Supervised Learning Based Peripheral Vision System for Immersive Visual Experiences for Extended Display

Muhammad Ayaz Shirazi 1,2, Riaz Uddin 2,3 and Min-Young Kim 1,4,*

1 School of Electronics Engineering, IT College, Kyungpook National University, 1370 Sankyuk-dong, Buk-gu, Daegu 41566, Korea; mshirazi@ee.knu.ac.kr
2 Haptics, Human-Robotics and Condition Monitoring Lab (National Center of Robotics and Automation), NED University of Engineering and Technology, Karachi 75270, Pakistan; riazuddin@neduet.edu.pk
3 Department of Electrical Engineering, NED University of Engineering and Technology, Karachi 75270, Pakistan
4 Research Center for Neurosurgical Robotic System, Kyungpook National University, 1370 Sankyuk-dong, Buk-gu, Daegu 41566, Korea
* Correspondence: minykim@knu.ac.kr

Abstract: Video display content can be extended to the walls of the living room around the TV using projection. The problem of providing appropriate projection content is hard for the computer and we solve this problem with deep neural network. We propose the peripheral vision system that provides the immersive visual experiences to the user by extending the video content using deep learning and projecting that content around the TV screen. The user may manually create the appropriate content for the existing TV screen, but it is too expensive to create it. The PCE (Pixel context encoder) network considers the center of the video frame as input and the outside area as output to extend the content using supervised learning. The proposed system is expected to pave a new road to the home appliance industry, transforming the living room into the new immersive experience platform.

Keywords: augmented video; human vision; immersion; large field of view; spatial augmented reality; video extrapolation; neural network; AI

1. Introduction

The television (TV) is considered as one of the major sources of the living room entertainment with the huge improvement in the visual and audio quality of the display over the past few years. However, the TV display has two limitations [1].

1. It only provides the screen with the limited content and field of view as the peripheral vision system lacks due to the limited screen size and the image formats.
2. It provides the visual experience restricted to the virtual scene boxed in the frame of the display and the physical environment surrounding the user is neglected.

The solution to these problems is the automatic generation of the new content beyond the existing screen size and image formats, and their projection onto the periphery of TV screen using projector [2]. This augmentation leads to immersive, interactive, and entertaining visual experiences for the user. Though the user may manually create the plausible content for the existing TV screen, it is too expensive to create it.

Peripheral projection is an efficient technique to enhance the visual experiences of the user, but its employment is limited in the common display applications due to the difficult problem of the context image generation [3]. Context (or context image) refers to the image or Video frame projected on the wall around the traditional TV screen to create immersive experiences [3]. Illumiroom [1] explained the concept of the context including the full image content, edges, partial image content and other special effects to enhance the visual experiences of the user. The proposed work utilized the F + C Full context [1] as the context image (see Table 1). The problem of context image generation was solved by...
the approaches either based on computer vision algorithm or deep neural network in the literature [3–6].

Table 1. Merits and demerits of extended vision design schemes [1].

| S. No. | Focus + Context Scheme | Merits | Demerits |
|--------|------------------------|--------|----------|
| 1.     | F + C Full             | The user pays attention to the LED TV screen and the visual experience is enhanced by the peripheral projection on the whole background. | It uses a non-flat, non-white projection surface with radiometric compensation. Limited ability to compensate the existing surface color. |
| 2.     | F + C Edges            | Robustness to the ambient light in the room with enhanced optical flow. | Projection of the black and white edge information instead of colored content. |
| 3.     | F + C Seg              | This scheme allows the projection on the specific area of the background, such as rear flat wall surrounding the television. | This scheme does not cover the whole background by peripheral projection. |
| 4.     | F + C Sel              | This scheme allows certain elements to escape the TV screen creating feelings of surprise and immersion. | This scheme does not cover the whole background by peripheral projection. |

In this paper, we propose the peripheral vision system that has the capability to produce immersive display projecting the content extended by deep neural network around the TV screen. The goals of the system are as follows:

(1) The design of the system to augment the area surrounding the TV with projected context image creating immersive visual experiences.

(2) Replacing the traditional video display with the immersive visual experiences for the user at low cost using the components of the TV and projector easily available everywhere.

(3) Quantitative and qualitative evaluation of our system with the State of the Art approaches

The proposed system is expected to pave a new road to the home appliance industry by transforming the living room into the new immersive experience platform. The proposed research provides immersive visual experiences to the user relaxing in living room and finds applications in gaming and for broadcasting video.

This paper is organized as follows: Section 2 presents a comparative overview of the existing related research work. Section 3 provides the brief overview of the proposed system and the methodology to extend the video content using deep neural network. Section 4 presents the experiments and results for the context-image generation and extended display application. Section 5 includes the conclusion and future research.

2. Related Prior Research

The concept of merging the physical and virtual worlds through video projection is termed as Spatial Augmented Reality (SAR) [7]. The idea to augment the colors of the physical objects with video projection was first developed by Raskar et al. [8] treating physical objects as displays. This work put forward a 3D model-based approach where the geometric information of the projectors and the object surfaces is assumed to be known. The employment of projectors was carried out in several projects for making the workspace environment immersive [9–11]. Several other projects enabled gaming experiences with physical objects (wooden blocks) and the game is created by the arrangement of the
blocks \[11,12\]. Several projects included the demonstration of treating the curtain (or building façade) as the ideal projection surface \[13,14\]. The physical paintings can be covered with the projected virtual content and the user may increase the contrast of the printed image by superimposing the projected content \[13,15\].

If the context image already exists, such as in games, PC work, and movies recorded with dedicated equipment, the user does not need to generate that content \[16\]. The user needs to generate the context images only from video data if the content is recorded without dedicated equipment. The simplest implementation of the context image generation is the TV with immersive viewing experience \[17\]. The TV is provided with the series of the colored LED strips on its edges and the colors of the LED strips change dynamically to match the TV content. In another project of the TV with immersive display \[17\], the projector projected the enlarged image on the wall behind the TV screen to demonstrate the immersive visual experiences for the user. The problem of the context image generation was tackled by using computer vision algorithms and deep neural networks in the literature \[3–6,18,19\].

Infinity-By-Nine \[6\] is similar to CAVE \[18\] with three projectors and screens around the TV. The method of the context image generation is based on optical flow, color analysis and heuristics. Aides et al. \[5\] employed the PatchMatch algorithm, which is implemented from the original research \[19\] for the extrapolation of the video frame to an image’s peripheral area. This method produced the extended videos close to the original content depending on the scene with high processing time of few minutes per frame. Turban et al. \[4\] proposed a lighter algorithm than the PatchMatch algorithm for real-time performance, but the generated results were more artificial than those generated by PatchMatch. In short, the approaches \[4–6,18,19\] used the computer vision algorithms to extend the context, while the proposed approach employs a deep learning-based approach to generate the context.

Image completion approaches based on convolutional neural network (CNN) \[20–27\] were also proposed in the literature to fill the hole in the image and recover the lost region. The approach for the deep learning-based context generation \[3\] satisfied the properties of the real-time performance and the naturalness at the same time, but the flickering problem disturbed the viewer’s immersion. A similar approach \[25,27\] removed the flickering problem proposing deep neural network and extended the research \[3\] to HMD videos. A recent approach based on Two-stage conditional GANS (generative adversarial network) \[26\], was presented in the literature to extend the image completion with the 360-degree image data. The computer vision-based context generation methods did not fulfill the properties of the real-time performance and the naturalness of the scene \[3\] simultaneously. These approaches lack the method to generate sufficient context images with the consideration of processing time and naturalness as compared to deep learning-based methods.

The human fovea comprises 1% of the retina. It consists of a large number of photoreceptors for high resolution vision, while the density of the photo receptors reduces outside it, resulting in the coarser vision. Due to the property of deep neural networks (DNN) to generate plausible content, deep learning was utilized for the context generation as it has an analogy with human vision being acute in narrower region \[3\]. The research proposed by Iizuka et al. \[20\] successfully generated the natural image in such a way that the user could not recognize it as ‘filled’ in first glance. The user can fill the peripheral imagery by applying the mask to the peripheral portion and considering it as ‘hole’ \[3\]. In this paper, we presented the peripheral vision system having the capability to create immersive visual display via extension of the video content using deep learning. The hardware of the peripheral vision system consists of the LED TV and the projector (see Table 2), while the software to extend the video content is PCE \[22\]. The original high-resolution video is played on LED TV, while the extended output of the deep neural network is projected using Digital projector to provide the immersive visual experiences to the user. The proposed research is different from the approaches \[3,25\] as we employed PCE network for extending the video frame, while the ExtVision and Deep dive used the pix2pix and 3D convolution-based
network, respectively. Our approach and the ExtVision tackled only the spatial correlation, while the deep dive also considered the temporal correlation in the deep neural network. Our approach provided the qualitative and quantitative evaluation of the extended content from DNN, while the ExtVision and Deep dive only performed the qualitative evaluation.

Table 2. Some of the specifications of the LED TV and DLP projector used for experiments.

| S. No. | Device           | Specifications                                                                 |
|-------|------------------|-------------------------------------------------------------------------------|
| 1.    | LED TV           | Display Type: LED<br>Resolution: 3840 × 2160<br>Display Format: 4K UHD 2160p<br>Diagonal Size: 54.6 inches<br>Refresh Rate: True Motion 120 (Refresh Rate 60 Hz) |
| 2.    | Digital Projector| Projector Type: DLP projector<br>Resolution: 1920 × 1080<br>Brightness: 3500 ANSI lumens<br>Projection ratio: 1.48 to 1.62:1<br>Size (W × H × D): 314 × 224 × 100 mm |

Our work is closely related to the ExtVision [3] which has already reported video projection and context image generation using DNN network “pix2pix”. The research [3,25] did not compare the testing results of their networks with the ground truth images in term of RMSE (Root Mean Square Error) or PSNR (Peak Signal to Noise Ratio) values. We have utilized PCE network for peripheral vision system (PCE was not previously used for peripheral vision system-based Video display application). We have performed the quantitative evaluation of the testing results with the ground truth data in term of RMSE and PSNR values. We investigated the effect of different sizes of the extended area on the network’s accuracy and concluded that larger size of the extended area results in less PSNR value (see Table 3).

Table 3. The results for RMSE and PSNR for testing video frames.

| S. No. | Extended Region (pixel) | RMSE (1 fps) | RMSE (30 fps) | PSNR (1 fps) | PSNR (30 fps) |
|--------|-------------------------|--------------|---------------|--------------|---------------|
| 1.     | 32                      | 13.63        | 13.62         | 25.43        | 25.44         |
| 2.     | 48                      | 17.96        | 18.12         | 23.04        | 22.96         |
| 3.     | 64                      | 22.51        | 22.68         | 21.08        | 21.01         |

This paper is an extended version of Shirazi et al. [2], which presented the work on small video dataset of the Ocean, while this article utilized large Ocean video dataset. Further, we have extended our research to know the validity of our proposed network on other data sets, i.e., Forest dataset. In [2], we did not provide the extensive details about the architecture of DNN and did not include the comparison with the state-of-the-art research of ExtVision and Deep dive. We have made the following contributions:

1. Utilization of the PCE network for peripheral vision system (PCE was not previously used for peripheral vision system).
2. We performed the investigation of the effect of different sizes of the extended area on the network’s accuracy and reported that larger size of the extended area results in less PSNR value (see Table 3).

3. Proposed Approach

3.1. Focus + Context Design Scheme

Illumiroom [1] focused on gaming application and included four design schemes. These schemes include Focus + Context Full (F + C Full), Focus + Context Edges (F + C Edges),
Focus + Context Segmented (F + C Seg) and Focus + Context Selective (F + C Sel). The merits and demerits of these schemes are recorded in the Table 1 as follows:

According to the approach [1], Illumiroom can also be utilized to create immersive visual experiences for the cinema and television. One of the related research, ExtVision [3], is also closest to the F + C Full of the Illumiroom. From the Table 1, it is evident that F + C Full is the most suitable Focus + Context Scheme for cinema and TV based display application as it covers the whole background by peripheral projection. The demerit of F + C Full (Table 1) can be eliminated by flat white projection surface when targeting cinema and TV application. So, we designed the approach similar to F + C Full system in which the extended content is generated using deep neural network and white flat screen was employed as projection surface.

3.2. Proposed Peripheral System

The proposed approach is a system-based approach. The proposed system consists of the integration of the hardware and context-image generation software. The peripheral hardware consists of LED TV and projector to enhance the visual experiences of the user by generating context-images using PCE network and their projection on the wall around the TV screen as shown in Figure 1. The original high-resolution video played on LED TV is synchronized with the extended video projected by the Digital projector. We have used LED TV and Digital projector in the peripheral vision system. Some of the specifications of the LED TV and projector are shown in Table 2, which we also used for the experiments of our system. The PCE supports the aspect ratio of 1:1 (input videos 256 × 256). The aspect ratios of the TV and projector both are 1.77 (16:9). We have some pre-processing steps to prepare the video frame of aspect ratio 1:1 to feed into PCE and then the output of the trained DNN is post-processed matching the aspect ratio of the TV and projector as shown in Figure 2. Any other TV screen can be used if the aspect ratio of TV and projector matches.

![Figure 1. Overview of the proposed peripheral system and its interaction with the user.](image-url)
Figure 2. The general overview of the proposed approach for the generation of the context images or video frames.

We call the proposed system as “the peripheral vision system” because human vision is acute in the fovea and the vision is coarser outside the fovea. The concept of peripheral vision has analogy with the Human vision. Our system has a central part (original high-resolution video), while the coarser content is generated by DNN [3] for peripheral projection.

3.3. PCE Architecture

In this research, we used the pixel context encoder (PCE) [22] for the context image generation. The PCE is based on PatchGAN, GAN (Generative Adversarial Network) which focuses on the structure in local patches, relying on the structural loss to ensure correctness of the global structure. The generator in the PCE architecture consists of encoder and decoder. There are two downsampling layers (discrete convolutions with a stride of 2) and four dilated convolutional layers in the encoder. The encoded image is then fed into the decoder block which consists of three discrete convolutional layers. The last two layers of the decoder upsamples the image or video frame to the original resolution using nearest neighbor interpolation. The PatchGAN discriminator consists of five layers of filters with spatial dimension of 3 × 3 using LeakyRelu activation function. The general architecture of PCE with the generator and discriminator architectures is shown in Figure 3.

Figure 3. Pixel Context Encoder (PCE) (a) General Architecture of PCE (b) Encoder decoder-based Generator Architecture and (c) Discriminator Architecture.
3.4. Mathematical Expression for Loss function

Suppose that the PCE architecture [22] denoted by function ‘F’ takes an image ‘x’ and a binary mask ‘M’. The binary mask contains ‘one’ for masked pixels and ‘zero’ for the provided pixels to generate plausible content for the masked image F(x, M). The PCE training consists of the optimization of the two loss functions: a $L_1$ loss and a GAN loss. The $L_1$ loss has its masking in such a way that it has non-zero values only inside the corrupted region. The mathematical expression for $L_1$ loss is given as follows:

$$L_{L1} = \| M \odot (F(x, M) - x) \|_1$$

(1)

where $\odot$ is the element-wise multiplication operation.

The general mathematical expression for PatchGAN [20] is as follows:

$$L_{GAN} = \min_F \max_D E_{x \sim P_{data}} [\log(D(x, M)) + \log(1 - D(F(x, M), M))]$$

(2)

where the discriminator ‘D’ differentiates between the real and fake images, while the generator ‘G’ tries to befool the discriminator using generated samples. In our case, we do not have any random mask. Finally, the GAN loss [20] is defined as follows:

$$L_{GAN} = \min_F \max_D E_{x \sim P_{data}} [\log(D(x, M)) + \log(1 - D(F(x, M), M))]$$

(3)

The PCE architecture utilized the discriminator having same functionality as the global discriminator [20] except that the PCE architecture restores the ground truth pixels before processing the generated image with the discriminator. Thus, the generated region remains consistent with the context. The overall loss is mathematically defined as follows:

$$L = \lambda L_{L1} + (1 - \lambda) L_{GAN}$$

(4)

where $\lambda$ has the value of 0.999 for all experiments following [21]. We used the PCE architecture [22] with the default settings. PCE network was used for image inpainting using binary mask and the extrapolation problem can be tackled [22], inverting the same mask.

4. Experiments and Results

4.1. PCE for Ocean and Forest Dataset

In this paper, we proposed the system that generates the extended video content using deep learning treating the center of video as input and the outside area as output using different epoch models trained on the training dataset. In this research, we used seven videos of the Ocean scene at 30 fps. We used six videos as training dataset and one as testing dataset. The training and testing videos were resized to the resolution of $256 \times 256$. We extracted the frames from the training and testing videos at 1 fps and 30 fps respectively. We trained the PCE architecture and the time for training the PCE architecture for 10 epochs was found to be 458 s. There was no risk of overfitting, and this is likely due to the low number of model parameters. The size of the minibatches varied depending on the memory capabilities of the graphics card [22]. The parameters, steps, and conditions for this experiment are the same as for PCE architecture [22].

For $256 \times 256$ videos, we considered three cases of the pixel band around the video center, i.e., 32, 48, and 64 pixels, respectively, to investigate the effect of different size of the extended area on the network’s accuracy. These pixel bands yield the video centers to be $(192 \times 192)$, $(160 \times 160)$, and $(128 \times 128)$, respectively, shown in Figure 4. For different pixel bands, we separately trained and tested PCE. The purpose of this experiment is to compare the PSNR values for the three cases. This experiment concludes that larger the pixel band results in less PSNR value. One similar experiment is reported in the related research [26]. We trained the PCE architecture for 10 to 1000 epochs for three cases and the testing result on the test video for different epochs in term of RMSE and PSNR are shown in Figure 5. The best values of RMSE and PSNR for 32, 48, and 64 pixels band are recorded.
at 120, 360, and 860 epochs, respectively. These results are recorded in Table 3 for test video frames at 1fps and 30 fps, respectively. Table 3 concludes that larger size of the extended area results in less PSNR value (see Table 3). Figure 6 shows the input, generated, and the original frame of the training and testing videos for the third case. The testing result shows that the extended area is filled with the related content of the seabed. This coincides with the fact that the DNN may generate the plausible content, but somewhat inaccurate context [3].

Figure 4. Input images for PCE network with central area as original content and white space shows the extended area (will be replaced with extended content) (a) 32 pixels band (b) 48 pixels band and (c) 64 pixels band.

Figure 5. Testing of the test video frames extracted at 30 fps using different epoch models trained on the training dataset.
In another experiment, we used two different categories of datasets, i.e., Ocean and forest datasets for PCE with only 64 pixels band (extended area). We used four videos of Ocean dataset (94 K frames extracted at 2 fps) for training and the PCE is trained for 753 epochs (1 epoch = 4700 steps). We then tested the trained PCE model on the two videos of Ocean dataset (frames extracted at 2 fps). We repeated the training of PCE with Forest dataset with seven videos (123 K frames extracted at 2 fps) and the PCE is trained for 1000 epochs (1 epoch = 6155 steps). We tested the trained PCE model on the two videos of Forest dataset (frames extracted at 2 fps). The testing results for PCE for the Ocean and forest data sets are shown in Figure 7. The testing result of the Ocean dataset shows that the extended area is filled with the related content of the seabed. In Figure 7b, the generated image contains the environment of sea plants as compared to the rock or sand in Figure 7f. We performed the testing of the trained model for the frames extracted at 30 fps. So, it seems that Figure 7 shows the same input image twice, but these images (Figure 7g,h are different. Similarly, Figure 7c,d contains the rock content (in the middle of the bottom) as compared to the grass (plant) in Figure 7g,h. These results with 64 pixel band are similar to those reported in the related work [26]. The research [26] reports the generation of the tree-like objects despite the input being building and the reason for this result was small input area (large extended area). We also performed Demos in the testbed with the peripheral vision system for the Ocean and Forest dataset. Figure 8 shows the Demos in term of the images, while the Demo videos can be found in the links.
4.2. Comparison with the State of the Art

The related works [3,25] are similar to our work. We compared our research with the work reported by Kimura and Rekimoto [3]. They used the popular deep neural network, pix2pix [23,24] for extending image and video content. As Kimura et al. [3] did not perform any qualitative evaluation of the results with the ground truth, we trained the pix2pix network on our Ocean data and compared the testing result of the trained PCE and pix2pix networks. The testing results of the trained PCE and pix2pix networks in terms of images are shown in Figure 9. This result shows that both the PCE and pix2pix networks generated the extended content close to the Ocean environment.

Figure 7. Testing of the PCE trained network for the Ocean and Forest datasets (a–d) First row shows the generated frames using trained PCE network, (e–h) Second row shows the ground truth frames for qualitative evaluation.

Figure 8. The snap shots of the Demos conducted in the Testbed for Ocean and Forest datasets with video links. (https://drive.google.com/open?id=1uUsAVfdWiqesX7eY1cmZyTCFEGhjmxP (accessed on 12 May 2021), https://drive.google.com/open?id=1p5-EJUKb3V3s-LoNzEy4Vo0uqIHvR-JD (accessed on 12 May 2021)).
We have also compared our results with the related works [3,25] using the Avatar 2009 video. We trained the PCE architecture with the Avatar dataset consisting of 199 K frames extracted from the video at 30 fps. We trained the PCE architecture with 70% data and we assigned 30% for testing PCE network. We trained the PCE architecture and compared the testing results with the ExtVision and the Deep dive [3,25] as shown in Figure 10. These results depict that the trained PCE network captured the overall features, and the generated content is closely related to the central portion. Since the Deep drive [25] captures the temporal features, we compare PCE network both in term of generated features and the variation of features due to frame motion. The PCE network and ExtVision research generate sufficient features closely related to the central image, while the result of the Deep dive has some blurring effect. The deep dive research captures the fire in consecutive frames of the video, which is missing in the case of PCE network and ExtVision.

4.3. User Study

We have performed user study for the investigation of the quality of experience (QoE) of the peripheral projection of context images. Our user study is similar to the one conducted in the ExtVision [3] and it consists of six questions on the 7 Likert-type for the “Effect 1” and “Effect 2” for Forest and Ocean datasets. “Effect 1” was a video without its context-image (extended content), i.e., a simulated traditional visual experience (ocean video playing on LED TV and light background projected around the TV). “Effect

Figure 9. The comparison of the trained PCE and pix2pix architectures on the Ocean data video during testing (a) The input image, (b) The ground truth image, (c) The output image or frame generated by PCE network, and (d) The output image or frame generated by pix2pix network.

Figure 10. The comparison of the PCE testing result with the ExtVision and Deep dive research in the same way as reported in ExtVision.
2” was a video with a peripheral projection of its context-image (ocean video playing on LED TV and extended ocean video projected around the TV). To prevent creating a bias when switching from peripheral light to no peripheral light, a light background image was projected around the television. The scenes for the default background, Effect 1 and Effect 2 are shown in Figure 11. Default background indicates the absence of both the effects, Effect 1 and Effect 2 (Figure 11a). Figure 11b depicts the scene of the LED TV with high resolution ocean video, while the periphery consists of light background (Effect 1). Finally, the scene of the Effect 2 is shown in Figure 11c with the LED TV playing ocean video and the extended ocean video content is also projected by the projector. We have recruited 20 male and female participants of ages 20–40 for the user study. The viewers were 300 cm from the screen for viewing experience. To exclude fatigue and boredom, the videos were reduced to less than 90 s scenes. We conducted the experiments in about 30 to 40 min following ITU recommendations [3]. Each person saw the set of total 4 videos, one video each for Effect 1 and Effect 2 for the two data sets, i.e., Forest and Ocean.

Figure 11. User study (a) Default background (b) Scene of Effect 1 and (c) Scene of Effect 2.
We have followed 7 Likert-type for the “Effect 1” and “Effect 2” for two videos, i.e., Forest, and Ocean. This Likert-type consists of Neutral (N), Agree (A), Strong Agree (SA), Slightly Agree (SLA), Disagree (D), Strong Disagree (SD), and Slightly Disagree (SLD). A scale of 1 to 7 is assigned from SD to SA, where N has a score of 4. The results of the user study are shown in Figure 12 for Forest and Ocean datasets in percentages. We have used the method of the mean distribution for Likert-type and the formula is given as follows:

\[
M = \frac{\sum (\text{scale} \times \text{response})}{n}
\]

- \(M\) = Mean score
- Scale = 1 to 7 (SD = 1, D = 2, SLD = 3, N = 4, SLA = 5, A = 6, SA = 7)
- Response = number of responses from strong disagree to strong agree
- \(n\) = number of participants.

We have reported the mean for the Forest and Ocean data set for each effect in Table 4. The questions, i.e., Q1 to Q6 are taken from ExtVision [3] which are as follows:
- Q1: Did you enjoy watching?
- Q2: Did you feel sick during viewing?
- Q3: Were you impressed by watching the video?
- Q4: Did you feel present?
- Q5: Was the experience comfortable?
- Q6: Did you feel that the inside and outside of the TV were connected?

We also conducted interviews from the participants, and we observed that there was a difference of opinion among the participants about the proposed system. For the case of Q1 (enjoyment) and Q2 (sickness), different opinions were recorded as follows:
- “I really enjoyed the videos. This effect can be used in cinemas and living room to enhance the visual experiences.” (positive comment),
- “The videos give dizziness or sickness feeling, either it’s going too fast or something else is an issue.” (negative comment).

Similarly, we observed different comments for the case of Q3 (Emotion) and Q4 (Immersion) as follows:
- “Colors are vibrant and stimulating. The illumination seems great making visual experiences immersive and entertaining.” (positive comment),
- “The projector resolution seems low disturbing the feeling of emotion and immersion.” (negative comment).

For Q5 (Comfort), the positive and negative comments are as follows:
- “The effects are enchanting and attractive.” (positive comment),
- “The visual experience causes some kind of flicker movement.” (negative comment).

Consistency (Q6) is considered to be very important consideration for this kind of user study. Negative and positive opinions were recorded for the case of consistency, which are as follows:
- “The contrast as well as the resolution seems enhanced. The bigger screen immerses the user fully in the scene, while the smaller screen focuses on the content and delivering the information.”
- “There is a synchronization problem between the video on TV and projected content as the projected video is slower than the one playing on TV, it’s hard to relate to both at a time.”

The analysis from the interviews of the participants reveals some kind of problems, such as sickness, flicker, inadequate projector resolution, and some synchronization problems. The experience can be enhanced by using Nvidia’s DLSS (Deep Learning Super Sampling) approach to provide high resolution data to the projector to tackle low resolution problem. Flicker problem can be removed by including temporal correlation in the PCE architecture.

![Figure 12. Cont.](a) Forest Effect 1)
Figure 12. Cont.
Questions for user study.

Table 4. The user study results for the Forest and Ocean Videos with Effect 1 and Effect 2.

| Measurement | Dataset       | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 |
|-------------|---------------|----|----|----|----|----|----|
| M           | Forest Effect 1 | 3  | 3  | 4  | 5  | 4  | 2  |
| E           | Forest Effect 2 | 5  | 3  | 5  | 5  | 5  | 5  |
| A           | Ocean Effect 1  | 5  | 2  | 5  | 5  | 5  | 2  |
| N           | Ocean Effect 2  | 5  | 3  | 5  | 5  | 5  | 5  |

Table 4 shows the results of the user study for the questions, Q1 to Q6. For Q1 and the forest data set, the mean score changes from 3 to 5 from Effect 1 to Effect 2, which shows that the Effect 2 was enjoyable. In the case of Ocean data set, the mean score remains same from Effect 1 to Effect 2, which shows that both the effects were equally enjoyed by the user. For Q2 and the forest data set, the mean score remains same from Effect 1 to Effect 2, showing disagreement to the sickness during viewing. In case of Ocean data set, the mean score varied from 2 to 3 from Effect 1 to Effect 2, which depicts the overall disagreement to the sickness. For Q3 and the forest data set, the mean score changes from 4 to 5 from Effect 1 to Effect 2, which depicts that the user was impressed by watching the Effect 2 video. Furthermore, the mean score for the Ocean dataset remains same showing that user was impressed equally watching the videos of both the effects. For Q4 and both the data sets, the mean score remains unchanged from Effect 1 to Effect 2, which indicates that the user felt present watching both the effect videos. For Q5 and the forest data set, the mean score changes from 4 to 5 from Effect 1 to Effect 2, which shows that the Effect 2 experience was comfortable. In case of Ocean dataset, the mean score remains constant indicating that the both the effects were comfortable. For Q6 and the two datasets, the mean score changes from 2 to 5 from Effect 1 to Effect 2, showing that the inside and outside of the TV were connected for Effect 2.

We also conducted interviews from the participants, and we observed that there was a difference of opinion among the participants about the proposed system. For the case of Q1 (enjoyment) and Q2 (sickness), different opinions were recorded as follows:

- “I really enjoyed the videos. This effect can be used in cinemas and living room to enhance the visual experiences.” (positive comment),
- “The videos give dizziness or sickness feeling, either it’s going too fast or something else is an issue.” (negative comment).

Similarly, we observed different comments for the case of Q3 (Emotion) and Q4 (Immersion) as follows:
Colors are vibrant and stimulating. The illumination seems great making visual experiences immersive and entertaining.” (positive comment),
• “The projector resolution seems low disturbing the feeling of emotion and immersion.” (negative comment).
For Q5 (Comfort), the positive and negative comments are as follows:
• “The effects are enchanting and attractive.” (positive comment),
• “The visual experience causes some kind of flicker movement.” (negative comment).
Consistency (Q6) is considered to be very important consideration for this kind of user study. Negative and positive opinions were recorded for the case of consistency, which are as follows:
• “The contrast as well as the resolution seems enhanced. The bigger screen immerses the user fully in the scene, while the smaller screen focuses on the content and delivering the information.”
• “There is a synchronization problem between the video on TV and projected content as the projected video is slower than the one playing on TV, it's hard to relate to both at a time.”

The analysis from the interviews of the participants reveals some kind of problems, such as sickness, flicker, inadequate projector resolution, and some synchronization problems. The experience can be enhanced by using Nvidia’s DLSS (Deep Learning Super Sampling) approach to provide high resolution data to the projector to tackle low resolution problem. Flicker problem can be removed by including temporal correlation in the PCE architecture.

We only included the spatial correlation in the PCE network. The PCE network also causes some flicker problem [3], which can be removed using temporal correlation [25]. We will tackle the problem of temporal correlation for our future work to solve flicker and capture other motions adequately. The PCE network with temporal correlation may easily detect high frequency changes in the Videos.

5. Conclusions
In this research, we proposed the peripheral vision system that provides the immersive visual experiences to the user via automatic extension of the video content using deep learning. The PCE architecture generates the extended content according to the input video and the projector projects the extended content on the wall of the living room around the TV screen, thereby creating immersive visual experiences.

In this research, we used two different datasets, Ocean and Forest datasets for our experiments. We trained the PCE network on the videos of Ocean and Forest datasets and tested the network on the different videos of the similar category. The results of the extended video frames and the Demo videos demonstrate the feasibility of the proposed approach. We have compared our results with the related works [3,25]. We have also performed a user study to evaluate the system qualitatively. The proposed system is expected to pave a new road to the home appliance industry by transforming the living room into the new immersive experience platform.

For further improvements in the peripheral system, we will modify the PCE architecture to handle the temporal features of the video datasets. We may also modify the PCE architecture using Nvidia’s DLSS (Deep Learning Super Sampling) approach [28,29] to improve the performance of the peripheral vision based extended display. Since the last two layers of the decoder (generator) upsamples the image or video frame to the original resolution using nearest neighbor interpolation, the decoder’s architecture can be improved by replacing nearest neighbor interpolation with the DLSS approach. We have also proposed some post processing steps for the output rendered by PCE network, we may get high resolution video from DLSS technology for the projector to project around the TV screen and thus the visual experience can be further enhanced. We will also extend this research for other categories of datasets such as mountains, deserts, and concert. We
also found user interface related reflections [30,31] in the Demo videos and we will solve these problems in our future research. We will also extend our work towards Virtual reality and Augmented Reality in future. The next generation displays will use hybrid strategy of combining projection mapped AR with the array of transparent digital screens and create room-scale interactive surfaces as visualization framework for life-sized digital overlays into physical spaces [32]. Near-Eye Display and Tracking technologies for VR/AR are expected to bring revolution in entertainment, healthcare, communication, and manufacturing industries [33,34]. We will implement our approach for VR/AR based Head mounted displays [25,35,36] in the future.

Author Contributions: Conceptualization, M.A.S. and M.-Y.K.; methodology, M.A.S. and R.U.; software, M.A.S. and R.U.; validation, M.A.S., R.U. and M.-Y.K.; formal analysis, M.A.S. and R.U.; investigation, M.A.S. and M.-Y.K.; resources, R.U.; data curation, M.A.S.; writing—original draft preparation, M.A.S. and M.-Y.K.; writing—review and editing, M.A.S., visualization, M.A.S. and R.U.; supervision, M.-Y.K.; project administration, M.-Y.K.; funding acquisition, M.-Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the BK21 Plus project funded by the Ministry of Education, Korea under Grant 21A20131600111.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jones, B.R.; Benko, H.; Ofek, E.; Wilson, A.D. IllumiRoom: Peripheral projected illusions for interactive experiences. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 869–878.

2. Shirazi, M.A.; Cho, H.; Woo, W. Augmentation of Visual Experiences using Deep Learning for Peripheral Vision based Extended Display Application. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; pp. 1–2.

3. Kimura, N.; Rekimoto, J. ExtVision: Augmentation of Visual Experiences with Generation of Context Images for a Peripheral Vision Using Deep Neural Network. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–10.

4. Turban, L.; Urban, F.; Guillotel, P. Extrafoveal video extension for an immersive viewing experience. IEEE Trans. Vis. Comput. Graph. 2016, 23, 1520–1533. [CrossRef] [PubMed]

5. Aides, A.; Avraham, T.; Schechner, Y.Y. Multiscale ultrawide foveated video extrapolation. In Proceedings of the IEEE International Conference on Computational Photography (ICCP), Pittsburgh, PA, USA, 8–10 April 2011; pp. 1–8.

6. Novy, D.E. Computational Immersive Displays. Ph.D. Thesis, Department of Architecture MIT, Cambridge, MA, USA, 2013.

7. Bimber, O.; Raskar, R. Spatial Augmented Reality: Merging Real and Virtual Worlds; AK Peters Ltd.: New York, NY, USA, 2005; pp. 1–392.

8. Raskar, R.; Welch, G.; Low, K.L.K.; Bandyopadhyay, D. Shader lamps: Animating real objects with image-based illumination. In Proceedings of the Eurographics Workshop on Rendering Techniques, London, UK, 25–27 June 2001; pp. 89–102.

9. Pinhanez, C. The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces. In Proceedings of the ACM UbiComp, Atlanta, GA, USA, 30 September–2 October 2001; pp. 315–331.

10. Raskar, R.; Welch, G.; Cutts, M.; Lake, A.; Stesin, L.; Fuchs, H. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In Proceedings of the ACM SIGGRAPH, Orlando, FL, USA, 19–24 July 1998; pp. 179–188.

11. Wilson, A.; Benko, H.; Izadi, S.; Hilliges, O. Steerable Augmented Reality with the Beamatron. In Proceedings of the ACM UIST, Cambridge, MA, USA, 7–10 October 2012; pp. 413–422.

12. Jones, B.; Sodhi, R.; Campbell, R.; Garnett, G.; Bailey, B.P. Build Your World and Play in It: Interacting with Surface Particles on Complex Objects. In Proceedings of the IEEE ISMAR, Seoul, Korea, 13–16 October 2010; pp. 165–174.

13. Bimber, O.; Coriand, F.; Kleppe, A.; Bruns, E.; Zollmann, S.; Langlotz, T. Superimposing pictorial artwork with projected imagery. IEEE Multimed. 2005, 12, 16–26. [CrossRef]

14. Bimber, O.; Emmerling, A.; Klemmer, T. Embedded entertainment with smart projectors. IEEE Comput. 2005, 38, 48–55. [CrossRef]

15. Flagg, M.; Rehg, J.M. Projector-Guided Painting. In Proceedings of the ACM UIST, Montreux, Switzerland, 15–18 October 2006; pp. 235–244.
16. Baudisch, P.; Good, N.; Stewart, P. Focus plus context screens: Combining display technology with visualization techniques. In Proceedings of the ACM UIST, Orlando, FL, USA, 11–14 November 2001; pp. 31–40.
17. Philips, T.V. Experience Ambilight | Philips. Available online: https://www.philips.co.uk/c-m-so/tv/p/ambilight (accessed on 14 August 2018).
18. Cruz-Neira, C.; Sandin, D.J.; DeFanti, T.A. Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the C.A.V.E. In Proceedings of the SIGGRAPH, Anaheim, CA, USA, 2–6 August 1993; pp. 135–142.
19. Avraham, T.; Schechner, Y.Y. Ultrawide Foveated Video Extrapolation. *IEEE J. Sel. Top. Signal Process.* 2011, 5, 321–334. [CrossRef]
20. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* 2017, 36, 107–120. [CrossRef]
21. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
22. Van Noord, N.; Postma, E. Light-weight pixel context encoders for image inpainting. Computing Research Repository. arXiv 2018, arXiv:1801.05585. Available online: https://arxiv.org/abs/1801.05585 (accessed on 1 January 2021).
23. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
24. Phillipi. philippi/pix2pix. December 2017. Available online: https://github.com/philippi/pix2pix (accessed on 17 August 2019).
25. Kimura, N.; Kono, M.; Rekimoto, J. Deep dive: Deep-neural-network-based video extension for immersive head-mounted display experiences. In Proceedings of the 8th ACM PerDis ’19, Palermo, Italy, 12–14 June 2019; pp. 1–7.
26. Naofumi, A.; Kasai, S.; Hayashi, M.; Aoki, Y. 360-Degree Image Completion by Two-Stage Conditional Gans. In Proceedings of the IEEE ICIP, Taipei, Taiwan, 22–25 September 2019; pp. 4704–4708.
27. Kimura, N.; Kono, M.; Rekimoto, J. Using deep-neural-network to extend videos for head-mounted display experiences. In Proceedings of the ACM VRST, Tokyo, Japan, 28 November–1 December 2018; pp. 1–2.
28. Kölüş, Ç.; Başçıftçı, F. The Future Effects of Virtual Reality Games. Computer Engineering: 47. Available online: https://academicworks.livredelyon.com/cgi/viewcontent.cgi?article=1014&context=engineer_sci#page=57 (accessed on 23 April 2021).
29. Nvidia. NVIDIA DLSS 2.0: A Big Leap in AI Rendering. 2020. Available online: https://www.nvidia.com/en-us/geforce/news/nvidia-dlss-2-0-a-big-leap-in-ai-rendering/ (accessed on 23 April 2021).
30. Gruenefeld, U.; El Ali, A.; Boll, S.; Heuten, W. Beyond Halo and Wedge: Visualizing out-of-view objects on head-mounted virtual and augmented reality devices. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, Barcelona, Spain, 3–6 September 2018.
31. Biener, V.; Schneider, D.; Gesslein, T.; Otte, A.; Kuth, B.; Kristensson, P.O.; Ofek, E.; Pahud, M.; Grubert, J. Breaking the Screen: Interaction Across Touchscreen Boundaries in Virtual Reality for Mobile Knowledge Workers. *IEEE Trans. Vis. Comput. Graph.* 2020, 26, 3490–3502. [CrossRef] [PubMed]
32. Almagro, M. Key Trends in Immersive Display Technologies and Experiences. Available online: https://www.soundandcommunications.com/key-trends-immersive-display-technologies-experiences/ (accessed on 25 May 2020).
33. Koulieris, G.A.; Aksit, K.; Stengel, M.; Mantiuk, R.K.; Mania, K.; Richarid, C. Near-Eye Display and Tracking Technologies for Virtual and Augmented Reality. *Comput. Graph. Forum* 2019, 38, 493–519. [CrossRef]
34. Aksit, K.; Chakravarthula, P.; Rathinavel, K.; Jeong, Y.; Albert, R.; Fuchs, H.; Luebke, D. Manufacturing Application-Driven Foveated Near-Eye Displays. *IEEE Trans. Vis. Comput. Graph.* 2019, 25, 1928–1939. [CrossRef] [PubMed]
35. Friston, S.; Ritschel, T.; Steed, A. Perceptual rasterization for head-mounted display image synthesis. *ACM Trans. Graph.* 2019, 38, 1–14. [CrossRef]
36. Tursun, O.T.; Arabadzhiyska-Koleva, E.; Wernikowski, M.; Mantiuk, R.; Seidel, H.-P.; Myszkowski, K.; Didyk, P. Luminance-contrast-aware foveated rendering. *ACM Trans. Graph.* 2019, 38, 1–14. [CrossRef]