Emotional Landscape Image Generation Using Generative Adversarial Networks

Chanjong Park[0000−0001−8872−7906] and In-Kwon Lee[0000−0002−1534−1882]

Dept. of Computer Science Yonsei University, Republic of Korea
{cjprist, iklee}@yonsei.ac.kr

Abstract. We design a deep learning framework that generates landscape images that match a given emotion. We are working on a more challenging approach to generate landscape scenes that do not have main objects making it easier to recognize the emotion. To solve this problem, deep networks based on generative adversarial networks are proposed. A new residual unit called emotional residual unit (ERU) is proposed to better reflect the emotion on training. An affective feature matching loss (AFM-loss) optimized for the emotional image generation is also proposed. This approach produced better images according to the given emotions. To demonstrate performance of the proposed model, a set of experiments including user studies was conducted. The results reveal a higher preference in the new model than the previous ones, demonstrating the production of images suitable for the given emotions. Ablation studies demonstrate that the ERU and AFM-loss enhanced the performance of the model.

1 Introduction

Computer vision and graphics applications, such as image classification [1–10], object detection [11–16], and image transformation [17–21], are effectively using deep learning techniques. Also, there have been ongoing studies on image generation using deep learning in recent years [22–29]. These are studies that produce images that match a given condition, for example, images that match the content of a given sentence or word [30–32]. On a higher level, recent studies have shown that machine learning effectively recognizes the emotions expressed in images [33–37]. However, studies that create images from scratch that reveal input emotions are rare due to the inherent ambiguity and abstraction of emotion.

Research on emotion-based image creation has mainly focused on image creation, including objects that express some specific emotions. In particular, studies related to the transformation or generation of human facial expressions based on given input emotions have been successful [38–42]. However, if objects such as people are not clearly present in the image, we must recognize emotions in the feeling and landscape of the whole image. For example, when the scenery in the image is night time, we can feel calm. Daytime images can have energy and excitement. However, the perception of the emotions felt in the image varies from viewer to viewer. As a result, the process of understanding the emotions of an image without an explicit object is complicated and confusing for both people and computers.

Thus, generating landscape images have the advantage of being able to express emotion with the entire image itself, unlike creating an image that includes a specific object. Even if the specified object exists in front of the background, we can change the feeling of the image by replacing the landscape behind the object. Besides, landscape images representing emotions can be used for behavioral therapy and can be used for psychological research, such as investigating how well people perceive emotions or analyzing brain waves from the landscape images [43–48]. So, we will deal with the creation of landscape images that represent emotions. Although there have been studies on how to create a landscape image [49–51], this work will be the first to create a landscape image from emotions.

There are several ways to represent emotions. One way is by categorizing them into classes such as happy, sad, angry, and relaxed, which is the most widely used method and has the advantage of expressing emotion through intuition. The drawback of this method, however, is that the emotions are classified into several other categories that cannot define various emotions in detail, and the criterion for judging a particular emotion is ambiguous. Osgood *et al.* [52] proposed a dimensional representation called the VA model for representing emotions with two variables, V (valence) and A (arousal). Valence represents the level of pleasure. The lower value of valence indicates a negative emotion, and the higher value indicates a positive emotion. Arousal is a level of excitement. The smaller the arousal value, the calmer the emotion. The larger the value, the more active the sensation. We are using this dimensional representation in describing the emotions that are the input conditions in this study.

In this paper, we propose a deep learning-based model using generative adversarial networks (GAN) [53] for generating landscape images from a given emotion. We design the model in a gradually increasing form according to the training process based on Karras *et al.* [54]. The proposed model also contains new residual units and a new loss function to understand the emotional concepts and generate the image based on that emotion. The former is an Emotional Residual Unit (ERU), and the latter is Affective Feature Matching loss (AFM-loss). The ERU and AFM-loss gradually change features in networks in the training process so that generated images are close to the target emotions.

We conducted various experiments to ensure that the output image sufficiently reflects the given emotions. These experiments compare different results when varying the structure of the network and the arrangement of units of ERU. The experiment also includes user surveys and emotional measurements using the trained emotional prediction model to measure the proposed model's performance.

The contributions of the paper can be summarized as follows:

- We propose a novel deep learning-based approach that can generate landscape images fitting to target given emotions.
- We propose a new residual unit that can train a deep neural network to adjust emotions.

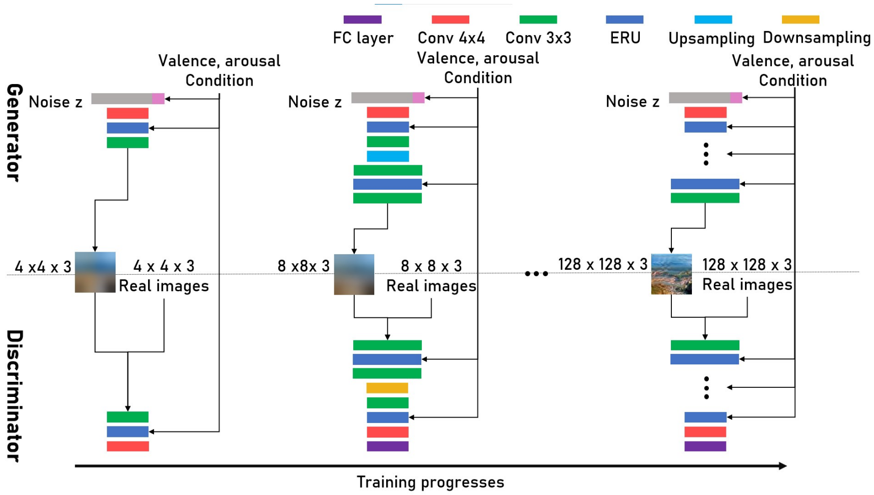


Fig. 1: The overall structure of emotional landscape image generation network. Based on the progressive structure, it shows the process of gradually training from low resolution to high resolution. Incrementally increases from 4×4 structures to 128×128 images. Emotional residual unit (ERU) is inserted into generator and discriminator, which is responsible for accepting emotional conditions and training according to emotion.

- We propose a GAN based generation approach to expressing the emotions.
- We propose an affective feature matching loss to express emotions on the generated landscape images effectively.

In this paper, we focus on creating an image representing a particular emotion from scratch, rather than transforming an image that already exists to fit a specific emotion. When we feel an emotion, it is natural to close our eyes and recall an image of that emotion. For human nature, this process of thinking of an image representing the emotions we feel can be more natural than transforming an existing image to match the emotions we feel. Likewise, if artificial intelligence is drawing art, it will be more valuable to create a new image that does not exist by taking emotion into account. Therefore, our work can be used for artificial intelligence that can express emotions. Our work will be a more critical step in that artificial intelligence fundamentally understands emotions than image transformation.

2 Methods

2.1 Overview of Landscape Image Generation Network

The proposed network (Fig. 1) is based on a progressive structure. The whole structure is divided into three parts: generator, discriminator, and ERU. The generator takes a noise vector as input and outputs an image with the target emotion according to the VA value representing the target emotion. The discriminator takes the output image of the generator and the real image as the input with the emotion values so that it can learn to determine if the image is real. ERUs are inserted in the middle of the generator and discriminator. The embedded ERUs play an important role in helping the generator and discriminator, allowing the entire model to produce images that match the target emotions. It should be noted that, as the network structure and size are increased, more ERUs are added. That is the number of ERUs doubles as the resolution of the output image of the network doubles.

2.2 Emotional Residual Unit (ERU)

Long short-term memory (LSTM) [55] and Gated recurrent unit (GRU) [56] continue to transfer features of the previous state to the next state through the cell state structure. Then, the cell state adds or removes features using gate elements with a refined structure. These gates are devices that allow the selected features of the previous state to flow into the next state. When they pass through the gate, features judged significant are retained, and those judged meaningless are discarded.

We apply the gate structure of LSTM and GRU to our model and propose a new unit. The ERU, the new unit, is designed for emotion-based landscape image generation whose structure is shown in Fig. 2. Let $[\cdot, \cdot]$ denote concatenation, $Conv(x)$ denote convolution on x, $\otimes$ denote element-wise multiplication, $\oplus$ denote element-wise addition, and $f(x)$ be a activation function. X is a feature map given as input to ERU, which is fed from the generator or discriminator these are from the layer before entering ERU.

Single-channel valence and arousal maps, V and A, are generated, whose width and height are equal to input X. The two maps are fed with valence and arousal values representing the current target emotion. Then, V and A are concatenated with X channel-wise, respectively. After passing through the convolution layer and sigmoid activation function in turn, the two feature maps can be represented as v and a, respectively. At this time, the sum of each channel of v and a is set to 1.0 like the soft attention method experimentally, as follows:

$$v = f(Conv([V, X])) \text{ and } a = f(Conv([A, X])). \tag{1}$$

For instance, let us assume that the valence and arousal values of the current target emotion are 3.6 and 6.7, respectively, and the size of X given as input is $4 \times 4 \times 32$. Then we create the maps of valence and arousal with dimensions of $4 \times 4 \times 1$ and fill the maps with 3.6 and 6.7, respectively. And then we concatenate X with the maps of valence and arousal respectively to generate the two feature maps v and a of size $4 \times 4 \times 33$ (see Eq. (1)). After the two feature maps pass through the convolution and activation layers, respectively, their sizes become equal to the input feature map X.

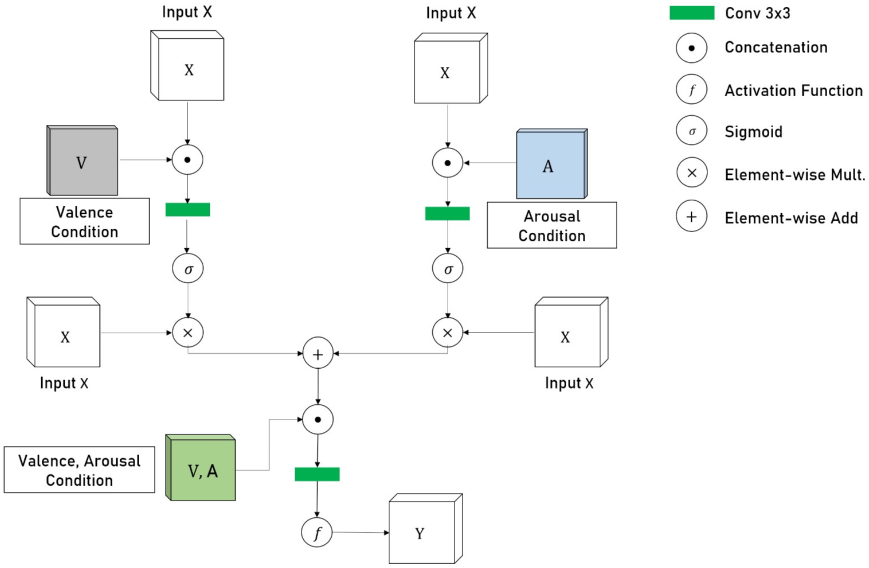


Fig. 2: Structure of the emotional residual unit (ERU). After concatenation of the VA emotional maps to the input feature X of the ERU, respectively, 3×3 convolution and sigmoid function are performed, followed by the element-wise multiplication with input X. After adding the results element-wise, we concatenate the VA map again to the result, and finally, get the output Y after the convolution and activation functions.

Now, v and a are element-wise multiplied with input X, respectively, and the results are combined using element-wise addition as follows:

$$m = (v \otimes X) \oplus (a \otimes X). \tag{2}$$

The combined feature map m in Eq. (2) is concatenated with $[V, A]$, the concatenation of the emotional maps of valence and arousal, and then passed through the convolution layer followed by tanh activation function in turn. Finally, we get Y, the output of ERU, as follows:

$$Y = f(Conv([m, [V, A]])). \tag{3}$$

2.3 ERU In Landscape Image Generation Network

Our model, which is modeled using the progressive structure as the underlying infrastructure, has features that allow the generator and discriminator to gradually evolve. Because of the image resolution of the data set, the training process is repeated after iterating up to 128 × 128 resolution. Since GAN tends to capture only the subset of variation found in the training data, we use the minibatch discrimination technique proposed by Salimans *et al.* [57], which adds a minibatch layer to the end of the discriminator. Additionally, to disallow the scenario where the magnitudes of the generator and the discriminator spiral out of control as a result of competition, we use the pixel-wise feature normalization, instead of the batch normalization commonly used to normalize the generator.

In Fig. 1, we can see that ERUs exist as the intermediate unit in the generator and discriminator. In the case of the generator, there is one ERU in the 4×4 resolution training process, and as the resolution increases, the number of ERUs also increases. In other words, there are $k-1$ ERUs in the generator trained using images with resolution $2^k \times 2^k$, $(2 \leq k \leq 7)$. At the same time, the dimensions of the input and output feature maps of ERUs also depend on the image resolution currently being processed. Note that because the discriminator has the inverted structure of the generator, the arrangement of ERUs in the discriminator also follows a reverse order. Since the proposed model grows gradually, it has the advantage of the progressive structure, which reflects both global and local features.

In our dataset, valence and arousal values are on a scale of 1 to 9, respectively. When the values are assigned to the ERU, the values are normalized to between 0 and 1. Each time an image is trained, the valence and arousal values of the image are given. At this point, the generator's ERU lets the generator learn in the direction of generating an image that expresses a given emotion. Similarly, the ERU in the discriminator is trained to determine if the generated image matches the conditions of the given emotion.

2.4 Affective Feature Matching Loss

Wang *et al.* [20] suggested the feature-matching loss that minimizes the statistical difference between the generated image and the ground truth image by minimizing the difference between the corresponding convolution maps (i.e., feature maps) of discriminator when the generated image and the ground truth image pass through the discriminator. In this work, we propose a new feature-matching loss function called affective feature-matching loss (AFM-loss) by enhancing the existing method to emphasize the reflection of emotional features in the landscape image generation process.

Suppose a generated image and a ground truth image in the dataset have the same target emotion value. The AFM-loss computes the element-wise difference between the feature values (v, a) (i.e., the results of the sigmoid function in Fig. 2) of ERU obtained by Eq. (1) of the generator's output image and the image in the dataset, where the two images are both passing through the convolution layer inside the discriminator's ERU block. The proposed method compares only the features of the convolution layer, where the information of V and A are concatenated. In other words, it does not just make the resulting image similar to the ground truth image but makes the emotion-affected features in the resulting image as similar to the corresponding features in the ground truth image as possible. As a result, the resulting image can be gradually transformed into an image having an emotion close to the target emotion.

Let x be a ground truth image, c_v and c_a be constant valence and arousal values, respectively, and z be an input noise vector. $G(z, c_v, c_a)$ represents the image generated from the generator module. The affective feature matching losses L_v and L_a for valence and arousal are respectively defined by:

$$\mathcal{L}_v = \frac{1}{k} \sum_{i=1}^{k} |D_c(G(z, c_v, c_a), c_v) - D_c(x, c_v)|_i, \tag{4}$$

$$\mathcal{L}_a = \frac{1}{k} \sum_{i=1}^{k} |D_c(G(z, c_v, c_a), c_a) - D_c(x, c_a)|_i, \tag{5}$$

where D_c is the feature of the layer to be compared in the ERU block of the discriminator. The average of all differences between the features in k ERUs becomes the respective feature matching loss for V and A. We take the average experimentally instead of the minimum and maximum.

2.5 Objective Function

The total objective function of the proposed model is defined by:

$$\begin{aligned} \mathcal{L}_{WGAN}(D) = \mathbb{E}_{x \backsim P_{image}}&[D(x, c_v, c_a)] \\ -\mathbb{E}_{z \backsim P_z}&[D(G(z, c_v, c_a))] \\ &+\lambda(L_v + L_a), \end{aligned} \tag{6}$$

$$\mathcal{L}_{WGAN}(G) = \mathbb{E}_{z \backsim P_z}[D(G(z, c_v, c_a))]. \tag{7}$$

In Eq. (6) and (7), we use the Wasserstein GAN [58] loss to stabilize the training. x and z refer to image and noise as in Eq. (4) and (5), respectively. Note that the c_v, c_a values in the loss functions are given as the emotional maps in the ERU. λ is a hyper parameter that balances the AFM-loss and GAN loss. For this experiment, we use $\lambda = 100$.

3 EXPERIMENTS

3.1 Experimental Data

The CGnA10766 is an emotional image dataset built by Kim et al. [37], where 10,766 images are labeled with emotion values, V and A, through the user study. The original CGnA10766 dataset has many images containing objects such as people, animals, and cars. We selected only natural landscape images from the original CGnA10766 dataset, excluding images that contain objects that can have a significant impact on our emotions. Images containing objects that did not significantly affect emotions, such as small boats or bicycles, were classified as valid natural landscape scenes. We obtained a total of 1,453 images (we call this reduced dataset V-A2) through this classification of the original CGnA10766 dataset (see Fig. 3). We collected more images because there were not enough

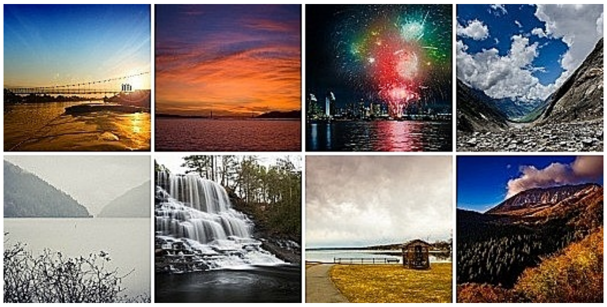


Fig. 3: Sample images of natural outdoor scenes without objects such as people, animals, or cars selected from the CGnA10766 dataset.

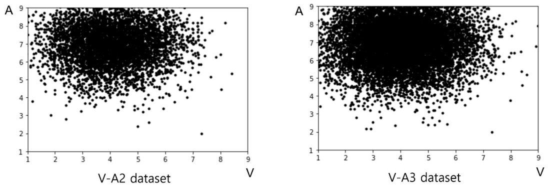


Fig. 4: The valence-arousal distribution of V-A2 dataset (left) and V-A3 dataset (right). The horizontal and vertical axes represent the valence and arousal values, respectively.

images in the V-A2 data set, and obtained their VA values using the emotion predictor trained by Kim et al. [37]. We collected 1,204 images from [59], 1,523 images from [60], and 5,039 images from [61], resulting in 10,453 images. We call this final dataset V-A3. The valence-arousal distributions of the V-A2 and V-A3 dataset are shown in Fig. 4. In the figure, the images in our dataset have a slightly lower valence, and slightly higher arousal value than the center, which is clearly shown by the V-A2 dataset's distribution based only on user studies. We guess that the distribution of the dark colors in the landscape images is probably causing a negative and active feeling rather than positive and calm. Fortunately, both datasets have relatively even distributions of emotion values between the minimum and maximum values, which is one of the conditions that a dataset for machine learning must-have.

3.2 Experimental Settings

We trained our model using both the V-A2 and the V-A3 dataset with 100,000 iterations for the first 4×4 resolution and 200,000 iterations for other resolutions. We set the batch size to 16. After the input noise vector of size 126 was generated, the valence and arousal values were concatenated with the noise vector, and eventually, the input size was 128. In the final stage of generating images of size

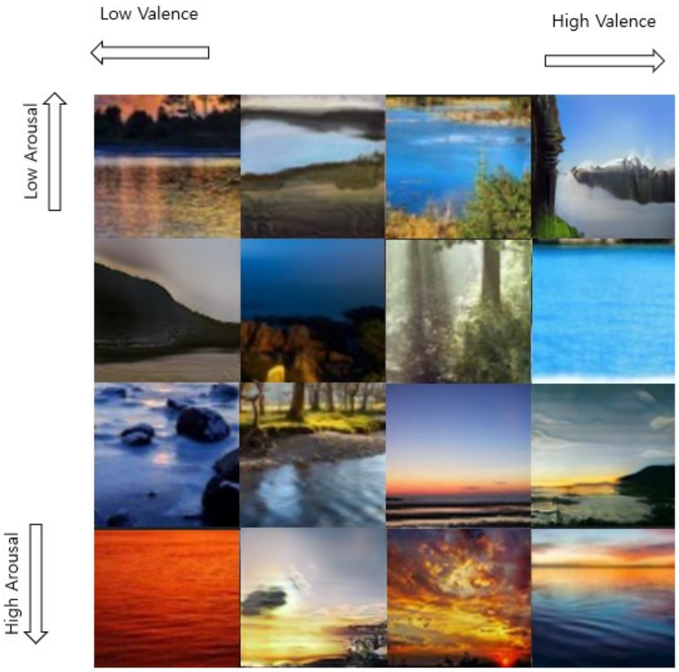


Fig. 5: Example results of proposed model with V-A3 dataset. Valence is small on the left and large on the right. Arousal is small on the top and large on the bottom.

128×128, the batch size was set to 8, to prevent memory out. The learning rate was set to 0.001, and leaky ReLU [62] was used as the activation function of the generator. When concatenating the input feature map X and the emotional map in the ERU, the former is concatenated behind the valence or arousal map. Additionally, in the second concatenation in the ERU, a combined valence and arousal map is concatenated forward. In ERU, we used soft attention instead of hard attention because when compared with the Frechet inception distance of the training results with the V-A3 dataset, the soft attention method was as good as compared to the hard attention method. We performed the training with one Nvidia GTX 1080 Ti for a total of (approximately) 10 days.

3.3 Results

Fig. 5 shows the example results of the images generated by the trained model. We aligned the resulting images in order of successive valence and arousal values. The results with lower valence have broader sea areas with darker brightness than those with higher valence. On the contrary, high valence results are relatively

Table 1: Ablation study in terms of FID scores: Non-ERU and Non-AFM loss represent the model without both of ERU and AFM-loss, and AFM-loss, in the proposed model, respectively. E-FID is an averaged FID score computed only from images with similar emotion values.

Dataset	Method	FID	E-FID
V-A2	Non-ERU	4.07	3.53
	Non-AFM loss	3.02	2.45
	Proposed method	**2.78**	**2.11**
V-A3	Non-ERU	3.82	3.30
	Non-AFM loss	2.82	2.30
	Proposed method	**2.51**	**1.87**

bright. The results with average valences have both low and high valence characteristics. The resulting images with higher arousal have darker colors and are more versatile than with lower arousal, consistent with the emotions expressed by the high arousal representing more active emotions. In Fig. 5, the resulting images on the 4^{th} rows are considered to represent a red sky, which means that an image containing the ruddy glow in the sky usually has high arousal. Additionally, images with low arousal include the blue sea, which is thought to be due to the calmness of the blue sea. Although the results have some artifacts, the proposed model is demonstrably well-controlled according to the conditions of arousal and valence.

To evaluate the results, we measured the FID score [63]. Also, we conducted an ablation study to verify the ERU and AFM-loss performance of the proposed model (see Table 1). In both the V-A2 and V-A3 datasets, the proposed model with both ERU and AFM-loss shows the lowest FID score (see FID column in the table) based on the total data, which means that the output of the proposed model (with both ERU and AFM-loss) is most statistically similar to the original dataset. Let us consider a 2D space with the mean of valence and arousal values in the dataset as origin and valence and arousal as two axes. The E-FID column in Table 1 shows the average value of the four FID scores computed using only the images in the dataset and output images belonging to each quadrant of the 2D space. In other words, it can be said that the proposed model produces outputs with a feature distribution statistically similar to images in a dataset with target emotions.

3.4 User study

To evaluate the results objectively, a series of user studies were conducted on the Amazon Mturk platform [64]. In these experiments, images were given to the subjects, and then they were instructed to select the valence and arousal values they considered appropriate. Fig. 6 shows how the valence and arousal were

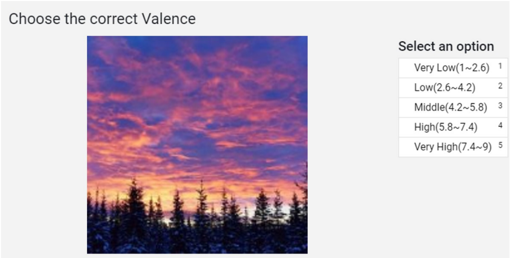

Choose the correct Valence

Fig. 6: Example of a scene during a user study

selected in the user experiments. The valence and arousal values were divided into five ranges, from low to high, respectively, and the subject had to choose one. Selecting the interval was easier, more intuitive than numbering valence and arousal, for the subjects who did not have a good understanding of the VA model. We use the results of our model that was trained with both V-A2 dataset and V-A3 dataset in the user study.

Before the evaluation phase, the user first learned about the VA representation through the training phase. We presented a randomly selected image from the dataset, and the subject was instructed to select VA values of the image. Users who answered outliers that were very different from the average answer were excluded from the next evaluation phase. In the evaluation phase, 200 resulting images were used respectively for the V-A2 and the V-A3 dataset, including 60 results from the model with neither ERU nor AFM-loss, 60 results with ERU only, and 80 results from the complete proposed model. Approximately 10,000 evaluations were performed on 400 images by nearly 1000 subjects. In other words, an image was rated by an average of 25 users. The evaluations were performed for VA values in one of five intervals, each with $[1, 2.6)$, $[2.6, 4.2)$, $[4.2, 5.8)$, $[5.8, 7.4)$, and $[7.4, 9]$. The average of user answers was then calculated using the median of the selected intervals. The error is the difference between the average of evaluated emotion value and the target emotion. Outlier answers were removed using the method of inter-quartile range (IQR). Table 2 shows average of error margins with the target emotion. The complete proposed method was measured to show the least error with the target emotion for both datasets. Considering that the size of one interval is 1.6, the figure is quite low. Besides, the valence error of most images was within 2, as was with the case of arousal. This result demonstrates that our model generates appropriate images for the condition of the target emotion.

Additionally, we conducted an ablation study to verify the ERU and AFM-loss performance of the proposed model. As shown in Table 2, the objects to be compared are the results of the proposed model, the non-ERU model, and the non-AFM model. The non-ERU model eliminates the ERU from the model, and the non-AFM model eliminates AFM-loss in the loss function. In the case of non-ERU models, AFM-loss cannot be applied because there is no ERU. As a result of the statistical analysis of the average error difference between the

Table 2: Comparison of average error between target emotion and results of the user study.

Dataset	Method	Valence-error	Arousal-error
V-A2	Non-ERU	2.73	2.51
	Non-AFM loss	1.92	1.71
	Proposed method	**1.67**	**1.45**
V-A3	Non-ERU	2.62	2.54
	Non-AFM loss	1.81	1.68
	Proposed method	**1.56**	**1.38**

Table 3: Performance comparison of the three models

Method	The Emotional GAN	MHingeGAN	Our method
Valence-error	2.58	1.81	**1.56**
Arousal-error	2.56	1.77	**1.38**
Preference	15.7	34.5	**49.8**
FID	4.85	2.53	**2.51**
E-FID	4.14	1.88	**1.87**

non-ERU model and the complete model with the independent samples' z-test, a significant difference was found in the 95% confidence interval. The average error difference between the non-AFM loss model and the complete model was not statistically significant under the same conditions but showed numerically noticeable differences.

3.5 Comparison with Other Model

We trained the V-A3 dataset that produced better results than the V-A2 dataset to produce results for comparison with the Emotional GAN [65], which is the baseline for comparing the proposed method. The V-A3 dataset has a continuous emotional condition, however, the Emotional GAN receives this condition as a categorized discrete value. Therefore, only the images in the VA plane with emotion values within 0.5 of the difference in both the valence and arousal values, representing anger, anxiety, disgust, etc., were used. For instance, according to Kuperman *et al.* [66], anger has values of 2.5 and 5.65 in the VA plane. Thus, we classified images with valence between $[2.0, 3.0]$ and arousal between $[5.15, 6.15]$ as anger. We also compared the results of our model with Kavalerov *et al.* (MHingeGAN) [67] model having the best performance among the conditional GANs. Because our model was not a method of transforming an existing image, but a method of creating a new image with a target emotion from noise, our method was compared only with other works that create an image from noise.

Table 4: The averages of predicted VA errors using the machine learning model of Kim *et al.* [37]

Dataset	Method	Valence-error	Arousal-error
V-A2 dataset	Non-ERU	2.74	3.05
	Non-AFM loss	1.94	1.68
	Proposed method	**1.74**	**1.59**
V-A3 dataset	Non-ERU	2.5	2.8
	Non-AFM loss	1.87	1.75
	The Emotional GAN	2.35	2.68
	MHingeGAN	1.85	1.79
	Proposed method	**1.69**	**1.51**

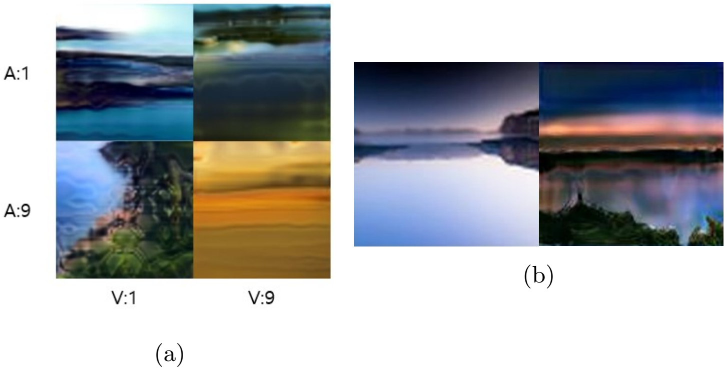
The comparison results between the different methods are shown in Table 3. The average of user preference and VA value errors were collected through user studies. In particular, the average of errors was investigated in the same way as in Section 3.4. The meaning of FID and E-FID is as in Table 1. In the case of FID and E-FID scores, the proposed model showed similar performance to MHingeGAN. However, there were significant differences in the error for the target emotion VA values, and our model recorded nearly 1.5 times more preference values.

3.6 Comparison using Machine Learning Model

To conduct a quantitative method of the ablation study rather than a qualitative one, we measured the emotion in the resulting images of both models trained from V-A2 and V-A3 dataset using the emotion predicting deep learning model of Kim *et al.* [37]. In the case of using V-A3 dataset, we additionally measured the emotion in resulting images of the Emotional GAN and MHingeGAN. For each method, 100 result images were used. Table 4 lists the valence and arousal error values for each method. The proposed method showed the lowest errors which imply that our method of using ERU and AFM-loss is the most appropriate also about the perspective of the machine learning model.

4 CONCLUSIONS

In this work, we designed a machine learning framework with a novel structure and generated emotion-based landscape scene images. To create an image that fits well with a given target emotion, we proposed a new structure called ERU that includes a unique concatenation structure. This structure had a significant and positive effect on emotional conditioning. We also presented a new feature matching loss that could highlight emotion-related features. We demonstrated that this model could generate landscape images that have target emotions. The



(a)

(b)

Fig. 7: The limitations of our model. The target emotion was given very large or very small Valence and Arousal values. The arousal changes vertically, with 1 on top and 9 on the bottom. The valence changes horizontally, with 1 on the left and 9 on the right (a) semantic location failure case (b)

suggested model had the limitation of generating images that contained artifacts or did not match the target emotion when valence and arousal were so small or so large (see Fig. 7(a)). When people watch natural landscape scenes, they usually do not feel extremely small or large arousal or valence. As a result, there are not enough natural landscape scene images in the data set that incorporates these extreme emotions. Another limitation is the case that the semantic position of the object in the resulting image is wrong, as in the left part of the figure (see the left image in Fig. 7(b)). In the figure, dark seas and gloomy skies were included in the results to represent depressed emotions with low valence and arousal values, with the positions of the sea and sky swapped up and down. In some cases, colors that are not typically seen in natural landscapes appear in the resulting image (see the right image in Fig. 7(b)). These cases appear because the model focuses only on expressing specific emotions and fails to set the correct semantic position or misses natural colors. It is not easy to strike a balance between expressing the target emotion and creating a realistic image.

We studied only natural scenes without objects in this work. In future works, we will be able to study how to create emotion-based images with scenes in which objects and backgrounds perfectly synchronize in terms of the given emotion. In addition to images, we can also apply the method in this work to video domains.

Acknowledgement This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2018-0-01419) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C2014622).

References

1. Yang, J., She, D., Sun, M.: Joint image emotion classification and distribution learning via deep convolutional neural network. In: IJCAI. (2017) 3266–3272
2. Shi, C., Pun, C.: Multiscale superpixel-based hyperspectral image classification using recurrent neural networks with stacked autoencoders. IEEE Transactions on Multimedia (2019) 1–1
3. Lyu, F., Wu, Q., Hu, F., Wu, Q., Tan, M.: Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks. IEEE Transactions on Multimedia **21** (2019) 1971–1981
4. Dong, L., He, L., Mao, M., Kong, G., Wu, X., Zhang, Q., Cao, X., Izquierdo, E.: Cunet: A compact unsupervised network for image classification. IEEE Transactions on Multimedia **20** (2018) 2012–2021
5. Wu, S., Ji, Q., Wang, S., Wong, H.S., Yu, Z., Xu, Y.: Semi-supervised image classification with self-paced cross-task networks. IEEE Transactions on Multimedia **20** (2018) 851–865
6. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3156–3164
7. Huang, G., Liu, Z., Maaten, L.v.d., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
9. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
11. Fu, K., Zhao, Q., Gu, I.Y.: Refinet: A deep segmentation assisted refinement network for salient object detection. IEEE Transactions on Multimedia **21** (2019) 457–469
12. Chen, C., Ling, Q.: Adaptive convolution for object detection. IEEE Transactions on Multimedia **21** (2019) 3205–3217
13. Tang, Y., Wu, X.: Scene text detection using superpixel-based stroke feature transform and deep learning based region classification. IEEE Transactions on Multimedia **20** (2018) 2276–2288
14. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2017) 2961–2969
16. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014)
17. Chen, L., Wu, L., Hu, Z., Wang, M.: Quality-aware unpaired image-to-image translation. IEEE Transactions on Multimedia **21** (2019) 2664–2674
18. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1501–1510

19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 2223–2232
20. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8798–8807
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. Lecture Notes in Computer Science (2016) 694–711
22. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015)
23. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network (2016)
24. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning. (2017) 214–223
25. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary equilibrium generative adversarial networks (2017)
26. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
27. Guo, Y., Chen, Q., Chen, J., Wu, Q., Shi, Q., Tan, M.: Auto-embedding generative adversarial networks for high resolution image synthesis. IEEE Transactions on Multimedia **21** (2019) 2726–2737
28. Xu, W., Keshmiri, S., Wang, G.R.: Adversarially approximated autoencoder for image generation and manipulation. IEEE Transactions on Multimedia (2019) 1–1
29. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2018)
30. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1219–1228
31. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7986–7994
32. Tan, F., Feng, S., Ordonez, V.: Text2scene: Generating compositional scenes from textual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6710–6719
33. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.S., Sun, X.: Exploring principles-of-art features for image emotion recognition. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 47–56
34. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, ACM (2015) 443–449
35. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM (2015) 435–442
36. Wu, B., Jia, J., Yang, Y., Zhao, P., Tang, J., Tian, Q.: Inferring emotional tags from social images with user demographics. IEEE Transactions on Multimedia **19** (2017) 1670–1684
37. Kim, H.R., Kim, Y.S., Kim, S.J., Lee, I.K.: Building emotional machines: Recognizing image emotions through deep neural networks. IEEE Transactions on Multimedia **20** (2018) 2980–2992

38. Zhou, Y., Shi, B.E.: Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE (2017) 370–376
39. Lu, Y., Tai, Y.W., Tang, C.K.: Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 282–297
40. Song, L., Lu, Z., He, R., Sun, Z., Tan, T.: Geometry guided adversarial facial expression synthesis. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 627–635
41. Ding, H., Sricharan, K., Chellappa, R.: Exprgan: Facial expression editing with controllable expression intensity. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
42. Yeh, R., Liu, Z., Goldman, D.B., Agarwala, A.: Semantic facial expression editing using autoencoded flow (2016)
43. Lang, P.J.: Imagery in therapy: An information processing analysis of fear. Behavior therapy **8** (1977) 862–886
44. Zhang, Q., Lee, M.: Emotion development system by interacting with human eeg and natural scene understanding. Cognitive Systems Research **14** (2012) 37–49
45. Bradley, M.M., Sabatinelli, D., Lang, P.: Emotion and motivation in the perceptual processing of natural scenes. MIT Press: Cambridge, MA (2014)
46. Simola, J., Le Fevre, K., Torniainen, J., Baccino, T.: Affective processing in natural scene viewing: Valence and arousal interactions in eye-fixation-related potentials. NeuroImage **106** (2015) 21–33
47. Zhao, S., Ding, G., Huang, Q., Chua, T.S., Schuller, B.W., Keutzer, K.: Affective image content analysis: A comprehensive survey. In: IJCAI. (2018) 5534–5541
48. Zhao, S., Yao, H., Gao, Y., Ding, G., Chua, T.S.: Predicting personalized image emotion perceptions in social networks. IEEE transactions on affective computing **9** (2016) 526–540
49. Karacan, L., Akata, Z., Erdem, A., Erdem, E.: Learning to generate images of outdoor scenes from attributes and semantic layouts (2016)
50. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1125–1134
51. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2337–2346
52. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning. Number 47. University of Illinois press (1957)
53. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
54. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
55. Sak, H., Senior, A., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth annual conference of the international speech communication association. (2014)
56. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling (2014)
57. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans (2016)

58. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. (2017) 5767–5777
59. Geisler, W.S., Perry, J.S.: Statistics for optimal point prediction in natural images. Journal of Vision **11** (2011) 14–14
60. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE (2010) 3485–3492
61. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 1452–1464
62. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
63. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. (2017) 6626–6637
64. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science **6** (2011) 3–5
65. David, A.M., Amores, J.: The emotional gan : Priming adversarial generation of art with emotion. In: NIPS 2017 Workshop. (2017)
66. Kuperman, V., Estes, Z., Brysbaert, M., Warriner, A.B.: Emotion and language: Valence and arousal affect word recognition. Journal of Experimental Psychology: General **143** (2014) 1065
67. Kavalerov, I., Czaja, W., Chellappa, R.: cgans with multi-hinge loss. arXiv preprint arXiv:1912.04216 (2019)