H-015

画像変換ネットワークによる透明物体への材質変換
Material transformation to transparent objects using an image transformation network

西長 紗奈† 増田 康希‡ 入山 太嗣‡ 小室 孝‡ 小川 賀代†
Sana Nishinaga Koki Masuda Taishi Iriyama Takashi Komuro Kayo Ogawa

1. はじめに

近年、拡張現実感(AR：Augmented Reality)のコンテンツが増加している。AR において仮想物体を現実空間に違和感なく融合させることは重要であり、これを実現させるためには、背景画像に合わせた映像重畳が必要不可欠である。

先行研究において増田ら、画像変換ネットワークを用いて、背景画像および重畳させる仮想オブジェクトの法線画像を元に、視覚的に違和感のない映り込みを生成する手法を提案した。その結果、未知周囲環境に対し視覚的に自然な映り込みを生成し、金属の質感を再現できることを確認した。[1]。

本研究ではより複雑な光学現象を有する透明物体に着目し、画像変換ネットワークを用いて、透過性のない金属、陶磁器から透過性のあるガラスへの材質変換を検討する。

2. 画像変換ネットワークによる材質変換

本研究で用いた画像変換ネットワークは Isola らが提案した pix2pix[2]に pixel shuffle と知覚損失を組み込んだものである。pix2pix は訓練画像である入力と正解のペアの画像を真偽判定することで学習を行う GAN の一種である。

2.1 ネットワーク

pix2pix の生成器には Ronneberger ら[3] が提案した U-Net を用いる。U-Net は Encoder-Decoder 構造のネットワークで、ダウンサンプリング層、アップサンプリング層、スキップコネクションからなる。逆畳み込みは画素ごとに参照される回数に差が出ることから、ノイズが生成されてしまうため、アップサンプリング層では入力特徴マップの各ピクセルを並び替えて高解像度特徴マップ生成する pixelShuffle[4] を使用する。またネットワークの識別器は、 pix2pix で用いられている識別器を使用する。

2.2 損失関数

既存の pix2pix は、生成された出力画像と正解画像のピクセル単位での差異を測定する L1 損失を採用しているが、正解画像と異なる位置に、反射や映り込みが再現されると、損失が大きくなるため、質感が失われるという問題がある。

Johnson ら[5]は画像変換において、知覚的損失がピクセル間の損失よりも優れていることを示した。そこで本研究では L1 損失の代わりに、知覚損失を採用した。特徴を抽出するためのネットワークは ImageNet で事前学習された VGG19[6]を用いる。VGG19 の指定された層(relul_2, relu3_4, relu4_4, relu5_4)における出力画像と正解画像の差の絶対値を知覚損失L_{vgg}とすると、ネットワーク全体の損失関数は式(1)のようになる。ただし、λは重みを示す。

$$L_P = \min_G \max_D L_{cGAN}(G,D) + \lambda L_{vgg} \quad (1)$$

$L_{cGAN}(G,D)$は敵対的損失であり、式(2)のように定義される。ただし、$D(x,y)$は識別器が入力と正解のペア画像を入力と正解のペアと判定する確率、$D(x,G(x,z))$は識別器が入力と出力のペア画像を入力と正解のペアと判定する確率を示す。

$$L_{cGAN}(G,D) = E_{x,y}[\log D(x,y)] + \lambda E_{x,y}\left[\log\left(1 - D(x,G(x,z))\right)\right] \quad (2)$$

3. 実験

3.1 データセットの生成

データセットは、3DCG 制作ソフトである Blender と BlenderProc[7]を用いて作成した。各データは、背景に 1 種類の全天球画像を用いて環境マップを作成し、3D オブジェクトを配置することで、真の映り込みと質感を与えている。これを正解画像として用いる。3D オブジェクトは The Stanford 3D Scanning Repository[8] で取得した、Stanford Bunny, Happy Buddha, Dragon, Armadillo, Lucy の 5 種類を用いた。また材質は Blender の Principled BSDF を用いて、ガラス、金属、陶磁器の各 3 種類を作成した。ガラスは roughness パラメータを 0、transmission パラメータを 1、ior パラメータを 1.45 に設定した。金属は roughness パラメータを 0.2、metallic パラメータを 1、ior パラメータを 1.35 に設定した。陶磁器は roughness パラメータを 0.2 に設定した。カメラの位置は半球面上で、3D オブジェクトとの距離が一定の範囲に収まるランダムな位置とし、画像のサイズは 256×256 に設定した。レンダリングエンジンは Cycles を使用した。3D オブジェクトと材質を組み合わせた画像に対し、共通のランダムな 5000 視点からレンダリングすることで各 5000 枚、合計 25000 枚作成した。データセットの作成の概要を図 1 に示す。

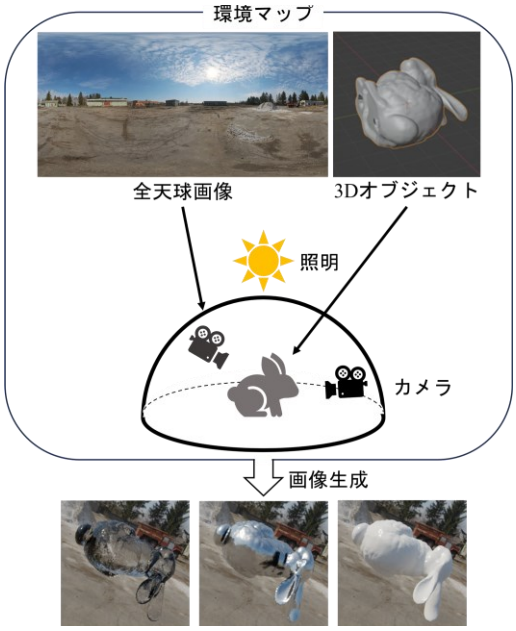

図 1　データセットの作成

3.2 画像変換ネットワークの学習

画像変換ネットワークを用いて、金属からガラスに変換する実験Ⅰと陶磁器からガラスを変換する実験Ⅱを行った。バッチサイズ 1 で 100 エポック学習を行った。最適アルゴリズムは生成器の学習率$\alpha_g = 2 \times 10^{-4}$、識別器の学習率$\alpha_d = 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.999, eps = 10^{-8}$である。また、識別器の入力画像に対してドロップアウトを適用した。ドロップアウトは 0.5 である。学習に使用した CPU は、Intel(R) Core(TM) i7-8700、GPU はNVIDIA GeForce GTX 1070である。

実験Ⅰの訓練データはテストデータに用いたもの以外の 4 種類の 3D オブジェクトの金属とガラスの画像のペアの画像 5000 セット、合計 20000 セットを使用した。テストデータの入力画像は作成したい 3D オブジェクトの金属の画像 2500 枚を使用した。よって出力画像も 2500 枚である。同様に全種類の 3D オブジェクトの結果を得るために、条件毎に pix2pix を学習させた。実験Ⅱは、実験Ⅰで用いた金属の画像を陶磁器の画像に置き換え、学習させた。

4. 生成結果

実験Ⅰ(金属からガラスへの変換)と実験Ⅱ(陶磁器からガラスへの変換)における仮想オブジェクト(a)の Dragon と仮想オブジェクト(b)の Armadillo を図 2 と図 3 に示す。いずれの出力画像において、地面や背景を透過しているとわかる。また、未知の周囲環境に対し、仮想オブジェクトの下面に空が表れ、視覚的に自然な映り込みを再現している。よって、透過や屈折など透明物体特有の光学現象が反映されており、“出力画像はガラスの画像である”と判断できる。また、5 種類全ての 3D オブジェクトに対し、出力画像がガラスの画像ように作成されており、汎用性を確認した。

実験Ⅰと実験Ⅱを比較すると、実験Ⅰのほうが、反射や透過が鮮明であり、よりガラスの質感を再現しているようにみえる。これは陶磁器より金属のほうが法線による輝度変化が大きいからであると考察した。

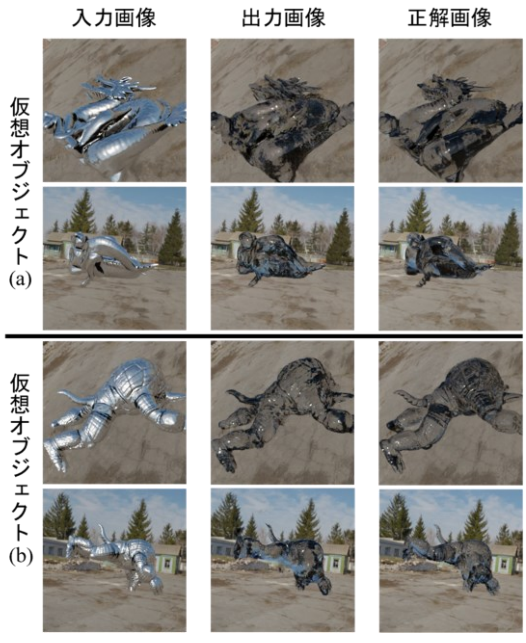

図 2　実験Ⅰ(金属からガラスへの変換)の結果

図 3　実験Ⅱ(陶磁器からガラスへの変換)の結果

5. まとめと今後の課題

本研究では複雑な光学現象を有する透明物体に着目し、画像変換ネットワークを用いて材質変換を行う手法を提案した。提案手法を用いて、視覚的に自然な映り込みを作成したガラスの質感を再現することができた。

今後の課題として、一部の画像に発生したもや状のアーチファクトを抑えるために、適切な学習の回数を検討する必要がある。また、1 種類だった背景の種類を増やすことで、未知の周囲環境に対し、汎用的な映り込みなどが再現できるかを検討する。

参考文献

[1] 増田 康希, 入山 太嗣, 小室 孝, “画像スタイル変換を用いた映り込みによる金属の質感再現”, 第 26 回 画像の認識・理解シンポジウム (MIRU 2023) Extended Abstract集, IS2-65 (2023)
[2] Phillip I, Jun-Yan Z, Tinghui Z, Alexei A. E, “Image-To-Image Translation with Conditional Adversarial Networks”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p1125–1134 (2017)
[3] Olaf R, Philipp F, Thomas B, “U-net: Convolutional networks for biomedical image segmentation”, in International Conference on Medical image computing and computer-assisted intervention, Springer, p234–241 (2015)
[4] Shi W, Caballero J, Huszár F, Totz J, Aitken A. P, Bishop R, Rueckert D, Wang Z, “Realtime Single Image and Video Super-resolution Using an Efficient Sub-pixel Convolutional Neural Network”, IEEE Conference on Computer Vision and Pattern Recognition, p1874–1883 (2016)
[5] Johnson J, Alahi A, Fei-Fei L, “Perceptual Losses for Real-time Style Transfer and Super-resolution”, Computer Vision ECCV 2016: 14th European Conference, Proceedings, Part II 14, Springer, p. 694-711 (2016)
[6] Simonyan K, Zisserman A, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, International Conference on Learning Representations (2015)
[7] D Denninger, I Lambourne, A Handa, W Yin, N Fox, A. Davison, "BlenderProc: Reducing the Reality Gap with Photorealistic Rendering", https://github.com/DLR-RM/BlenderProc, (Accessed: June 9, 2024)
[8] Stanford University Computer Graphics Laboratory, "The Stanford 3D Scanning Repository", http://graphics.stanford.edu/data/3Dscanrep/, (Accessed: June 9, 2024)

† 日本女子大学大学院理学研究科　Graduate School of Science, Japan Women’s University
‡ 埼玉大学大学院理工学研究科　Graduate School of Science and Engineering, Japan Saitama University