Enhanced Perspective Generation by Consensus of NeX neural models

Marcos Sergio Pacheco dos Santos Lima Junior
Lenguajes y Ciencias de la Computación
University of Málaga
Curitiba, Brazil
marcos.sergio.pacheco@gmail.com

Jose David Fernández Rodríguez
Lenguajes y Ciencias de la Computación
University of Málaga
Málaga, Spain
josedavid@uma.es

Juan Miguel Ortiz de Lazcano-Lobato
Lenguajes y Ciencias de la Computación
University of Málaga
Málaga, Spain
jmortiz@lcc.uma.es

Ezequiel López-Rubio
Lenguajes y Ciencias de la Computación
University of Málaga
Málaga, Spain
ezeqlr@lcc.uma.es

Enrique Domínguez
Lenguajes y Ciencias de la Computación
University of Málaga
Málaga, Spain
enriqued@lcc.uma.es

***Abstract*—Neural rendering is a relatively new field of research that aims to produce high quality perspectives of a 3D scene from a reduced set of sample images. This is done with the help of deep artificial neural networks that model the geometry and color characteristics of the scene. The NeX model relies on neural basis expansion to yield accurate results with a lower computational load than the previous NeRF model. In this work, a procedure is proposed to further enhance the quality of the perspectives generated by NeX. Our proposal is based on the combination of the outputs of several NeX models by a consensus mechanism. The approach is compared to the original NeX for a wide range of scenes. It is found that our method significantly outperforms the original procedure, both in quantitative and qualitative terms.**

***Index Terms*—deep learning, convolutional neural networks, neural rendering, consensus model**

This work is partially supported by the Ministry of Science, Innovation and Universities of Spain under grant RTI2018-094645-B-I00, project name Automated detection with low-cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA18-FEDERJA-084, project name Detection of anomalous behavior agents by deep learning in low-cost video surveillance intelligent systems. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA20-FEDERJA-108, project name Detection, characterization and prognosis value of the non-obstructive coronary disease with deep learning. All of them include funds from the European Regional Development Fund (ERDF). It is also partially supported by the University of Malaga (Spain) under grants B1-2019_01, project name Anomaly detection on roads by moving cameras, and B1-2019_02, project name Self-Organizing Neural Systems for Non-Stationary Environments. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPUs. The authors also thankfully acknowledge the grant of the Universidad de Málaga and the Instituto de Investigación Biomédica de Málaga - IBIMA.

I. Introduction

Computer vision applications such as virtual reality and augmented reality have been increasingly used due to the technological development of devices such as smartphones and tablet computers, which have at least a built-in single or dual camera. In these types of applications it is essential to address the problem of photorealistic view synthesis in real time, which involves a procedure to build a visual representation of the scene from a finite set of input sparse images and a method to generate new images that correspond to views of the scene different from those with which the images of the original set were taken.

Novel view synthesis is not an easy task that has to cope with object occlusion, thin structures and complex surface reflectance properties such as the rainbow reflections on a CD and the refraction through glass objects. Most of the works presented in recent years make use of deep neural networks that contribute to implicitly represent the scene once they are properly trained. In [1] a multi-layer perceptron learns the scene 3D properties for each spatial location. On the other hand, Neural Radiance Fields (NeRF) [2] and its extensions [3] [4] train a fully connected neural network so that it receives the spatial location and the viewing direction of each pixel and returns the corresponding RGB color and volume density. Although they are able to model view-dependent effects, the number of input images needed for training and the computational cost required for optimizing their neural model parameters make them not suitable for applications that require real time performance.

Another very common approach to the task is based on Multiplane Images (MPI), where a set of parallel semi-transparent planes placed at different depths from the same reference viewpoint is used to represent the scene. The length of the set of input data ranges from a single image [5], where the main challenge is the lack of information to infer 3D geometrical properties, especially those of occluded regions, to multiple input photos [6] [7] [8]. In [8] the new scene poses are produced by blending the MPIs generated by a Convolutional Neural Network (CNN) for each input view. DeepView [6] builds a MPI in a few iterations thanks to a CNN that learns the gradient updates, which allows the method to avoid the overfitting that would occur when predicting the gradient directly. Finally, [7][17] combines several MPIs to model scenes whose appearance varies over time. However, the approach of modeling the view-dependent effects as a blend of multiple view-independent MPIs is limited and it does not always work appropriately.

One of the most recent approaches is NeX [9], which attempts to overcome the MPI-based model difficulty in representing non-Lambertian surfaces. For that purpose, the color of each pixel is considered to be dependent on the viewing direction and is approximated as a linear combination of a fixed number of spherical basis functions learned from neural networks. Furthermore, since implicit representation of the scene by means of only neural networks tends to blur the images and may lose fine details, NeX proposes a hybrid parameter modeling strategy in which some reflectance parameters such as the one corresponding to the base color are optimized separately and saved explicitly for each one of the MPI planes. The performance of NeX, as well as other view synthesis methods which are based on scene implicit neural network representation, relies on the features learned by its core neural system after training. If the neural network is not able to adequately learn the scene features, specially of those regions that are not present in the input images, then flickering effects or even artifacts may appear in the rendered output.

Neural network ensemble is a learning paradigm where several neural networks are trained for the same task. The combination of predictions of the distinct neural networks is expected to improve the overall generalization ability of the neural network system [10]. It has already been effectively applied to areas as diverse as face recognition [11], medical diagnosis [12] [13], and seismic signals classification [14] and fault detection [15].

The proposal that is presented in this paper consists in ensembling a set of NeX neural networks with the aim of achieving better rendered images. A moderate number of NeX networks are trained on the same input images but with different pseudorandom seeds. When an image from a new pose is generated, the color and transparency of each one of its pixels are obtained by consensus. As a result of that, the artifacts and other undesired visual effects in the output image that would be produced by individual NeX networks are less likely to appear due to the compensation and correction made by the other ensemble components.

The rest of the paper is organized as follows: Section II describes the methodology. Section III is devoted to the experiments that have been carried out and the analysis of their results, Finally, the main conclusions are summarized in Section IV.

II. Methodology

Let us note $\mathcal{T} = \{(\mathbf{x}_1, \mathbf{Y}_1), ..., (\mathbf{x}_M, \mathbf{Y}_M)\}$ a training set with M patterns, each composed by an input pose $\mathbf{x}$ and an output image (perspective) $\mathbf{Y}$. Then we may note $F_{\mathcal{T},i}$ the NeX network trained with the set $\mathcal{T}$ using a pseudorandom seed $i \in \mathbb{N}$. It must be noted that, even though the training set is the same, we obtain different networks $F_{\mathcal{T},i}$ by varying i.

We propose to build an ensemble of N NeX networks $F_{\mathcal{T},i}$, for $i \in \{1, ..., N\}$. After the networks are trained, given a test input pose $\mathbf{x}$, a consensus output perspective can be obtained as follows:

$$\mathbf{Y} = \varphi_{G,N}(\mathbf{x}) = G(\{F_{\mathcal{T},1}(\mathbf{x}), ..., F_{\mathcal{T},N}(\mathbf{x})\}) \qquad (1)$$

where G is a suitable aggregation function. In this work, we have considered two alternatives: $G = \text{mean}$ and $G = \text{median}$.

In the case of $G = \text{mean}$, by the law of large numbers we know that:

$$\lim_{N \to \infty} \varphi_{\text{mean},N}(\mathbf{x}) = E_{i \in \mathbb{N}}[F_{\mathcal{T},i}(\mathbf{x})] \qquad (2)$$

where E stands for the mathematical expectation operator. Moreover,

$$\text{var}[\varphi_{\text{mean},N}(\mathbf{x})] = \frac{1}{N}\text{var}_{i \in \mathbb{N}}[F_{\mathcal{T},i}(\mathbf{x})] \qquad (3)$$

where var stands for variance, computed separately for each RGB color component of each pixel of the perspective. Also, for $G = \text{median}$, it is known that the distribution of the sample median is asymptotically normal with variance proportional to $\frac{1}{N}$ [16], [17].

The above results imply that, provided that the variance of $F_{\mathcal{T},i}(\mathbf{x})$ is low, we can expect that the value of $\varphi_{G,N}(\mathbf{x})$ will converge for relatively small values of N. This reduces the computational load of implementing (1).

III. Experiments

This work evaluated quantitatively and qualitatively the proposed method using an arrangement similar to the described by NeX paper [9], targeting the achievement of experimental results that could prove the robustness of our strategy.

A. Test setup

The Shiny dataset was chosen so that the same scenes used in the NeX article could be analyzed in our work. This dataset is composed of 8 scenes that were conceived to test the network under challenging view-dependent effects such as reflections, thin-film interference, refraction through non-planar glassware and magnifying glass [9].

The setup was arranged so that for each scene, ensembles from $N = 2$ up to $N = 20$ were computed where both scenarios with G = mean and G = median were considered. It was accomplished firstly by calculating the consensuses of validation images for each ensemble using both aggregation functions and secondly by measuring the metrics PSNR, SSIM [18] and LPIPS [19] of the consensus images. This process was repeated for each 25 epochs during the training of CD and Lab scenes and for each 10 epochs for the remaining scenes. The results of our approach was compared to the original NeX model, which has been noted $N = 1$ as it consists of a single NeX network with no consensus.

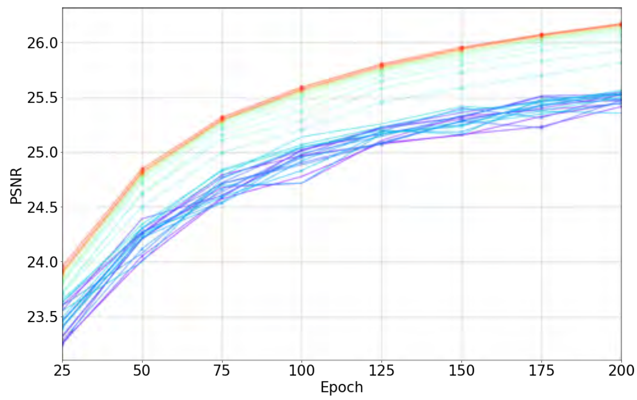
When training several NeX networks for a given scene, the resulting performance will not be exactly the same, but a given degree of variability is expected. In our case, the variability is originated from two main sources: the sequence of pseudo-random numbers used internally in the training process, and the differences between training and validation datasets. For the purposes of building our consensuses, we want to allow the former while suppressing the latter. Accordingly, modifications were done to the NeX source code to ensure a different and repeatable pseudo-random sequence for each different training (specifying the seed to the sequence), while also having a consistent 50%-50% split of the scene images into training and validation subsets across all trainings. While it is usual in deep learning for the validation dataset to be substantially smaller than the training one, we committed to a 50%-50% split because most NeX scenes have just a few tens of images, and we are measuring relatively small differences across very similar images, so having a relatively large validation subset helps to strengthen the case for our proposed technique.

Our experiments were conducted at lower resolutions than the NeX article, with the images of each scene resized to a common width of 400 pixels, modifying the image height to keep the aspect ratio of the images. To be more specific, the images of the Tools, Crest, Seasoning, Food, Giants and Pasta scenes were resized to 400x300 pixels, while the images of the Lab and CD scenes were resized to 400x225 pixels.

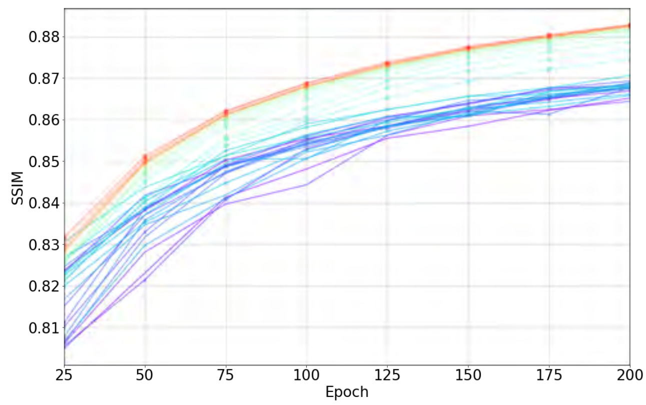
B. Quantitative results

Each of the scenes from Shiny dataset was trained with 20 different seeds for 200 epochs and it was possible to observe that in general there was a clear improvement by using our strategy over the original NeX model. In Fig. 1 and Fig. 2, we show an example of the performance of the ensembles in CD scene from $N = 2$ up to $N = 20$ for respectively G = mean and G = median. In both figures the performance of the metrics of the ensembles were plot alongside with the 20 networks trained with individual seeds ($F_{\mathcal{T}_{\text{CD}},1}, F_{\mathcal{T}_{\text{CD}},2}, ..., F_{\mathcal{T}_{\text{CD}},20}$). All the metrics present a consistent improvement of the consensuses over the individual networks for both aggregation functions specially for consensuses of higher N, although in some cases for specific seeds the individual networks show better performance in some isolated epochs.

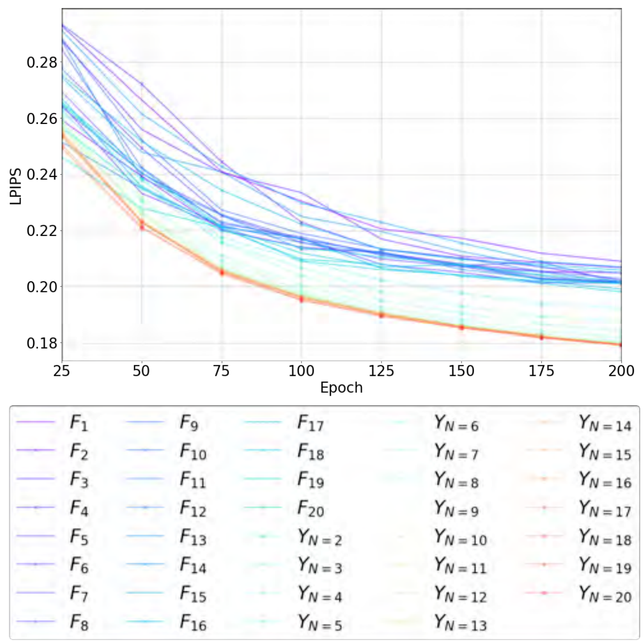
It is also noticeable that the performance improves for when N increases while the rate of improvement decreases. This statement is more clear when looking to Fig. 3, where the improvement of the metrics in relation to the original NeX



(a) PSNR, G = mean

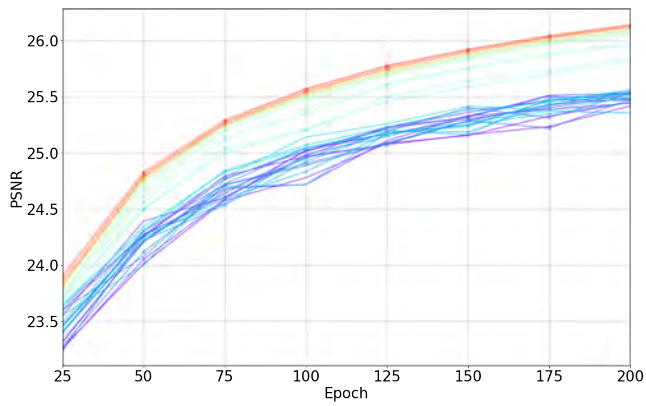


(b) SSIM, G = mean

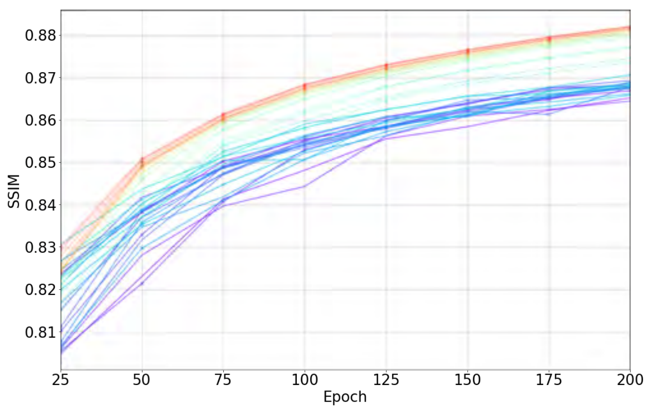


(c) LPIPS, G = mean

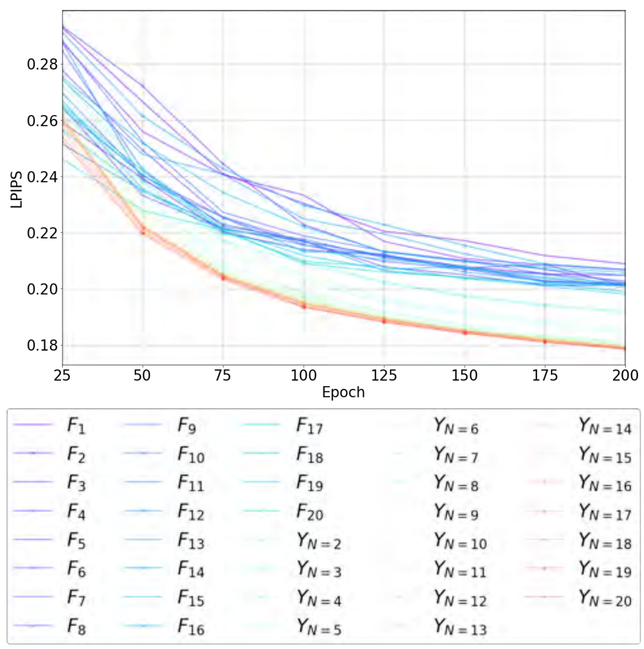
Fig. 1: Evolution of the metrics in CD scene with G = mean: (a) PSNR; (b) SSIM; (c) LPIPS



(a) PSNR, G = median



(b) SSIM, G = median



(c) LPIPS, G = median

Fig. 2: Evolution of the metrics in CD scene with G = median: (a) PSNR; (b) SSIM; (c) LPIPS

TABLE I: N at which the best performance was achieved in epoch 200

	Mean			Median		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
CD	20	20	19	20	20	20
Crest	16	20	6	16	20	18
Food	20	20	5	20	20	8
Giants	16	16	20	16	19	20
Lab	20	15	14	20	20	14
Pasta	20	20	20	20	20	20
Seasoning	20	20	4	20	20	5
Tools	17	17	6	17	17	10

TABLE II: Best metric achieved by training until epoch 200

	Mean			Median		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
CD	**26.174**	**0.883**	0.179	26.143	0.882	**0.179**
Crest	**24.895**	**0.837**	0.161	24.867	0.837	**0.160**
Food	**22.328**	**0.789**	0.217	22.299	0.788	**0.216**
Giants	**27.984**	**0.934**	0.111	27.954	0.933	**0.110**
Lab	**27.233**	**0.914**	0.141	27.195	0.913	**0.140**
Pasta	**23.655**	**0.899**	0.125	23.619	0.898	**0.125**
Seasoning	**30.216**	**0.961**	0.151	30.184	0.960	**0.150**
Tools	**25.413**	**0.925**	0.103	25.346	0.924	**0.101**
Mean	**25.987**	**0.893**	0.149	25.951	0.892	**0.147**

model ($N = 1$) is presented in percentage according to the following:

$$R_{METRIC} = \frac{METRIC}{METRIC_{N=1}} \tag{4}$$

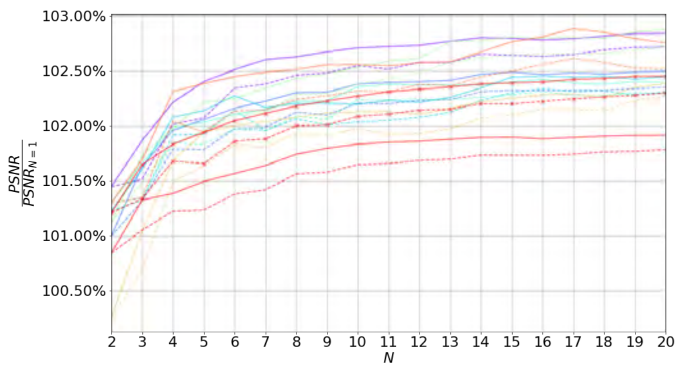
Where R is the relative measurement of a given metric in relation to the original NeX model trained until the same epoch. Thus, $R > 100\%$ means improvement over the original NeX model for PSNR and SSIM, while for LPIPS improvement is achieved when $R < 100\%$.

Figure 3 and Table I show that in epoch 200 all the consensuses behave better when $N > 1$, were it is possible to see a clear tendency of improvement as N increases for PSNR and SSIM, while in some cases for LPIPS the optimum performance is achieved at a lower value of N.

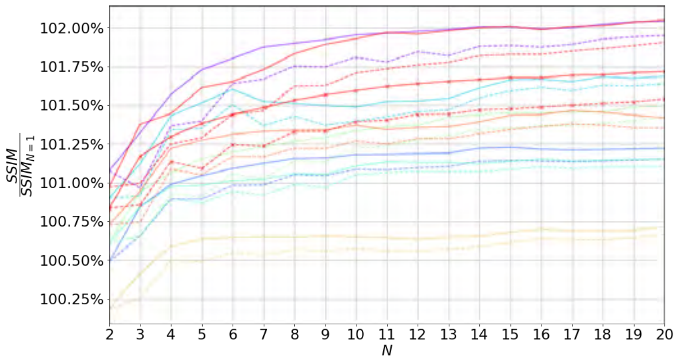
Regarding the aggregation functions, a more detailed comparison can be accomplished by analyzing Table II, where G = mean achieves a better performance for PSNR and SSIM in all cases. In opposition, G = median is more efficient when considering LPIPS as a metric of reference.

C. Qualitative results

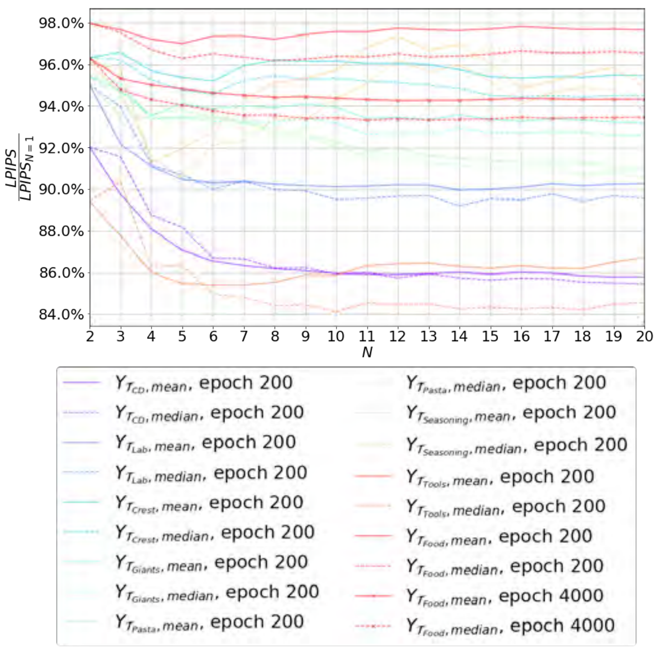
In addition to the metrics, other comparisons might be done using some of the images outputted by the original NeX network and the ensembles with different values of N so that the visual impact of our strategy become more clear. The first effect our strategy produced was the attenuation of erroneous renderizations generated in some scenes, specially in some corners and regions close to the borders of the images.



(a) R_{PSNR}



(b) R_{SSIM}



(c) R_{LPIPS}

Fig. 3: Improvement of the metrics in relation to the original NeX model ($N = 1$): (a) PSNR; (b) SSIM; (c) LPIPS.

However, in spite of the improvement, in most cases these visual artifacts were not completely fixed by our strategy. An example of this behavior is shown in Fig. 4, where the imperfections of the renderization of the table cloth are improved by the consensuses with the cost of adding some blur to that region of the image.

Another impact the consensus has over the original NeX model was observed in metallic surfaces, where it could improve the sharpness of contours of objects. This effect is visible in the spoon on Fig. 4, where it is also noticeable an improvement in the detail of the frame present along the middle of the cable of the spoon. This frame is almost not perceptible in $F_{\mathcal{T}_{\text{Food}},1}$ image.

A third characteristic was noted in some scenes, where both mean and median strategies were able to reduce the noise on some surfaces. Fig. 5 shows an example of such effect, where a reduction of the perceived noise may be observed as N increases. A similar effect is noted in the magnifying glass present in Fig. 6. Although, in this case, neither the original NeX nor any of the consensuses were able to accurately reproduce the optical behavior of the lens.

Nevertheless, the visual effects caused by our strategy were not always effortlessly perceived by naked eye. In some scenes, even though there was a clear improvement in the metrics, they were difficult to be detected.

D. Other experiments

After investigating the effect of our strategy in all scenes until epoch 200, the Food scene was chosen to be trained further to check if the usage of NeX ensembles would be beneficial even when the training does not improve anymore. For this experiment, the number of epochs was chosen to be 4000, the same number used in NeX original work. Fig. 7 and Fig. 8 show the results of this analysis, where it is possible to see that around epoch 1400 the networks achieved the best performance and kept it almost constant until epoch 4000. The abrupt modification in the performance of the networks in epoch 1400 happens because in epoch 1333 the learning rate is multiplied by a factor of 0.1 [9] and as our resolution is 100 epochs, it is only perceptible in epoch 1400. In the end the consensuses proved to outperform the individual networks during the entire training.

The comparison of epochs 200 and 4000 of Food scene in Fig. 3 provides a deeper understanding of the effect of N in this experiment. For PSNR and SSIM the best choice continued to be $N = 20$ in epoch 4000. On the other hand, the best number of networks in the consensus when considering LPIPS showed to be different than for 200 epochs, passing from $N = 5$ to $N = 12$ for $G = \text{mean}$ and from $N = 8$ to $N = 11$ for $G = \text{median}$.

IV. Conclusions

In this work, a procedure to enhance the quality of the perspectives generated by the NeX neural rendering model has been proposed. The procedure is based on the training of

(a) $F_{\mathcal{T}_{\text{Food}},1}$ (b) $Y_{\mathcal{T}_{\text{Food}},\,\text{mean},\,N=20}$ (c) $Y_{\mathcal{T}_{\text{Food}},\,\text{median},\,N=20}$ (d) Ground truth

Fig. 4: Effect of our strategy in Food scene where the consensus improved the details of the spoon and the texture of the table cloth: (a) Original NeX output; (b) Mean consensus of 20 networks; (c) Median consensus of 20 networks; (d) Ground truth image.

several NeX models, whose outputs are subsequently combined by a suitable consensus mechanism. Two consensus mechanisms have been developed, namely the mean and the median consensus. Computational experiments have been conducted on a variety of 3D scenes to compare the perspective generation performance of our proposal, as compared to the application of a single NeX model. Our approach yields a consistently better quality, both in quantitative and qualitative terms, in all experiments. Erroneous renderizations produced by the NeX network are attenuated, in particular in specific corners and regions close to the borders of the generated images. Noise is reduced, and the sharpness of metallic surfaces is also improved. These results prove the relevance and utility of our proposal.

REFERENCES

[1] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," in *Advances in Neural Information Processing Systems*, 2019.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 405–421.

[3] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," *NeurIPS*, 2020.

[4] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *CVPR*, 2021.

[5] R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, "Deepview: View synthesis with learned gradient descent," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2362–2371.

[7] Z. Li, W. Xian, A. Davis, and N. Snavely, "Crowdsampling the plenoptic function," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 178–196.

[8] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Trans. Graph.*, vol. 38, no. 4, jul 2019. [Online]. Available: https://doi.org/10.1145/3306346.3322980

[9] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, and S. Suwajanakorn, "Nex: Real-time view synthesis with neural basis expansion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[11] V. Mohanraj, S. Sibi Chakkaravarthy, and V. Vaidehi, "Ensemble of convolutional neural networks for face recognition," in *Recent Developments in Machine Learning and Data Analytics*, J. Kalita, V. E. Balas, S. Borah, and R. Pradhan, Eds. Singapore: Springer Singapore, 2019, pp. 467–477.

[12] K. Thurnhofer-Hemsi, E. López-Rubio, E. Domínguez, and D. A. Elizondo, "Skin lesion classification by ensembles of deep convolutional networks and regularly spaced shifting," *IEEE Access*, vol. 9, pp. 112 193–112 205, 2021.

[13] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S.-F. Chen, "Lung cancer cell identification based on artificial neural network ensembles," *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25–36, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365701000094X

[14] Y. Shimshoni and N. Intrator, "Classification of seismic signals by integrating ensembles of neural networks," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1194–1201, 1998.

[15] Z. Wang, B. Li, N. Liu, B. Wu, and X. Zhu, "Distilling knowledge from an ensemble of convolutional neural networks for seismic fault detection," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[16] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ: Wiley, 2003.

[17] E. López-Rubio, "Superresolution from a single noisy image by the median filter transform," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 82–115, 2016.

[18] Z. e. a. Wang, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

(a) Ground truth (b) $Y_{\mathcal{T}_{\text{CD}},\,\text{mean},\,N=5}$ (c) $Y_{\mathcal{T}_{\text{CD}},\,\text{mean},\,N=10}$ (d) $Y_{\mathcal{T}_{\text{CD}},\,\text{mean},\,N=20}$

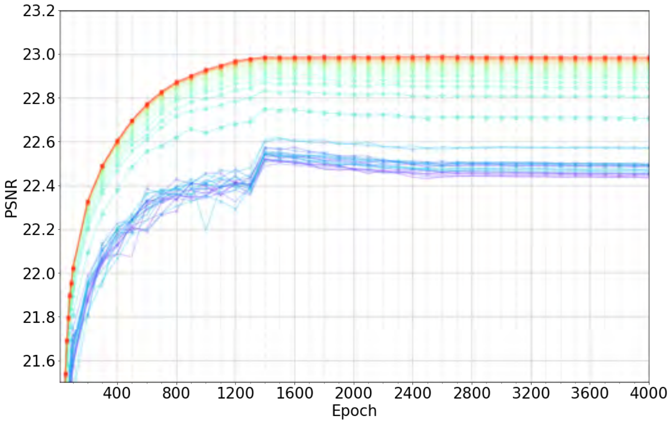
(e) $F_{\mathcal{T}_{\text{CD}},1}$ (f) $Y_{\mathcal{T}_{\text{CD}},\,\text{median},\,N=5}$ (g) $Y_{\mathcal{T}_{\text{CD}},\,\text{median},\,N=10}$ (h) $Y_{\mathcal{T}_{\text{CD}},\,\text{median},\,N=20}$

Fig. 5: Comparison of details in CD scene where the noise reduction is emphasized in yellow and in red an improvement of the shape of the CD is noted for G = median as N increases: (a) Ground truth image; (b) Mean consensus of 5 networks; (c) Mean consensus of 10 networks; (d) Mean consensus of 20 networks; (e) Original NeX output; (f) Median consensus of 5 networks; (g) Median consensus of 10 networks; (h) Median consensus of 20 networks.

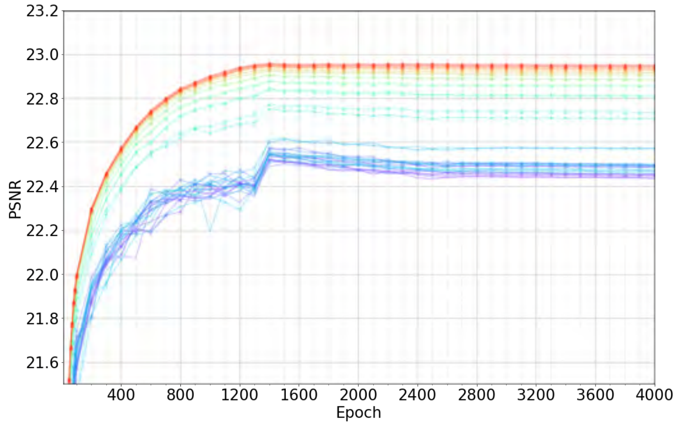
(a) Ground truth (b) $Y_{\mathcal{T}_{\text{Tools}},\,\text{mean},\,N=5}$ (c) $Y_{\mathcal{T}_{\text{Tools}},\,\text{mean},\,N=10}$ (d) $Y_{\mathcal{T}_{\text{Tools}},\,\text{mean},\,N=20}$

(e) $F_{\mathcal{T}_{\text{Tools}},1}$ (f) $Y_{\mathcal{T}_{\text{Tools}},\,\text{median},\,N=5}$ (g) $Y_{\mathcal{T}_{\text{Tools}},\,\text{median},\,N=10}$ (h) $Y_{\mathcal{T}_{\text{Tools}},\,\text{median},\,N=20}$

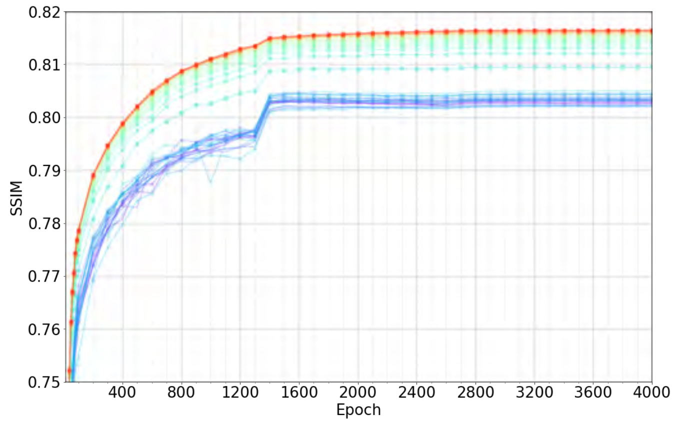
Fig. 6: Zoom in the improvement of the magnifying glass with higher N in Tools scene. Even though the image noise is reduced, the ground truth remains different from all the images presented. (a) Ground truth image; (b) Mean consensus of 5 networks; (c) Mean consensus of 10 networks; (d) Mean consensus of 20 networks; (e) Original NeX output; (f) Median consensus of 5 networks; (g) Median consensus of 10 networks; (h) Median consensus of 20 networks.



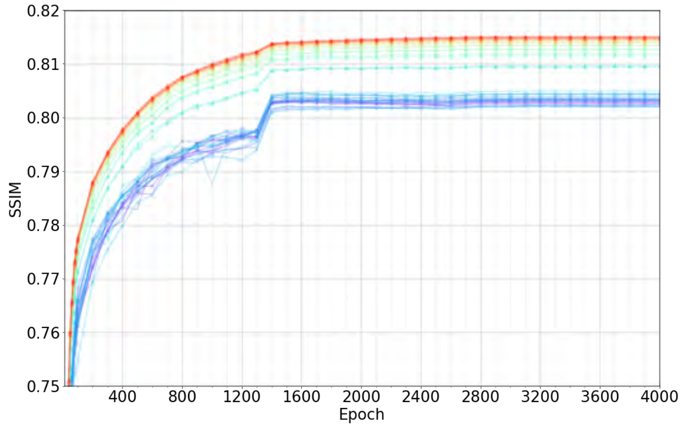
(a) PSNR, $G =$ mean



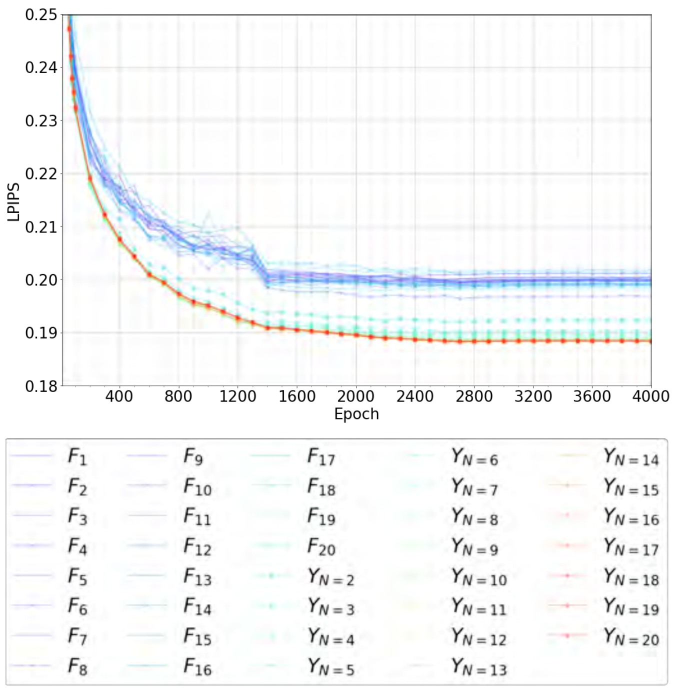
(a) PSNR, $G =$ median



(b) SSIM, $G =$ mean

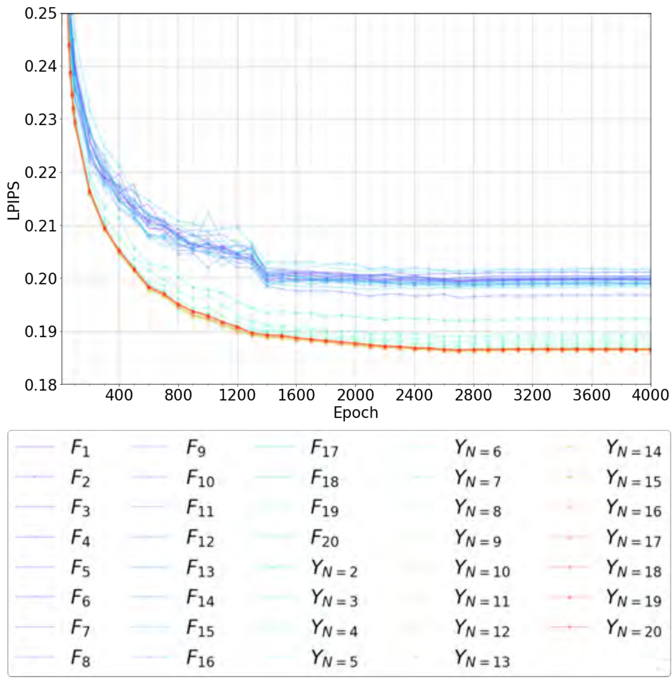


(b) SSIM, $G =$ median



(c) LPIPS, $G =$ mean

Fig. 7: Evolution of the metrics in Food scene until epoch 4000 with $G =$ mean: (a) PSNR; (b) PSNR; (c) SSIM.



(c) LPIPS, $G =$ median

Fig. 8: Evolution of the metrics in Food scene until epoch 4000 with $G =$ median: (a) PSNR; (b) PSNR; (c) SSIM.