symmetry property: If $\forall S \subseteq N \setminus \{i, j\}, v(S \cup \{i\}) = v(S \cup \{j\})$, then $\phi(i|N) = \phi(j|N)$.

(4) Efficiency property: Overall utility can be assigned to all patches, $\sum_{i \in N} \phi(i|N) = v(N) - v(\emptyset)$.

E PROOF OF EQ. (4)

We have proven that the term $\Delta f(i, L)$ in the Shapley value can be re-written as the sum of interaction utilities of some visual concepts.

$$\Delta f(i, L) = \sum_{L' \subseteq L} I(S' = L' \cup \{i\}). \quad (14)$$

• *Proof:*

$$\begin{aligned}
\text{right} &= \sum_{L' \subseteq L} I(S' = L' \cup \{i\}) \\
&= \sum_{L' \subseteq L} \left[\sum_{K \subseteq L'} (-1)^{|L'|+1-|K|} f(K) + \sum_{K \subseteq L'} (-1)^{|L'|-|K|} f(K \cup \{i\}) \right] \quad \% \text{ Based on Eq. (1)} \\
&= \sum_{L' \subseteq L} \sum_{K \subseteq L'} (-1)^{|L'|-|K|} [f(K \cup \{i\}) - f(K)] \\
&= \sum_{L' \subseteq L} \sum_{K \subseteq L'} (-1)^{|L'|-|K|} \Delta f(i, K) \\
&= \sum_{K \subseteq L} \sum_{P \subseteq L \setminus K} (-1)^{|P|} \Delta f(i, K) \quad \% \text{ Let } P = L' \setminus K \\
&= \sum_{K \subseteq L} \left(\sum_{p=0}^{|L|-|K|} (-1)^p C_{|L|-|K|}^p \right) \Delta f(i, K) \quad \% \text{ Let } p = |P| \\
&= \sum_{K \subsetneq L} 0 \cdot \Delta f(i, K) + \sum_{K=L} \left(\sum_{p=0}^{|L|-|K|} (-1)^p C_{|L|-|K|}^p \right) \Delta f(i, K) \\
&= \Delta f(i, L) = \text{left}
\end{aligned}$$

F DISCUSSIONS ON THE ENUMERATION OF CONTEXTS

In this section, we analyze reasons for only enumerating contexts L consisting of $1 \leq r \leq 0.5n$ patches in Eq. (4), Eq. (5), and Eq. (6). In summary, there are two reasons.

First, a large r hurts the sparsity assumption. In this paper, we assume that the marginal utility $\Delta f(i, L)$ is very sparse among all contexts $\{L\}$. In other words, only a small ratio of contexts $\{L\}$ have significant impacts $\|\Delta f(i, L)\|_2$ on the feature representation, and most contexts $\{L\}$ make negligible impacts $\|\Delta f(i, L)\|_2$. If r is larger, according to Eq. (4), $\Delta f(i, L)$ contains more contexts $\{L\}$. For example, if $r = n$, there are 2^n contexts $\{L\}$ in total. It is more difficult to maintain $\Delta f(i, L)$ sparse among these massive tuples $\{(i, L)\}$. It is because, based on Eq. (4), $\Delta f(i, L)$ is represented as the sum of interaction utilities of massive visual concepts, and then we cannot guarantee only a small ratio of contexts $\{L\}$ have significant impacts $\|\Delta f(i, L)\|_2$ on the feature representation. In this way, a large r hurts the sparsity assumption.

Second, a large r boosts the difficulty of the image revision. It is because contexts L consisting of massive patches (a large r) are usually encoded as interactions between middle-scale concepts or between large-scale concepts (e.g. background), rather than patch-level collaborations. In this way, it is difficult to revise images at the patch level by penalizing such middle/large-scale concepts. Whereas, contexts consisting of a few patches are mainly encoded as interactions between small-scale concepts at the patch level, such as a local texture or a local object. Thus, punishing small-scale concepts is more likely to revise the image in detail at the patch level than punishing large-scale concepts.

G MORE VISUALIZATION RESULTS

In experiments, we revised the hue, saturation, brightness, and sharpness of an image to test its cognitive difficulty from different perspectives. Specifically, given an original image x_{ori} , the image x_{cog}^+ was revised from x_{ori} by strengthening cognitive concepts, *i.e.* $\min_{\theta} \text{Loss}_2(\theta)$ defined in Eq. (6), where $\theta = \{k^{(c,\text{hue})}, k^{(c,\text{sat})}, k^{(c,\text{bright})}, k^{(c,\text{blur/sharp})} | c \in N_{\text{center}}\}$. The image x_{cog}^- was revised from x_{ori} by discarding cognitive concepts, *i.e.* $\max_{\theta} \text{Loss}_2(\theta)$. The image \hat{x} was revised from x_{ori} by strengthening cognitive concepts and discarding non-cognitive concepts, *i.e.* $\min_{\theta} (\text{Loss}_1(\theta) + \text{Loss}_2(\theta))$. The image x^- was revised from x_{ori} by by weakening concepts, *i.e.* $\min_{\theta} (\text{Loss}_1(\theta) - \text{Loss}_2(\theta))$.



Figure 7: Visualization of the original image x_{ori} , the image x_{cog}^+ revised by strengthening cognitive concepts ($\min_{\theta} \text{Loss}_2(\theta)$), the image x^- revised by weakening concepts ($\min_{\theta} \text{Loss}_1(\theta) - \text{Loss}_2(\theta)$). Here, we revised images from the ImageNet dataset and the MIT-Adobe 5K dataset based on the VGG-16 model and the ResNet-18 model, respectively.

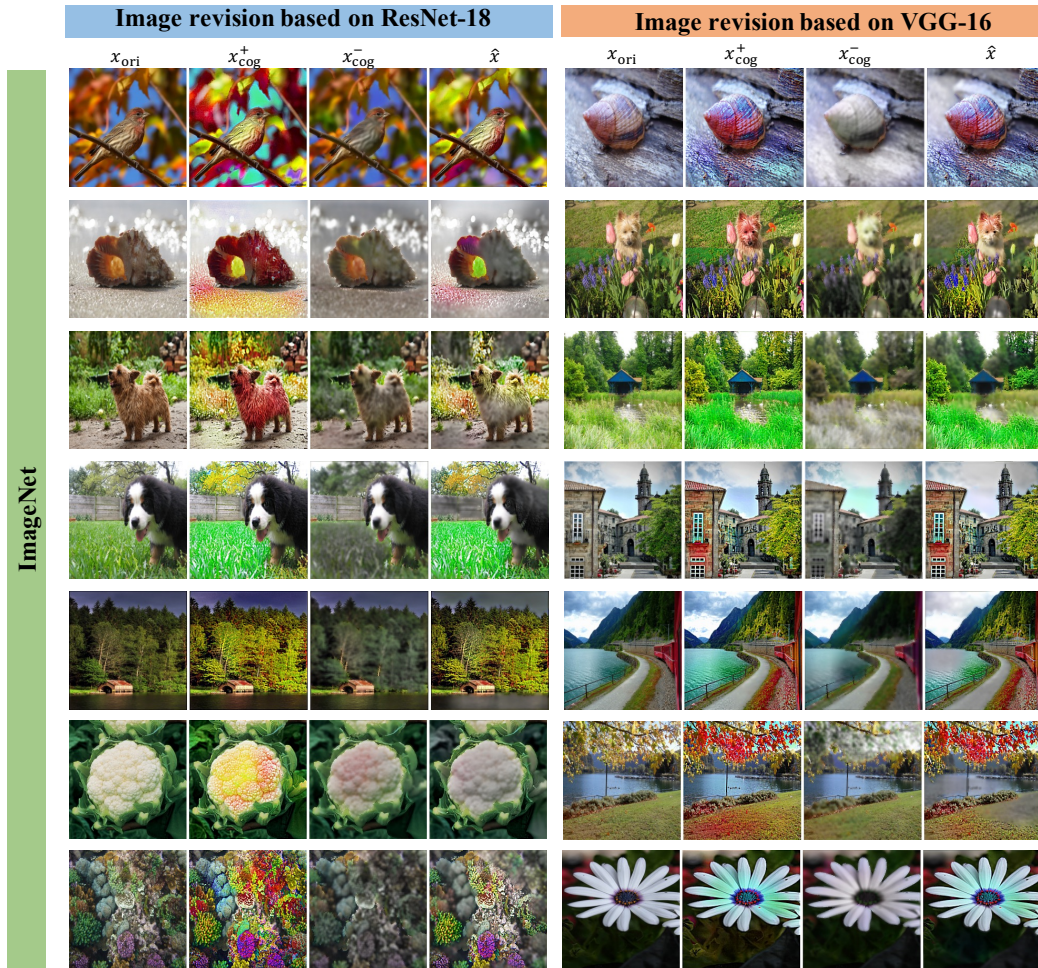


Figure 8: Comparisons between original images x_{ori} , images x_{cog}^+ revised by strengthening cognitive concepts ($\min_{\theta} \text{Loss}_2(\theta)$), images x_{cog}^- revised by discarding cognitive concepts ($\max_{\theta} \text{Loss}_2(\theta)$), and images \hat{x} revised by strengthening cognitive concepts and weakening non-cognitive concepts ($\min_{\theta} \text{Loss}_1(\theta) + \text{Loss}_2(\theta)$). We revised images from the ImageNet dataset based on the VGG-16 model and the ResNet-18 model, respectively.

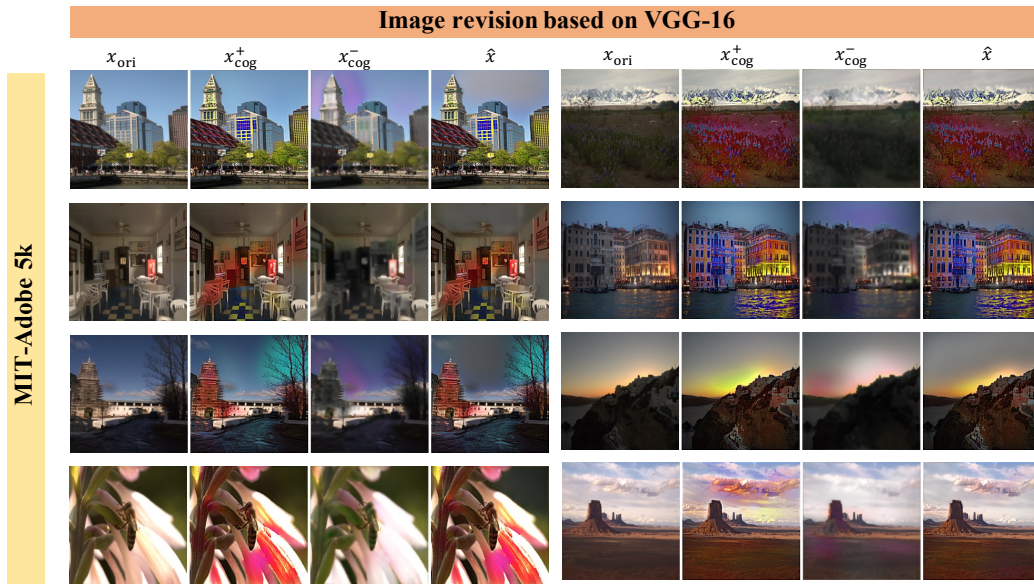


Figure 9: Comparisons between original images x_{ori} , images x_{cog}^+ revised by strengthening cognitive concepts ($\min_{\theta} \text{Loss}_2(\theta)$), images x_{cog}^- revised by discarding cognitive concepts ($\max_{\theta} \text{Loss}_2(\theta)$), and images \hat{x} revised by strengthening cognitive concepts and weakening non-cognitive concepts ($\min_{\theta} \text{Loss}_1(\theta) + \text{Loss}_2(\theta)$). We revised images from the MIT-Adobe 5K dataset based on the VGG-16 model .

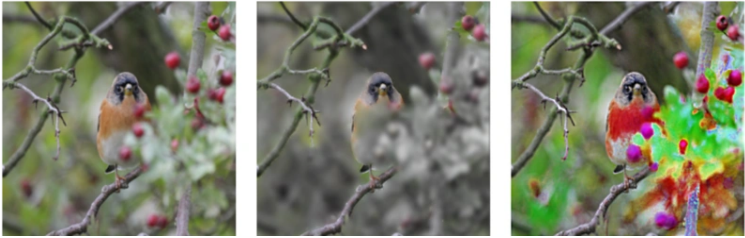
H EXAMPLE FOR THE HUMAN MEASURE OF THE COGNITIVE DIFFICULTY

In order to examine whether strengthening cognitive concepts could decrease the cognitive difficulty of an image, we conducted an experiment with 383 human participants. In this experiment, we showed each participant a group of images, including one original image x_{ori} , the image x_{cog}^+ revised from x_{ori} by strengthening cognitive concepts, and the image x_{cog}^- revised from x_{ori} by weakening cognitive concepts. Then, as Fig. 10 shows, we asked each participant to sort the cognitive difficulty of these three images, and each participant was shown 20 groups of images in total.

Human measure for cognitive difficulty


Here, the cognitive difficulty of an image is referred to as whether this image is easy to recognize.

1.Q: please sort the cognitive difficulty of the following three images.



*

2.Q: please sort the cognitive difficulty of the following three images.



*

Figure 10: An example for the human measure of the cognitive difficulty. We asked each human participant to sort the cognitive difficulty of three images, including one original image, the image revised by strengthening cognitive concepts, and the image revised by weakening cognitive concepts.

I DIFFERENCE IN ENCODING OF CONCEPTS BETWEEN HIGHLY COGNITIVE IMAGES AND LOW COGNITIVE IMAGES

To verify the hypothesis, we considered the cognitive difficulty of an image from two aspects, *i.e.* the beauty of image contents and the noise level. We considered that images with beautiful contents were usually more cognitive than images without beautiful contents, and assumed to contain more cognitive concepts. It is because images with beautiful contents often had bright colors and clear contours, thereby being more cognitive. On the other hand, we considered strong noises in images boosted the cognitive burden, and the smoothing operation of an image could decrease its cognitive difficulty. In this way, we assumed that non-cognitive concepts in the smoothed images were weakened.

Beauty of image contents. We used the VGG-16 (Simonyan & Zisserman, 2015) model trained on the ImageNet dataset (Krizhevsky et al., 2012) to evaluate images from the the Aesthetic Visual Analysis (AVA) dataset (Murray et al., 2012) and the CUHK-PhotoQuality (CUHK-PQ) dataset (Luo et al., 2011), respectively. All aesthetic images and not so aesthetic images had been annotated in these two datasets³.

For verification, we computed the metric $\tau^{(r)}(\Delta f) = P^{(r)}(\Delta f|X_{\text{aesthetic}}) - P^{(r)}(\Delta f|X_{\text{unaesthetic}})$, where $X_{\text{aesthetic}}$ and $X_{\text{unaesthetic}}$ referred to a set of massive aesthetic images with beautiful contents and a set of not so aesthetic images without beautiful contents, respectively. $P^{(r)}(\Delta f|X_{\text{aesthetic}})$ denoted the value

³In particular, we considered images in the AVA dataset with the highest aesthetic scores as aesthetic images and regarded images with the lowest aesthetic scores as not so aesthetic images.

distribution of $\Delta f = \|\Delta f(i, L)\|_2$ among all patches $\{i \in N\}$ and all contexts $\{L|L \subseteq N \setminus \{i\}, |L| = r\}$ contained by massive aesthetic images $x \in X_{\text{aesthetic}}$. Similarly, $P^{(r)}(\Delta f|X_{\text{unaesthetic}})$ represented the value distribution of Δf among all patches $\{i\}$ and all contexts $\{L\}$ contained by not so aesthetic images $x \in X_{\text{unaesthetic}}$. If $\tau^{(r)}(\Delta f) > 0$ for large values of Δf , it indicated that aesthetic images contained more cognitive concepts than not so aesthetic images. In experiments, f was implemented as the feature of the first fully-connected (FC) layer of the VGG-16 model. Fig. 11 (a) shows that aesthetic images usually included more cognitive concepts than not so aesthetic images.

Noises in the image. To verify the assumption that non-cognitive concepts in the smoothed images were significantly weakened, we conducted the following experiments. We used the Gaussian blur operation to smooth images to eliminate noises. The metric $\kappa^{(r)}(\Delta f) = P^{(r)}(\Delta f|X_{\text{smoothed}}) - P^{(r)}(\Delta f|X_{\text{original}})$ was calculated for verification, where X_{smoothed} and X_{original} referred to a set of the smoothed images and a set of the corresponding original images, respectively. $P^{(r)}(\Delta f|X_{\text{smoothed}})$ represented the value distribution of $\Delta f = \|\Delta f(i, L)\|_2$ among all pixels $\{i \in N\}$ and all contexts $\{L|L \subseteq N \setminus \{i\}, |L| = r\}$ contained by the smoothed images $x \in X_{\text{smoothed}}$. Accordingly, $P^{(r)}(\Delta f|X_{\text{original}})$ denoted the value distribution of Δf among all patches $\{i\}$ and all contexts $\{L\}$ contained by the original images $x \in X_{\text{original}}$. Fig. 11 (b) shows non-cognitive concepts in the smoothed images were weakened, *i.e.* $\kappa^{(r)}(\Delta f) > 0$ for small values of Δf .

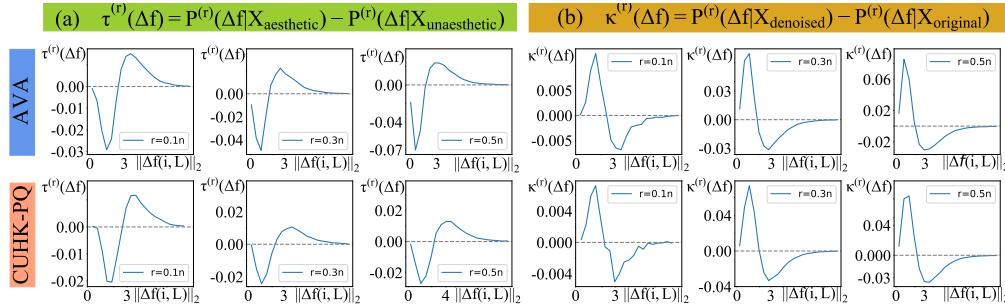


Figure 11: (a) Aesthetic images usually contained more cognitive concepts than not so aesthetic images. (b) non-cognitive concepts in the smoothed images were weakened.