

- (51) International Patent Classification:
G06T 15/20 (2011.01) *G06N 3/02* (2006.01)
- (21) International Application Number:
PCT/US2021/065595
- (22) International Filing Date:
30 December 2021 (30.12.2021)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
202110014129.8 06 January 2021 (06.01.2021) CN
- (71) Applicant: **MICROSOFT TECHNOLOGY LICENSING, LLC** [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).
- (72) Inventors: **WEI, Yanan**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **ZHANG, Zheng**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **LIANG, Yaobo**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

ZHANG, Xiao; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **TANG, Jie**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). **XU, Tao**; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(74) Agent: **CHATTERJEE, Aaron C.** et al.; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: END-TO-END 3D SCENE RECONSTRUCTION AND IMAGE PROJECTION

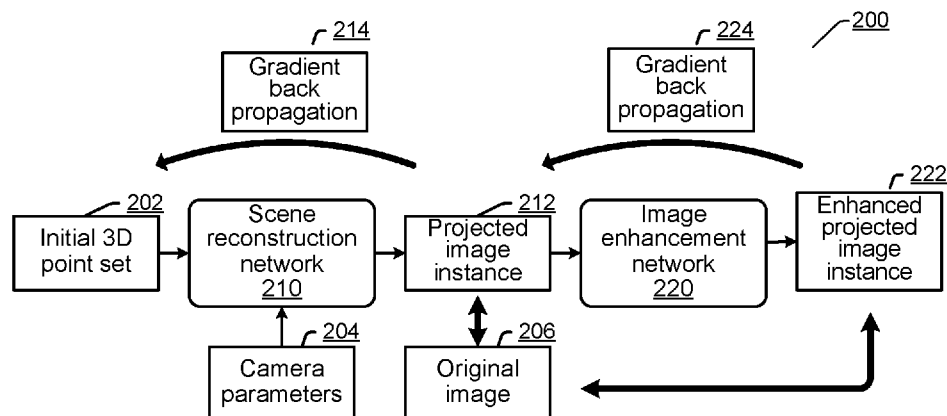


FIG 2

(57) Abstract: The present disclosure provides methods and apparatuses for end-to-end three-dimension (3D) scene reconstruction and image projection. A set of original images shot by a set of cameras may be obtained. A 3D scene may be reconstructed based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network. A target viewpoint may be obtained. A projected image corresponding to the target viewpoint may be generated with the 3D scene through the scene reconstruction network. The projected image may be updated to an enhanced projected image through the image enhancement network.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

END-TO-END 3D SCENE RECONSTRUCTION AND IMAGE PROJECTION

BACKGROUND

[0001] 3D scene reconstruction may refer to the process of establishing a 3D
5 mathematical model, suitable for computer representing and processing, for a scene in the
objective world, which is a key technique for establishing virtual reality that expresses the
objective world in a computer. For example, in an image-based 3D scene reconstruction,
3D information may be reconstructed and a 3D scene may be reconstructed, with a
plurality of scene images shot from different angles and through a predetermined
10 algorithm. The 3D scene reconstruction has been widely applied for, e.g., industrial
measurement, architectural design, medical imaging, 3D animation games, virtual reality
(VR), etc.

SUMMARY

[0002] This Summary is provided to introduce a selection of concepts that are further
15 described below in the Detailed Description. It is not intended to identify key features or
essential features of the claimed subject matter, nor is it intended to be used to limit the
scope of the claimed subject matter.

[0003] Embodiments of the present disclosure propose methods and apparatuses for
end-to-end 3D scene reconstruction and image projection. A set of original images shot by
20 a set of cameras may be obtained. A 3D scene may be reconstructed based at least on the
set of original images and camera parameters of the set of cameras, through joint
optimization of a scene reconstruction network and an image enhancement network. A
target viewpoint may be obtained. A projected image corresponding to the target
viewpoint may be generated with the 3D scene through the scene reconstruction network.
25 The projected image may be updated to an enhanced projected image through the image
enhancement network.

[0004] It should be noted that the above one or more aspects comprise the features
hereinafter fully described and particularly pointed out in the claims. The following
description and the drawings set forth in detail certain illustrative features of the one or
30 more aspects. These features are only indicative of the various ways in which the
principles of various aspects may be employed, and this disclosure is intended to include
all such aspects and their equivalents.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] The disclosed aspects will hereinafter be described in conjunction with the

appended drawings that are provided to illustrate and not to limit the disclosed aspects.

[0006] FIG.1 illustrates an exemplary process of end-to-end 3D scene reconstruction and image projection according to an embodiment.

5 [0007] FIG.2 illustrates an exemplary process of performing joint optimization to a scene reconstruction network and an image enhancement network according to an embodiment.

[0008] FIG.3 illustrates an exemplary implementation of a scene reconstruction network according to an embodiment.

10 [0009] FIG.4 illustrates an exemplary implementation of a scene reconstruction network according to an embodiment.

[0010] FIG.5 illustrates an exemplary implementation of an image enhancement network according to an embodiment.

[0011] FIG.6 illustrates a flowchart of an exemplary method for end-to-end 3D scene reconstruction and image projection according to an embodiment.

15 [0012] FIG.7 illustrates an exemplary apparatus for end-to-end 3D scene reconstruction and image projection according to an embodiment.

[0013] FIG.8 illustrates an exemplary apparatus for end-to-end 3D scene reconstruction and image projection according to an embodiment.

DETAILED DESCRIPTION

20 [0014] The present disclosure will now be discussed with reference to several example implementations. It is to be understood that these implementations are discussed only for enabling those skilled in the art to better understand and thus implement the embodiments of the present disclosure, rather than suggesting any limitations on the scope of the present disclosure.

25 [0015] There are some existing techniques for performing image-based 3D scene reconstruction. These techniques need to collect a plurality of original images shot by a plurality of pre-deployed cameras firstly, and then reconstruct a 3D scene with these original images. Generally, in order to obtain a better 3D scene reconstruction effect, a large number of cameras need to be deployed, and these cameras may adopt expensive 3D
30 cameras, e.g., VR cameras, etc. In some applications, the reconstructed 3D scene may be further used for implementing image projection, so as to present a projected image associated with a specific space viewpoint to the user. Herein, a viewpoint or space viewpoint may refer to a point in space, which has attributes such as a specific position, a specific direction, etc. For example, when the user is watching at a specific viewpoint in

the virtual space corresponding to the 3D scene, a projected image corresponding to the viewpoint may be projected with the reconstructed 3D scene, and the projected image may be presented to the user. Therefore, if the user selects different viewpoints in the 3D scene, corresponding projected images may be presented to the user at these viewpoints, respectively. Accordingly, an experience effect that the user feels like being in a virtual space may be achieved. However, the viewpoints that the user can select are often restricted by the cameras that are pre-deployed during the image collecting process. For example, the user can only select a viewpoint corresponding to each pre-deployed camera, but cannot watch at other viewpoints in the 3D scene. Moreover, in these existing techniques, 3D scene reconstruction and image projection are two independent processes that are trained separately, and these two processes are only combined in the application phase.

[0016] Embodiments of the present disclosure propose end-to-end 3D scene reconstruction and image projection. The embodiments of the present disclosure also propose an image enhancement mechanism for projected images to generate enhanced projected images with higher quality. In the embodiments of the present disclosure, a 3D scene reconstruction process, an image projection process, and an image enhancement process may be concatenated together and optimized or trained together. For example, end-to-end joint optimization may be performed to a scene reconstruction network used for the 3D scene reconstruction and an image enhancement network used for the image enhancement mechanism. Through the joint optimization, the scene reconstruction network and the image enhancement network may be coupled more closely and effectively, and may adapt to each other more accurately, thereby generating more realistic images and improving the user experience accordingly.

[0017] The embodiments of the present disclosure may achieve high-quality 3D scene reconstruction for the whole space associated with a 3D scene, through at least the end-to-end joint optimization of the scene reconstruction network and the image enhancement network. Through modeling the whole space, high-quality image projection may be performed at any viewpoint in the 3D scene, without being restricted by the cameras pre-deployed during the image collecting process. For example, it may be simulated that, during the user is walking arbitrarily in the 3D scene, images corresponding to any viewpoints are presented to the user simultaneously. Therefore, the user's interaction freedom in the 3D scene may be significantly improved, and the user experience may be improved accordingly. In contrast, since the existing techniques do not perform the joint

optimization involved in the embodiments of the present disclosure, the 3D scene reconstructed by the existing techniques can only achieve effective image projection at viewpoints corresponding to the cameras used for the image collection, but cannot support high-quality image projection at any other viewpoints.

5 [0018] The embodiments of the present disclosure may utilize limited camera resources for performing 3D scene reconstruction. For example, compared with the existing techniques, the embodiments of the present disclosure may utilize less number of cameras. Moreover, the cameras adopted in the embodiments of the present disclosure are not limited to 3D cameras, and any other types of ordinary camera for shooting 2D images
10 may also be adopted. Therefore, the embodiments of the present disclosure may greatly reduce the collection cost of original images and improve the convenience of the original image collecting process. For example, through the end-to-end joint optimization of the scene reconstruction network and the image enhancement network, the embodiments of the present disclosure may achieve good 3D scene reconstruction and further high-quality
15 image projection only with original images shot by a small number of cameras.

[0019] The embodiments of the present disclosure may be deployed in any known or potential applications. For example, in a VR live for, e.g., a concert, through the embodiments of the present disclosure, a viewer may move arbitrarily in the 3D scene of the concert and watch the performance at any viewpoint. For example, in a 3D video
20 conference involving, e.g., a picture of a conference room, through the embodiments of the present disclosure, a participant may move arbitrarily in the 3D scene of the conference room and watch the conference scene at any viewpoint. Only some exemplary applications are given above, and the embodiments of the present disclosure may also be deployed for any other applications. Moreover, the embodiments of the present disclosure
25 are not limited to applications involving streaming live or applications for playing recorded contents, that is, original images may be shot in real time, shot in advance, etc.

[0020] FIG.1 illustrates an exemplary process 100 of end-to-end 3D scene reconstruction and image projection according to an embodiment.

[0021] According to the process 100, a set of original images 104 shot by a set of
30 cameras 102 may be obtained first. The set of cameras 102 may be pre-deployed in the actual scene. Taking a concert scene as an example, a plurality of cameras may be deployed at different locations such as the stage, auditorium, passages, etc., so that these cameras may shoot images from different shooting angles. In an implementation, the set of original images 104 may be shot by the set of cameras 102 at the same time point.

Accordingly, the set of original images 104 corresponding to the same time point may be used for reconstructing the 3D scene at that time point through the process 100. It should be understood that the set of original images 104 may be shot by the set of cameras in 102 in real time, and thus the process 100 may be performed for e.g., applications involving streaming live, or the set of original images 104 may be previously shot by the set of cameras 102, and thus the process 100 may be performed for e.g., applications involving playing recorded contents.

[0022] The set of cameras 102 may comprise a total of K cameras. Each camera may have corresponding camera parameters. For example, camera parameters of the k -th camera may be represented as $CAM_k = \{x_k, y_k, z_k, d_k, fx_k, fy_k\}$, wherein $1 \leq k \leq K$, x_k, y_k, z_k are space position coordinate of the k -th camera in the real space, d_k is a direction or orientation of the k -th camera, and fx_k, fy_k are field of view (FOV) parameters of the k -th camera. An original image I_k shot by the k -th camera is composed of a set of pixels, and may be represented as:

$$I_k = \{(\hat{x}_i, \hat{y}_i, c_i)\} \quad \text{Equation (1)}$$

wherein $1 \leq i \leq N$, N is the number of pixels included in the original image I_k , $\hat{x}_i, \hat{y}_i$ are position coordinate of the i -th pixel in the original image I_k , and c_i is an appearance property of the i -th pixel, e.g., RGB value, etc. It should be understood that in the case of adopting ordinary cameras for shooting 2D images, a shot image may be directly represented by Equation (1), while in the case of adopting 3D cameras or VR cameras for shooting images with depth-of-field information, depth-of-field information obtained in the shooting process may be ignored, and a shot image may still be represented by Equation (1). It can be seen that the embodiments of the present disclosure may even only adopt images shot by ordinary cameras without requiring the use of more expensive 3D cameras, thereby reducing the collection cost of original images and improving the convenience of the original image collecting process.

[0023] At 110, 3D scene reconstruction may be performed. For example, a 3D scene may be reconstructed based at least on the set of original images 104 and the camera parameters of the set of cameras 102.

[0024] An actual 3D scene S may be represented as:

$$S = \{(x_i, y_i, z_i, c_i)\} \quad \text{Equation (2)}$$

wherein $1 \leq i \leq M$, M is the number of points or voxels included in the actual 3D scene S , x_i, y_i, z_i are space position coordinate of the i -th point in the actual 3D scene S , and c_i is

an appearance property of the i -th point, e.g., RGB value, etc.

[0025] As the theoretical basis of 3D scene reconstruction, the following relationship may be established between the original image I_k shot by the k -th camera and the actual 3D scene S :

$$I_k = \mathcal{M}(CAM_k)S \quad \text{Equation (3)}$$

wherein $\mathcal{M}$ is a transformation model for projecting the actual 3D scene S into a 2D image corresponding to the camera parameters CAM_k . $\mathcal{M}(CAM_k)$ may be referred to as camera information of the k -th camera, which is obtained through applying the transformation model $\mathcal{M}$ to the camera parameters CAM_k of the k -th camera. $\mathcal{M}$ may be implemented through various approaches. For example, in an implementation, $\mathcal{M}$ may be a hybrid transformation matrix which is used for performing projection transformation, affinity transformation, rendering transformation, etc. Equation (3) shows that I_k may be represented by performing transformation to the actual 3D scene S based on camera parameters or camera information. Accordingly, I_k may be used for reconstructing a 3D scene through a variant of Equation (3). For example, the actual 3D scene S may be reconstructed with a combination of camera parameters or camera information and corresponding original images.

[0026] The 3D scene reconstruction at 110 may implement joint optimization or training of a scene reconstruction network 112 and an image enhancement network 114 through concatenating a 3D scene reconstruction process, an image projection process, and an image enhancement process. Through training the scene reconstruction network 112, a reconstructed 3D scene may be obtained. In an implementation, the scene reconstruction network 112 may be constructed based on an approach of explicitly representing a 3D scene, which may generate an explicit 3D scene representation. In an implementation, the scene reconstruction network 112 may be constructed based on an approach of implicitly representing a 3D scene, which may be used for obtaining an implicit 3D scene representation. A 3D scene reconstructed with the scene reconstruction network 112 may be used for performing image projection. Moreover, during training, the image enhancement network 114 may be used for performing image enhancement to a projected image instance output by the scene reconstruction network 112, in order to improve image quality, e.g., to make the image clearer, to make the image look more realistic, etc. The image enhancement network 114 may be constructed based on various approaches, e.g., a Generative Adversarial Network (GAN). It should be understood that the 3D scene

reconstruction at 110 actually performs end-to-end joint optimization to the processes of 3D scene reconstruction, image projection, image enhancement, etc. Further details of this joint optimization will be discussed later in connection with FIG.2.

[0027] It should be understood that, through the above joint optimization, a good 3D scene reconstruction may be achieved even if only utilizing original images shot by a small number of cameras. Therefore, compared with the existing techniques, the embodiments of the present disclosure may utilize a smaller number of cameras, thereby reducing the collection cost of original images and improving the convenience of the original image collecting process.

[0028] After the 3D scene is reconstructed, the process 100 may obtain a target viewpoint 106 at 120. The target viewpoint 106 may be, e.g., designated by a user, or automatically detected based on the user's behavior. The target viewpoint 106 may indicate at what space position, in what direction, etc. the user wants to watch in the 3D scene. The target viewpoint 106 may be represented in an approach similar to camera parameters, e.g., it may be represented through at least one of space position coordinate, direction, field of view parameters, etc. It should be understood that since the 3D scene reconstruction is performed at least through the above joint optimization, the reconstructed 3D scene can effectively and fully characterize any point in the whole space, and thus may be used for performing the subsequent image projection process for any target viewpoint. Accordingly, the target viewpoint 106 may actually correspond to any space position in the 3D scene.

[0029] At 130, an image projection process may be performed. For example, a projected image corresponding to the target viewpoint 106 may be generated with the reconstructed 3D scene, through the trained scene reconstruction network 112.

[0030] At 140, an image enhancement process may be performed. For example, the projected image generated at 130 may be updated to an enhanced projected image 108 corresponding to the target viewpoint 106, through the trained image enhancement network 114. The enhanced projected image 108 may be further presented to the user.

[0031] It should be understood that the process 100 may be repeatedly performed along with time. For example, assuming that the set of original images 104 is obtained at the time point t , accordingly, the 3D scene reconstruction at 110 is actually reconstructing a 3D scene at the time point t , and the scene reconstruction network 112 and the image enhancement network 114 are also trained for the time point t . When reaching the time point $t+1$, a new set of original images obtained at the time point $t+1$ may be used for

performing the 3D scene reconstruction at 110 again, and accordingly a new scene reconstruction network and image enhancement network may be obtained for producing a new enhanced projected image finally. The target viewpoint at the time point $t+1$ may be the same as or different from the target viewpoint at the time point t .

5 [0032] FIG.2 illustrates an exemplary process 200 of performing joint optimization to a scene reconstruction network and an image enhancement network according to an embodiment. The process 200 may be performed during the 3D scene reconstruction process at 110 in FIG.1. A scene reconstruction network 210 and an image enhancement network 220 may correspond to the scene reconstruction network 112 and the image enhancement network 114 in FIG.1 respectively.

[0033] In the process 200, an initial 3D point set 202 may be generated first. In an implementation, the initial 3D point set 202 may be a randomly initialized 3D point set. The initial 3D point set 202 may be represented as $H_S = \{(x_i, y_i, z_i, s_i)\}$, wherein $1 \leq i \leq M$, M is the number of points or voxels included in a 3D scene. Each item in H_S corresponds to a point in the 3D scene, and includes at least a space position coordinate of the point and a randomly initialized space information encoding representation. For example, x_i, y_i, z_i are pre-defined space position coordinate of the i -th point through, e.g., uniform sampling in the whole space, and s_i is a randomly initialized space information encoding representation of the i -th point. s_i is a randomly initialized vector, which may be regarded as a hidden variable that encodes 3D scene space information, and at least implicitly contains information related to appearance property and other possible information.

[0034] It is assumed that the process 200 is currently performed for an original image I_k 206 shot by the k -th camera with camera parameters CAM_k 204. A projected image instance $I_p(CAM_k)$ 212 may be generated based on the initial 3D point set H_S 202 and the camera parameters CAM_k 204, through the scene reconstruction network 210.

[0035] Gradient back propagation 214 may be generated based at least on the projected image instance $I_p(CAM_k)$ 212 and the original image I_k 206 shot by the k -th camera, to optimize the scene reconstruction network 210 and the initial 3D point set H_S 202. For example, the scene reconstruction network 210 and the initial 3D point set H_S 202 may be optimized by minimizing the difference between the projected image instance $I_p(CAM_k)$ 212 and the original image I_k 206. In an implementation, e.g., the per-pixel L1 loss may be adopted in the gradient back propagation.

[0036] In the process 200, the projected image instance 212 output by the scene reconstruction network 210 may be updated to an enhanced projected image instance 222 through the image enhancement network 220. Gradient back propagation 224 may be generated based at least on the enhanced projected image instance 222 and the original image 206, to optimize the image enhancement network 220. For example, the image enhancement network 220 may be optimized by minimizing the difference between the enhanced projected image instance 222 and the original image 206.

[0037] The joint optimization of the scene reconstruction network 210 and the image enhancement network 220 in the process 200 is based at least on both the gradient back propagation mechanism of the scene reconstruction network 210 (e.g., the gradient back propagation 214) and the gradient back propagation mechanism of the image enhancement network 220 (e.g., the gradient back propagation 224). For example, since the projected image instance 212 serves as both the output of the scene reconstruction network 210 and the input of the image enhancement network 220, when the scene reconstruction network 210 and the image enhancement network 220 are concatenated together in the approach shown in FIG.2 and are optimized or trained together, the influence of the gradient back propagation 224 will be further propagated to the gradient back propagation 214, thereby achieving end-to-end joint optimization of the scene reconstruction network 210 and the image enhancement network 220.

[0038] It should be understood that the process 200 may be repeatedly performed for each original image in a set of original images shot by a set of cameras (e.g., the set of original images 104 in FIG.1), so as to iteratively train or optimize the scene reconstruction network 210 and the image enhancement network 220. Through the joint optimization of the scene reconstruction network 210 and the image enhancement network 220 based on the process 200, a 3D scene may be reconstructed at the scene reconstruction network 210 more accurately and more effectively. Moreover, it should be understood that the embodiments of the present disclosure are not limited to any specific techniques for constructing the scene reconstruction network 210 and the image enhancement network 220.

[0039] As described above, according to the embodiments of the present disclosure, depending on different implementations of the scene reconstruction network, the reconstructing of a 3D scene may comprise generating an explicit 3D scene representation, obtaining an implicit 3D scene representation, etc.

[0040] FIG.3 illustrates an exemplary implementation of a scene reconstruction

network according to an embodiment. In this implementation, a scene reconstruction network 300 is constructed based on an approach of explicitly representing a 3D scene, which may generate an explicit 3D scene representation. The scene reconstruction network 300 is an example of the scene reconstruction network 210 in FIG.2.

5 [0041] An initial 3D point set 302 may correspond to the initial 3D point set 202 in FIG.2 and may be represented as $H_S = \{(x_i, y_i, z_i, s_i)\}$.

[0042] The scene reconstruction network 300 may comprise a randomly initialized deep learning model 310 which may be represented as $\mathcal{N}(\theta)$, wherein θ is a learnable network parameter. The deep learning model 310 may generate a decoded 3D point set 10 312 based on the initial 3D point set H_S 302. The decoded 3D point set 312 may be represented as $H_E = \{(x_i, y_i, z_i, c_i)\}$, wherein $1 \leq i \leq M$, M is the number of points or voxels included in a 3D scene, x_i, y_i, z_i are space position coordinate of the i -th point, and c_i is an appearance property of the i -th point. The deep learning model 310 may at least decode the space information encoding representation s_i in the initial 3D point set H_S 302 15 into the appearance property c_i in the decoded 3D point set H_E 312. Since the appearance property c_i explicitly represents parameters for presenting the i -th point in a 3D scene, e.g., RGB value, etc., the decoded 3D point set 312 may correspond to an explicit 3D scene representation of the 3D scene.

[0043] The scene reconstruction network 300 may comprise a transformation model 20 320 which may utilize camera parameters 304 for projecting the decoded 3D point set 312 into a projected image instance 322. In an implementation, the transformation model 320 may perform image projection according to Equation (3), e.g., $I_p(CAM_k) = \mathcal{M}(CAM_k)H_E$, wherein H_E represents the decoded 3D point set 312, CAM_k represents the camera parameters 304 of the k -th camera, and $I_p(CAM_k)$ represents the projected image instance 25 322.

[0044] As described above in connection with FIG.2, gradient back propagation may be generated based at least on the projected image instance $I_p(CAM_k)$ 322 and an original image I_k shot by the k -th camera. The gradient back propagation will optimize the scene reconstruction network 300 and optimize the initial 3D point set 302 and the decoded 3D point set 312. Accordingly, the optimized decoded 3D point set may be used as an explicit 30 3D scene representation.

[0045] After the optimization of the scene reconstruction network 300 is completed, the scene reconstruction network 300 may be used for performing image projection for a

target viewpoint at, e.g., 130 in FIG.1. It should be understood that during the image projection process, a projected image corresponding to the target viewpoint may be generated with the optimized decoded 3D point set through the transformation model 320 in the scene reconstruction network 300, wherein the target viewpoint may be represented in an approach similar to camera parameters and provided as an input to the transformation model.

[0046] FIG.4 illustrates an exemplary implementation of a scene reconstruction network according to an embodiment. In the implementation, the scene reconstruction network 400 is constructed based on an approach of implicitly representing a 3D scene, which may be used for obtaining an implicit 3D scene representation. The scene reconstruction network 400 is an example of the scene reconstruction network 210 in FIG.2.

[0047] An initial 3D point set 402 may correspond to the initial 3D point set 202 in FIG.2 or the initial 3D point set 302 in FIG.3, and may be represented as $H_S = \{(x_i, y_i, z_i, s_i)\}$.

[0048] The scene reconstruction network 400 may comprise a transformation model 410 which may obtain camera information corresponding to a camera based on camera parameters 404 of the camera. For example, the transformation model 410 may output camera information $\mathcal{M}(CAM_k)$ according to Equation (3), wherein CAM_k represents camera parameters of the k -th camera, and $\mathcal{M}$ is the transformation model.

[0049] The scene reconstruction network 400 may comprise a deep learning model 420 which may be represented as $\mathcal{R}(\theta)$, wherein θ is a learnable network parameter. The deep learning model 420 may generate a projected image instance $I_p(CAM_k)$ 412 based on the initial 3D point set H_S 402 and the camera information $\mathcal{M}(CAM_k)$ output by the transformation model 410.

[0050] As described above in connection with FIG.2, gradient back propagation may be generated based at least on the projected image instance $I_p(CAM_k)$ 412 and an original image I_k shot by the k -th camera. The gradient back propagation will optimize the scene reconstruction network 400 and optimize the initial 3D point set 402. In FIG.4, although the scene reconstruction network 400 does not generate an explicit 3D scene representation similar to the decoded 3D point set 312 in FIG.3, the optimized initial 3D point set will contain a space information encoding representation of a 3D scene, e.g., at least implicitly contain information related to appearance property and other possible

information, therefore, the optimized initial 3D point set may be used as an implicit 3D scene representation.

[0051] After the optimization of the scene reconstruction network 400 is completed, the scene reconstruction network 400 may be used for performing image projection for a target viewpoint at, e.g., 130 in FIG.1. It should be understood that during the image projection process, a projected image corresponding to the target viewpoint may be generated with the optimized initial 3D point set through the transformation model 410 and the deep learning model 420 in the scene reconstruction network 400, wherein the target viewpoint may be represented in an approach similar to camera parameters and provided as an input to the transformation model.

[0052] FIG.5 illustrates an exemplary implementation of an image enhancement network according to an embodiment. The image enhancement network 500 is an example of the image enhancement network 220 in FIG.2. In this implementation, the image enhancement network 500 is constructed based on GAN.

[0053] The image enhancement network 500 may comprise an enhancement model 510. The enhancement model 510 may generate an enhanced projected image instance 512 based on a projected image instance 502, wherein the projected image instance 502 may correspond to the projected image instance 212 in FIG.2, the projected image instance 322 in FIG.3, the projected image instance 412 in FIG.4, etc. During the training process, the enhancement model 510 aims to update a projected image instance to improve image quality.

[0054] The image enhancement network 500 may comprise a discriminator 520. The discriminator 520 may take the enhanced projected image instance 512 and an original image 504 as inputs, wherein the original image 504 may correspond to the original image 206 in FIG.2.

[0055] The enhancement model 510 may be trained to generate an image that is as similar as possible to a real image, e.g., the original image 504, and the discriminator 520 may be trained to distinguish between the image generated by the enhancement model 510 and the real image as accurately as possible.

[0056] As described above, gradient back propagation may be generated based at least on the enhanced projected image instance 512 and the original image 504 to optimize the image enhancement network 500.

[0057] After the optimization or training of the image enhancement network 500 is completed, the image enhancement network 500 may be used for performing image

enhancement at, e.g., 140 in FIG. 1. It should be understood that in the image enhancement process, the enhancement model 510 in the image enhancement network 500 may be used for updating a projected image, to obtain a high-quality enhanced projected image.

[0058] It should be understood that the embodiments of the present disclosure are not limited to constructing an image enhancement network with GAN, but may adopt any other technique for constructing an image enhancement network.

[0059] FIG. 6 illustrates a flowchart of an exemplary method 600 for end-to-end 3D scene reconstruction and image projection according to an embodiment.

[0060] At 610, a set of original images shot by a set of cameras may be obtained.

10 [0061] At 620, a 3D scene may be reconstructed based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network.

[0062] At 630, a target viewpoint may be obtained.

15 [0063] At 640, a projected image corresponding to the target viewpoint may be generated with the 3D scene through the scene reconstruction network.

[0064] At 650, the projected image may be updated to an enhanced projected image through the image enhancement network.

20 [0065] In an implementation, the joint optimization may be based at least on a gradient back propagation mechanism of the scene reconstruction network and a gradient back propagation mechanism of the image enhancement network.

[0066] In an implementation, the reconstructing a 3D scene may comprise: generating an initial 3D point set; generating a projected image instance based on the initial 3D point set and camera parameters of at least one camera in the set of cameras, through the scene reconstruction network; and producing gradient back propagation based at least on the projected image instance and an original image shot by the at least one camera, to optimize the scene reconstruction network and the initial 3D point set.

25 [0067] In an implementation, the reconstructing a 3D scene may comprise: generating an explicit 3D scene representation. The generating an explicit 3D scene representation may comprise: generating an initial 3D point set; generating a decoded 3D point set based on the initial 3D point set, through a deep learning model in the scene reconstruction network; projecting the decoded 3D point set to a projected image instance with camera parameters of at least one camera in the set of cameras, through a transformation model in the scene reconstruction network; and producing gradient back propagation based at least on the projected image instance and an original image shot by the at least one camera, to

optimize the scene reconstruction network, the initial 3D point set and the decoded 3D point set, wherein the optimized decoded 3D point set corresponds to the explicit 3D scene representation.

5 [0068] In an implementation, the reconstructing a 3D scene may comprise: generating an implicit 3D scene representation. The generating an implicit 3D scene representation may comprise: generating an initial 3D point set; obtaining camera information corresponding to at least one camera in the set of cameras based on camera parameters of the at least one camera, through a transformation model in the scene reconstruction network; generating a projected image instance based on the initial 3D point set and the camera information, through a deep learning model in the scene reconstruction network; and producing gradient back propagation based at least on the projected image instance and an original image shot by the at least one camera, to optimize the scene reconstruction network and the initial 3D point set, wherein the optimized initial 3D point set corresponds to the implicit 3D scene representation.

15 [0069] In an implementation, the method 600 may further comprise: updating a projected image instance, which corresponds to at least one camera in the set of cameras and is output by the scene reconstruction network, to an enhanced projected image instance through the image enhancement network; and producing gradient back propagation based at least on the enhanced projected image instance and an original image shot by the at least one camera, to optimize the image enhancement network. The image enhancement network is based on a GAN.

[0070] Each item in the initial 3D point set may correspond to a point in the 3D scene, and may comprise at least a space position coordinate and a randomly-initialized space information encoding representation of the point.

25 [0071] Each item in the decoded 3D point set may correspond to a point in the 3D scene, and may comprise at least a space position coordinate and appearance property of the point.

[0072] In an implementation, the camera parameters of the set of cameras may comprise a space position coordinate, a direction and field of view parameters of each camera.

[0073] In an implementation, the 3D scene may be reconstructed for the whole space associated with the 3D scene.

[0074] In an implementation, the target viewpoint may correspond to any space position in the 3D scene.

[0075] In an implementation, the set of original images may be shot at the same time point.

[0076] In an implementation, the set of original images may be shot in real time or shot in advance.

5 [0077] It should be understood that the method 600 may further comprise any step/process for end-to-end 3D scene reconstruction and image projection according to the above embodiments of the present disclosure.

[0078] FIG.7 illustrates an exemplary apparatus 700 for end-to-end 3D scene reconstruction and image projection according to an embodiment.

10 [0079] The apparatus 700 may comprise: an original image obtaining module 710, for obtaining a set of original images shot by a set of cameras; a 3D scene reconstructing module 720, for reconstructing a 3D scene based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network; a target viewpoint obtaining
15 module 730, for obtaining a target viewpoint; an image projecting module 740, for generating a projected image corresponding to the target viewpoint with the 3D scene through the scene reconstruction network; and an image enhancement module 750, for updating the projected image to an enhanced projected image through the image enhancement network.

20 [0080] Moreover, the apparatus 700 may further comprise any other modules that perform steps of the methods for end-to-end 3D scene reconstruction and image projection according to the above embodiments of the present disclosure.

[0081] FIG.8 illustrates an exemplary apparatus 800 for end-to-end 3D scene reconstruction and image projection according to an embodiment.

25 [0082] The apparatus 800 may comprise: at least one processor 810; and a memory 820 storing computer-executable instructions. When executing the computer-executable instructions, the at least one processor 810 may: obtain a set of original images shot by a set of cameras; reconstruct a 3D scene based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene
30 reconstruction network and an image enhancement network; obtain a target viewpoint; generate a projected image corresponding to the target viewpoint with the 3D scene through the scene reconstruction network; and update the projected image to an enhanced projected image through the image enhancement network. Moreover, the processor 810 may further perform any other steps/processes of the methods for end-to-end 3D scene

reconstruction and image projection according to the above embodiments of the present disclosure.

[0083] The embodiments of the present disclosure propose a computer program product for end-to-end 3D scene reconstruction and image projection, comprising a computer program that is executed by at least one processor for: obtaining a set of original
5 images shot by a set of cameras; reconstructing a 3D scene based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network; obtaining a target viewpoint; generating a projected image corresponding to the target viewpoint with the 3D
10 scene through the scene reconstruction network; and updating the projected image to an enhanced projected image through the image enhancement network. Moreover, the computer program may be further executed for implementing any other steps/processes of the methods for end-to-end 3D scene reconstruction and image projection according to the above embodiments of the present disclosure.

[0084] The embodiments of the present disclosure may be embodied in a non-transitory computer-readable medium. The non-transitory computer readable medium may comprise instructions that, when executed, cause one or more processors to perform any steps/processes of the methods for end-to-end 3D scene reconstruction and image
15 projection according to the above embodiments of the present disclosure.

[0085] It should be appreciated that all the operations in the methods described above are merely exemplary, and the present disclosure is not limited to any operations in the methods or sequence orders of these operations, and should cover all other equivalents under the same or similar concepts.

[0086] It should also be appreciated that all the modules in the apparatuses described
25 above may be implemented in various approaches. These modules may be implemented as hardware, software, or a combination thereof. Moreover, any of these modules may be further functionally divided into sub-modules or combined together.

[0087] Processors have been described in connection with various apparatuses and methods. These processors may be implemented using electronic hardware, computer
30 software, or any combination thereof. Whether such processors are implemented as hardware or software will depend upon the particular application and overall design constraints imposed on the system. By way of example, a processor, any portion of a processor, or any combination of processors presented in the present disclosure may be implemented with a microprocessor, microcontroller, digital signal processor (DSP), a

field-programmable gate array (FPGA), a programmable logic device (PLD), a state machine, gated logic, discrete hardware circuits, and other suitable processing components configured to perform the various functions described throughout the present disclosure. The functionality of a processor, any portion of a processor, or any combination of
5 processors presented in the present disclosure may be implemented with software being executed by a microprocessor, microcontroller, DSP, or other suitable platform.

[0088] Software shall be construed broadly to mean instructions, instruction sets, code, code segments, program code, programs, subprograms, software modules, applications, software applications, software packages, routines, subroutines, objects, threads of
10 execution, procedures, functions, etc. The software may reside on a computer-readable medium. A computer-readable medium may include, by way of example, memory such as a magnetic storage device (e.g., hard disk, floppy disk, magnetic strip), an optical disk, a smart card, a flash memory device, random access memory (RAM), read only memory (ROM), programmable ROM (PROM), erasable PROM (EPROM), electrically erasable
15 PROM (EEPROM), a register, or a removable disk. Although a memory is shown as being separate from the processor in various aspects presented in this disclosure, a memory may also be internal to the processor (e.g., a cache or a register).

[0089] The previous description is provided to enable any person skilled in the art to practice the various aspects described herein. Various modifications to these aspects will
20 be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other aspects. Thus, the claims are not intended to be limited to the aspects shown herein. All structural and functional equivalents to the elements of the various aspects described throughout the present disclosure that are known or later come to be known to those of ordinary skilled in the art are intended to be encompassed by the
25 claims.

CLAIMS

1. A method for end-to-end three-dimension (3D) scene reconstruction and image projection, comprising:

obtaining a set of original images shot by a set of cameras;

reconstructing a 3D scene based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network;

obtaining a target viewpoint;

generating a projected image corresponding to the target viewpoint with the 3D scene through the scene reconstruction network; and

updating the projected image to an enhanced projected image through the image enhancement network.

2. The method of claim 1, wherein

the joint optimization is based at least on a gradient back propagation mechanism of the scene reconstruction network and a gradient back propagation mechanism of the image enhancement network.

3. The method of claim 2, wherein the reconstructing a 3D scene comprises:

generating an initial 3D point set;

generating a projected image instance based on the initial 3D point set and camera parameters of at least one camera in the set of cameras, through the scene reconstruction network; and

producing gradient back propagation based at least on the projected image instance and an original image shot by the at least one camera, to optimize the scene reconstruction network and the initial 3D point set.

4. The method of claim 2, wherein the reconstructing a 3D scene comprises:

generating an explicit 3D scene representation.

5. The method of claim 4, wherein the generating an explicit 3D scene representation comprises:

generating an initial 3D point set;

generating a decoded 3D point set based on the initial 3D point set, through a deep learning model in the scene reconstruction network;

projecting the decoded 3D point set to a projected image instance with camera parameters of at least one camera in the set of cameras, through a transformation model in the scene reconstruction network; and

producing gradient back propagation based at least on the projected image instance and an original image shot by the at least one camera, to optimize the scene reconstruction network, the initial 3D point set and the decoded 3D point set, wherein the optimized decoded 3D point set corresponds to the explicit 3D scene representation.

6. The method of claim 2, wherein the reconstructing a 3D scene comprises:
generating an implicit 3D scene representation.

7. The method of claim 6, wherein the generating an implicit 3D scene representation comprises:

generating an initial 3D point set;

obtaining camera information corresponding to at least one camera in the set of cameras based on camera parameters of the at least one camera, through a transformation model in the scene reconstruction network;

generating a projected image instance based on the initial 3D point set and the camera information, through a deep learning model in the scene reconstruction network;
and

producing gradient back propagation based at least on the projected image instance and an original image shot by the at least one camera, to optimize the scene reconstruction network and the initial 3D point set, wherein the optimized initial 3D point set corresponds to the implicit 3D scene representation.

8. The method of claim 2, further comprising:

updating a projected image instance, which corresponds to at least one camera in the set of cameras and is output by the scene reconstruction network, to an enhanced projected image instance through the image enhancement network; and

producing gradient back propagation based at least on the enhanced projected image instance and an original image shot by the at least one camera, to optimize the image enhancement network

9. The method of claim 8, wherein

the image enhancement network is based on a generative adversarial network (GAN).

10. The method of claim 5 or 7, wherein

each item in the initial 3D point set corresponds to a point in the 3D scene, and comprises at least a space position coordinate and a randomly-initialized space information encoding representation of the point.

11. The method of claim 5, wherein

each item in the decoded 3D point set corresponds to a point in the 3D scene, and

comprises at least a space position coordinate and appearance property of the point.

12. The method of claim 1, wherein

the 3D scene is reconstructed for the whole space associated with the 3D scene.

13. The method of claim 1, wherein

the target viewpoint corresponds to any space position in the 3D scene.

14. An apparatus for end-to-end three-dimension (3D) scene reconstruction and image projection, comprising:

at least one processor; and

a memory storing computer-executable instructions that, when executed, cause the at least one processor to:

obtain a set of original images shot by a set of cameras,

reconstruct a 3D scene based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network,

obtain a target viewpoint,

generate a projected image corresponding to the target viewpoint with the 3D scene through the scene reconstruction network, and

update the projected image to an enhanced projected image through the image enhancement network.

15. A computer program product for end-to-end three-dimension (3D) scene reconstruction and image projection, comprising a computer program that is executed by at least one processor for:

obtaining a set of original images shot by a set of cameras;

reconstructing a 3D scene based at least on the set of original images and camera parameters of the set of cameras, through joint optimization of a scene reconstruction network and an image enhancement network;

obtaining a target viewpoint;

generating a projected image corresponding to the target viewpoint with the 3D scene through the scene reconstruction network; and

updating the projected image to an enhanced projected image through the image enhancement network.

1/4

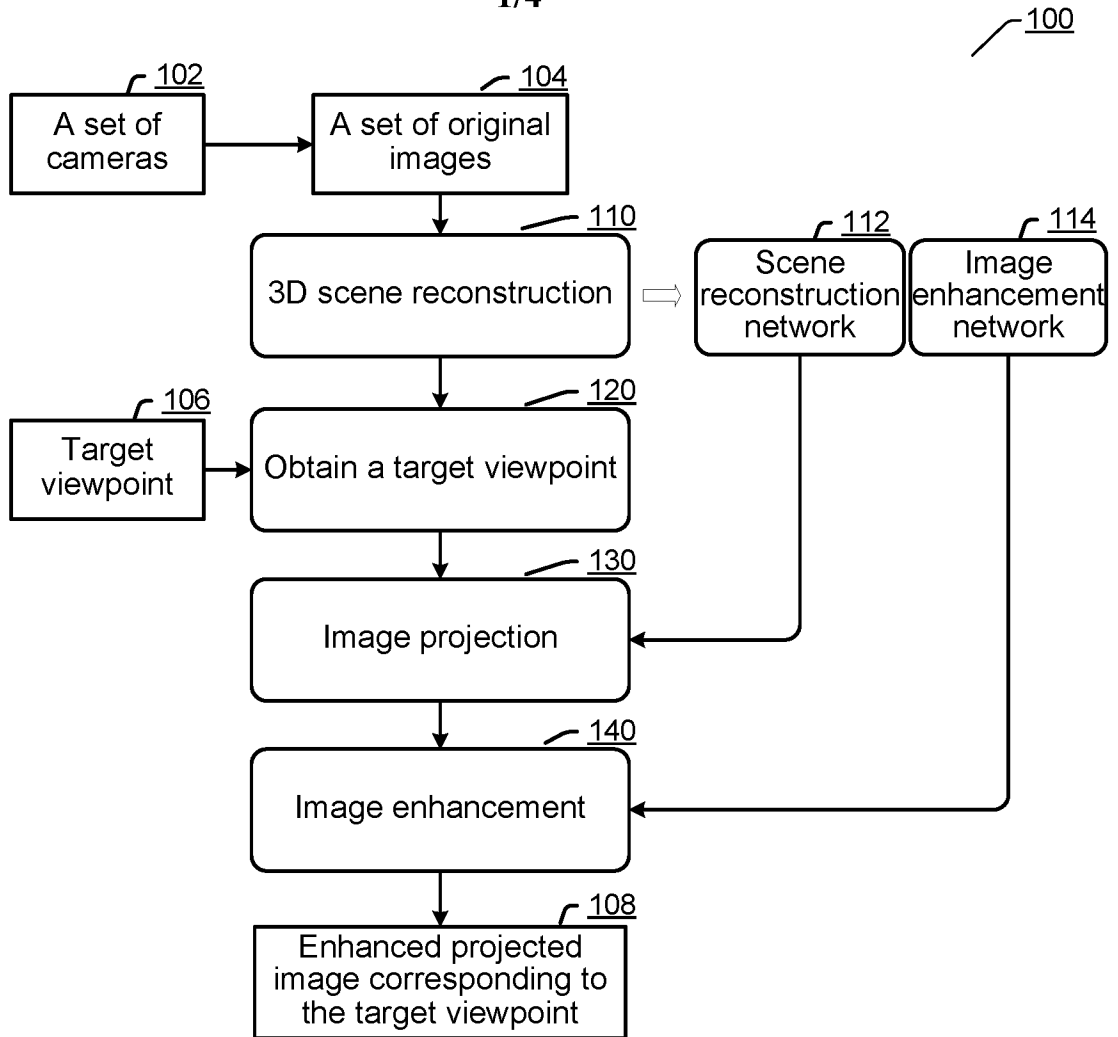


FIG 1

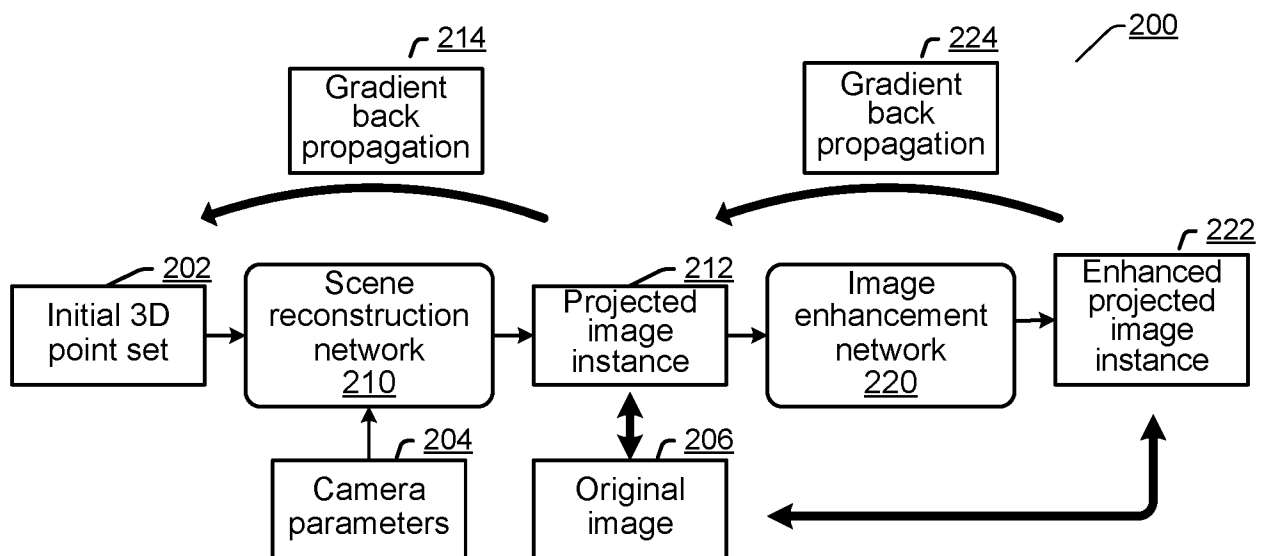
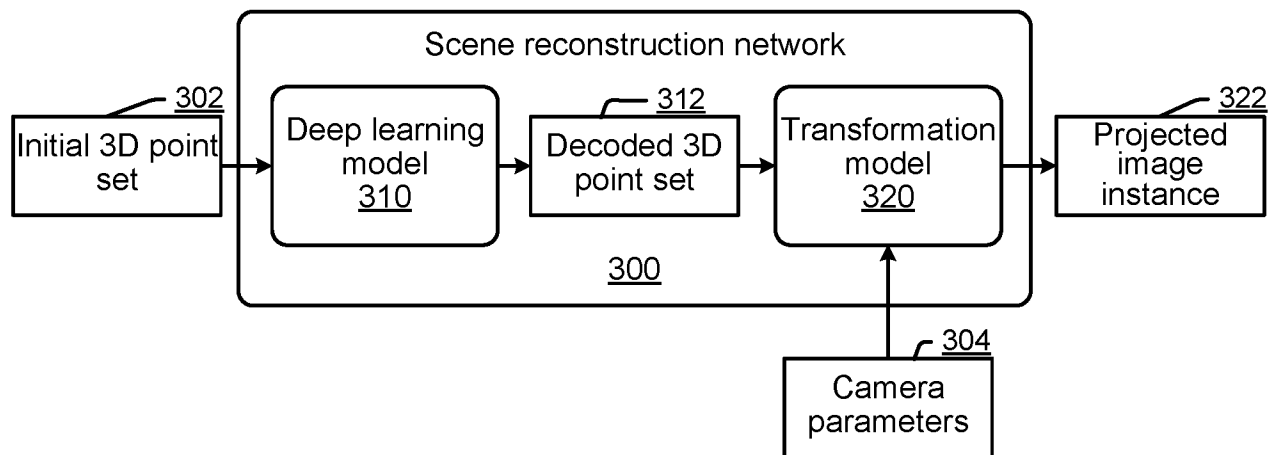
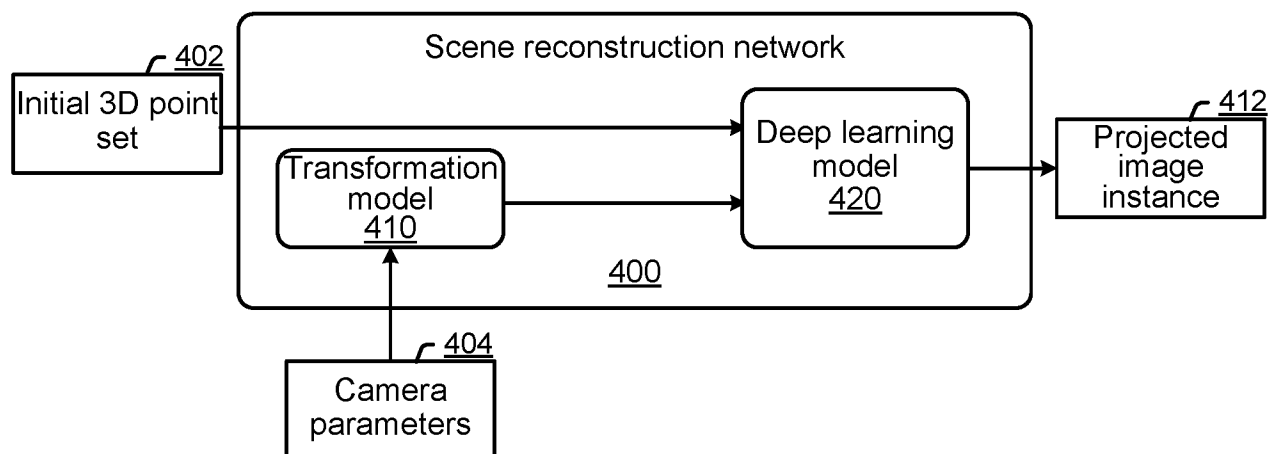
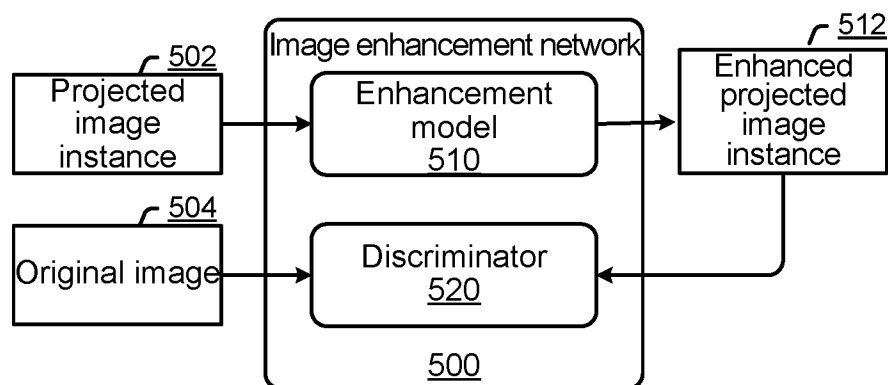
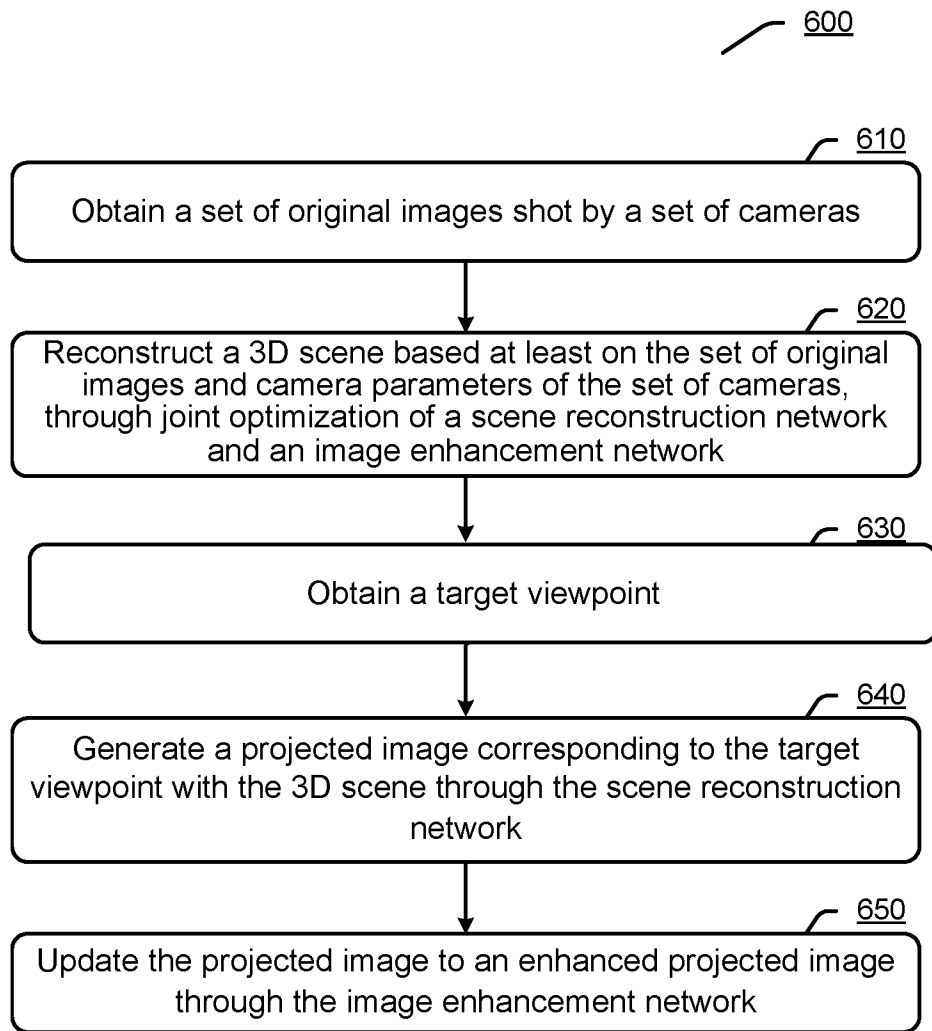


FIG 2

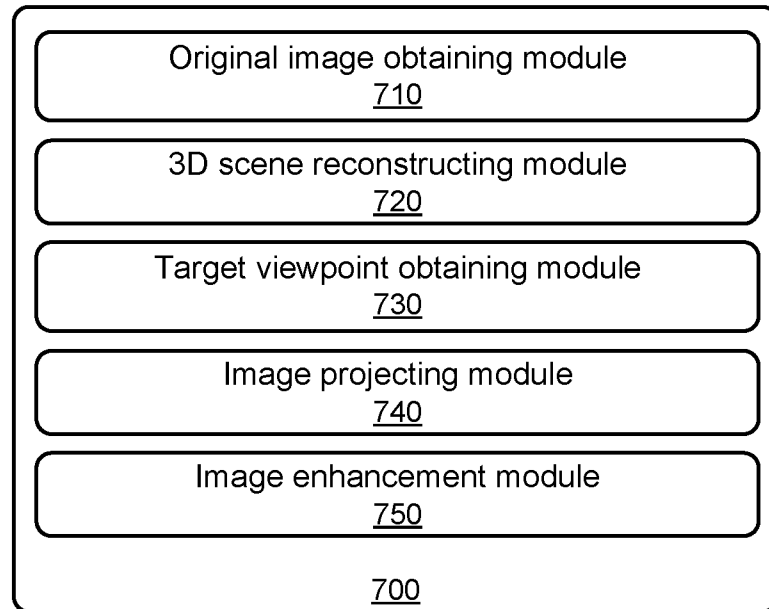
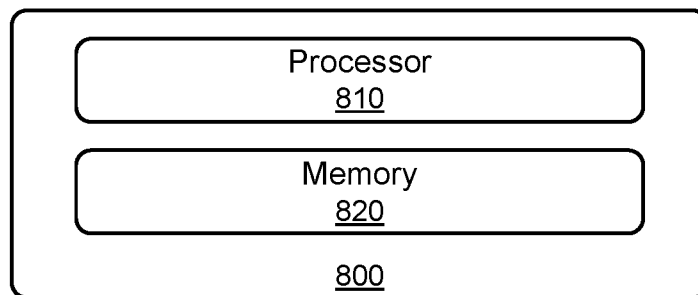
2/4

**FIG 3****FIG 4****FIG 5**

3/4

**FIG 6**

4/4

**FIG 7****FIG 8**

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2021/065595

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06T15/20 G06N3/02
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06T G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WILES OLIVIA ET AL: "SynSin: End-to-End View Synthesis From a Single Image", 2020 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), IEEE, 13 June 2020 (2020-06-13), pages 7465-7475, XP033805337, DOI: 10.1109/CVPR42600.2020.00749 [retrieved on 2020-08-03] abstract; p.7466, left col., par.3; p.7467, left col., par.2; p.7467, right col., par.2; fig.1-3 ----- -/-	1-15

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

28 April 2022

Date of mailing of the international search report

17/05/2022

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Ellerbrock, Thomas

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2021/065595

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MESHRY MOUSTAFA ET AL: "Neural Rerendering in the Wild", 2019 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), IEEE, 15 June 2019 (2019-06-15), pages 6871-6880, XP033687320, DOI: 10.1109/CVPR.2019.00704 [retrieved on 2020-01-08] abstract; p.6880, "3.1. Neural rerendering framework"; p.6881, left col., fig.3 -----	1-15
A	WO 2020/242170 A1 (SAMSUNG ELECTRONICS CO LTD [KR]) 3 December 2020 (2020-12-03) p.13, par.2; p.14; p.16, L5-6; p.17, L6-12 -----	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2021/065595

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2020242170	A1	03-12-2020	NONE
