観光地のリアルタイム状況説明システムの検討

Real-Time Situation Description System for Tourist Spots

河中 昌樹 † 冨田 周作 †⋆ 中村 優吾 ‡ 諏訪 博彦 †⋆ 安本 慶一 †⋆
Masaki Kawanaka Shusaku Tomita Yugo Nakamura Hirohiko Suwa Keiichi Yasumoto

1. はじめに

近年，スマートフォンなどのモバイル端末の普及に伴う情報通信技術 (ICT: Information and Communication Technology) の活用は，日常生活の様々な場面に利便性をもたらしている．観光分野でも，スマートフォンを用いて様々な情報を提供する観光支援を行うシステム [1, 2] が提案されている．このようなシステムでは，リアルタイムの観光地情報が必要であり，観光地の情報収集を行うゲーミフィケーションを用いた参加型センシング手法 [3] が提案されている．この参加型センシングにおいては，個人が撮影した写真などをそのまま共有する手法が採用されている．しかしながら，この場合，プライバシ保護の観点からサーバに個人的な写真や音声などのデータを共有することとなりプライバシー保護の観点からリスクを伴う．そこで，プライバシ保護をしながらをリアルタイムに情報共有するために，画像データを直接共有するのではなく，画像に含まれるコンテキストのみを共有することを検討する．

検討技術として，一次データを直接共有するのではなく，個々に学習された機械学習モデルの重みパラメタのみを交換・統合する Federated Learning に注目する [4]．冨田ら [5] は，観光ビッグデータの収集を目的に，観光地において撮影した画像に写っているオブジェクト群のリストを出力するモデルをユーザの端末間の Federated Learning により学習する手法を提案している．このモデルにより，観光客は自身の写真を提供することなく，写真をモデルに入力して得られる出力（オブジェクトのリスト）のみを参加型センシングプラットフォームに提供することで、プライバシーアウェアな情報提供が可能になる．

本研究では，観光客が撮影したリアルタイム画像からコンテキストを抽出し，そのコンテキストから生成される画像を各観光地のベース写真に重畳して表示する手法を提案する．提案手法では，ユーザが撮影した観光地の写真に冨田らの研究 [5] で構築されたモデルを適用することで得られる観光地のオブジェクトリストからコンテキストを抽出する．さらに，環境センサから得られた情報も追加したうえで，観光地を代表するテンプレート画像の画像変換を行い，リアルタイムな観光地の状況を説明する画像を生成する．具体的には，テンプレート画像にセマンティックセグメンテーションを適用することで，テンプレート画像に存在する物体ごとにクラス分けを行う．その後，それぞれのクラスごとに観光地のオブジェクト群や環境センサの情報を用いて画像変換を行う．

本稿の構成は以下の通りとする．第 2 章では，画像変換に必要な技術であるセマンティックセグメンテーションや画像変換技術について述べる．第 3 章では，提案システムの目的と想定環境，アプローチを述べる．第 4 章では画像変換アルゴリズムについて検討し，第 5 章では本稿のまとめと今後の展望について述べる．

2. 関連研究

本研究では，テンプレート画像の物体ごとに適切な画像変換を行うことで，観光地を代表するテンプレート画像から現在の状況を説明する画像を生成する．そこで，本章では，画像内の特徴や意味が類似した部分領域に画像を分割するセグメンテーションと画像変換技術について説明を行う．その後，関連研究の課題について述べる．

2.1 セグメンテーション

画像内の特徴や意味が類似した部分領域に画像の分割を行う処理はセグメンテーションと呼ばれる．セグメンテーションは，セマンティックセグメンテーション (Semantic Segmentation)[6]，インスタンスセグメンテーション (Instance Segmentation)[7]，パノプティックセグメンテーション (Panoptic Segmentation)[8] に分類される．セマンティックセグメンテーションは，同じクラスの物体が隣接する場合に，それらの物体を個別に認識することができない．一方，インスタンスセグメンテーションは，同じクラスの物体が隣接する場合に，それらの物体を個別に認識することが可能である．パノプティックセグメンテーションは，上記の 2 つのセグメンテーションを組み合わせたものである．本研究では，同じクラスの場合には同様の画像変換を行うため，セマンティックセグメンテーションを利用する．

2.2 セマンティックセグメンテーション

機械学習を用いたセマンティクセグメンテーションとして最も広く認知されているものに Fully Convolu-

† 奈良先端科学技術大学院大学，Nara Institute of Science and Technology
‡ 九州大学，Kyushu University
⋆ 理化学研究所, RIKEN

tional Network (FCN) がある [6]．FCN は，全結合層を持たずに畳み込み層のみで構成されており，入力画像の pixel 単位でどのクラスに属するかの分類を行う．その後，Encoder-Decoder 構造を用いた FCN である Segnet[9] や FCN で有効性が示された skip connection を更に用いた U-net[10] を用いたセマンティックセグメンテーションが提案されている．

2.3 GAN を用いた画像変換

Generative Adversarial Networks (GAN) は入力画像に対応する出力を生成する Generator とその入出力のペアが本物かの判別を行う Discriminator を競わせながら学習を行う手法である [11]．GAN は，従来の画像生成モデルと比較して鮮明で本物らしい画像を生成することが可能な手法である．GAN は，入力画像を任意の画像に変換することを目的とした画像変換にも利用されている [12, 13, 14]．

2.4 pix2pix

Isola らが提案した pix2pix は，ペアとなる訓練データの対応関係を学習する汎用的な Conditional GAN である [12]．Conditional GAN は，生成したい画像のラベル情報を入力に加えることで出力する情報を制御する方法である．pix2pix の Generator は U-net 構造であり，Discriminator は PatchGAN である．PatchGAN は入出力のペアを N×N に分割し，分割した画像の入出力のペアごとに本物かを判別する．pix2pix は，従来の画像生成モデルの問題点であった出力画像がぼやける点を解消している．

pix2pix の目的関数 G^* は，Genertor を G，Discriminator を D とすると以下のように表される．

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{1}$$

ここで，$\mathcal{L}_{cGAN}(G, D)$ は Adversarial Loss，$\mathcal{L}_{L1}(G)$ は L1 損失，λ は重み係数を示す．入力画像を x，教師画像を y，ランダムノイズを z とすると $\mathcal{L}_{cGAN}(G, D)$ と $\mathcal{L}_{L1}(G)$ は以下のように表される．

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) =& \mathbb{E}_{x,y}[\log D(x, y)] + \\ & \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))] \end{aligned} \tag{2}$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right] \tag{3}$$

ここで，$\mathbb{E}[\cdot]$ は期待値，$\|\cdot\|_1$ は L_1 ノルムを示す．

2.5 StarGAN

Choi らによって提案された StarGAN は，異なる複数のドメインに画像変換を行う GAN であり，入力画像の男性の髪色や表情などを変化させることが可能である [14]．StarGAN は異なるドメインに画像変換を行う CyclyGAN[13] と Conditional GAN を組み合わせた手法である．入力画像とドメイン情報を含むラベルベクトルを取り込むことで 1 つの Generator を用いて画像の変換を行う．

StarGAN の Discriminator と Generator の目的関数 L_D と L_G はそれぞれ以下のように表される．

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls} \tag{4}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{\text{rec}}\,\mathcal{L}_{\text{rec}} \tag{5}$$

ここで，$\mathcal{L}_{adv}$ は Wasserstein GAN の目的関数に Gradient Penalty を適用した損失関数，$\mathcal{L}^r_{cls}$ は Discriminator の Domain Classification Loss，$\mathcal{L}^f_{cls}$ は Generator の Domain Classification Loss，$\mathcal{L}_{\text{rec}}$ は Reconstruction Loss，λ_{cls} と λ_{rec} は重み係数を示す．ドメインラベルを c，x の元のドメインラベルを c' とすると，$\mathcal{L}_{adv}$，$\mathcal{L}^r_{cls}$，$\mathcal{L}^f_{cls}$，$\mathcal{L}_{\text{rec}}$ はそれぞれ以下のように表される．

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_x\left[D_{src}(x)\right] - \mathbb{E}_{x,c}\left[D_{src}(G(x, c))\right] \\ & - \lambda_{gp}\mathbb{E}_{\hat{x}}\left[\left(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1\right)^2\right] \end{aligned} \tag{6}$$

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}\left[-\log D_{cls}\left(c' \mid x\right)\right] \tag{7}$$

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c}\left[-\log D_{cls}(c \mid G(x, c))\right] \tag{8}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}\left[\|x - G\left(G(x, c), c'\right)\|_1\right] \tag{9}$$

ここで，$\|\cdot\|_2$ は L_2 ノルムを示す．

2.6 関連研究の課題

セマンティックセグメンテーションを行う際の問題点として，観光地に特化したクラス分けが行われていない点が挙げられる．公開されているセマンティックセグメンテーション用のデータセットの多くは，クラスの分類が細分化されていることが多い．そのため，本研究のように同じ傾向で画像変換を行う場合には，事前に同じ画像変換を行う物体でクラスを統合する必要がある．また，画像変換のデータセットとして，全ての観光地にカメラを設置し，日時や混雑度などが違う画像として収集し，アノテーションを行うことは現実的ではない．そのため，既存の画像変換に用いられているデータセットを用いて画像変換を行う必要があるが，テンプレート画像から現在の情報を一度に反映させる画像変換を行うためのデータセットは存在しない．そのため，観光地のテンプレート画像から現在の状況を説明する画像を生成するためには，セマンティックセグメンテーションによって得られたクラスごとに画像変換を行う機械学習モデルを構築し，それらの画像を最終的に統合する必要がある．

3. 提案システムの概要と想定環境

本章では，提案システムの目的と想定環境について述べる．提案するシステムの概要図を図 1 に示す．

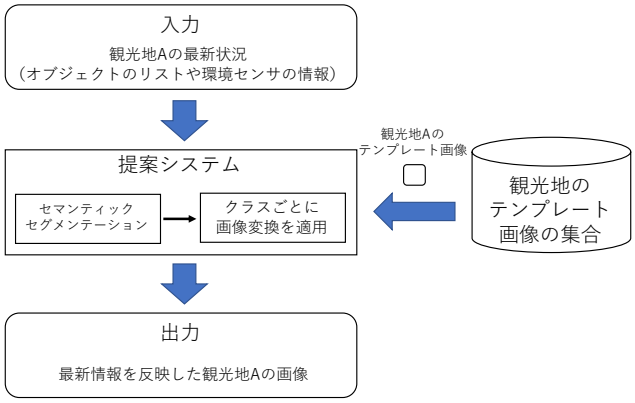

図 1: 本システムの概要図

3.1 提案システムの概要

提案システムは，観光オブジェクト群や環境センサから得られた情報を用いて，テンプレート画像を画像変換することで観光地のリアルタイムな状況を示す画像の生成を行う．従来の観光地の情報を提供するシステム [15] では，事前に撮影された写真を用いて観光地について説明を行っている．しかし，実際に観光地を観光すると提供された情報との乖離が発生する可能性がある．事前に撮影された画像の代わりに，ユーザが撮影した写真を用いる方法が考えられるが，プライバシーの問題がある．そこで，提案システムでは，ユーザが撮影した観光地の写真からユーザの端末内で自動生成した観光地のオブジェクトリスト（冨田らの手法 [5] を使用）を用いて現在の状況を示す画像を生成することで，提供する写真と実際に観光した際の状況の乖離を小さくする．

3.2 想定環境

本研究では，観光地を代表するテンプレート画像を画像変換することで観光地の現在の状況を説明する画像の生成を行うことを想定する．そのため，観光地を代表するテンプレート画像の他に，現在の状況を説明する観光地のオブジェクトのリストや環境センサから得られる情報が必要である．テンプレート画像は観光地ごとに事前に用意する必要がある．観光オブジェクトのリストには，樹木などの固定されたオブジェクトの情報や群衆や動物や雲などの移動するオブジェクトが含まれる．環境センサでは，天気や群衆などのリアルタイムの情報を得ることが可能であると想定する．

3.3 提案システムのアプローチ

本研究のアプローチを以下に示す．

1. 観光エリアを代表するテンプレート画像にセマンティックセグメンテーションを適用
2. 観光オブジェクトのリストと環境センサの情報を取得
3. これらの情報と GAN を用いて画像を変換

まず，観光エリアを代表するテンプレート画像にセマンティックセグメンテーションを適用し，特徴や意味が類似した複数のクラスに画像を分割する．ここでのクラスは，適切な画像変換を行うことが可能になるように画像の分割を行う．例えば，樹木や芝生はセグマンティックセグメンテーションでは別のクラスに分類されることが多いが，本研究では，樹木も芝生は季節などによって同じ傾向で変化するため，同じクラスとしてクラス分類を行い，同様の画像変換を適用する．次に，観光オブジェクト認識モデル [5] を用いて，ユーザが撮影した写真から観光オブジェクトのリストを取得する．また，環境センサから混雑度や天気などの情報も同様に取得する．最後に，観光エリアを代表するテンプレート画像のクラスごとに適切な観光オブジェクトのリストと環境センサの情報と GAN を用いて画像変換を行う．

4. アルゴリズムの検討

本章では，3.3 のアプローチを実現するためにセマンティックセグメンテーションとそれぞれのクラスごとに適用する GAN を用いた画像変換について検討を行う．

4.1 セマンティックセグメンテーションにおけるパネル分類

本章では，奈良県を想定環境とし，奈良公園，春日大社，東大寺を観光エリアとする．観光エリアを代表するテンプレート画像を図 2 に示す．図 2 に示す画像に対して，セマンティックセグメンテーションを適用することで，特徴や意味が類似したクラスごとに画像を分割する．データセットには，Stanford Background Dataset [16] を利用した．本データセットは，715 枚の画像で構成され，それぞれのラベルは Amazon's Mechanical Turk を用いてラベル付けが行われている．画像のピクセルごとに，空，木，道，草，水，建物，山，前景のオブジェクトのラベルが付与されている．本システムでは，これらのラベルを用いて 4 つのパネルを作成し，それぞれに画像変換を行う．ここで示すパネルは，同様の方法を用いて画像変換を行うクラスをまとめたものとする． 1 つ目のパネルは，天気パネルであり，空ラベルを利用する．2 つ目のパネルは，季節パネルであり，木，草，水，山ラベルを利用する．3 つ目のパネルは混雑度ラベルであり，道ラベルを利用する．最後のパネルは物体パネルであり，建物ラベルと前景のオブジェクトラベルを利用する．機械学習のモデルは U-net，目的関数はクロスエントロピー，最適化手法は Adam[17] を利用した．また，バッチサイズは 16 とし，学習回数は 100 epoch とした．図 2 の画像をセマンティックセグメンテーションを用いて変換した結果の一部を図 3 に示す．図 3 より，天気

（a）奈良公園

（b）春日大社

（c）東大寺

図 2: 奈良県を代表する観光エリアの画像

パネルは概ね正しく分類できていることが分かる．しかし，季節パネルでは，芝生部分を物体パネルと誤認識していることが分かる．また，混雑パネルでは，道路の模様によっては物体パネルと誤認識している．最後に，物体パネルは，建物については概ね分類できているが，鹿などの前景のオブジェクトは正しく分類できていないことが分かる．これは，学習用のデータセットに鹿が存在しないため，何と分類すれば良いか分からないためだと考えられる．本実験では，データセットの量が少なかったため，学習が十分に行われていない可能性が考えられる．この問題に対しては，鹿のデータの追加や画像の回転やノイズを注入するなどの Data Augmentation[18] を用いることで分類精度が向上する可能性がある．

4.2 テンプレート画像の画像変換手法の検討

本研究では，テンプレート画像のパネルごとに画像変換を行うことで観光地の現在の状況を説明する画像を生成する．冨田らの構築したモデル [5] を用いて，ユーザが撮影した写真から得られる観光地のオブジェクトリストには，観光地の風景や写っている観光者の人数から，現在の天気，季節，混雑度などの情報を得ることができることを想定する．これらの情報と環境センサから得られる現在の情報を用いて，現在の状況を説明する画像の生成を行う．

(1) **天気パネル:** 天気パネルでは，晴れ，曇りの 2 つの状態を変化させることで現在の観光地の天気に近い画像の生成を行う．学習用のデータセットには，晴れと曇りの画像がそれぞれ 5000 枚ずつで構成されている Weather Image Dataset[19] を利用する．本データセットは，対応するペア画像は存在しないため，pix2pix を用いて学習することが困難である．そのため，異なるドメインに画像変換を行う CyclyGAN と太陽光センサや観光オブジェクト群の雲の情報を用いて画像変換を行う．

(2) **季節パネル:** 季節パネルでは，春，夏，秋，冬の 4 つの状態を変化させることで現在の観光地の季節に近い画像の生成を行う．文献 [13] では，夏と冬の画像の変換を目的とした画像変換用のデータセットとして，Flickr API を用いて，yosemite と datataken をタグ指定し画像のダウンロードを行っている．本システムでは，夏と冬の画像だけではなく，春と秋のデータについても datataken を変更することで入手し，構成されたデータセットを用いてモデルの学習を行う．その後，異なる複数のドメインに画像変換を行う必要があるため，StarGAN と観光オブジェクト群から植物の色の情報を用いて画像変換を行う．

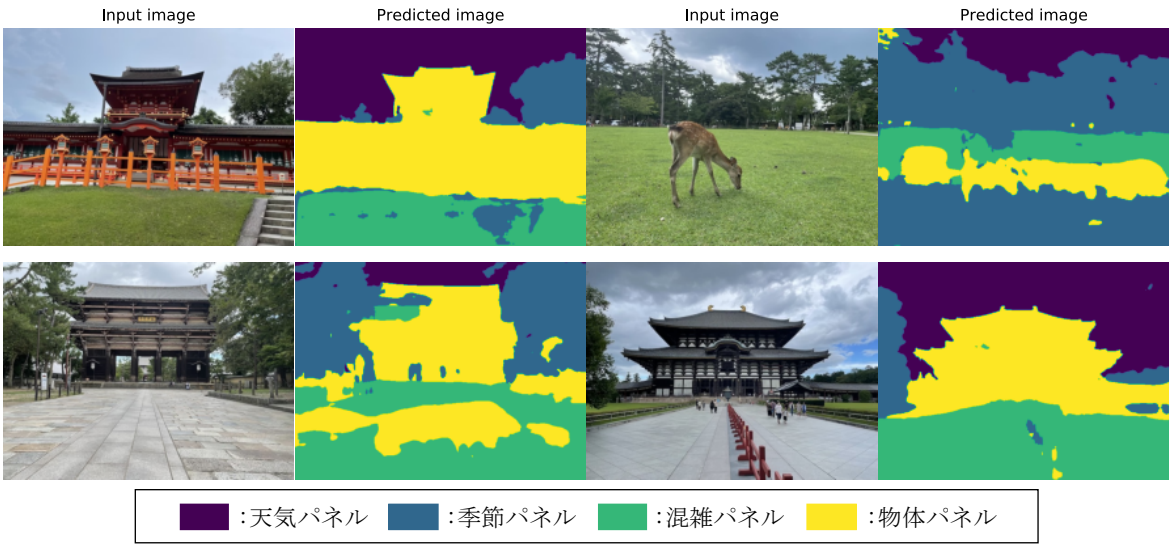

図 3: セマンティックセグメンテーションの結果

(3) **混雑度パネル:** 混雑度パネルでは，人間の人数を変化させることで現在の観光地の混雑度に近い画像の生成を行う．学習用のデータセットは，画像内の物体を除去する深層学習 [20] を用いて，人間の人数が異なる対応あるデータセットを作成を行い，pix2pix の学習を行う．その後，pix2pix と観光オブジェクト群と環境センサの混雑度データを用いて混雑度を表現するように画像変換を行う．

(4) **物体パネル:** 物体パネルは，時間の流れによって急激に変化するものではないため，本システムでは画像変換を行わない．

(5) **全体パネル:** 全体パネルでは，すべてのパネルに共通する明るさについて画像変換を行う．学習用のデータセットには，101 個のウェブカメラで撮影された，時間の経過によって明るさの異なる画像が含まれる 8571 枚のデータセットを利用する [21]．本システムでは，明るさが異なる対応あるデータを用いて pix2pix の学習を行う．その後，pix2pix と環境センサの光情報を用いて適切な明るさに画像変換を行う．

5. おわりに

本稿では，観光オブジェクト認識モデルが構築されていると仮定し，観光地の現在の状況を説明するシステムの検討を行った．想定環境として，奈良県を対象とし，その中で春日大社，奈良公園，東大寺を対象観光エリアとし，セマンティックセグメンテーションを用いたパネル分類や観光エリアごとのテンプレート画像の画像変換手法について検討を行った．今後の研究では，セマンティックセグメンテーションの性能の向上や画像変換手法の実装を行い，観光地の現在の状況を説明する画像の生成が可能であるか検討する予定である．

参考文献

[1] Shogo Isoda, Masato Hidaka, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Timeliness-aware on-site planning method for tour navigation. *Smart Cities*, Vol. 3, No. 4, pp. 1383–1404, 2020.

[2] Masato Hidaka, Yuki Kanaya, Shogo Kawanaka, Yuki Matsuda, Yugo Nakamura, Hirohiko Suwa, Manato Fujimoto, Yutaka Arakawa, and Keiichi Yasumoto. On-site trip planning support system based on dynamic information on tourism spots. *Smart Cities*, Vol. 3, No. 2, pp. 212–231, 2020.

[3] Shogo Kawanaka, Yuki Matsuda, Hirohiko Suwa, Manato Fujimoto, Yutaka Arakawa, and Keiichi Yasumoto. Gamified participatory sensing in tourism: An experimental study of the effects on tourist behavior and satisfaction. *Smart Cities*, Vol. 3, No. 3, pp. 736–757, 2020.

[4] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 54, , 2017.

[5] Shusaku Tomita, Yugo Nakamura, Hirohiko Suwa, and Keiichi Yasumoto. 観光オブジェクト認識モデルのユーザ参加型構築手法の提案. *DICOMO*, 2021.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[7] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4438–4446, 2017.

[8] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9396–9405, 2019.

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 12, pp. 2481–2495, 2017.

[10] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Vol. 9351 of *LNCS*, pp. 234–241. Springer, 2015.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, 2018.

[15] 阪口大弥, 泉朋子, 仲谷善雄. 場の雰囲気にもとづく散策観光支援システム. 情報処理学会 全国大会 講演論文集, 2015.

[16] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1–8, 2009.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

[18] Shorten C. and Khoshgoftaar T.M. A survey on image data augmentation for deep learning. *J Big Data*, 2019.

[19] Cewu Lu, Di Lin, Jiaya Jia, and Chi-Keung Tang. Two-class weather classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 12, pp. 2510–2524, 2017.

[20] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Computer Vision – ECCV 2020*, pp. 1–17. Springer International Publishing, 2020.

[21] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Trans. Graph.*, Vol. 33, No. 4, 2014.