Paper

視点補間ネットワークによる複雑な光学特性を持つ物体の材質固有外観再現

星澤知宙, 入山太嗣, 小室孝

概要本研究では、自由視点画像生成ネットワークを用いて、様々な光学特性を持つ物体の材料外観の再現を試みる。ネットワークは、4つの特定の視点で撮影されたRGB画像と深度画像を入力とし、その内部の中間視点で画像を生成する。RGB画像は、画像ワーピングにより、出力視点から見た画像に幾何学的に変換される。次に、U-Netベースの画像変換ネットワークを用いて、輝度を補間した画像を生成する。我 々は、光学的に正しい出力を得るのではなく、材料固有の外観を生成するために敵対的損失を使用する。実験では、周囲の環境を反映した金属材料、光を透過・屈折させるガラス材料、表面下散乱を伴う材料を使用した。その結果、敵対的損失の使用は、人間の知覚に近い画質評価指標であるLPIPSにおいても、人間の参加者による評価においても、これらすべての材料でより良い結果を与えることが示された。

Key words: 自由視点映像, 視点合成, 画像変換

1. まえかき

近年、インターネットショッピングが普及しているが、多くの場合、ユーザーには数枚の商品写真しか提示されず、資料の感覚を得られない場合がある。物質の視覚的感覚は、反射、透過、散乱などの物体と相互作用する光が人間の目に入射したときに提供される。特に金属やガラスなどの物体では、視点による輝度の変化が素材の知覚に大きく寄与していることがわかる。したがって、少数の入力画像から輝度変化のある自由視点画像を生成する技術は、そのような物体の材質的外観を再現するのに有用であると考えられる。

自由視点画像生成には大きく分けて2つのアプローチがある。入力画像から物体やシーンの3次元形状と反射率の両方を推定する[1)2)3)4)]。鏡面反射のパラメータを推定することで、視点による輝度変化を再現することができる。しかし、透過率、屈折率、表面下散乱による輝度変化は再現されず、一般に画像からこれらの光学特性を推定することは困難である。

もう一つは、入力画像を補間して自由視点画像を生成することである[5)6)]。このアプローチは、3D形状や光学パラメータを明示的に推定するのではなく、異なる視点で直接画像を生成する。ディープニューラルネットワークを用いた画像補間は、単に画素値を補間するのではなく、構造情報を補間できることが示されている。しかし、様々な光学現象による輝度変化を視点別に再現することが可能かどうかは検証されていない。

本研究では、視点による輝度変化を再現する自由視点画像生成ネットワークを用いて、様々な光学特性を持つ物体の物質知覚を提供する。金属やガラスなどの物体は視点によって外観を大きく変化させるため、人間が自分の材質を認識するのに役立つが、光学的に複雑な物体に対して正しい補間結果を得ることは困難である。そこで、敵対的損失[7)]を用いて、光学特性を明示的にモデル化することなく、経験的に材料固有の外観を生成する。金属、ガラス、表面下散乱材料を用いた実験により、材料の外観生成に対する敵対的損失の影響を検証する。

2. 関連研究

2. 1 3次元形状および反射率の推定キャプチャ画像または画像から物体またはシーンの3次元形状および反射特性を推定する多くの研究が行われており、異なる視点から見た画像を再現することができる。これらのほとんどは、反射特性としてBlinn-Phongモデルや Cook-Torranceモデルのような単純なパラメトリックモデルを採用しており、

Received January 16, 2025; Revised February 21, 2025; Accepted March 10, 2025
† Saitama University (Saitama, Japan)

Paper

Reproducing material-specific appearances of optically complex objects using a view interpolation network

Chihiro Hoshizawa †, Taishi Iriyama †, Takashi Komuro (member)†

Abstract In this study, we attempt to reproduce material appearance of objects with various optical characteristics using a free viewpoint image generation network. The network takes RGB and depth images captured at four specific viewpoints as input and generates an image at an intermediate viewpoint inside them. The RGB images are geometrically transformed into images seen from the output viewpoint by image warping. Then, an image with interpolated luminances is generated using a U-Net based image transformation network. We use adversarial loss to generate material-specific appearances rather than obtaining optically correct outputs. In our experiments, we used metal materials that reflect the surrounding environment, glass materials that transmit and refract light, and materials with sub-surface scattering. The results showed that the use of adversarial loss gave better results for all these materials both in LPIPS, an image quality assessment metric that is close to human perception, and evaluations by human participants.

Key words: free viewpoint image, view synthesis, image transformation.

1. Introduction

Internet shopping has become widespread in recent years, but in many cases, users are presented with only a few photographs of products, which may not give them a sense of the materials. The visual sense of materials is provided when light that interacts with an object such as reflection, transmission, and scattering, enters human eyes. Especially for objects such as metal and glass, changes in luminance depending on the viewpoint contribute significantly to the perception of materials. Therefore, a technology for generating free viewpoint images with the luminance changes from a few input images would be useful for reproducing the material apparances of such objects.

There are two major approaches to free viewpoint image generation. One is to estimate both 3D geometry and reflectance of an object or a scene from input images[1)2)3)4)]. By estimating the parameters of specular reflection, luminance change by viewpoint can be reproduced. However, luminance change due to transmission, refraction, and sub-surface scattering are not reproduced, and it is generally difficult to estimate these optical characteristics from images.

The other is to generate free viewpoint images by interpolating input images[5)6)]. This approach does not explicitly estimate 3D geometry or optical parameters, but directly generates images at different viewpoints. It has been shown that image interpolation using deep neural networks can interpolate structural information rather than simply interpolating pixel values. However, it has not been verified whether it is possible to reproduce luminance change by viewpoint due to various optical phenomena.

In this study, we use a free viewpoint image generation network that reproduces changes in luminance depending on the viewpoint to provide material perception of objects with various optical characteristics. Objects such as metal and glass greatly change their appearance by viewpoint, helping humans to perceive their material, but it is difficult to obtain correct interpolation results for optically complex objects. Therefore, we use adversarial loss[7)] to generate material-specific appearances empirically without explicitly model optical characteristics. We verify the effect of adversarial loss on material appearance generation by the experiment using metal, glass, and subsurface scattering materials.

2. Related work

2.1 3D geometry and reflectance estimation

Many studies have been conducted to estimate the three-dimensional shape and reflection properties of objects or scenes from a captured image or images, which can reproduce images viewed from different viewpoints. Most of them employ simple parametric models such as

Received January 16, 2025; Revised February 21, 2025; Accepted March 10, 2025

† Saitama University
(Saitama, Japan)

拡散アルベド、鏡面アルベド、粗さを推定している。物体や領域ごとに一様な反射特性を仮定するもの[8)9)]もあれば、空間的に変化する反射特性を推定するもの[1)2)3)]もあり、後者はピクセルごとに反射率パラメータを推定する。

いくつかの研究では、複数の異なる視点から撮影された画像から、3D形状と反射特性を推定している。Biらは、6つの異なる視点から撮影された画像から、法線、拡散アルベド、鏡面アルベド、粗さ分布を推定する方法を提案した[10)]。深度マップはまず各入力画像から推定され、次にワープした入力画像から反射特性が推定される。小野らは、マルチビューステレオ(MVS)技術[4)]によって取得された物体の3次元形状に基づいて、反射率特性を表す関数であるBRDFを推定する方法を提案した。彼らの方法は、推定に必要な測定回数を減らすために、BRDFデータセットから抽出された主成分を使用することで、より少ないパラメータでBRDFを表現する。これらの方法は、画像から3次元的な形状や反射特性を推定し、異なる視点から画像を再現することができる。しかし、物体表面の反射のみを推定し、他の光学現象は対象としていない。

近年、画像からの高精度3次元再構成法として、NeRFと呼ばれる手法が注目されている[11)12)13)]。この方法では、シーンは視点依存の輝度と体積密度を出力する5次元連続関数として表現され、ニューラルネットワークによって学習される。新しい視点画像はボリュームレンダリングによって合成され、反射だけでなく透過も再現される。しかし、NeRFはボリュームレンダリングの性質上、屈折や表面下散乱を扱うことができず、多数の入力画像を必要とする。

2.2 画像の構造的補間について

画像は2次元の情報であるが、画像中の物体の3次元的な形状や位置関係などの高次特徴は、画像の構造情報と呼ばれることもある。ニューラルネットワークを用いて構造情報を補間することで、自由視点画像を生成する試みがある。

Oringらは、ニューラルネットワークの深い表現[5)]を補間することで、物体の3次元形状を反映する画像補間法を提案した。オートエンコーダの深い層は、入力画像から圧縮されたより多くの構造情報を表現していることが知られており、この深い表現は、画像の意味をよりよく捉える画像変換を可能にする。

Kalantariらは、ニューラルネットワーク[6)]を用いた画像補間により、自由視点画像を生成する方法を提案した。本手法では、2次元の視点配列の四隅で撮影された画像を入力として、他の視点の画像を生成する。最初のニューラルネットワークは、入力画像から特徴を抽出し、視差マップを生成する。視差マップと入力画像から、新しい視点から見たワープ画像を生成し、これらの画像を2番目のニューラルネットワークに入力し、新しい視点からの画像を得る。

RGB画像と深度マップの組み合わせ以外にも、MPI(Multiplane images)と呼ばれる複数の半透明画像レイヤーで3Dシーンを表現する方法がある。ニューラルネットワーク[14)15)16)]を用いてMPIを推定し、自由視点画像を生成する方法がいくつか提案されている。MPIは半透明の画像で表現されるため、光の透過を扱うことができる。また、鏡に映った画像を鏡の後ろにある層情報として推定することで、完全な鏡面反射を管理することができる。しかし、MPIから画像を合成するだけでは、粗い鏡面反射、屈折、表面下散乱を扱うことは困難である。

Mildenhallらは、入力画像の各視点に対してMPIを推定し、各MPIから新しい視点画像を生成し、生成された全ての画像を加重和を用いて結合し、最終出力[17)]を得る方法を提案した。これにより、視点移動による輝度変化の再現が可能となる。しかし、線形補間であるため、シャープなハイライトの変化が再現されない可能性がある。

構造情報を補間するためにニューラルネットワークを使用することで、様々な材料固有の光学現象を再現できる可能性がある。しかし、我々の知る限り、そのような調査は行われていない。

3. ビュー補間ネットワーク

図1に本研究で使用したネットワークを示す。ネットワークは、4つの特定の視点で撮影されたRGB画像と深度画像を入力とし、その内部の中間視点で画像を生成する。

まず、RGB画像は、奥行き画像を用いた画像ワーピングにより、出力視点から見た画像に幾何学的に変換される。画像ワーピング変換

the Blinn-Phong or Cook-Torrance models as reflection properties, estimating diffuse albedo, specular albedo, and roughness. Some assume uniform reflection properties per object or region[8)9)], while others estimate spatially-varying reflection properties[1)2)3)], and the latter estimates reflectance parameters pixel by pixel.

Some studies estimate 3D shape and reflection properties from images taken from several different viewpoints. Bi et al. proposed a method to estimate normal, diffuse albedo, specular albedo, and roughness distributions from images taken from six different viewpoints[10)]. A depth map is first estimated from each input image, and then the reflection properties are estimated from warped input images. Ono et al. proposed a method to estimate the BRDF, a function that represents reflectance properties, based on the 3D shape of an object acquired by Multi-View Stereo (MVS) technology[4)]. Their method represents BRDFs with fewer parameters by using principal components extracted from a BRDF dataset to reduce the number of measurements necessary for estimation. These methods can estimate three-dimensional shape and reflection properties from images and reproduce images from different viewpoints. However, they estimate only reflections on object surfaces and do not target other optical phenomena.

In recent years, a method called NeRF has attracted much attention as a high-precision 3D reconstruction method from images[11)12)13)]. In this method, the scene is represented as a 5D continuous function that outputs viewpoint-dependent radiance and volume density, which is learned by a neural network. New viewpoint images are synthesized by volume rendering, which reproduces not only reflections but also transmissions. However, NeRF cannot handle refraction and sub-surface scattering due to the nature of volume rendering, and it requires a large number of input images.

2.2 Structural interpolation of images

An image is two-dimensional information, but high-level features such as the three-dimensional shape and positional relationship of objects in the image are sometimes referred to as the structural information of the image. There are attempts to generate free viewpoint images by interpolating structural information using neural networks.

Oring et al. proposed an image interpolation method that reflects the three-dimensional shape of an object by interpolating the deep representation of a neural network[5)]. It is known that the deeper layers of an autoencoder represent more structural information compressed from the input image, and this deep representation enables image transformation that better captures the meaning of the image.

Kalantari et al. proposed a method for generating free viewpoint images by image interpolation using a neural network[6)]. In this method, images captured at the four corners in a two-dimensional array of viewpoints are used as input to generate images at the other viewpoints. The first neural network extracts features from the input images and generates a disparity map. From the disparity map and the input images, warped images seen from a new viewpoint are generated, and these images are input to the second neural network to obtain images from a new viewpoint.

Besides the combination of an RGB image and a depth map, there is another way of representing a 3D scene with multiple translucent image layers called Multiplane images (MPI). Several methods have been proposed for generating free viewpoint images by estimating MPI using neural networks[14)15)16)]. Since MPI is represented by translucent images, it can handle light transmission. It can also manage perfect specular reflection by estimating the image reflected in a mirror as layer information positioned behind the mirror. However, it is difficult to handle coarse specular reflection, refraction, and sub-surface scattering by simply synthesizing images from MPI.

Mildenhall et al. proposed a method that estimates the MPI for each viewpoint of input images, generates a new viewpoint image from each MPI, and combines all the generated images using a weighted sum for the final output[17)]. This enables reproduction of luminance changes due to viewpoint shifts. However, since it is a linear interpolation, sharp highlight changes may not be reproduced.

By using a neural network to interpolate structural information, it may be possible to reproduce a variety of material-specific optical phenomena. However, to the best of our knowledge, no such investigation has been conducted.

3. View interpolation network

Figure 1 shows the network used in this study. The network takes RGB and depth images taken at four specific viewpoints as input and generates an image at an intermediate viewpoint inside them.

First, the RGB images are geometrically transformed into images seen from the output viewpoint by image warping using the depth images. Image warping trans-

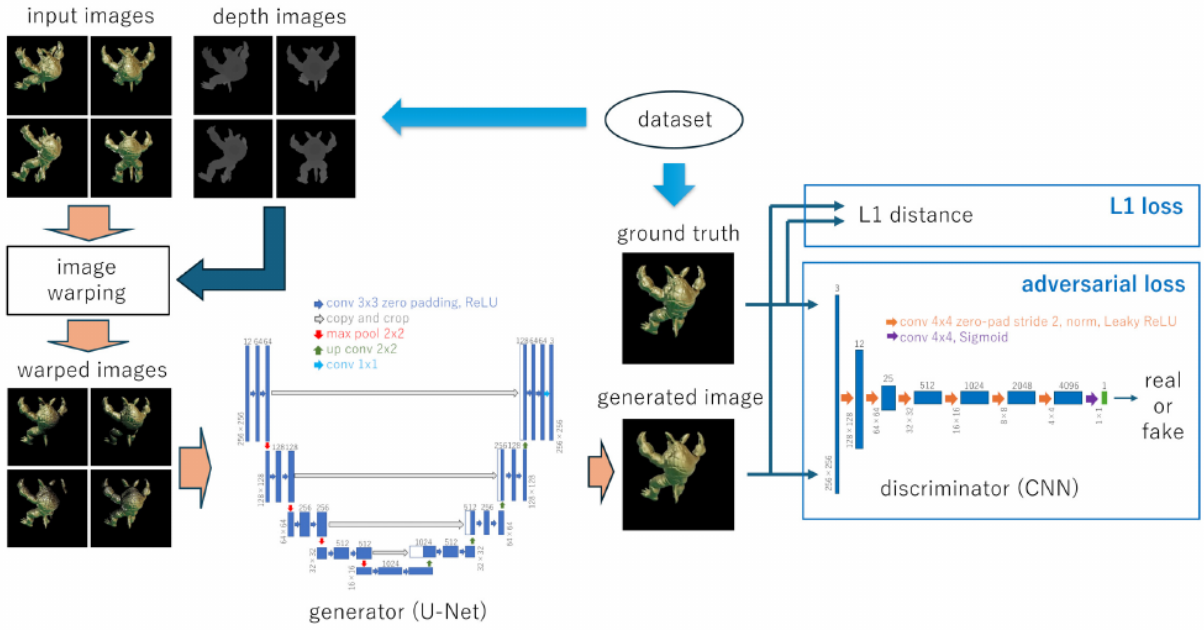


図1:ネットワークアーキテクチャ

は、図2に示すように、入力視点から撮影した画像を、順方向ワーピングを用いて出力視点から見た画像に形成する。入力画像中の画素の位置とその位置の深度値を用いて、入力視点からの3次元座標を求め、出力視点から見た座標に変換し、画像座標に投影する。この処理を入力画像内の全画素に対して行い、ワープ画像を得る。

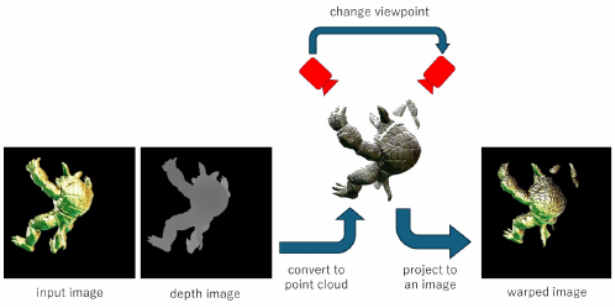


図2:フォワードワーピング

ワープされた4つの入力画像はジェネレータに入力され、ジェネレータは補間された画像を出力するように学習される。生成器は、出力画像をよりグランドトゥルースに近づけるだけでなく、生成的敵対ネットワーク(GAN)のメカニズムを用いて、特定のカテゴリの画像を生成するように訓練される。

この方法は、GANベースの画像変換ネットワークpix2pix[18]で使用されているものと似ているが、生成器の入力と出力の組を識別器に与えるpix2pixとは異なり、生成器の出力のみを識別器に与える。これは、本研究の目的が、ある素材の具体的な外観を再現することであるためである。

U-Net[19]を生成器として使用し、4つのワープしたRGB画像をチャンネル方向に連結して12チャンネルを形成し、生成器に入力する。出力は1枚のRGB画像であり、ネットワークは4つの入力画像の補間画像を生成するように学習される。ワープした入力画像をチャンネル方向に連結するというアイデアは、Kalantariらの研究[6]で用いられており、本手法でも用いられている。入力画像の数は4枚に設定されているが、これはカメラの位置が2次元で表現され、各次元に対して少なくとも2つの方法が必要であるため、彼らの研究と同じである。彼らの研究では視点補間に畳み込みニューラルネットワーク(CNN)を使用しているが、我々は画像変換にpix2pixで使用されているU-Netを使用している。

敵対的損失の計算に用いる識別器は、典型的な生成的敵対的ネットワーク(GAN)[7]の識別器と同じである。生成された画像または真の画像のいずれかが識別器に入力され、識別器は本物か偽物かを推定した結果を出力する。

生成器の学習に用いる損失関数は、L1損失L_1と敵対的損失L_{adv}の組み合わせである。L1損失は、

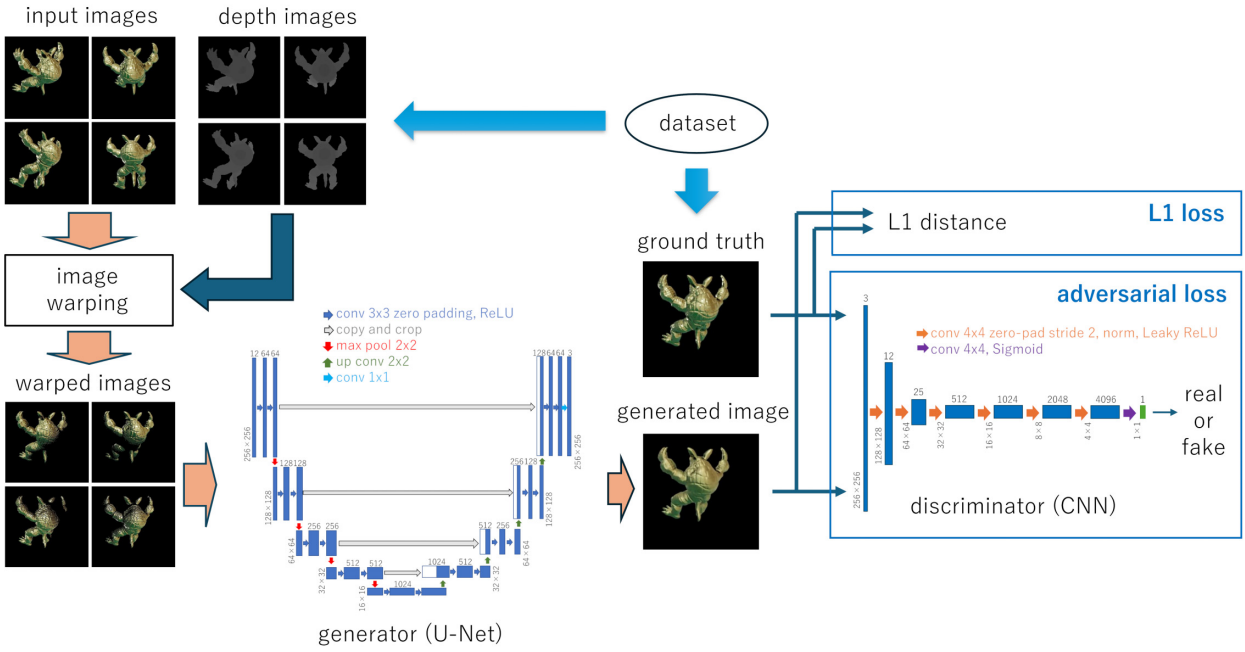


Fig. 1: Network architecture

forms an image taken from the input viewpoint into an image seen from the output viewpoint using forward warping, as shown in Fig. 2. Using the position of a pixel in the input image and the depth value at that position, the 3D coordinates from the input viewpoint are obtained, converted to the coordinates seen from the output viewpoint, and projected to the image coordinates. This process is performed for all pixels in the input image to obtain the warped image.

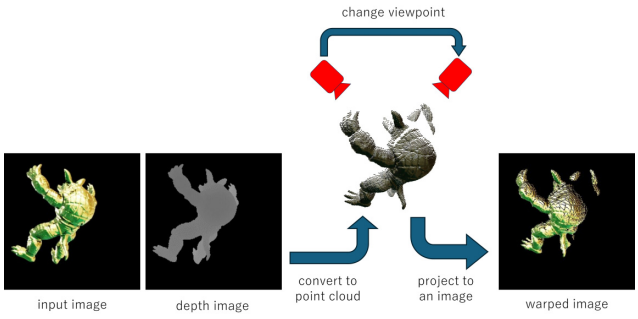


Fig. 2: Forward warping

The warped four input images are input to the generator, and the generator is trained to output an interpolated image. The generator is trained not only to make the output image closer to the ground truth, but also to generate an image of a specific category using the mechanism of a generative adversarial network (GAN). This method is similar to that used in pix2pix[18], a GAN-based image transformation network, but unlike pix2pix, which gives the pair of generator input and output to the discriminator, our method gives only the generator output to the discriminator. This is because the purpose of this study is to reproduce the specific appearance of a certain material.

U-Net[19] is used as the generator and the four warped RGB images are concatenated in the channel direction to form twelve channels and are input to the generator. The output is a single RGB image and the network is trained to generate the interpolated image of the four input images. The idea of concatenating the warped input images in the channel direction was used in Kalantari et al.'s study[6], and is also used in our method. The number of input images is set to four, which is also the same as in their study, because the camera positions are expressed in two dimensions, and at least two ways are necessary for each dimension. Although their study uses a convolutional neural network (CNN) for viewpoint interpolation, we use U-Net which is used in pix2pix for image transformation.

The discriminator used to calculate the adversarial loss is the same as that of a typical generative adversarial network (GAN)[7]. Either a generated image or a true image is input to the discriminator, which outputs the result of estimating whether it is real or fake.

The loss function used to train the generator is a combination of L1 loss L_1 and adversarial loss L_{adv}. The L1 loss is calculated as the L1 norm of the difference be-

生成画像とグランドトゥルースの差のL1ノルムとして計算され、敵対的損失は、式(2)に示すように、バイナリクロスエントロピーによって計算される。

$$L = L_1 + \lambda \min_G \max_D L_{adv}(G, D), \quad (1)$$

$$L_{adv}(G, D) = E_y[\log D(y)] + E_x[\log(1 - D(G(x)))], \quad (2)$$

ここで、λは重み、D(y)は真の画像が入力されたときの識別器の出力、D(G(x))は生成画像が入力されたときの識別器の出力である。

4. データセットの作成

3Dモデルをレンダリングして得られたCG画像を用いてデータセットを作成した。BlenderProcは、3DソフトウェアBlenderで機械学習用データを作成するためのパッケージである。スタンフォード3Dスキャンリポジトリからダウンロードした7つの3Dモデルを使用した *.学習には6つのモデル(スタンフォード・バニー、ハッピー・ブッダ、ドラゴン、ルーシー、アジア・ドラゴン、タイ・タメ)を用い、テストには1つのモデル(アルマジロ・モデル)を用いた。3次元物体の材料は、BlenderのBSDFパラメータ設定を用いて、金属、ガラス、表面下散乱材料とし、色とその他のパラメータはランダムに設定した。環境マップが周囲の環境の反射と透明度を再現するために、Poly Haven ** から取得した HDRI 画像、トレーニング用 500 枚、テスト用 136 枚を使用した。3次元物体の大きさ、位置、向き、環境マップの向き、光源の方向は、各画像においてランダムに設定した。

カメラは3Dオブジェクトの近くにある球体の上に置かれ、球体の中心に向かって配置された。4台の入力視点カメラは球面四角形の頂点に設置され、その位置はオイラー角で2π/9の角度差を持つように設定された。20台の出力視点カメラを配置し、その位置を入力視点の球面四角形の内側にランダムに設定した。図3は、オブジェクトとカメラの配置の例である。赤のカメラは入力視点、白のカメラは出力視点を表している。

レンダリング画像の背景は、ネットワークが背景ではなく物体に集中できるように、図4のゾーンのように黒く塗られた。

*_{http://graphics.stanford.edu/data/3Dscanrep} **_{https://polyhaven.com/}.

図3: オブジェクトとカメラの配置例

(a) 除去前　(b) 除去後

図4:バックグラウンド除去

学習データとして、2,000個のランダムな材料パラメータ、6個の3Dモデル、20個のランダムな出力視点を持つ、各材料タイプについて240,000セットの入出力画像を生成した。テストデータとして、200個のランダムな材料パラメータ、1個の3Dモデル、20個のランダムな出力視点を持つ、各材料タイプについて4,000セットの入出力画像が生成された。画像サイズは256×256ピクセルとした。

5. 実験

5.1 セットアップ

本実験では、作成したデータセットを学習データとテストデータに用い、敵対的損失がある場合とない場合のネットワークの学習を行った。

損失関数の重みλは0.001に設定した。識別器は事前に訓練されていない。λの値を大きくしすぎると、L1損失とのバランスが崩れ、モード崩壊を引き起こしたが、値が小さくなると安定した学習が得られた。最適化アルゴリズムとしてAdamを用い、α = 10 -4, β_1 = 0.9, β_2 = 0.999, = 10 -8で両世代の学習を行った。

tween the generated image and the ground truth, and the adversarial loss is calculated by the Binary Cross-Entropy as shown in Eq. (2).

$$L = L_1 + \lambda \min_G \max_D L_{adv}(G, D), \quad (1)$$

$$L_{adv}(G, D) = E_y[\log D(y)] + E_x[\log(1 - D(G(x)))], \quad (2)$$

where λ is a weight, $D(y)$ is the output of the discriminator when a true image is input, and $D(G(x))$ is the output of the discriminator when a generated image is input.

4. Dataset creation

We used CG images obtained by rendering 3D models to create a dataset. BlenderProc, a package for creating data for machine learning with the 3D software Blender, was used. We used seven 3D models downloaded from the Stanford 3D Scanning Repository*. Six models (Stanford Bunny, Happy Buddha, Dragon, Lucy, Asian Dragon, and Thai Statue) were used for training, and one model (Armadillo model) was used for testing. The materials of the 3D objects were set to be metal, glass, and sub-surface scattering materials using BSDF parameter settings of Blender, while the color and other parameters were set randomly. For the environment maps to reproduce the reflection and transparency of the surrounding environment, we used HDRI images taken from Poly Haven **, 500 images for training and 136 images for testing. The size, position, and orientation of the 3D object, the orientation of the environment map, and the direction of the light source were set randomly in each image.

The cameras were placed on a sphere centered near the 3D object, and were oriented toward the center of the sphere. The four input viewpoint cameras were placed at the vertices of a spherical quadrangle, and their positions were set to have an angular difference of $2\pi/9$ in Euler angles. Twenty output viewpoint cameras were placed, and their positions were randomly set inside the spherical quadrangle of input viewpoints. Figure 3 shows an example of object and camera placement. The red camera represents the input viewpoints and the white camera the output viewpoints.

The background of the rendered images was painted black as shon in Fig. 4 so that the network can focus on the object rather than the background.

* http://graphics.stanford.edu/data/3Dscanrep
** https://polyhaven.com/

Fig. 3: Example of object and camera placement

(a) before removal (b) after removal

Fig. 4: Background removal

As training data, 240,000 sets of input/output images were generated for each material type, with 2,000 random material parameters, six 3D models, and 20 random output viewpoints. As testing data, 4,000 sets of input/output images were generated for each material type, with 200 random material parameters, one 3D model, and 20 random output viewpoints. The image size was set to 256×256 pixels.

5. Experiment

5.1 Setup

In this experiment, we used the created dataset for training and testing data, and the network was trained with and without adversarial loss.

The weight λ in the loss function was set to 0.001. The discriminator was not pre-trained. When the value of λ was made too large, it was out of balance with the L1 loss and caused mode collapse, but a smaller value resulted in stable learning. Adam was used as the optimization algorithm, with $\alpha = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ for training both the genera-

トールと識別器。

5.2 出力例

学習済みネットワークへの入力として与えられた画像の例を図5に、敵対的損失なしと敵対的損失ありの手法の4つの入力視点間の中心視点での生成画像と、グランドトゥルース画像を図6に示す。生成画像とグランドトゥルース画像については、各画像の横に中央部の拡大画像が配置されている。金属材料の場合、敵対的損失がある場合とない場合の生成画像は、いずれもグランドトゥルース画像に近い出力画像を生成する。しかし、敵対的損失を伴わない生成画像は若干ぼやけており、金属固有のシャープシェーディングは弱い。一方、敵対的損失で生成された画像は、シャープなシェーディングを再現することに成功した。ガラス材料については、敵対的な損失なしに生成された画像はかなりぼやけており、ガラスの材料外観はほとんど知覚できない。一方、敵対的損失を除いた生成画像は、境界が明瞭で、ガラスのような特徴を示すが、グランドトゥルース画像と比較すると完全ではない。表面下散乱物質については、敵対的損失がある場合とない場合の生成画像は、グランドトゥルース画像とほぼ同じように見え、人間の目には区別できない。

5.3 画像の再現性の評価

まず、生成された画像がどの程度グランドトゥルースに近いかを評価した。客観的評価には、PSNR、SSIM、LPIPSの3種類の評価指標を用いた。全テストデータに対する客観的評価指標の平均値を表1に示す。

PSNRは画素単位の誤差を利用した評価指標であり、SSIMは画像構造の類似性を評価する評価指標である。LPIPS(Learned Perceptual Image Patch Similarity)[20]は、生成画像とグランドトゥルースを入力として、学習済み画像分類ネットワークの中間層ベクトル間の距離をそれぞれ計算する。LPIPSは画像分類ネットワークによって抽出された特徴量を用いるため、PSNRやSSIMのような古典的な評価指標よりも人間の目の評価に近い結果を得ることができる。本実験では、学習済みのAlexNet[21]のモデルを用いてLPIPSを算出した。PSNRとSSIMが大きく、LPIPSが小さいほど良い結果が得られる。

PSNRとSSIMについては、どの材料においても、敵対的損失を用いた手法と用いない場合の間に有意差はなかったが、LPIPSについては、敵対的損失を用いた手法の方が、全てのテストデータに対する客観的評価指標の平均値である表1よりも良い結果を示した。w/o ALとw/ ALはそれぞれ敵対的損失なしと敵対的損失ありの手法を表す。SSSは表面下散乱を表す。

material	method	PSNR↑	SSIM↑	LPIPS↓
metal	w/o AL	**26.943**	**0.9320**	0.0897
	w/ AL	26.368	0.9299	**0.0534**
glass	w/o AL	24.746	0.8593	0.2497
	w/ AL	**25.167**	**0.8711**	**0.1111**
SSS	w/o AL	**35.109**	**0.9829**	0.0167
	w/ AL	34.648	0.9809	**0.0142**

敵対的な損失はなく、ガラス、金属、表面下の散乱材料の順でその差は大きくなった。

5．4 物質知覚の評価本研究の目的は、光学現象を忠実に再現することではなく、物質の知覚を正しく提供することである。そこで、ヒトを対象とした主観評価を実施した。20代の男女11名を募集した。参加者には、図7に示すような画面が提示され、画面の左側にグランドトゥルースが提示され、画面の右側に敵対的損失なしと敵対的損失ありの出力がランダムに配置され、物質的知覚の観点から、グランドトゥルースに近いと感じるものを選択するよう求められた。これを3つの材料（金属、ガラス、表面下散乱）それぞれについて順番に100回繰り返し、参加者1人につき合計300回繰り返した。

自然素材の外観を再現するためには、視点の動きに合わせて、ハイライトや反射が滑らかに変化する必要がある。そこで、視点がダイナミックに変化する映像を提示した。左上からの入力視点からスタートし、カメラは右下方向に移動して中央付近に移動し、右上の入力視点に時計回りに回転する。動画は15fpsで提示され、この動きは30フレーム（2秒）にわたって行われた。図8に、金属材料の場合のグランドトゥルース映像の例を示す。右上の視点に到達した後、カメラは同じ軌跡に沿って左上の視点に戻った。同じ動作を、参加者が答えを選ぶまで繰り返した。

各素材タイプについて、敵対的損失を伴う出力を選択した回答の割合を表2に示す。二項検定の結果、

tor and discriminator.

5.2 Output examples

Examples of images given as input to the trained network are shown in Figure 5, and the generated images at the center viewpoint between the four input viewpoints for the methods without and with adversarial loss, and ground truth images are shown in Figure 6. For the generated and ground truth images, an enlarged version of the central part is placed next to each image. For the metal material, the generated images with and without adversarial loss both produce output images close to the ground truth images. However, the generated images without adversarial loss are slightly blurred, and the metal-specific sharp shading is weak. In contrast, the generated images with adversarial loss successfully reproduce sharp shading. For the glass material, the generated images without adversarial loss are quite blurry and the material appearance of the glass is barely perceivable. In contrast, the generated images without adversarial loss have clear boundaries and exhibit some glass-like features, but they are not perfect compared to the ground truth images. For the sub-surface scattering material, the generated images with and without adversarial loss appear almost identical to the ground truth images and are indistinguishable to the human eye.

5.3 Evaluation of image reproducibility

First, we evaluated how close the generated images were to ground truth. We used three types of evaluation metrics, PSNR, SSIM, and LPIPS, for objective evaluation The average values of objective evaluation metrics for all test data are shown in Table 1.

PSNR is an evaluation metric that uses per-pixel errors, while SSIM is a metric that evaluates the similarity of image structures. LPIPS(Learned Perceptual Image Patch Similarity)[20] calculates the distance between the intermediate layer vectors of a trained image classification network with input of the generated image and ground truth, respectively. Since LPIPS uses features extracted by an image classification network, it gives results that are closer to human eye evaluation than classical evaluation metrics such as PSNR and SSIM. We used the trained model of AlexNet[21] to calculates LPIPS in this experiment. The larger the PSNR and SSIM, and the smaller the LPIPS, the better the results.

For PSNR and SSIM, there was no significant difference between the methods with and without adversarial loss for any material, but for LPIPS, the method with adversarial loss showed better results than the method without adversarial loss, and the differences were larger in the order of glass, metal, and sub-surface scattering materials.

Table 1: Average values of objective evaluation metrics for all test data. w/o AL and w/ AL stand for the methods without and with adversarial loss, respectively. SSS stands for sub-surface scattering.

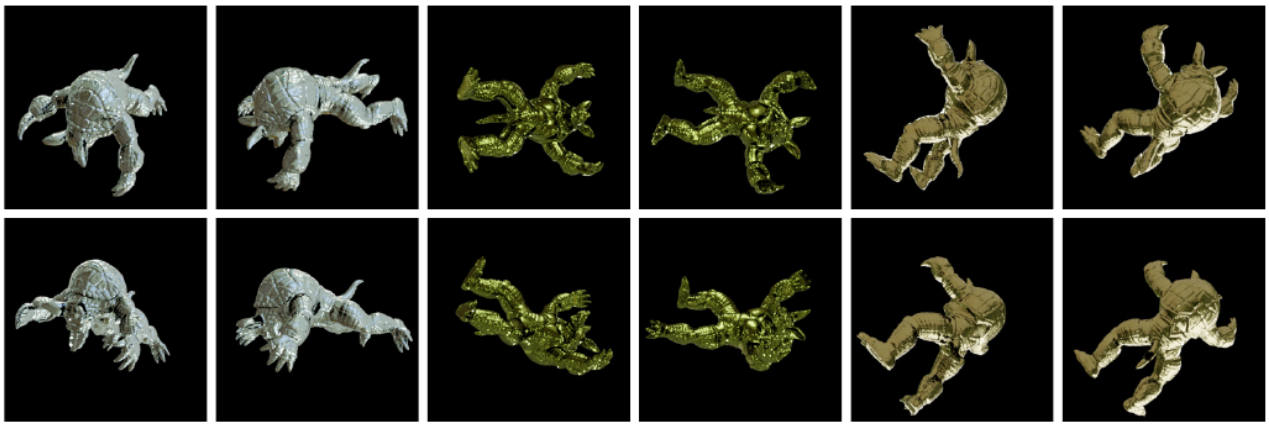
material	method	PSNR↑	SSIM↑	LPIPS↓
metal	w/o AL	**26.943**	**0.9320**	0.0897
	w/ AL	26.368	0.9299	**0.0534**
glass	w/o AL	24.746	0.8593	0.2497
	w/ AL	**25.167**	**0.8711**	**0.1111**
SSS	w/o AL	**35.109**	**0.9829**	0.0167
	w/ AL	34.648	0.9809	**0.0142**

5.4 Evaluation of material perception

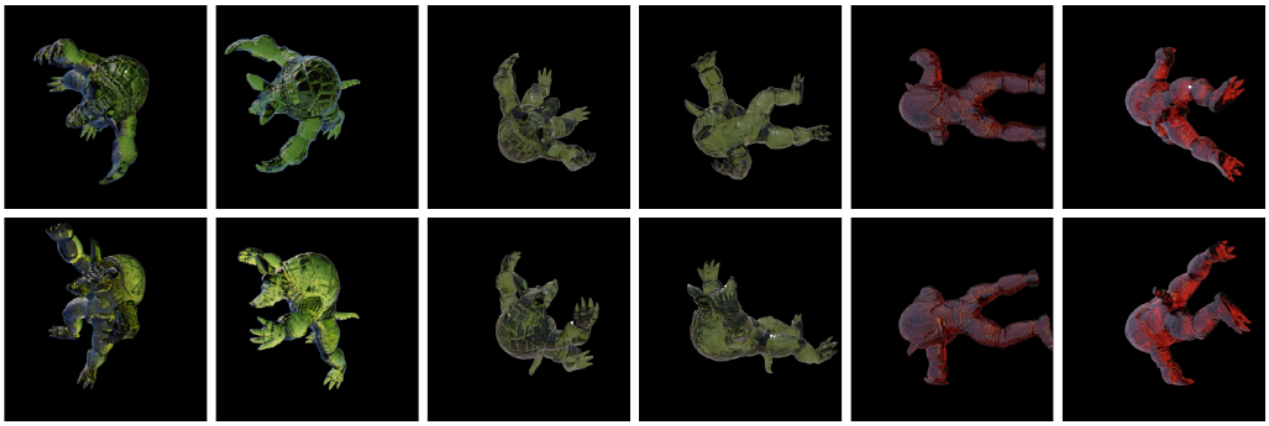
The objective of this study is not to faithfully reproduce optical phenomena but to correctly provide the perception of materials. Therefore, we conducted a subjective evaluation by human participants. We recruited eleven male and female participants in their twenties. The participants were presented with the screen shown in Fig. 7 , where ground truth was presented on the left side of the screen and the outputs without and with adversrial loss were presented in a random arrangement on the right side of the screen, and were asked to choose the one that they felt was closer to ground truth in terms of material perception. This was repeated 100 times for each of the three materials (metal, glass, and sub-surface scattering) in turn, for a total of 300 times per participants.

In order to reproduce natural material appearances, highlights and reflections should change smoothly along with viewpoint movement. Therefore, we presented videos in which the viewpoint changes dynamically. Starting from the upper-left input viewpoint, the camera moved in the lower-right direction to near the center, then rotated clockwise to the upper-right input viewpoint. The videos were presented at 15 fps, and this movement took place over 30 frames (2 seconds). An example of a ground truth video in the case of a metal material is shown in Fig. 8. After reaching the upper-right viewpoint, the camera returned to the upper-left viewpoint along the same trajectory. The same movement was repeated until the participants chose the answer.

The rate of the answers who chose the output with adversarial loss for each material type is shown in Table 2. The result of a binomial test showed that the



(a) Metal



(b) Glass

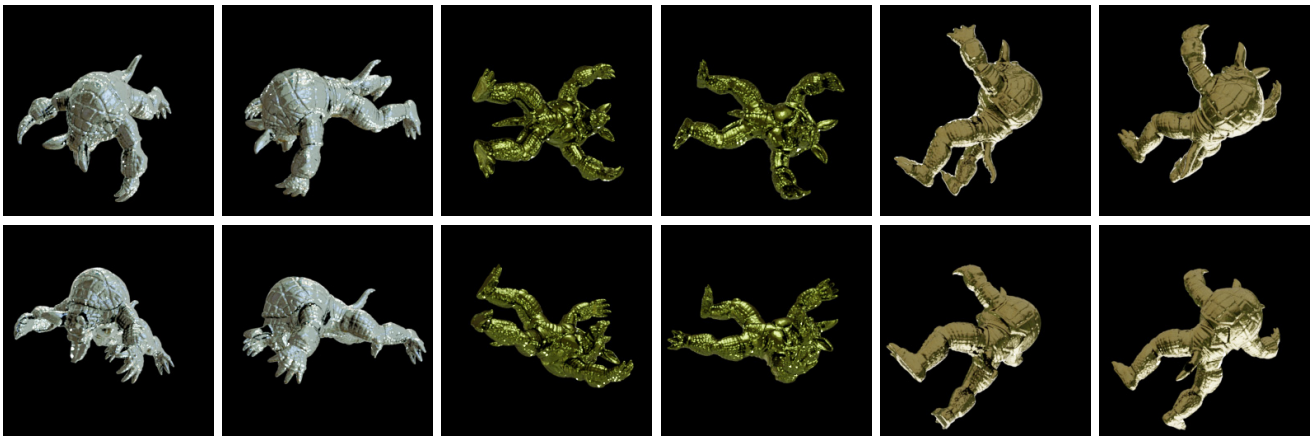
（c）表面下散乱

Fig. 5: Examples of input images

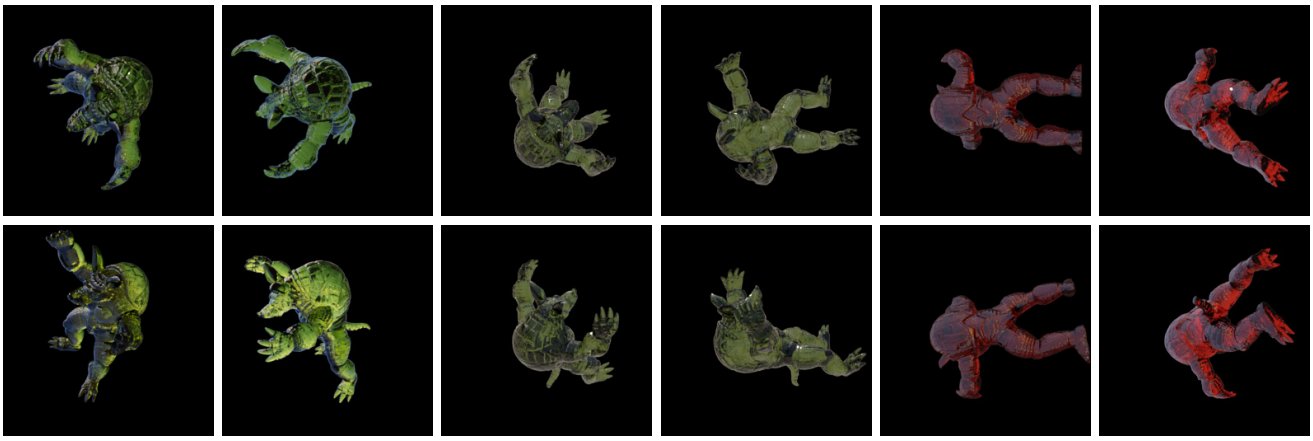
すべての素材のp値は有意水準1%より小さいことが示され、敵対的損失なしと敵対的損失ありの出力の間に有意な選択肢の違いがあることが示された。

6. ディスカッション

実験の結果、敵対的損失を用いた手法は、LPIPSにおいても、人間による主観評価においても、PSNRとSSIMにおいて、敵対的損失を用いた手法と用いない場合とで、あまり差がないことが示された。本研究の目的は、光学的に正しい出力を得ることではなく、物体の材質的外観を再現することである。正確な補間を困難にする複雑な光学的特性を持つ物体に対しても、材料固有の外観を再現するために逆説的損失が導入される。



(a) Metal



(b) Glass

(c) Sub-surface scattering

Fig. 5: Examples of input images

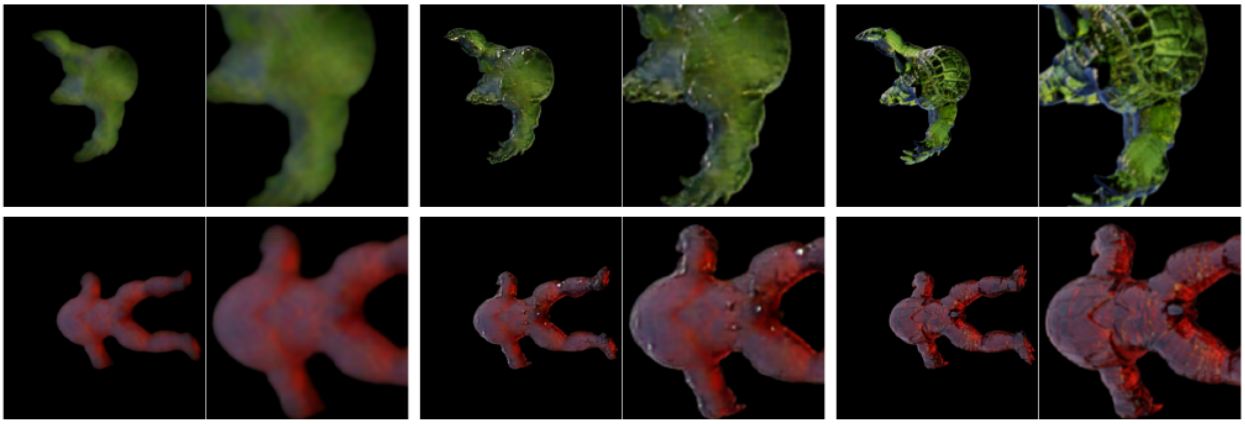
p-values for all materials are smaller than the significance level of 1%, indicating that there was a significant difference in choices between the outputs without and with adversarial loss.

6. Discussion

The experimental results showed that the method with adversarial loss performed better both in LPIPS and in the subjective evaluation by humans, while there was not much difference in PSNR and SSIM between the methods with and without adversarial loss. This study aims not to obtain optically correct output but to reproduce the material appearance of objects. Adversarial loss is introduced to reproduce the material-specific appearance, even for objects with complex optical characteristics that make accurate interpolation difficult. Therefore, the lack of significant differences in PSNR and SSIM, which make pixel-by-pixel com-

w/o adversarial loss　　w/ adversarial loss　　ground truth

(a) Metal



w/o adversarial loss　　w/ 敵対的損失　　ground truth

(b) Glass

w/o adversarial loss　　w/ 敵対的損失　　ground truth

（c）表面下散乱

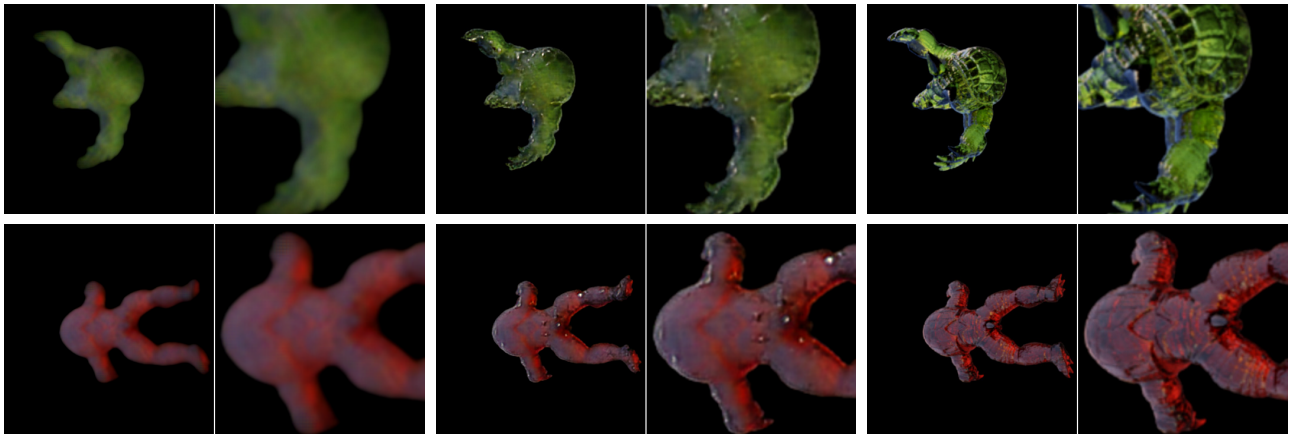
図6:4つの入力視点間の中央視点での生成画像。

したがって、画素ごとの比較を行うPSNRとSSIMに有意差がないことは予想通りであり、敵対的損失を用いた手法が人間の知覚特性に近いLPIPSにおいて有意に良好な結果を示したことは、本研究の目的が達成されたことを示している。主観評価の結果もLPIPSと同じ傾向を示している。

敵対的損失を用いない方法は、L1-Norm項のみを持つ損失関数を用いるため、画素毎の誤差を最小化することができる。

w/o adversarial loss | w/ adversarial loss | ground truth

(a) Metal



w/o adversarial loss | w/ adversarial loss | ground truth

(b) Glass

w/o adversarial loss | w/ adversarial loss | ground truth

(c) Sub-surface scattering

Fig. 6: Generated images at the center viewpoint between the four input viewpoints.

parisons, is as expected, and the fact that the method with adversarial loss showed significantly better results in LPIPS, which is closer to human perceptual characteristics, indicates that the objective of this study was achieved. The results of subjective evaluation also show the same trend as for LPIPS.

The method without adversarial loss uses a loss function with only an L1-Norm term, which means that it

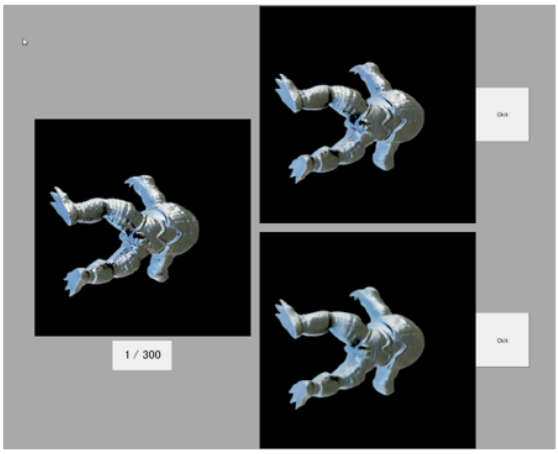


図7:参加者に見せる画面イメージ

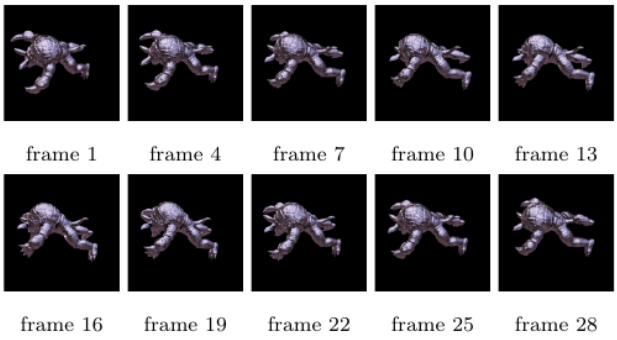


図8:実験に使用したビデオの例(金属、グランドトゥルース)

表2:主観的評価の結果である。率は、敵対的損失を持つ出力を選択した回答の割合を示す。各素材タイプについて、回答総数は1,100件であった。SSSは表面下散乱を表す。

material	rate (w/ AL)	p-value
metal	99.0 %	≈ 0
glass	99.6 %	≈ 0
SSS	55.9 %	0.0000494

PSNRとSSIMは、ピクセル値またはその統計量をグランドトゥルースと比較することで評価され、ピクセル値の類似性はこれらのスコアを最大化する。このため、金属および表面下散乱材料のPSNRとSSIMスコアがわずかに向上した可能性がある。

しかし、ガラス材料については、敵対的損失を用いた手法の方が、PSNRとSSIMのスコアがわずかに優れていた。これはおそらく、ガラス材料の外観を再現することが、金属や表面下の散乱材料に比べて特に困難であるためであろう。

敵対的損失を用いない方法は、明らかにぼやけた画像を生成し、材質的な外観を再現することができなかった。ガラス材料は、透過や屈折のため、他の材料に比べて視点の変化に伴う輝度変化が大きく、不規則になる傾向がある。したがって、このような複雑なガラス材料の輝度変化をネットワークが推定することは困難であり、正確な出力を得ることはできなかったであろう。敵対的損失の使用は、ガラスのユニークな外観を再現するのに役立ち、画素ごとの誤差の減少にもつながったようである。

LPIPSが画像分類ネットワークの中間層ベクトルから計算されるため、敵対的損失を用いた手法が全ての材料でLPIPSでより良い性能を示した理由は、敵対的損失を用いた手法によって生成された画像が、画像分類の点でよりグランドトゥルースに近いと認識されたためと考えられる。LPIPSは人間の目による評価と同様の結果を示し、実験結果もLPIPSと主観評価で同じ傾向を示している。これらの結果は、敵対的損失が人間の目で見た自然な物質の外観を再現するのに有効であることを示唆している。
定性的な分析として、敵対的損失を用いた方法は、金属材料に対して敵対的損失を用いない方法よりも、物体表面の凹凸をより細かく再現した。このことが、参加者の主観的な評価結果につながった可能性がある。ガラス素材では、敵対的損失なしと敵対的損失ありの生成画像の外観の違いが特に顕著であった。敵対的損失のない方法でも、金属材料ではハイライトがある程度再現されたが、ガラス材料では敵対的損失なしのハイライトはほとんど再現されなかった。この結果は、ガラス材料の外観を再現することの難しさを明確に示している。表面下散乱材料については、どちらの方法もかなり正確な出力を生成し、人間の目にはその違いを区別することは難しい。しかし、参加者による主観評価では、金属材料やガラス材料ほどではないものの、有意な差が見られた。評価のために視点が動くビデオ画像を見せたので、表面下散乱における敵対的損失の影響はより容易に見ることができただろう。

7. むすび

本研究では、数枚の入力画像から新しい視点画像を生成する視点補間ネットワークを用いて、複雑な光学特性を持つ材料の外観の再現を試みた。

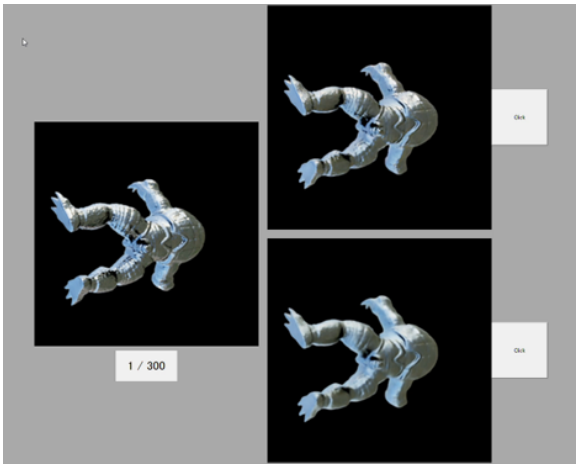


Fig. 7: A screen image shown to the participants

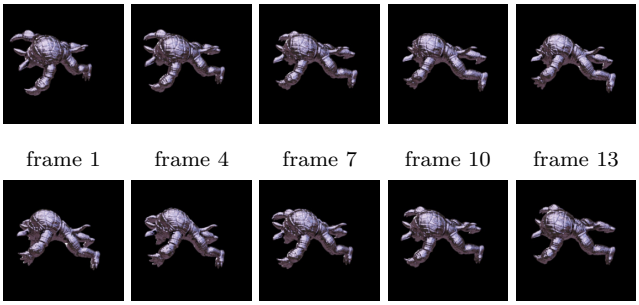


Fig. 8: An example of a video used in the experiment (metal, ground truth)

Table 2: The result of subjective evaluation. The rate indicates the percentage of the answers that chose the output with adversarial loss. The total number of answers were 1,100 for each material type. SSS stands for sub-surface scattering.

material	rate (w/ AL)	p-value
metal	99.0 %	≈ 0
glass	99.6 %	≈ 0
SSS	55.9 %	0.0000494

minimizes per-pixel errors. PSNR and SSIM are evaluated by comparing pixel values or their statistics with ground truth, and similarity of pixel values maximizes these scores. This may have led to slightly better PSNR and SSIM scores for metal and sub-surface scattering materials.

For the glass material, however, the method with adversarial loss achieved slightly better PSNR and SSIM scores. This is probably because reproducing the appearance of glass materials is particularly challenging compared to metal and sub-surface scattering materials. The method without adversarial loss obviously produced blurred images and failed to reproduce the material appearance. Glass materials tend to have larger and more irregular luminance variations with changes in viewpoint than other materials because of transmission and refraction. Therefore, it would have been difficult for the network to estimate luminance changes for such complex glass materials, and accurate output could not be obtained. It appears that the use of adversarial loss helped reproduce the unique appearance of glass, which also led to the reduction of per-pixel errors.

The reason why the method with adversarial loss performed better in LPIPS for all materials may be that the images generated by the method with adversarial loss were recognized as closer to ground truth in terms of image classification, since LPIPS is computed from the middle layer vectors of an image classification network. LPIPS gives results similar to the evaluation by the human eye, and the experimental results show the same tendency in LPIPS and subjective evaluation. These results suggest that adversarial loss is effective in reproducing natural material appearance as seen by the human eye.

As a qualitative analysis, the method with adversarial loss reproduced the unevenness of the object surface more finely than the method without adversarial loss for metal materials,. This may have led to the results of subjective evaluation by the participants. For glass materials, the difference in appearance between the generated images without and with adversarial loss was particularly noticeable. Even with method without adversarial loss, highlights were reproduced to some extent for metal materials, but for glass materials, highlights were hardly reproduced without adversarial loss. This result clearly shows the difficulty of reproducing appearance of glass materials. For sub-surface scattering materials, both methods produced fairly accurate outputs, and it is difficult for the human eye to distinguish the difference. However, the subjective evaluation by the participants showed a significant difference, although not as large as those for metal and glass materials. Since the participants were shown video images with moving viewpoints for evaluation, the effect of adversarial loss in sub-surface scattering would have been more easily seen.

7. Conclusion

In this study, we attempted to reproduce the appearance of materials with complex optical characteristics

敵対的損失を導入することで、グランドトゥルースにできるだけ近い画像を生成するのではなく、素材に特化した外観を持つ画像を再現することを目指した。金属、ガラス、表面下散乱材料を用いた実験を行い、あらゆる種類の材料について、人間の知覚に近い客観的な指標と、人間の参加者による評価の両方において、外観を再現する際の敵対的損失の有効性を確認した。

今後の課題としては、より一般的な視点補間を行う。本研究では、入出力視点は球面上の視点に限定し、入力視点は4つの特定の位置に固定した。視点配置を拡張して柔軟にすることで、視点補間による素材の外観再現をより実用的なものにすることができるだろう。入力画像数が出力画像の品質に与える影響についても調査する必要がある。

本研究では、ネットワークが物体領域に集中できるように、背景を除去した画像を入力として用いた。しかし、インターネットショッピングのような実際のアプリケーションでは、背景領域も補間する必要がある。ネットワークが物体領域と背景領域を区別するためには、物体領域をマスクとして明示的に与えることが効果的であろう。

既存の手法との比較も必要である。本研究は、従来の自由視点画像生成とは異なる目的を持ち、特殊なデータセットを用いている。また、奥行き画像は補足情報として提供されているため、既存の手法との公平な比較を行うことは困難である。したがって、比較の方法についても慎重に検討する必要がある。

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22K12088.

References

1) Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics*, Vol. 37, No. 4, pp. 1–15, 2018.
2) Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics*, Vol. 37, No. 6, pp. 1–11, 2018.
3) Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3982–3991, 2020.
4) Taishi Ono, Hiroyuki Kubo, Kenichiro Tanaka, Takuya Funatomi, and Yasuhiro Mukaigawa. Practical brdf reconstruction using reliable geometric regions from multi-view stereo. *Computational Visual Media*, Vol. 5, No. 4, pp. 325–336, 2019.
5) Alon Oring, Zohar Yakhini, and Yacov Hel-Or. Autoencoder image interpolation by shaping the latent space. *arXiv preprint arXiv:2008.01487*, 2020.
6) Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, Vol. 35, No. 6, pp. 1–10, 2016.
7) Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, Vol. 27, pp. 1–9, 2014.
8) Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 8, pp. 1932–1947, 2018.
9) Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6315–6324, 2018.
10) Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5960–5969, 2020.
11) Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of European Conference on Computer Vision*, pp. 405–421, 2020.
12) Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.
13) Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.
14) Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, Vol. 37, No. 4, pp. 1–12, 2018.
15) John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2367–2376, 2019.
16) Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 175–184, 2019.
17) Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, Vol. 38, No. 4, pp. 1–14, 2019.
18) Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, 2017.
19) Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 234–241, 2015.
20) Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.
21) Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Ima-

using a viewpoint interpolation network that generates a new viewpoint image from a few input images. By introducing adversarial loss, we aimed to reproduce images with material-specific appearance, rather than generating images as close as possible to ground truth. We conducted an experiment using metal, glass, and subsurface scattering materials and confirmed the effectiveness of adversarial loss in reproducing appearances for all types of materials both in an objective metric that is close to human perception, and evaluations by human participants.

Future work includes more general viewpoint interpolation. In this study, input and output viewpoints were limited to those on a sphere, and the input viewpoints were fixed to four specific locations. It would be possible to make reproduction of material appearance by viewpoint interpolation more practical by extending the viewpoint arrangement to be more flexible. The effect of the number of input images on output image quality should also be investigated.

In this study, images with the background removed were used as input to allow the network to focus on the object region. However, in real applications such as Internet shopping, it is necessary to interpolate the background region as well. In order for the network to distinguish between object and background regions, it would be effective to explicitly provide the object region as a mask.

Comparison with existing methods is also necessary. This study has a different objective from conventional free-viewpoint image generation and uses a specialized dataset. In addition, depth images are provided as supplementary information, making it difficult to conduct a fair comparison with existing methods. Therefore, the method of comparison should also be carefully considered.

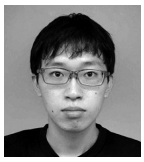
Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22K12088.

References

1) Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics*, Vol. 37, No. 4, pp. 1–15, 2018.
2) Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics*, Vol. 37, No. 6, pp. 1–11, 2018.
3) Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3982–3991, 2020.
4) Taishi Ono, Hiroyuki Kubo, Kenichiro Tanaka, Takuya Funatomi, and Yasuhiro Mukaigawa. Practical brdf reconstruction using reliable geometric regions from multi-view stereo. *Computational Visual Media*, Vol. 5, No. 4, pp. 325–336, 2019.
5) Alon Oring, Zohar Yakhini, and Yacov Hel-Or. Autoencoder image interpolation by shaping the latent space. *arXiv preprint arXiv:2008.01487*, 2020.
6) Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, Vol. 35, No. 6, pp. 1–10, 2016.
7) Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, Vol. 27, pp. 1–9, 2014.
8) Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Efstratios Gavves, Mario Fritz, Luc Van Gool, and Tinne Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 8, pp. 1932–1947, 2018.
9) Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6315–6324, 2018.
10) Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5960–5969, 2020.
11) Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of European Conference on Computer Vision*, pp. 405–421, 2020.
12) Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.
13) Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.
14) Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, Vol. 37, No. 4, pp. 1–12, 2018.
15) John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2367–2376, 2019.
16) Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 175–184, 2019.
17) Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, Vol. 38, No. 4, pp. 1–14, 2019.
18) Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, 2017.
19) Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, pp. 234–241, 2015.
20) Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.
21) Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Ima-

genet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1–9, 2012.

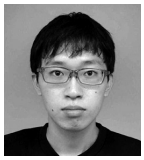


Chihiro Hoshizawa received the B.E. and M.E. degrees in information and computer sciences from Saitama University in 2022 and 2024, respectively. At present he works for IVIS Corp.

Taishi Iriyama received the B.E., M.E. and Ph.D. degrees in electronic information engineering from Tamagawa University, Tokyo, Japan, in 2017, 2019 and 2022, respectively. Currently, he is assistant professor of mathematics, electronics and informatics at Saitama University. His research interests include image processing, conputer vision and machine learning.

Takashi Komuro received the B.E., M.E., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo in 1996, 1998, and 2001, respectively. At present he is a professor of mathematics, electronics and informatics at Saitama University. His current research interests include computer vision and human-computer interaction.

genet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1–9, 2012.



Chihiro Hoshizawa received the B.E. and M.E. degrees in information and computer sciences from Saitama University in 2022 and 2024, respectively. At present he works for IVIS Corp.

Taishi Iriyama received the B.E., M.E. and Ph.D. degrees in electronic information engineering from Tamagawa University, Tokyo, Japan, in 2017, 2019 and 2022, respectively. Currently, he is assistant professor of mathematics, electronics and informatics at Saitama University. His research interests include image processing, conputer vision and machine learning.

Takashi Komuro received the B.E., M.E., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo in 1996, 1998, and 2001, respectively. At present he is a professor of mathematics, electronics and informatics at Saitama University. His current research interests include computer vision and human-computer interaction.