Hard-negative Sampling with Cascaded Fine-Tuning Network to Boost Flare Removal Performance in the Nighttime Images

Soonyong Song and Heechul Bae
Digital Convergence Research Laboratory
Electronics and Telecommunications Research Institute (ETRI)
{soony, hessed}@etri.re.kr

Abstract

When light passes through a camera lens, it creates a residue called "flare" due to the interaction between foreign substances on the lens surface and internal glasses. At night, images can be distorted by flare due to multiple light sources, and research has been conducted using neural networks to remove the flare and solve this problem. However, to our knowledge, research on this approach has only recently begun, and the results are still limited, with only a few models available for use. Further research is needed to determine if the existing models provide optimal results. As part of the mentioned research, we propose a cascaded neural network structure as a means of fine-tuning earlier models to improve their performance. We optimize the performance of the proposed model by constructing triplets using the outputs of two identical neural networks and applying contrastive learning. To demonstrate the superiority of the proposed method, we quantitatively evaluated it by measuring PSNR and SSIM. We also visually compared the differences in image details after removing the flare. Experimental results confirmed that the images reconstructed by the proposed model were superior in terms of PSNR and SSIM in streak regions, compared to the results generated by the reference model.

1. Introduction

Flare is a type of distortion that can occur in images captured by a camera when a bright light source is present in or around the image frame. Flare is a visual distortion that can appear in images captured by a camera, caused by light scattering or reflection due to optical interference among glass clusters inside the lens or damage on the surface of the lens, as reported in previous studies [17], [23]. Flare can manifest in various patterns, such as stripes, lines, spots, color saturation, ghosts, blurring, blotches, or haze [3], [28]. These patterns can reduce the details around the flare and even obscure important content in the image. The impact of flare becomes more severe, especially in low-light conditions at night, due to the influence of multiple artificial lights [12].

The conventional method for removing flare is to compensate for the optical components physically. Indeed, when using a hood or filter on the lens or a lens with a special coating material, it is possible to obtain a photo without flare by adjusting the shooting conditions appropriately [15]. However, when taking photos, there are limitations in the composition, and additional costs may be required for using auxiliary equipment. Furthermore, these methods have a critical disadvantage that they almost do not work, especially at night when there are multiple artificial light sources [3]. Post-processing methods for distortion compensation have been devised because it is physically impossible to completely overcome the phenomenon of lens flare. The post-processing methods can be divided into computational and learning-based approaches. Computational methods use mathematical models of the physical characteristics of flare to compensate for the distortion in the image caused by flare. The flare can be removed by using linear transformations, filtering, or thresholding methods, as reported in previous studies [4], [27], [32]. Computational methods work well for given specific problems, but since flares can occur due to various factors, they do not generalize well to various real-world images [23]. A learning-based approach is a method of removing flares from images using a trained neural network. Most learning-based methods train neural networks for flare removal using the supervised learning approach. The dataset for training flare removal neural networks consists of flare images and images without flares. Flare removal neural networks mainly use autoencoder structures such as U-Net. The learning-based method is known to be more successful in removing flares compared to conventional methods. However, creating a dataset composed of image pairs is a very difficult task, and obtaining ground truth from completely identical perspectives is difficult. In addition, obtaining a sufficient number of training samples to demonstrate good performance for

flare removal requires a significant amount of human labor. Recently, datasets have been developed using flare patterns as the base data to synthesize real images, due to the difficulty of obtaining a sufficient number of training samples for flare removal using real-world images [3].

Recent studies have proposed the use of advanced autoencoder structures such as Uformer, in addition to U-Net, for flare removal. In an autoencoder, the encoder takes an input image and encodes its main features into a latent space, which is then decoded by the decoder to reconstruct the desired output. Autoencoders are known to be optimized through Mean-Squared Error (MSE) loss if it is assumed that the latent variables follow a Gaussian distribution [7]. However, in a reconstruction task, using MSE loss can result in excessive regularization, which can lead to poor representation of image details [30], [8]. In the field of dehazing, contrastive learning has been applied to alleviate the problem of loss of details in image reconstruction [26]. This method sets the autoencoder output as the anchor, the target image as positive, and the source image as negative. The method trains the distance between the anchor and positive images in the embedding space to be close, while the distance between the anchor and negative images is trained to be far. To achieve this goal, a large number of negative samples with high similarity to the anchor are needed [18]. As the number of hard negative training samples increases, the neural network can distinguish features more precisely, leading to better performance [29].

This paper proposes a cascaded fine-tuning network to enhance flare removal performance in images captured at night. The proposed method can obtain triplet samples through a cascaded fine-tuning network and then optimize the network using contrastive learning. The cascaded fine-tuning network consists of two identical flare removal networks connected in series. The first network acts as a negative sampler. For the first network, pre-trained weights from a previously trained flare removal network will be applied, and the gradients will be fixed to ensure that coherent negative samples are produced. The second network acts as the anchor sampler. Each sample from the sampling networks will be a part of a triplet sample to train the proposed cascaded fine-tuning network. The proposed network makes it easy to obtain hard-negative samples, which can lead to successful contrastive learning. This structure is trained so that its output is far from the output of the first network, and close to the target. The proposed network will reconstruct output images close to the target images. Therefore, it can improve the performance of flare removal.

The contribution of this paper can be summarized as follows:

- We introduced a cascaded network structure for generating triplet samples, which makes it easy to obtain hard-negative samples to train on contrastive learning. This network improves the performance of flare removal.
- We demonstrated the effectiveness of the proposed network by providing quantitive experiment results on the flare removal dataset. Additionally, we provided to compare reconstructed details by visualization.

2. Related Works

2.1. Artifact Removal

The topic of directly researching artifact removal from images based on the characteristics of flares has only recently been discussed. However, topics dealing with artifact removal in images, such as deblurring [33], dehazing [2], denoising [1] [14], dust spot removal [9], image compression [21], [11], and super-resolution [24], have been researched in various directions even before the recent discussion on directly studying artifact removal based on the characteristics of flares. Zhou et al. [33] proposed STFAN proposes a new Filter Adaptive Convolutional (FAC) layer to address spatially variant blur for alignment and deblurring in a unified framework. Evaluation of benchmark datasets and real-world videos shows that STFAN performs favorably against state-of-the-art methods in terms of accuracy, speed, and model size. Chen et al. [2] proposed an end-to-end gated context aggregation network that directly restores haze-free images without traditional low-level or handcrafted image priors. Chang et al. [1] proposed a spatial-adaptive denoising network (SADNet) for efficient single-image blind noise removal. This method introduced a residual spatial-adaptive block, deformable convolution, and an encoder-decoder structure with a context block to capture multiscale information. Prakash et al. [14] proposed an architecture called Hierarchical DivNoising (HDN) based on a hierarchical variational Autoencoder. HDN learns an interpretable multi-scale representation of artifacts and removes image artifacts commonly occurring in microscopy data. Li et al. [9] proposed to detect attention maps to identify regions that need to be restored and use a flow completion module to hallucinate the flow of the background scene. Svoboda et al. [21] proposed to train large and deep convolutional neural networks (CNN) for JPEG compression artifacts reduction using residual learning, skip architecture, and symmetric weight initialization, networks with 8-layers can be trained in a single step with relatively short time. Most learning-based studies for removing image artifacts use autoencoders, and some use generative adversarial networks (GANs). In terms of image reconstruction, GANs have the disadvantage of being difficult to train due to the mode collapse problem and requiring long training times [20]. Autoencoders can be trained more stably compared to GANs because they converge well [22]. On the other hand, the reconstructed images from autoencoders are

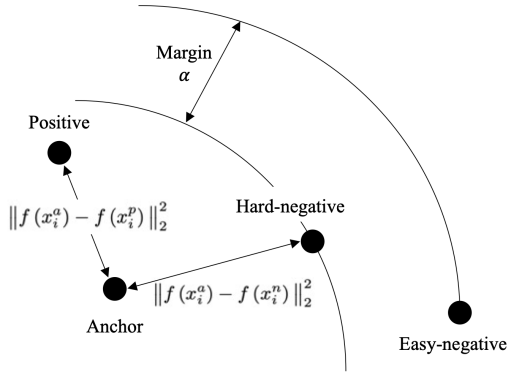


Figure 1. The geometric relationship of the triplet samples in an embedding space: anchor, positive, and negatives (hard and easy cases)

often blurry compared to GANs. However, several studies have been proposed to improve the image details using techniques such as perceptual loss, which were adopted in GANs, to enhance the image details [13]. This paper also uses neural networks with an autoencoder for artifact removal caused by lens flare.

2.2. Contrastive Learning

Contrastive learning is a neural network training technique that induces similar images to cluster together and dissimilar images to increase distance farther apart in the latent space of image embeddings. In contrastive learning, various types of loss functions are defined depending on the number of data composing one sample. First, a contrastive loss is used when a single sample is composed of two data points [5]. The two data are defined as positive if their similarity is high, and negative if their similarity is low. Next, a triplet loss is defined for a sample that consists of a positive, a negative, and an anchor [18]. The anchor is a data point that is similar to the positive data but dissimilar to the negative data in the triplet loss. The triplet loss is defined by the following equation:

$$L_{triplet}(x_i^a, x_i^p, x_i^n) = \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \quad (1)$$

where, x_i^a, x_i^p, and x_i^n are i-th anchor, positive, and negative samples, respectively. $f(\cdot)$ is embedding representation, α is margin, and $\|\cdot\|_2^2$ is L2-distance calculated by MSE.

During optimization for the model, triplet selection is known to affect the convergence speed and training performance. Especially, hard-negative mining is crucial since it affects those purposes. The hard-negative mining refers to the rule of selecting negative data that is as similar as possible to the anchor data, and training models according to this rule can improve the ability to distinguish features of the data [16]. When training a model using the data samples obtained by hard-negative mining as shown in Figure 1, it encourages the model to learn to be close to similar samples but be far apart from dissimilar samples in the latent space [34]. In this paper, the proposed cascaded fine-tuning network will be optimized using triplet loss. The proposed network conducts hard-negative mining using the intermediate outputs of the model as negative samples.

2.3. Dataset: Flare7K

Flare7K is a state-of-the-art nighttime flare dataset created by observing and statistically analyzing the lens flare phenomenon that occurs in nighttime environments [3]. This dataset consists of 5000 scattered flare images and 2000 reflected flare images. With the flare patterns in this dataset and the synthesis of flare-free images, virtual flare images can be produced. This approach makes nighttime flare images possible to automate the dataset creation process and obtain perfect ground truth. In this paper, flare removal is performed based on the Flare7K dataset. The subset provided in [32] is used as the 23,949 flare-free background images in the Nighttime Flare Removal competition.

3. Cascaded Fine-Tuning Network

3.1. Network Structure

This paper proposes a network structure for flare removal and a training method using triplet samples. The proposed method in this paper for triplet mining involves applying a cascaded model structure, as shown in Figure 2. This proposed network is designed to concatenate two identical neural network models in series. The first model is the negative sampling network, which is used to get hard-negative samples. The output of the first model should be similar to the positive or anchor image. As mentioned, contrastive learning groups negative samples similar to anchor samples when constructing triplets. This policy leads the model can learn to separate the distance between clusters in the latent space. We will use a pre-trained legacy flare removal model to obtain hard-negative samples. Since the legacy flare removal model is optimized to output images similar to the target image, the proposed network can be considered as providing sufficiently difficult negative samples that are challenging to distinguish from positive samples. The negative sampling network should be able to make high-quality negative samples consistently. To this end, in the proposed method, the gradients are fixed so that the parameters of the negative sampling network do not change. The second network is the anchor sampling network, which is used to get anchor samples. The anchor sampling network is trained to

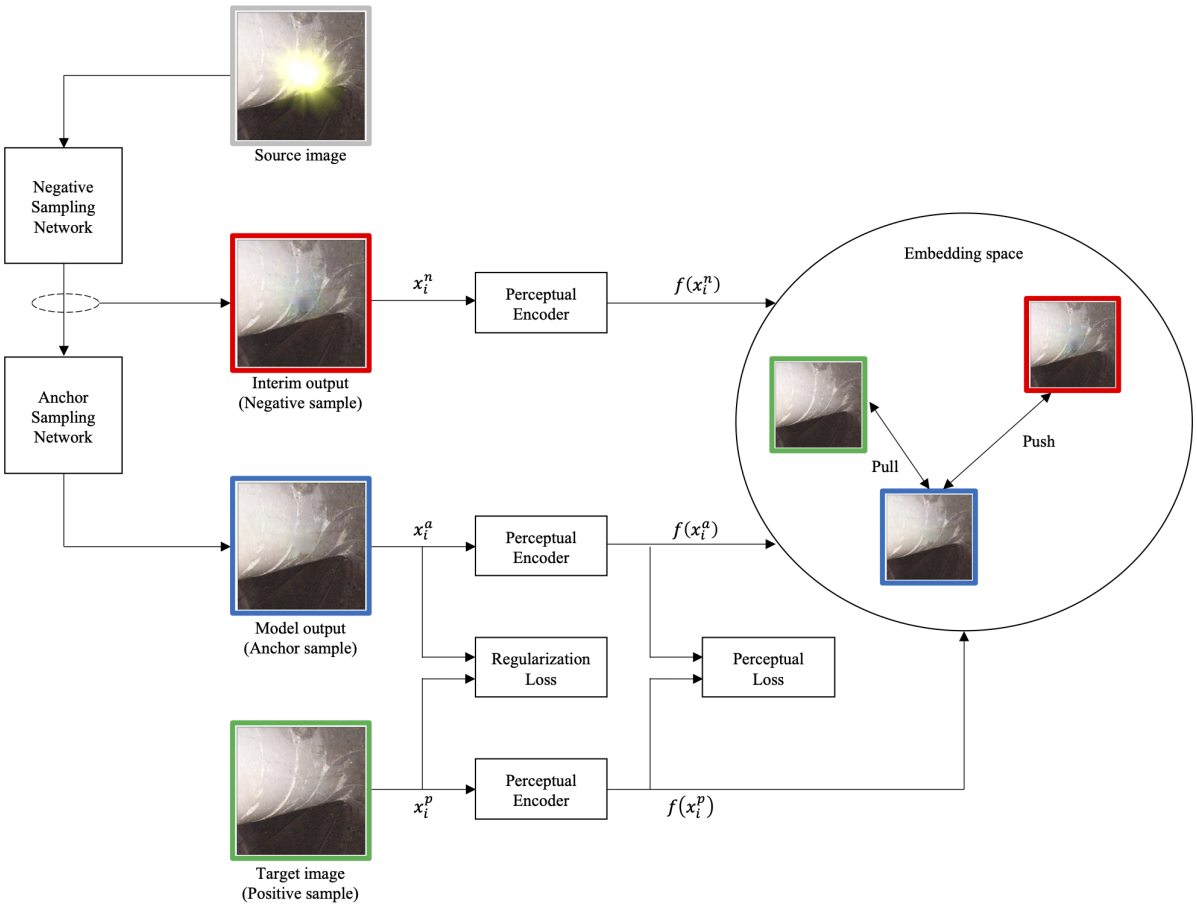


Figure 2. Structure of the proposed cascaded fine-tuning network for flare removal: The proposed design allows to makes easier hard-negative samples for contrastive learning.

output images that are similar to the positive image while being distinguishable from the negative image.

Three loss functions are used to train the proposed cascaded fine-tuning network. The first one is the regularization loss, denoted as L_{REG}. The pixel-wise error between the positive sample x_i^p and the anchor sample x_i^a of the i-th triplet is computed as follows:

$$L_{REG}(x_i^a, x_i^p) = \left\| x_i^a - x_i^p \right\|_k^k \tag{2}$$

where k is either 0 or 1, and if $k = 1$, it represents the L1 norm, whereas if $k = 2$, it represents the L2 norm.

The second one is the perceptual loss L_{PL}. Training neural networks with both MSE loss and perceptual loss has been found to have advantages in preserving image details in super-resolution tasks, as shown in [6], [8] and [31]. Perceptual loss utilizes the feature maps of the activation layer just before the output layer in a pre-trained neural network. Perceptual loss can be calculated using the Euclidean distance between the feature map $f\left(x_i^p\right)$ of a positive sample and the feature map $f\left(x_i^a\right)$ of an anchor sample.

$$L_{PL}(x_i^a, x_i^p) = \left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2. \tag{3}$$

The third one is the triplet loss $L_{triplet}$ defined in equation (1). $L_{triplet}$ calculates the embedding similarity of triplet samples. The embeddings of triplet samples can be obtained through the perceptual encoder. The similarity between images is evaluated by calculating the Euclidean distance in the latent space of the perceptual encoder. We trained the proposed network, which takes into account the weights of the three loss functions mentioned above. We will apply 0.001 for δ, as set in [8], and determine the value of λ experimentally in the later section.

$$\begin{aligned} &L_{prop}(x_i^a, x_i^p, x_i^n) \\ &= \lambda \times L_{REG}(x_i^a, x_i^p) + \delta \times L_{PL}(x_i^a, x_i^p) \\ &+ (1 - \lambda) \times L_{triplet}(x_i^a, x_i^p, x_i^n). \end{aligned} \tag{4}$$

3.2. Perceptual Encoding

Perceptual encoding is used to extract feature maps from images, and perceptual encoders use these extracted features to represent them in embedding space. In [8], a pre-trained VGG neural network was used as the perceptual encoder, but it is possible to use non-VGG neural networks as well [10]. It has been studied that using a neural network as the backbone of the perceptual encoder, which has superior classification or detection performance, results in good perceptual similarity judgment ability [31]. Many publicly available image classification neural networks are trained on the ImageNet dataset, and recent models have exceeded 88 % top-1 accuracy. In this paper, the ViT-Large model fine-tuned using large-scale weakly supervised learning was set as the backbone of the perceptual encoder [19].

4. Experiment

4.1. Evaluation Setup

We evaluate the performance of the proposed network in flare removal using PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) [25]. The PSNR quantitatively evaluates the difference between images by measuring the pixel-wise MSE between the target image and the flare-removed image. Another metric, SSIM, quantitatively evaluates the difference between two images by measuring luminance, contrast, and structure, similar to human visual perception. We calculated the PSNR and SSIM of the entire image area, streak area, and glare area [3] using the `peak_signal_noise_ratio` and `structural_similarity` modules implemented in the scikit-image package.

We describe the software and hardware setup for conducting experiments on the proposed network. All software packages used for the experiments are based on Python 3.8.10. We developed the training functionality of the proposed cascaded fine-tuning network using neural network modules implemented in PyTorch 1.14 and Torchvision 0.15. We applied pre-trained models provided in [3] as the initial weights for the negative and anchor sampling networks in the proposed cascaded fine-tuning network. Here, the negative sampling network was specified with the `requires_grad` attribute of its neural network parameters set to False to prevent training. Therefore, only the anchor sampling network undergoes the training process. The pre-trained model to be used in the experiment is implemented using the Uformer architecture.

For training, the background images were augmented using random crop, vertical flip, horizontal flip, and random erasing techniques. The resolution of the image was adjusted to be 128×128 or 512×512 by applying random crop or resize. For quantitative evaluation, images of size 128×128 were used, whereas images of size 512×512 were used for visual evaluation. The proposed model was trained using the SGD optimizer with a learning rate of 0.001. The software used for the experiment was executed on an environment with an AMD Threadripper 3975X CPU, 256 GB RAM, and NVIDIA Quadro RTX A6000 GPUs. To conduct training using multi GPUs, we used the horovod 0.26.1 package.

Method	PSNR (dB)	SSIM
Uformer [3]	23.74 / 30.28 / 27.79	0.8692 / 0.9646 / 0.9269
Ours		
$k = 1, \lambda = 1.0$	23.75 / 30.42 / **27.80**	0.8698 / 0.9651 / 0.9274
$k = 1, 0.8$	**23.76** / 30.39 / **27.80**	0.8693 / 0.9649 / 0.9270
$k = 1, 0.6$	23.69 / 30.46 / 27.73	0.8699 / 0.9653 / 0.9273
$k = 1, 0.4$	23.62 / 30.50 / 27.68	0.8706 / 0.9656 / 0.9277
$k = 1, 0.2$	23.72 / 30.44 / 27.78	0.8695 / 0.9651 / 0.9271
$k = 1, 0.0$	23.71 / 30.37 / 27.75	0.8679 / 0.9648 / 0.9263
$k = 2, 1.0$	23.42 / 30.58 / 27.50	**0.8710** / 0.9661 / 0.9282
$k = 2, 0.8$	23.09 / **30.60** / 27.20	0.8702 / **0.9665** / 0.9284
$k = 2, 0.6$	23.18 / **30.60** / 27.29	0.8706 / 0.9664 / **0.9285**
$k = 2, 0.4$	23.71 / 30.44 / 27.76	0.8692 / 0.9651 / 0.9269
$k = 2, 0.2$	23.71 / 30.38 / 27.76	0.8680 / 0.9648 / 0.9264
$k = 2, 0.0$	23.71 / 30.37 / 27.75	0.8679 / 0.9648 / 0.9263

Table 1. Quantitative evaluation results of the reference and the proposed networks in terms of PSNR and SSIM for the entire image, streak, and glare regions

4.2. Results

To confirm the performance benefits of the proposed model, we compared the PSNR and SSIM metrics of the entire image, streak, and glare areas using a reference model [3]. The flare removal images in [3] were obtained from the output images of a negative sampling network with a pre-trained model applied, and the performance was evaluated using a validation set of real flare-corrupted images that provided both ground truth and flare mask. To compare the performance of the proposed model with the reference model, we used 100 validation images and calculated the average PSNR and SSIM values for the entire image area and the flare-corrupted area. The results are presented in Table 1.

In the quantitative evaluation results of Table 1, it can be observed that the proposed method demonstrates good performance in most test cases. The PSNR of the streak region of the proposed model is generally better than that of the reference model under most conditions, with a maximum performance gain of 0.32 dB. However, the PSNR of the glare region shows poor performance compared to the reference model. Meanwhile, the proposed method shows a maximum performance gain of 0.0019 and 0.0016 in the streak and glare regions, respectively, according to SSIM. Based on these results, it can be anticipated that the proposed method will have an advantage in streak region PSNR per-

Figure 3. Comparing the two output samples from negative(reference) and anchor(proposed) networks: quantitative results of the negative and the anchor samples in terms of PSNR and SSIM (negative/anchor) for synthetic flare dataset: (a) 29.40 / 30.12, 0.9895 / 0.9883 (b) 33.88 / 34.37, 0.9598 / 0.9689 (c) 17.44 / 16.66, 0.9058 / 0.9000 (d) 22.37 / 21.94, 0.9731 / 0.9751

formance. Therefore, we trained our flare removal model using the condition of $k = 2$ and $\lambda = 0.6$, which we expect to exhibit better performance than the others in that region.

Since quantitative evaluations cannot perfectly assess image quality, it is necessary to visually confirm the results. We visualized flare-removed images obtained from the proposed and the reference networks to confirm whether the proposed network has a superior ability to remove flare in practice. We used two types of datasets provided by [3] for visualization purposes. The first type is synthetic data that combines actual images with patterns from the flare dataset, while the other type is real data where actual flares occurred. First, we compared the flare removal results of the existing model and our proposed model on synthetic data in Figure 3. We have marked the areas where Flare has been removed with a red box, and the positions that appear abnormal in the Anchor sample with a green box. Figure 3 (a) shows an image where the light source, streak, and glare regions coexist, and both the negative and anchor samples appear to have flare removed well. While the anchor sample showed superior results in terms of PSNR, the negative sample showed superior results in terms of SSIM. Figure 3 (b) is an image where streaks and glare are present, with relatively strong streaks appearing in the image. When examining the streak area of the image, it appears that flare has been well removed in the anchor sample, and both PSNR and SSIM results are measured to be superior. Figure 3 (c) shows a strong light source, streak, and glare. It is a condition similar to Figure 3 (b), but both PSNR and SSIM are measured to be superior in the negative sample. This is considered to be due to the proposed model excessively correcting for flare. Figure 3 (d) also shows a similar phenomenon to Figure 3 (c). While the restoration of the streak area showed good results, the problem of excessively attempting to remove flare was identified. In Figure 3, since the flare mask for synthetic data was not provided, PSNR and SSIM were

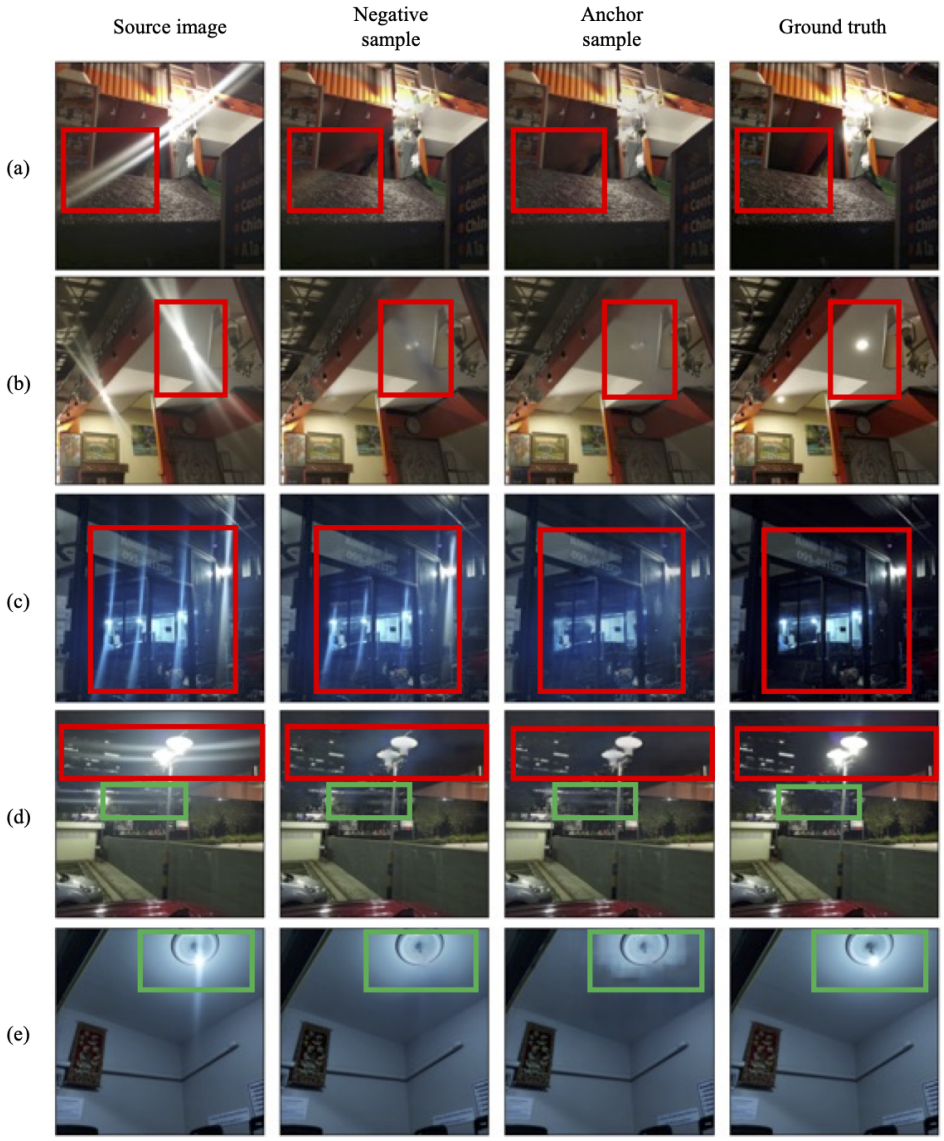
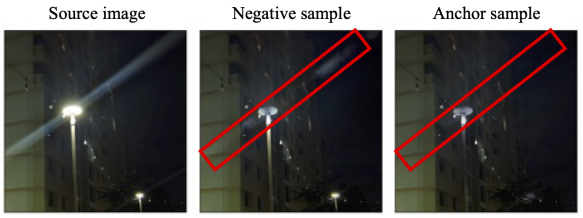


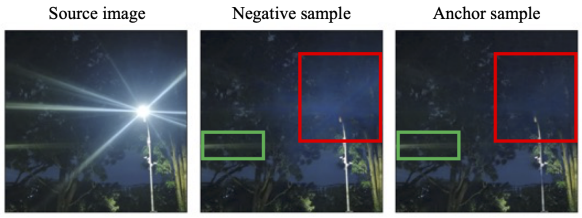
Figure 4. Comparing the two output samples from negative(reference) and anchor(proposed) networks: quantitative results of the negative and the anchor samples for real flare dataset (streak PSNR, glare PSNR, streak SSIM, and glare SSIM): (a) 29.34 / 29.96, 28.87 / 28.10, 0.9670 / 0.9774, 0.9508/0.9556 (b) 30.94 / 32.47, 29.73 / 30.04, 0.9661 / 0.9745, 0.9438 / 0.9484 (c) 19.54 / 24.54, 17.72 / 19.18, 0.8900 / 0.9145, 0.7514 / 0.7848 (d) 31.45 / 30.62, 28.84 / 26.92, 0.9783 / 0.9773, 0.9553 / 0.9520 (e) 38.94 / 34.18, -/-, 0.9984 / 0.9973, -/-

calculated for the entire area of the image. When testing the flare removal performance on synthetic data, the average PSNR of the reference model and the proposed model was measured to be 27.63dB and 26.59dB, respectively, with the reference model being higher. However, the average SSIM was measured to be 0.9615 and 0.9633, respectively, with the proposed model being higher.

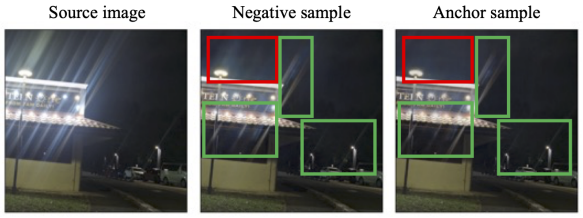
Next, Figure 4 shows the visualization results for the real data, which included a flare mask that allowed for quantitative evaluation of flare removal performance in the flare regions. In Figures 4 (a) and (b), the anchor samples exhibit natural-looking results in areas where flare has been removed. Figure 4 (c) confirms that flare has been effectively removed in the streak area of the anchor sample, even in situations where multiple light sources are present. The anchor samples in Figures 4 (a), (b), and (c) all displayed excellent PSNR and SSIM performance. However, in Figure 4 (d), while the anchor sample removed strong light source flares effectively, it left residual flare compared to the negative sample in areas with weaker light sources. The negative sample showed better performance in quantitative results. Additionally, in Figure 4 (e), the anchor



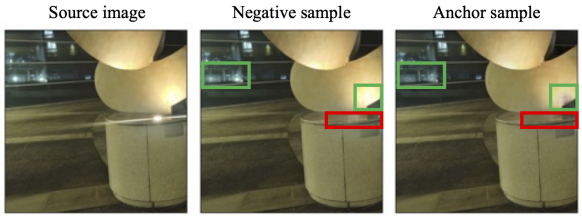
(a) Flare caused by a strong single light source



(b) Flare caused by both a strong and a weak light source



(c) Flare caused by multiple light sources



(d) Widely spread light source and excessive correction

Figure 5. Checking the results of two networks (the reference and the proposed network) using a set of provided test images in the Nighttime Flare Removal competition

sample excessively removed reflection regions, resulting in an unnatural-looking image and inferior PSNR and SSIM measurements. When testing the flare removal performance on real data in terms of the average streak PSNR, glare PSNR, streak SSIM, and glare SSIM, the reference model was measured as 34.10dB, 28.93dB, 0.9843, and 0.9450, respectively. On the other hand, the proposed model was measured as 34.30dB, 28.41dB, 0.9863, and 0.9482, respectively. Based on these measurements, it was confirmed that the proposed model has superior performance in removing flare in streak regions.

Lastly, we visually inspected the test images of the challenge through the network we learned. In Figure 5a, 5b, and 5c, we can visually confirm that the network we proposed improves flare removal performance compared to the reference. The subfigures in Figure 5 are arranged in the order of the test image, negative sample image, and anchor sample image provided in the Nighttime Flare Removal competition. In Figure 5a, the image contains flare caused by a strong single light source, and we can see that the proposed model removes the flare almost completely while still looking natural. In Figure 5b, the image contains flare caused by both a strong and a weak light source, and we can observe that the proposed model removes the flare caused by the strong light source well, but has weaknesses in removing the flare caused by the weak light source. In Figure 5c, the image contains flare caused by multiple light sources, and we can see that the proposed model does not perform well in removing flare in images where multiple flares overlap. In Figure 5d, we can see that the proposed model recognizes a widely spread light source as flare and excessively corrects the image.

5. Conclusion

In this paper, we proposed a method to improve image reconstruction performance by fine-tuning a network for artifact removal caused by nighttime lens flare. The proposed fine-tuning network has a cascaded structure that concatenates the existing flare removal models, and the actual training is performed in the second neural network. We constructed triplet samples using the intermediate and final outputs of the connected networks and trained the proposed network using contrastive learning to make it closer to the features of the target image. When compared to the reference network, we confirmed that the images reconstructed with the proposed network generally had higher PSNR and SSIM values. These results were also confirmed by visual analysis. However, it was also observed that the proposed network tends to excessively correct the image, leading to a decrease in the quality of the output results. We confirmed that these characteristics are maintained even in environments where the actual nighttime flare is mixed. While it is true that the proposed network performs better than the reference network, there is still a limitation in that various comparative experiments have not been conducted. These limitations will be addressed through further research.

Acknowledgment

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. Also, thank you to the reviewers who provided helpful comments. [23ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways]

References

[1] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 171–187. Springer, 2020. 2

[2] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1375–1383. IEEE, 2019. 2

[3] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7k: A phenomenological nighttime flare removal dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 2, 3, 5, 6

[4] Jinwei Gu, Ravi Ramamoorthi, Peter Belhumeur, and Shree Nayar. Removing image artifacts due to dirty camera lenses and thin occluders. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10. 2009. 1

[5] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 4

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[8] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2, 4, 5

[9] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V Sander. Let's see clearly: Contaminant artifact removal for moving cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2011–2020, 2021. 2

[10] Yifan Liu, Hao Chen, Yu Chen, Wei Yin, and Chunhua Shen. Generic perceptual loss for modeling structured output dependencies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5424–5432, 2021. 5

[11] Danial Maleki, Soheila Nadalian, Mohammad Mahdi Derakhshani, and Mohammad Amin Sadeghi. Blockcnn: A deep network for artifact removal and image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2555–2558, 2018. 2

[12] Laura FB Marangoni, Thomas Davies, Tim Smyth, Airam Rodríguez, Mark Hamann, Cristian Duarte, Kellie Pendoley, Jørgen Berge, Elena Maggi, and Oren Levy. Impacts of artificial light at night in marine ecosystems—a review. *Global Change Biology*, 28(18):5346–5367, 2022. 1

[13] Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki. Improving image autoencoder embeddings with perceptual loss. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 3

[14] Mangal Prakash, Mauricio Delbracio, Peyman Milanfar, and Florian Jug. Interpretable unsupervised diversity denoising and artefact removal. *arXiv preprint arXiv:2104.01374*, 2021. 2

[15] Xiaotian Qiao, Gerhard P Hancke, and Rynson WH Lau. Light source guided single-image flare removal from unpaired data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4177–4185, 2021. 1

[16] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 3

[17] Eden Sassoon, Yoav Y Schechner, and Tali Treibitz. Flare in interference-based hyperspectral cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10174–10182, 2019. 1

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 3

[19] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022. 5

[20] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017. 2

[21] Pavel Svoboda, Michal Hradis, David Barina, and Pavel Zemcik. Compression artifacts removal using convolutional neural networks. *arXiv preprint arXiv:1605.00366*, 2016. 2

[22] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-gan: An improved gan using distance constraints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–385, 2018. 2

[23] Patricia Vitoria and Coloma Ballester. Automatic flare spot artifact detection and removal in photographs. *Journal of Mathematical Imaging and Vision*, 61(4):515–533, 2019. 1

[24] Tao Wang, Kaihao Zhang, Xuanxi Chen, Wenhan Luo, Jiankang Deng, Tong Lu, Xiaochun Cao, Wei Liu, Hongdong Li, and Stefanos Zafeiriou. A survey of deep face restoration: Denoise, super-resolution, deblur, artifact removal. *arXiv preprint arXiv:2211.02831*, 2022. 2

[25] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 5

[26] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10551–10560, 2021. 2

[27] Tai-Pang Wu and Chi-Keung Tang. A bayesian approach for shadow extraction from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 480–487. IEEE, 2005. 1

[28] Yicheng Wu, Qiurui He, Tianfan Xue, Rahul Garg, Jiawen Chen, Ashok Veeraraghavan, and Jonathan T Barron. How to train neural networks for flare removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2239–2247, 2021. 1

[29] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 126–142. Springer, 2020. 2

[30] Ronald Yu. A tutorial on vaes: From bayes' rule to lossless compression. *arXiv preprint arXiv:2006.10273*, 2020. 2

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 5

[32] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 1, 3

[33] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2482–2491, 2019. 2

[34] Ruixi Zhu, Li Yan, Nan Mo, and Yi Liu. Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:72–89, 2019. 3