深層生成モデルを用いた画像の周辺補完
～通常画角から 360 度へ～

2022 年度

秋 本　直 郁

学位論文　博士(工学)

深層生成モデルを用いた画像の周辺補完
～通常画角から360度へ～

2022年度

慶應義塾大学大学院理工学研究科

秋本 直郁

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究背景

スマートフォンを始めとする電子デバイスの進歩と普及によって，今や誰もがカメラを携帯し，日々多くの写真や動画が撮影されている．このような撮影デバイスの多様化に加え，それらを表示する表示デバイスもまた多様化している．例えば，撮影デバイスはスマートフォン，ミーラーレス一眼カメラ，アクションカメラなどで，表示デバイスはスマートフォン，テレビ，タブレット，PC モニター，ヘッドマウントディスプレイなどである (図1.1)．多様化された表示デバイスにおいて最適に映し出される画像の縦横比はそれぞれ異なる．例えば，ミラーレス一眼は 3:4 の写真が一般的な比率である一方で，PC モニターは 16:9 が一般的な比率であるため，その写真を見切れないように表示しつつ，その写真で PC モニターの画面を覆い尽くすことはできない．広告分野でのデジタルサイネージでは，無駄なスペースが発生すると存在のアピールが弱まり，エンタメ分野でのヘッドマウントディスプレイでは，表示されない領域の存在は没入感覚の阻害要因となる．したがって，これらの課題を回避するために，画像に対してその幅を調整することが頻繁に行われる．

画像の幅を調整する方法は，次の 3 つのアプローチに大別することができる．異なる画像で空スペースを埋める方法，リサイズによって幅を調整する方法，そして，画像の周辺を補完する方法である．まず，異なる画像で空スペースを埋める方法は，例えばテレビ番組が古い写真を放送するときに，その写真を拡大しブラーをかけて画面を埋める方法に相当する．しかし，不足領域を補っていることが明白にわかる．一方，リサイズによって幅を調整する方法は，単純なスケールや画像内部のピクセ

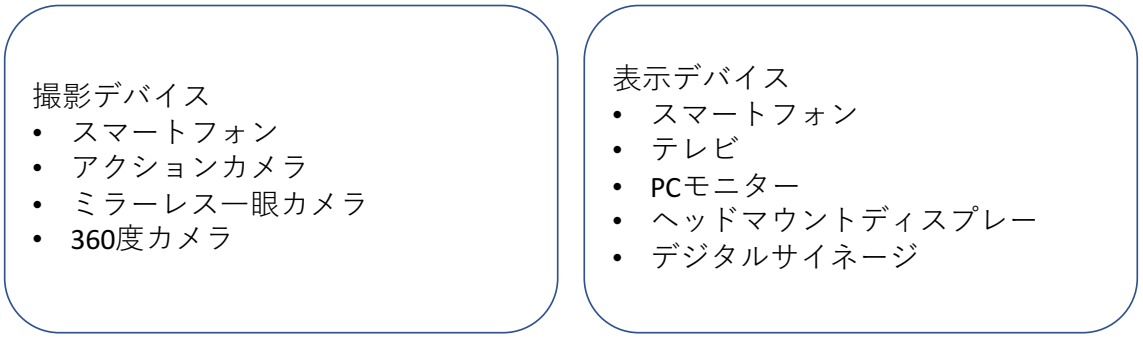


図 1.1 多様な撮影デバイスと多様な表示デバイス.

ルを補間する手段での実現があるが，画像に映った内容物 (コンテンツ) に影響を与えてしまう．最後に，画像の周辺を補完するとは，元のコンテンツに一切の変更を加えずに，画像を拡張することに相当する．

本研究では，深層生成モデルを用いた画像の周辺補完による画像拡張に取り組む．具体的には，与えられた画像に対して，整合性のあるピクセルを画像の端に生成するという問題である．この画像の周囲のピクセルを補完する問題は Image outpainting と呼ばれる問題設定である．この問題の応用例には，表示デバイスに合わせたコンテンツの自然な表示や，広告等のデザイナーに対する制作の支援がある．

画像の周囲を補完することとは，撮影者を中心として，見えている範囲を拡大した画像を得ることである．図1.2のように，見える範囲の拡大を繰り返すと，最終的には撮影者を中心として 360 度の周囲が映し出されることになる．つまり，画像の周辺補完の終着点は 360 度の景観を補完することである．

したがって本研究の後半では，深層生成モデルを用いて入力画像の周辺を補完し，360 度画像を生成する問題にも取り組む．具体的には，与えられた一枚の通常画角相当の画像に対して，整合性のあるピクセルを生成し，通常画角の画像の周辺を補完することによって 360 度画像を得る問題である．この問題の応用例には，Virtual Reality (VR) の通信容量の削減や環境マップの効率的な制作がある．

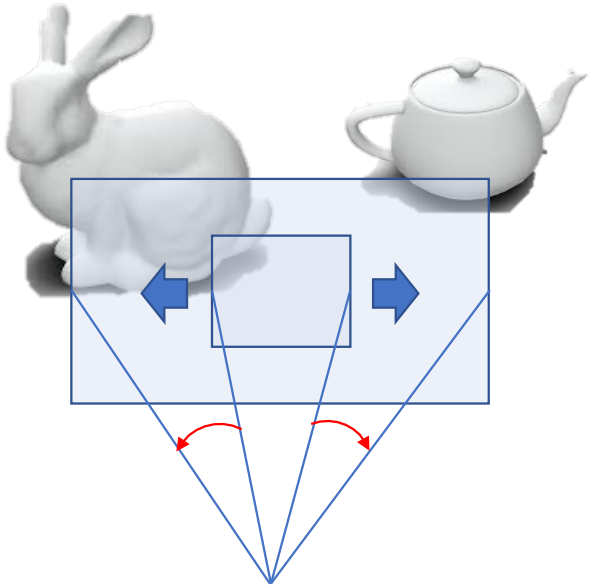


図 1.2 画像拡張と 360 度画像補完の関係性．画像拡張は視野を広げることに相当し，繰り返すと最終的には 360 度の景観を映し出すことになる．

1.2 問題設定・研究目的

本論文では，Image outpainting による画像拡張と 360 度画像補完に取り組む．いづれの問題においても，実利用可能な生成方法と出力結果を得られることが最大の目標である．以下では，実利用に近づけるという方向性に沿って，本研究での各問題に対する目的を示す．

1.2.1 Image outpainting による画像拡張

Image outpainting による画像拡張は，入力された画像の外側のピクセルを補完することで画像を拡張する，という問題である．以下では，理想的状況，既存研究の課題，そして目的設定について述べる．

入力画像と整合性を持つこと，そして自然な見た目のピクセルが補完されることが必要である．また，デザイン工程での利用を考慮すると，デザイナーの意図が反映されるという意味で，生成されるコンテンツをコントロール可能であることが理想である．

既存の研究では，深層生成モデルを利用して画像内の一部の領域を補う画像補完 (Image inpainting) が研究されてきたが，この問題は画像の外側を補う外挿問題である点が異なる．そのため，Image outpainting に特化した手法が提案されている．例えば，生成される領域のピクセルが破綻しないように，学習安定化のためにセマンティックロスを利用し，その生成ピクセルの質を向上させた手法 [1] がある．より長い画像拡張を行うための機構を提案した手法 [2] では，リピート感の強いピクセルが生成されるため，リピート感があっても違和感の少ない山脈風景画像に限定される．

以上のように，Image outpainting による画像拡張は未だ十分に検討されておらず，課題の原因も明らかになっていない．したがって，本研究の目的を，(1) 既存手法の問題点の考察と，(2) 自然な見た目かつコントロール可能な生成，と設定する．

1.2.2 360 度画像の Image outpainting

Image outpainting による 360 度画像補完は，通常画角の画像に相当する 360 度画像の一部分からその全体を生成する，という問題である．以下では，理想的状況，既存研究の課題，そして目的設定について述べる．

Three-dimensional Computer Graphics (3DCG) の環境マップの効率的な制作という応用事例を想定すると，任意の解像度が扱え，また，多様な出力が得られることが望ましい．多様な出力によって，ユーザーは制作に利用する環境マップの候補を複数得ることができる．さらに，入力画像と整合性があり，自然な見た目であることも必要である．

既存の 360 度画像補完の手法は，特定の画像解像度に過度に適合してしまい，学習時と異なる解像度での生成結果の質が下がるという課題がある．また，一つの入力画像に対して，一つの出力を得るという決定的な手法 [3, 4] か，Variational Autoencoder (VAE) からシンメトリーのタイプをサンプリングする手法 [5] を採用しており，多様な結果を得ることができないという課題がある．

したがって，本研究の目的を，(1) 尤もらしい見た目の結果を得る生成と，(2) 任

意の解像度の入力画像で機能し，(3) 多様な結果を得る生成を実現すること，による改善を通して実利用へ近づけることと設定する．

1.3 本論文の構成

第 1 章では，まず初めに本研究の背景について述べた．次に，扱う 2 つの問題設定について述べた．第 2 章では，それらの関連研究と前提となる技術についての説明を行う．第 3 章では，Image outpainting による画像拡張について述べる．これの主な内容は，文献 [6] で発表したものである．第 4 章では，360 度画像の Image outpainting について述べる．これの主な内容は，文献 [7] で発表したものである．この文献 [7] は，従前の取り組みである文献 [3, 8] の発展アプローチの検証に相当する．第 5 章では，本論文の結論について述べる．

第2章　関連研究・基礎技術

2.1　Generative Adversarial Networks

Generative Adversarial Networks (GANs) [9] は，ディープニューラルネットワークの一種である. 特に，Generator と Discriminator の二つのネットワークを使用する構成が特徴である．この GANs が適用される領域は様々であり，動画像や音声や言語と多岐にわたる．その中でも特に，画像生成分野の進展に GANs は大きく貢献している．その代表的な手法の分類を図2.1に示し，以下では，条件なし設定，条件付き設定，画像間変換について概説する．

条件なし設定 (Unconditional setting）は，GANs に畳み込み層を組み込んだ DCGAN [10] や，図2.2のようにリアリスティックな生成を行える StyleGAN [11] のような，ノイズベクトルから画像を生成する設定が分類される．GANs は，真のデータを $p_r(x)$ とし，そこからサンプリングされるデータ x と，ノイズ $p_z(z)$ からサンプリングされる z を用いて，

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.1}$$

と表される目的関数を最適化する．その結果，真のデータ分布に一致する分布を与える Generator G を学習する. これに対して，Discriminator D を一種の損失関数と見なすことで，これらは単に

$$y = G(z) \tag{2.2}$$

と表現することができる．ノイズ z は，正規分布からサンプリングされることが多

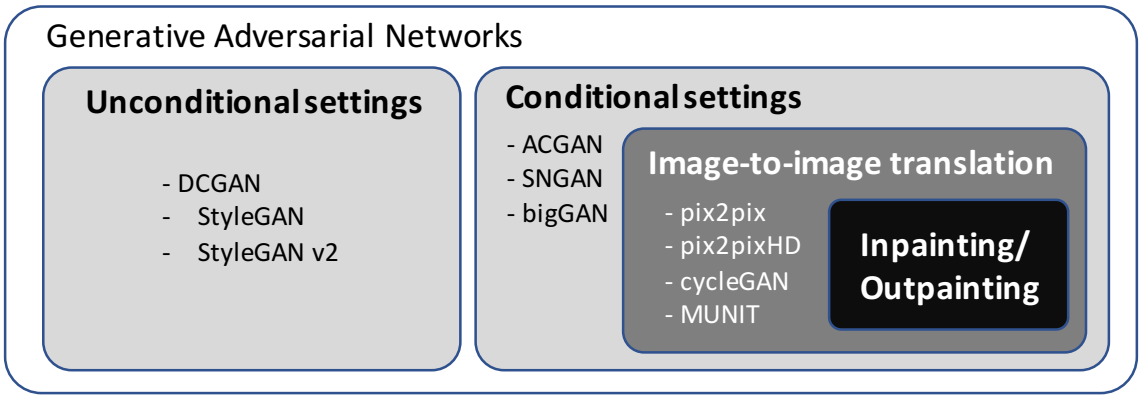


図 2.1 GANs の分類.

図 2.2 StyleGAN による生成例．1024×1024 の解像度で，非実在の顔である．[11] より引用.

く，y は生成画像を意味する．

条件付き設定（Conditional setting）は，物体のカテゴリークラスを表す符号やヒントとなる画像を条件として画像を生成する ACGAN [12], SNGAN [13], bigGAN [14] などが分類される．条件を c とすると，

$$y = G(z|c) \tag{2.3}$$

である．

画像間変換 (Image-to-image translation) は，この条件付き設定の一種である．入出力のペア画像を学習データに用いる pix2pix [15]，pix2pixHD [16] や，ペア無しで変換器を学習する手法に CycleGAN [17] がある．これらは入力画像 x を出力画像 y に

変換するものであるから,

$$y = G(x) \tag{2.4}$$

と表される.

このペア有り画像間変換が適用される問題設定の一つが画像補完である.

2.2 Transformer

Transformer [18] は，自然言語処理の一つである機械翻訳で圧倒的な性能を発揮したモデルとして提案されたディープニューラルネットワークの一種である．そしてTransformer は今や，物体認識問題 [19, 20] をはじめとするコンピュータビジョンの様々な領域において利用されている.

図2.3のように，Transformer の基本の構成は Encoder と Decoder である．機械翻訳の問題を例にすると，Encoder への入力が英語のシーケンス (単語列)，Decoder への入力が翻訳中の日本語のシーケンスである．Encoder の基本要素は，Self-attention と Feed Forward Network (FFN) である．一方で，Decoder の基本要素は，Self-attention，Cross-attention，FFN である.

以下では，まず，Transformer の基本要素である Attention について2.2.1節で述べる．次に，その Attention の一種である Self-attention について定義を2.2.2節で述べる．続けて，Attention 以外の，Transformer に含まれる性能向上ためのテクニックを2.2.3節で紹介する．2.2.4節では，言語モデリングとしての Transformer の応用方法について，2.2.5節では，画像変換としての Transformer の応用方法について解説する.

2.2.1 Attention

Attention を利用するメリットの一つは，非局所的に情報を集約できることである．図2.4のように，CNN は，固定長の局所の情報のみを後段に伝播するのに対して，Transformer は，全ての位置から後段に情報を伝播することができる．もう一つ

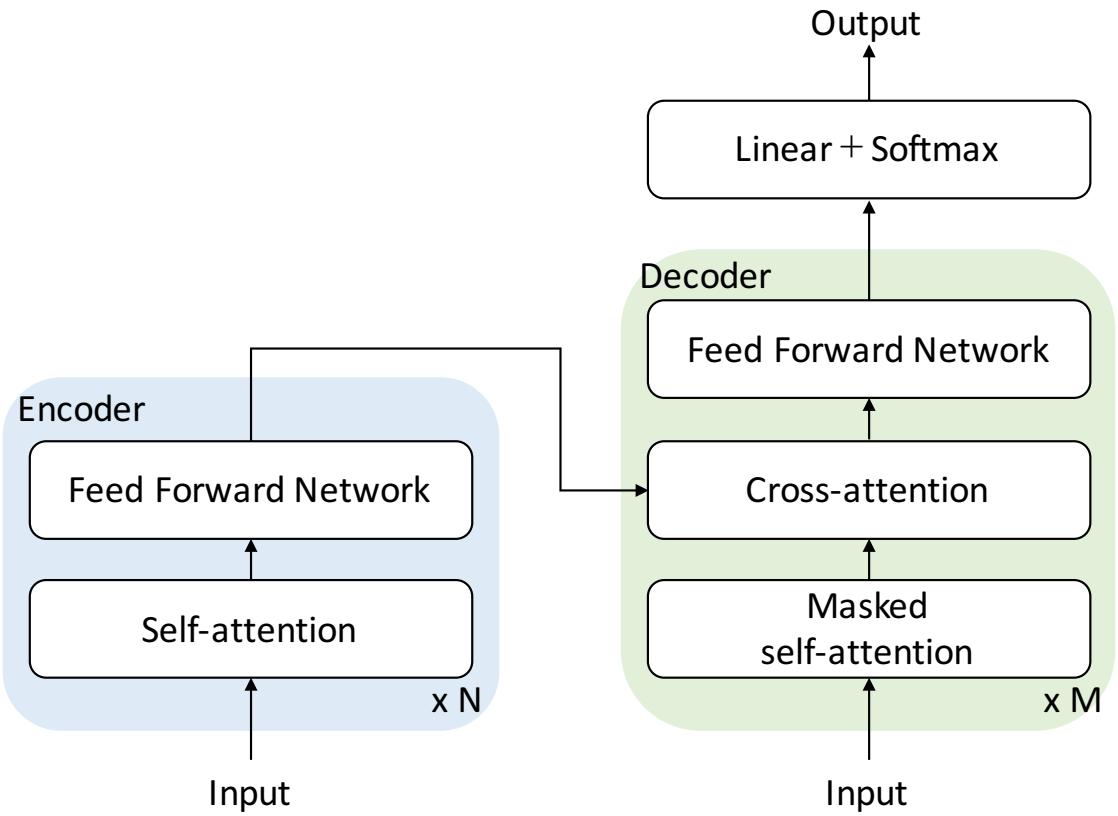


図 2.3 Transformer の構成．Transformer は Encoder と Decoder から成る．図はそれぞれの主な構成要素のみを示している．

のメリットは，シーケンスを扱う際に，RNN のように時刻 t のために $t-1$ を得ておく必要はなく，同時に求めることができ，効率的である．さらに，単純な機構，単純なネットワーク構造でありながらも，性能が高い点も挙げられる．

ここでは Attention の仕組みについて述べる．Attention は，クエリに応じて，メモリから情報を取り出す操作と言うことができる．ここでは，クエリを query，メモリを value(key) とする．それらのベクトルを複数個合わせて行列として表現したものが Q, K, V である．メモリから情報を取り出すとは，V の中から value を取り出すということに相当する．取り出すために，どの位置の value を取り出せば良いかを決める必要がある．その方法が，query と key の内積によってベクトルの類似度を計算して，クエリに関連するメモリの位置を決めるという方法である．そして，その位

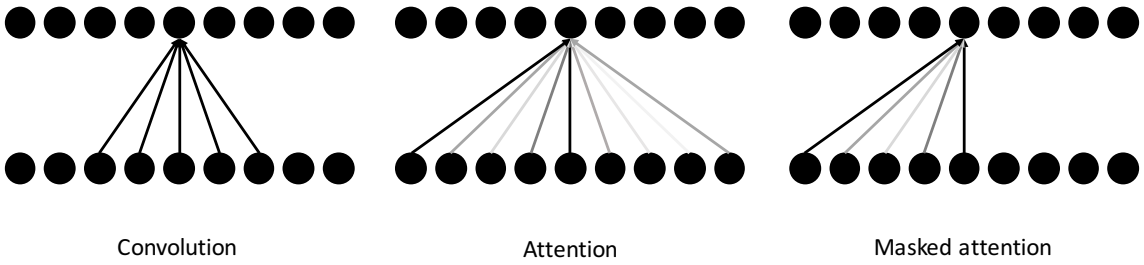


図 2.4 Convolution と Attention の違い．ネットワークにおいて前の層から情報を受け取る際の対象範囲が異なる．Masked attention はマスクにより Attention の範囲を制限した方法.

置に対応する value を V から取り出す．以上の内容から Attention の基本式は,

$$\text{Attention} = \text{Softmax}(QK^T)V \tag{2.5}$$

となる．ここでは，query ベクトルが複数あるとして，行列 Q で表現し，V からベクトルを取り出す操作を重み付き和として表現するために softmax 関数を利用した.

2.2.2 Self-attention

Transformer で使われる Attention には，Self-attention と Cross-attention がある．Self-attention は，一つの特徴量を別々に射影変換して K, Q, V を得たものである．一方で，Cross-attention は，図2.3のように，Q が Encoder から得られる場合である．いずれも式2.5と同じ仕組みで動作する.

Decoder での Self-attention にはマスク付きの Self-attention が使用される．特にこれはシーケンス予測を行うために必要な方法である．シーケンス予測の推論時には，自己回帰的に時刻 $t-1$ までのシーケンスから時刻 t の単語を予測したい．そのためには学習時にも同じような状況を用意する必要がある．RNN のような t を計算するために，$t-1$ の結果を待つ必要がある場合，学習に時間がかかってしまうという問題がある．Transformer では学習を効率的に行うために，全ての単語を同時に入力し，同時に出力を得る．ただし，予測されるべき単語が予測前に Attention により参照されないようにする必要がある．以上の理由で，Attention の一部がマスクによって制

限された Self-attention を Decoder の学習に利用する．

2.2.3 性能向上のテクニック

そのほかにも，Transformer には，性能を向上させるためのテクニックがいくつも導入されている．Positional encoding は，単語の相対位置もしくは絶対位置を特徴量に追加する処理である．FFN は，Feedforward ネットワークで，学習で更新される，ベクトル長方向の線型変換である．Layer normalization は，ベクトル長方向への正規化処理である．Multi-head attention は，複数の線型射影を用意することで，Attention の結果を複数得て，アンサンブルの効果で性能を向上させる工夫である．

2.2.4 応用

応用には，Transformer の Encoder 部分のみを使用した BERT [21] や Decoder 部分のみを使用した GPT [22, 23, 24] などのモデルがある．このように Transformer で提案された一部の構造のみを利用することも多い．Decoder のみの場合は，Encoder の出力を受け取るための Cross-attention は含まれない．以下の章では，Transformer と記述したとき，Transformer の Decoder を指すものとする．

学習方法．Transformer や GPT の学習は，言語モデリング (次単語予測) で行う．その学習は, ネットワーク θ の更新による，次の式2.6の尤度の最大化である．

$$U = \Sigma_t \log p(x_t | x_{t-k}, \cdots, x_{t-1}; \theta) \tag{2.6}$$

ここで，x をトークンと呼び，t は時刻またはシーケンス内の位置を表す．k は区間を表すために使用する．学習時には，k 個の入力トークンのシーケンスを全て入力し，つまり k 個の全ての時刻を入力し，各時刻での出力を得る．出力の正解は，入力トークンのシーケンスを右にシフトして作成される．図2.5に具体例を示す．簡単のためにトークンを単語として表現すると，<BOS>が文章の始まり (Beginning of

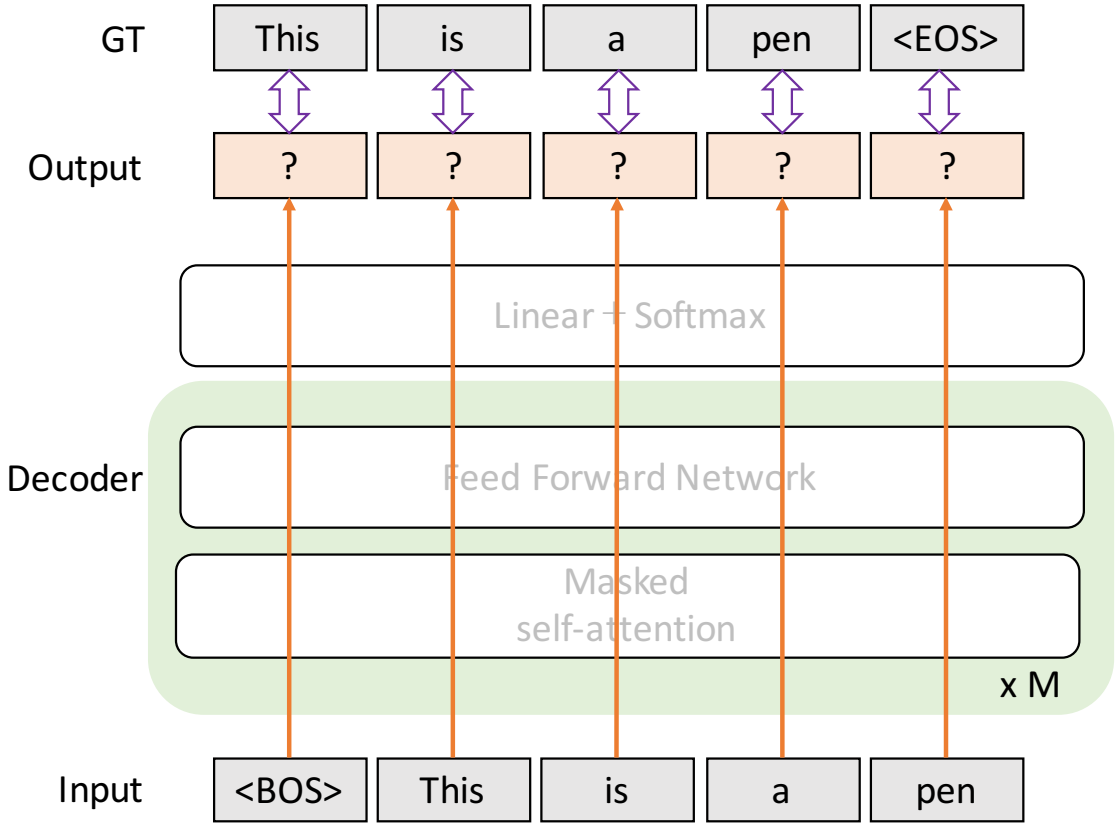


図 2.5 Transformer decoder の学習方法.

Sentence)，<EOS>が文章の終わり (End of Sentence) を表すとして，“<BOS> This is a pen <EOS>” という文章がある時，入力は，“<BOS> This is a pen” で，出力は，“This is a pen <EOS>” となる．出力はトークン種類ごとの logits であり，正解トークンとの Cross entropy loss で学習を行う．

推論方法．ネットワークの学習が完了した後，シーケンスを推論する際には，自己回帰推定を行う．すでに説明しているように，これは前のトークンのシーケンスから，次のトークンを得るという方法である．学習時とは異なり，予測結果の logits を多項分布 (Multinomial distribution) に当てはめ，サンプリングし，トークンの種類を決定する．これにより，単純に logits の中で最も大きい値のトークンの種類 (クラス) を選択するよりも，多様な出力を得ることができる．例えば，“This pen is” が続いた時，“black”, “red”, “long” など複数の候補があり得る．ネットワークの学習

後に，“This pen is” に続く単語の logits が “black” で最も高いと学習されていても，このサンプリングを実施することによって，他の種類の単語を多様に生成することが可能となる．

2.2.5 Transformer による画像変換

TT [25] は，Transformer を使って画像生成や画像間変換を行うことができる手法である．例えば，セグメンテーションマップから画像の変換や，Inpainting を行うことができる．

この手法の工夫は，Transformer が扱う対象を，ピクセルではなく，量子化された特徴量にした点である．ピクセルを Transformer に入力して画像生成を行う iGPT は，64×64 を超える解像度を扱うことが困難であるという限界があった．その原因は，画像のピクセル数は解像度に対して指数関数的に増えることである．TT はこの限界を解決するために，画像をエンコードして得られる特徴量を扱う．さらに，その特徴量をあらかじめ量子化してトークンとすることにより，文章と同様に Transformer で扱うことを可能にする．具体的には，256×256 の画像を 16×16 の特徴量とする．また，量子化を行うことで特徴量にインデックスを割り当てることができる．16×16 のインデックスを 1 次元の長さ 256 のシーケンスとすることで，文章のように見立てることが可能である．

量子化された特徴量を得る方法は，VQVAE [26] の量子化機構を利用している．これは，Encoder-Decoder のボトルネック部分の出力を取り出す方法であり，そのボトルネック部分に量子化機構がある．量子化機構では特徴量の固定数のクラスタの中心を決める仕組みとなっており，エンコードされた特徴量を中心の特徴量に置き換えることで量子化を行う．TT では損失関数として Adversarial loss を追加しており，VQVAE と区別するために VQGAN と呼んでいる．

画像間変換の学習方法について述べる．ここでは，変換対象の画像に対応するインデックスのシーケンスから変換後の画像に対応するシーケンスを得ることが目的と

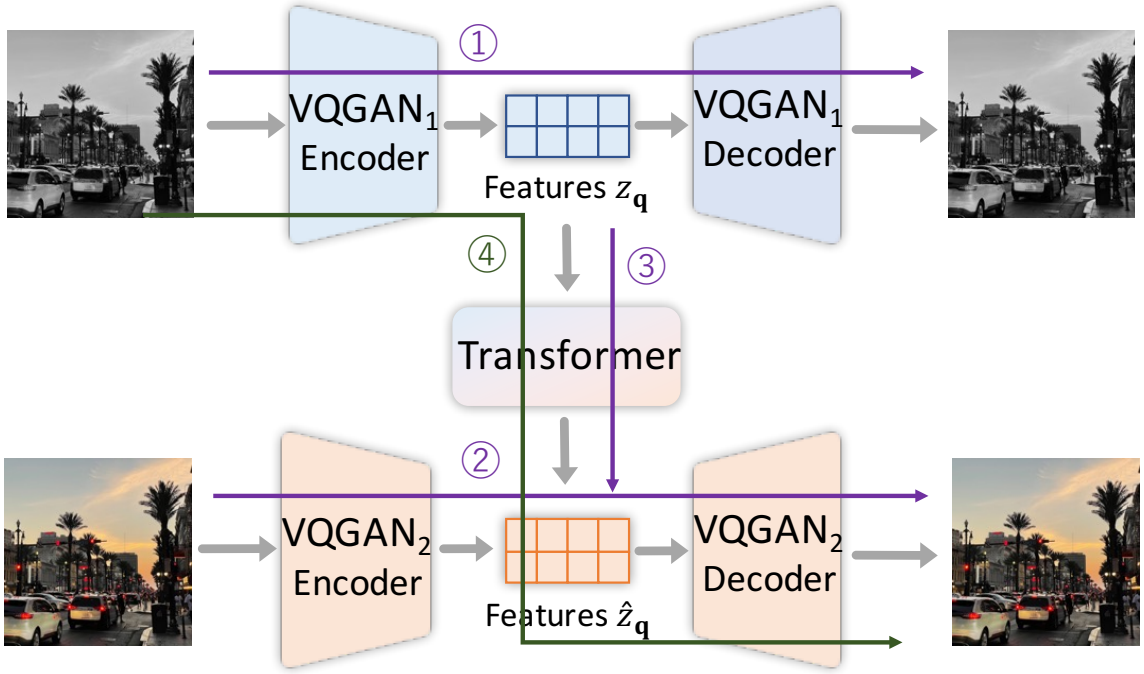


図 2.6 TT のネットワーク構成とフロー.

なる．そのために図2.6のように，2 種類のシーケンスを得るための，2 つの VQGAN を学習する．1 つは，入力画像のドメインで再構成学習をおこなった VQGAN，もう 1 つが出力画像のドメインで再構成学習をおこなった VQGAN である．例えば，セマンティックマップ用の VQGAN と，実画像用の VQGAN である．これらから，それぞれ個別に画像に対応するインデックスのシーケンスを得る．変換前のシーケンスを条件として，変換後にあるべきシーケンスを Transformer が学習する．

推論方法は，文章生成と同様に，Transformer の自己回帰推定で行う．変換前のシーケンスを条件として Transfomer に入力し，それに続くインデックスを予測を繰り返す．さらに，TT ではより大きな画像を扱うためのスライディングアテンションウィンドウを提案している．例えば，256×256 の変換で学習したモデルを，1024×1024 に適用する場合に使用する．学習で扱った 256 × 256 の領域をラスタオーダーでスライドさせるという方法である．

以上の方法で，TT は Transformer での画像間変換を行う方法を提案している．360 度画像の Outpainting では TT を使用しつつ，さらに 360 度画像の補完に合わせた拡

張を行う．特に Transformer の自己回帰推定の順序に対する新たな方法と 360 度画像用の損失関数を提案する．

2.3 画像補完

画像補完 (Image completion) とは，Image inpainting や Image outpainting の総称である．本論文においては，Inpainting とは Image inpainting を，Outpainting とは Image outpainting を表すものとする．

2.3.1 Image inpainting

Image inpainting は欠損領域に適切なピクセルを埋めるタスクで，ギズ修復や物体除去に使用される．パッチベースの Image inpainting 手法 [27, 28, 29, 30] は，低レベル特徴を使った手法である．近年では，CNN を大規模なデータセットで学習する深層学習ベースの Image inpainting 手法 [31, 32, 33] が盛んに研究されている．パッチベースの手法のアイデアを取り入れた手法も提案されていて，有望な結果を出している．Image inpainting 手法の進展を 3 つ分類すると，大域的な整合性つまりコンテクストが考慮できるようになること，高解像度が扱えるようになること，多様な出力が得られることである．従来のパッチベースの手法は特に，古い写真についた線を修復するという小さい領域に対する補完や，データベースにあるパッチを検索して貼り合わせるという手法である．それに対して，学習ベースの手法では，ディープニューラルネットワークの深い層から得られるハイレベルな特徴量を使い，大域的な情報やコンテクストの理解を活用し，大きな欠損領域でも補完できるようにしている [31, 32]．高解像度の画像を扱うためのアプローチは，補完を二段階に行うことである．一段階目に低周波の補完を行い，二段階目に高周波も含め補完するアプローチが一般的である．多様な出力に関しては，CNN の Encoder-decoder ネットワークを GAN で学習する手法では多くの場合，入力画像に対して一つの結果を出力

するという，決定的な出力であることが課題である．それに対して，PICnet [34] は CVAE [35] を採用することにより複数の結果を出力する．また，Transformer を使った画像補完も提案されている [36, 25, 37]．Transformer に含まれる Attention によって，グローバルな構造の生成やコンテキストの整合性が CNN に比べて非常に高まっている．TransFill [37] での検証では，CVAE よりも多様な補完であるため，多様な出力結果を得る目的では VAE の代わりに Transformer を用いることが有望な方法であると示されている．一方で，Transformer の弱みとして，大きな画像を扱う際には大きな計算コストが発生するということと，入力条件領域のピクセルの再サンプリングが原因で本来のピクセルとの整合性が失われること [25, 38] である．

2.3.2 Image outpainting

Image outpainting は，Inpainting 手法がより大きな領域が補完できるように発展したことによって，一層注目され出した問題である．入力画像の周辺を生成するという外挿問題であり，入力画像に対する補完領域の割合が大きい傾向にある．この問題の一種として，Image extension [39, 1] や，Novel view synthesis [40], Infinity landscapes synthesis [2, 41], Panorama generation [42, 43] も含めることができる．以下では，本論文で取り組む問題設定である，Image outpainting による画像拡張と，360 度画像の Image outpainting についてより詳細に述べる．

2.4 Image outpainting による画像拡張

画像拡張は，画像のサイズを変更することである．画像のサイズを変更するために利用される画像補間 (image interpolation) には，バイリニア法，バイキュービック法，ニアレストネイバー法，ランザック法などが含まれる．また，シームカービング [44] という方法もあり，画像の縮小だけでなく拡大にも利用することが可能である．Photoshop に搭載されているシームカービングは，拡大縮小の影響を与えたくな

い領域を選択することで，選択されたコンテンツの形状を保持しつつ，画像のサイズを変更することが可能である．しかしながらこれらの方法での画像拡張は，画像内部のピクセルに対する操作が行われるため，入力画像の変形を避けることができない．つまり画像内のコンテンツの見た目に影響を与えるという課題がある．

一方で，入力画像の周辺ピクセルの外挿によって，画像を拡大するアプローチが存在する．クラシカルな方法として，境界がマッチする画像をデータベースから検索する方法がある [45, 46, 47, 48]．データベースの画像に適切なパッチが含まれない場合では，ターゲット画像のコンテキストと不整合な結果になる．CNN を利用した学習ベースの方法では，その汎化性能によってその課題を解決する．GAN を利用した補完による画像拡張では，ターゲット画像周辺のセマンティックコンテンツの予測を行う [1, 2]．[49, 50] はテクスチャパターンを対象としたピクセルの外挿をする手法である．SinGAN [51] では一枚の画像だけでモデルを学習し，その画像のコンテンツを反映した様々なサイズの画像を生成できるが，別の画像に対して汎化しない．概して画像補完における外挿は内挿よりも難しい問題である．

2.4.1 問題設定

本研究では，入力画像に対して水平方向への拡張を行う．これは生成モデルによる入力画像 x から補完画像 y への変換であるため，式で表すと，

$$y = G(x) \tag{2.7}$$

である．また，出力画像 y においても入力画像 x のピクセルを利用するときには，入力領域を表すバイナリマスク M を用いて，

$$y = x \times M + G(x) \times (1 - M) \tag{2.8}$$

と表せる．左右方向や上下方向への拡張も同様に表現できる．

2.4.2 既存手法

Outpainting による画像拡張手法の既存研究について本研究と近いものついて説明し，それらとの違いを述べる．

Boundless [1]．Boundless の研究目的は，画像拡張に GAN を利用する方法の探索である．そのため GAN を使った Outpainting の先駆け的な研究となっている．Outpainting の課題として，GAN を用いた Inpainting 手法を画像拡張に適用すると性能が上がらないことを指摘している．それに対して画像拡張に適するネットワークの構成要素を検証した．そして，提案手法として GAN の Discriminator へセマンティックコンディションを導入し，学習の安定化を行った．その結果，画像拡張問題において，Inpainting 手法と比べて良い拡張結果を得ることを示した．

LVNS [2]．LVNS の研究目的は，Outpainting によって画像を非常に長く拡張することである．非常に長い画像へ拡張していく場合に入力画像とピクセル空間的に離れてしまうため，どのように入力画像の情報を伝達するかという点が課題になる．それに対し，空間の整合性とセマンティックの整合性を持つ画像を生成するため，Encoder から Decoder に情報を伝えるための Skip horizontal connection と，より水平方向へ情報を伝達していくためにボトルネック部分に Recurrent Content Transfer と呼ぶ LSTM 層を追加したネットワークを提案する．その結果，風景画像において，整合性を維持しつつ画像サイズを水平方向に 2 倍以上に拡張できること示した．

2.4.3 先行研究と本研究の違い

本研究は，Outpainting による画像拡張のために，Inpainting 手法を活用する方法を検討する点が先行研究と異なる．Boundless も LVNS も Outpainting のために新たな構造を提案しようとするが，本研究の提案手法は Inpainting ネットワークを利用することを考える．そのため，入力画像の処理部分に注目する．特に入力画像の情報を伝える方法として，ミラーするという手法を提案する．結果として，風景画像を

1.5 倍に拡張するというタスクで，良い結果を得る．

2.5 360 度画像の Image outpainting

2.5.1 問題設定

360 度画像の一部の領域からその周辺を補完することによって完全な 360 度画像を生成するという問題設定である．生成モデルによる画像変換問題であるので，入力画像 x から補完画像 y への変換,

$$y = G(x) \tag{2.9}$$

と表される．従来この問題は，Inverse rendering [52, 53] や Lighting estimation [54, 55, 56, 57] のサブタスクとして取り組まれている．これは照明を表現するために環境マップを推定するということに相当している．光源の位置を推定するという目的では，高周波のテクスチャを予測する必要がないため，画像解像度が小さい．一方で，背景画像として利用するために詳細なテクスチャを持つ 360 度画像を得るという目的でも取り組まれている [3, 5, 4].

2.5.2 360 度画像の種類

本研究では，360 度画像として Equirectangular projection (ERP) 画像を利用する．360 度の球面を 2 次元平面へ投影する方法には，ERP (正距円筒図法) 以外に様々な方法がある．例えば，キューブマップやメルカトル図法である．中でもコンピュータビジョンやコンピュータグラフィックス領域で特に使われるのが ERP とキューブマップである．本研究が ERP を利用するのは，CNN という矩形のフィルターを利用する点で，ERP も長方形の画像であるため扱いやすいからである．

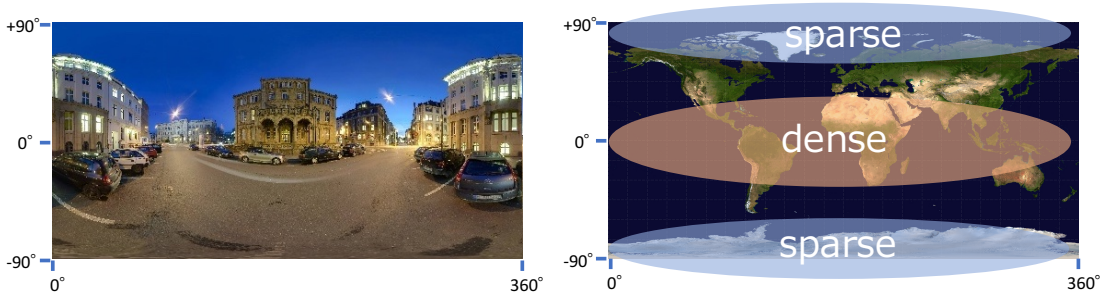


図 2.7 正距円筒図法 (ERP) による 360 度画像の例と情報量の差.

図2.7のようにERP画像では，緯度方向に情報量の差が発生する．画像の上下の領域はスパースであり，画像の中央はデンスである．ERPで表現される360度画像の特徴は，この緯度方向の情報量の差と，左右両端のつながりである．360度の画像であるため，ERP画像の両端は連続しているものであり，補完問題では出力結果が両端のつながりを持つかどうかが評価の対象となる．以下の文章において，360度画像とは，ERP画像のことを指し示すものとする．

2.5.3 既存手法

360IC [3]．360ICは，GANによる補完手法によって360度画像を生成することを目的した研究である．この研究では，ネットワーク構造の提案と入力画像の前処理を新規に提案している．具体的には，Pix2pixHDに対して，大きな穴を補完するために受容野を拡大するためのResidual blockを導入している．この受容野の拡大方法は，Dilationを徐々に大きくしたConvolution層を直列かつ並列に繰り返す方法である．また，ネットワークへ入力する前に入力画像を組み直す前処理を提案している，この前処理を行うことにより，両端のつながりを持つ結果を得られることを示している．図2.8は論文からの引用で，処理フローの概要を示している．

360ICを発展させた研究 [8] では，学習データ数を増加させた場合の実験を行なっている．360ICでは学習に利用した画像数の少なさが補完結果に影響を与えており，ステージ１のネットワークで過学習を起こしたため補完結果の修正用としてステー

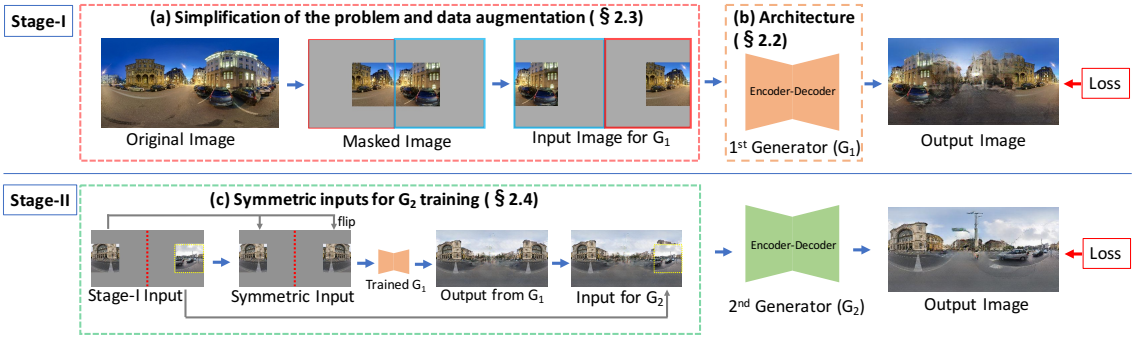


図 2.8 360IC の処理フローの概要．[3] から引用．

ジ 2 のネットワークが必要であった．一方で，学習データ数を増加させた場合 [8] には，ステージ 1 のみで 360IC の品質を超えることが可能であると明らかになった．

SIG-SS [5]．SIG-SS は，シンメトリーをコントロールして 360 度画像を補完することを目的にした研究である．Conditional VAE (CVAE) を利用しつつ，予め分類しておいた種類のシンメトリー性それぞれに対して，そのシンメトリー性の強度を VAE からサンプリングする．また，CNN で利用するパディングの方法として Circular padding を提案し，左右のつながりを持たせている．図2.9は論文からの引用で，処理フローの概要を示している．

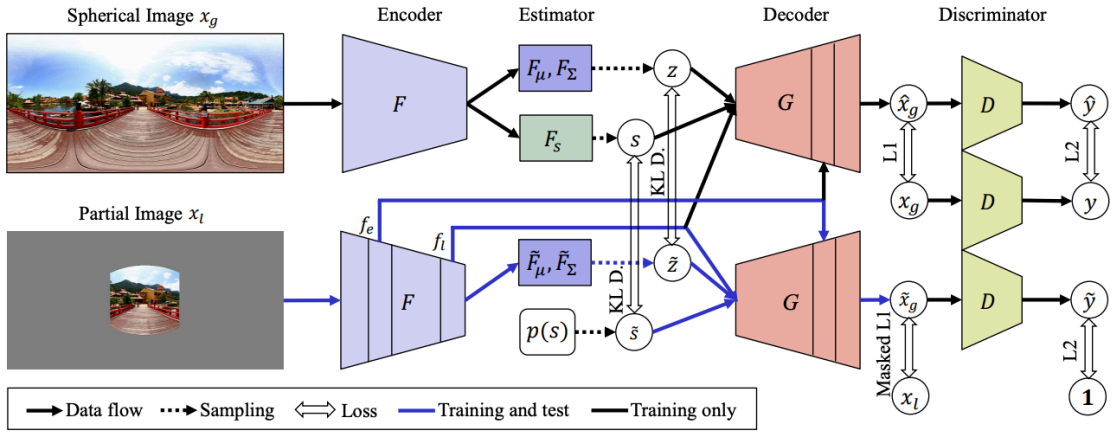


図 2.9 SIG-SS の処理フローの概要．[5] から引用．

EnvMapNet [4]．EnvMapNet は，HDR 環境マップをリアルタイムに生成することを目的した研究である．スマートフォン上でもリアルタイムに実行可能とするために，軽量な Encoder-Decoder ネットワークを利用する．学習を安定化するための損

失関数が主な提案であるが，ERP の緯度方向に沿った情報量の差を考慮して重み付けしたピクセルレベルの損失関数も提案する．図2.10は論文からの引用で，処理フローの概要を示している．

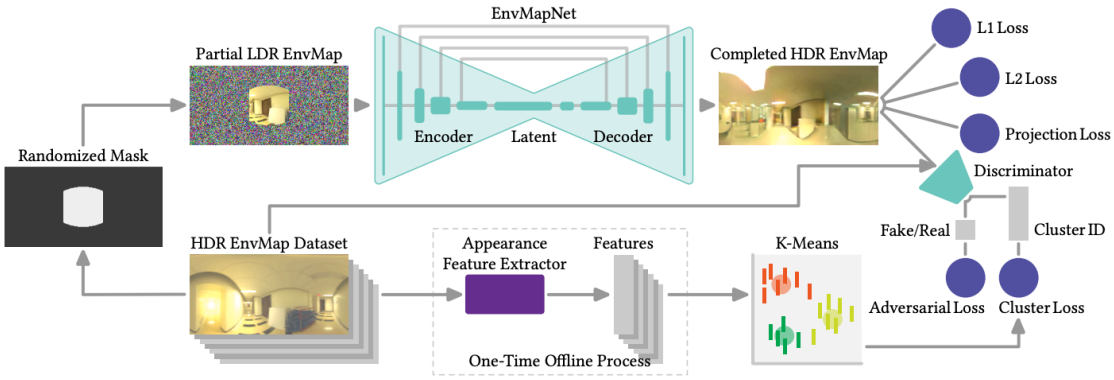


図 2.10 EnvMapNet の処理フローの概要．[4] から引用．

2.5.4 先行研究と本研究の違い

多様性．CVAE でシンメトリーの強さをサンプリングする SIG-SS を除いて，これらの手法は決定的な出力を得るネットワークである．つまり，一つの入力に対して一つの出力しか得られない．一方，本研究では，Transformer を導入することによって，一つの入力に対して多様な出力を生成する．

解像度．先行手法には，学習時の画像解像度に過剰適合するという課題がある．本研究では，補完画像との調整を行うことによって解像度に対する課題を解決し，任意の解像度での入出力を可能とする．さらに，先行手法 (256×128 や 512×256) より大きな解像度 (1024×512) を出力可能である．

見た目．先行手法の結果画像よりも圧倒的に自然で尤もらしい見た目の画像を得ることができる．実利用に最も近い結果と言える．

2.6 評価方法

この節は，本研究では定性評価を重視し，定量評価指標として FID スコアを主に利用することについて述べる．

多くの Image inpainting や Image outpainting の目的は，欠損領域の補完を違和感なく行うことであり，実世界の再現の優先度は低い．さらに，補完問題では一つの入力に対して，複数の解が存在し得る．このような問題設定の場合，オリジナルの画像や正解画像と言われる画像と補完結果とのピクセルレベルの差分の指標 (MAE, MSE，PSNR，SSIM) での評価は重要ではない．一方，参照画像を使う場合や動画内から物体を取り除くといった問題設定では，欠損領域がどのように補完されるべきかが一意に決まる場合があり，この場合にはそのような指標も利用される．本研究の補完問題は前者に相当し，結果画像の見た目での主観評価が重視される．多くの研究は，同じ入力画像に対して複数の手法の結果画像を並べることで比較しており，本研究でもそれに習い，定性的な評価を行う．

定量評価方法として，利用される指標が FID スコア [58] である．FID スコアは特徴量の分布に基づく距離によって，補完結果の品質を評価している．補完問題のみならず，画像生成問題の評価で広く利用されている指標である．以下では，FID スコアの算出方法について述べる．

FID は，学習に用いた実際の画像の特徴量の平均・分散と生成画像の特徴量の平均と分散を比較することで，生成画像が実際の画像にどれくらい近いかを数値化する方法である．より具体的には，学習済みネットワークである Inception-v3 [59] での画像の特徴表現が多変量正規分布に従うと仮定し，実際の画像群 X と生成画像群 Y の平均と共分散をそれぞれ，μ_X と μ_Y，Σ_X と Σ_Y とした時，

$$\mathrm{FID} = \|\mu_X - \mu_Y\|^2 + \mathrm{Tr}(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \tag{2.10}$$

で表される．ここでの Tr は行列のトレースを意味する．FID は学習済みの Inception-

v3 に依存しているという点，画像をそのネットワークに適する固定サイズ (299×299) にリサイズする必要があるという点に注意をする必要がある．

2.7 データセット

Places [60]．Places は，屋内と屋外の景観を対象にした画像データセットである．提案当時，CNN と ImageNet による学習で物体認識の精度が高まることが明らかになったが，獲得される CNN の特徴量はシーン認識には適していなかった．そのため，物体を中心にしたデータセット ImageNet とは異なり，シーン画像が中心という性質の異なるデータセットとして提案された．このデータセットには 476 シーンカテゴリの，7,076,580 枚の画像が含まれており，規模が大きい．図2.11は Places に含まれる画像の例である．

図 2.11 Places に含まれる画像例．

Scenery dataset [2]．Scenery dataset は，山脈風景を主とする景観画像データセットである．一部には海岸や川岸，雪や星空といったデータも含まれている．このデータセットが公開された論文で取り組まれた問題は水平方向への画像拡張であり，その問題に適した風景を選択して収集していると予想される．公開されている画像データの解像度は 128 × 256 pixels で，5000 枚の訓練用画像と 1000 枚の評価用画像から

なる．そのカテゴリの種別には Places ほどの多様性はなく，データセットの規模も比較すると小規模であると言える．図2.12は Scenery dataset に含まれる画像の例である．

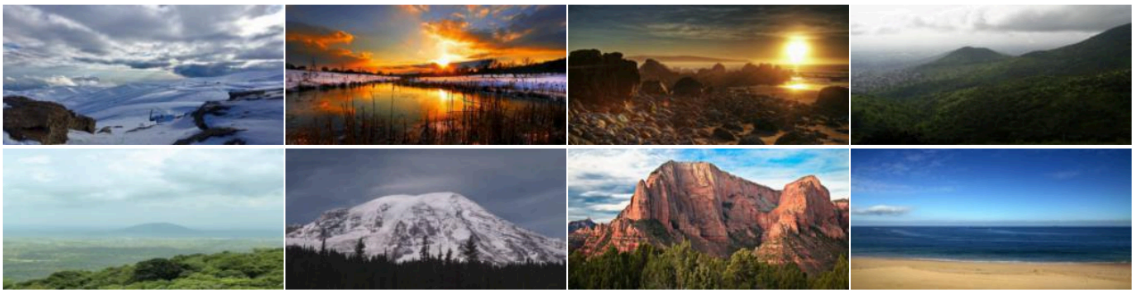


図 2.12 Scenery dataset に含まれる画像例.

SUN360 [61]．SUN360 は，屋内と屋外の景観を対象にした 360 度画像のデータセットである．これらの 360 度画像はインターネット上から収集されたのち，シーンカテゴリのアノテーションが行われている．そのデータ数は 67,583 枚であり，80 カテゴリに分けられる．図2.13は SUN360 に含まれる画像の例である．

図 2.13 SUN360 に含まれる画像例.

Laval Indoor Dataset [54]．Laval Indoor Dataset は，屋内の 360 度画像のデータセットである．大学敷地内を撮影することによって収集されているデータセットであり，1837 枚の訓練用画像と 289 枚の評価用画像から構成される．提案された論文の問題設定は，照明の推定であるため，High Dynamic Range Image (HDRI) であるという特徴もある．図2.14は Laval Indoor Dataset に含まれる画像の例である．

図 2.14 Laval Indoor Dataset に含まれる画像例.

第3章　画像拡張

3.1　概要

研究動機．画像拡張は画像素材のサイズを大きくするために頻繁に行われている．例えば，素材のアスペクト比を様々なデバイスに合わせて調整するために，図3.1に示すようなバイキュービック法やシームカービング [44] を用いたリサイズが頻繁に使われる．しかし，バイキュービック法ではターゲット画像を引き伸ばし，画像中のコンテンツが意図しない変形を起こしてしまう．さらに，シームカービングを使った Photoshop の一機能である `content-aware scale` では，選択したコンテンツの変形を避けることができるが，そのコンテンツの位置の変化を避けることができない．これらの望ましくない変更はどちらもターゲット画像の内のピクセルの変更に由来する．逆に，その内側のピクセルを変更せず，ターゲット画像の周辺を補完することでこれらの問題を避けることができる．図3.1最下段のように, ターゲット画像の周辺を補完する生成的な画像拡張は，本来のターゲット画像を保存しつつ全体の画像サイズを拡張できるという優位点を持っている．

既存研究の限界．既存の生成的な画像拡張の手法は，生成領域がターゲット画像の外側にあるという外挿問題を扱う．Boundless [1] は，GANs においてシーンの条件付けを考慮する Discriminator を学習に用いることにより，多様な種類のシーンの画像を拡張できることを示した．しかしながら，Boundless [1] の生成結果は外側のピクセルになるほど，生成の質が悪化する傾向にある．例えば，高周波成分を欠いたテクスチャやアーティファクトの発生が見られる．その理由の一つは生成しようとするピクセルの位置が入力画像のピクセルと離れているため，畳み込み層による

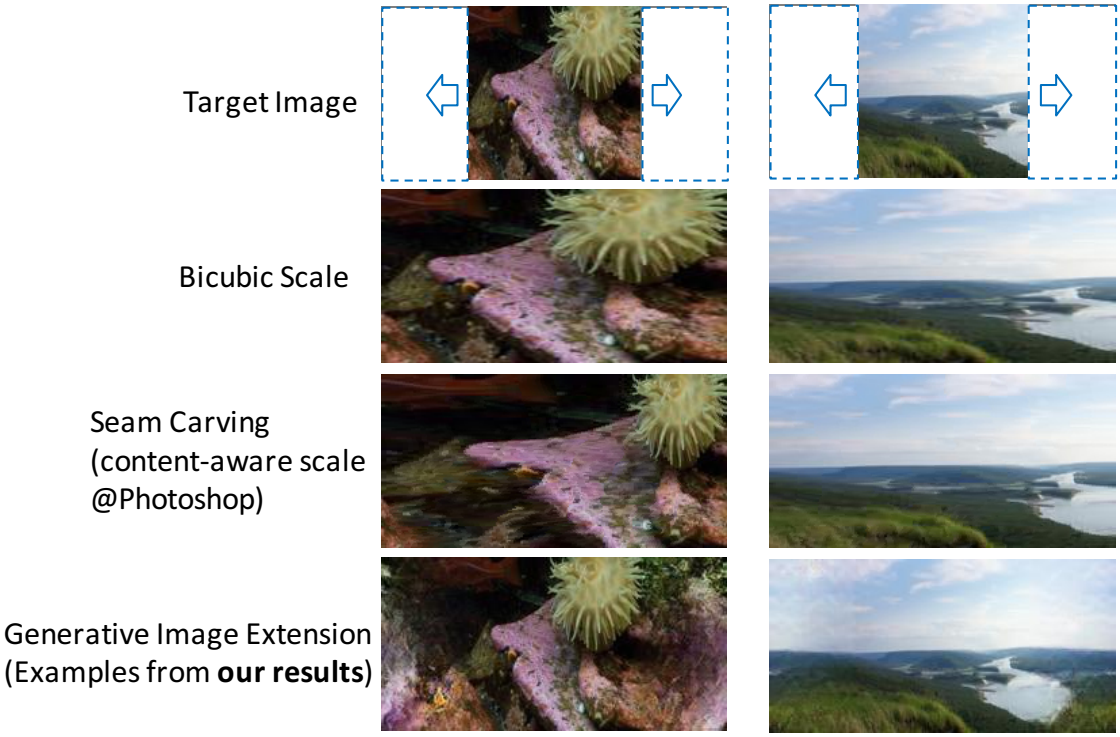


図 3.1 生成による画像拡張と，Bicubic scale や Seam carving による拡張との違い．Bicubic scale と Seam carving は，ターゲット画像の内側を変化させてしまい，期待していない変形や位置ずれを引き起こす．一方で，生成による画像拡張は，ターゲット画像に外挿し，ターゲット画像を維持しつつ画像を拡張する．

情報の伝達が十分にできていないからである．もう一つの理由は，拡張する領域の内，入力領域に接する片側だけが境界の制約を受けることである．前者の原因に対応するため，VLNS [2] は，水平方向に情報を効果的に伝搬させるためのモジュールを提案している．しかしながら，同一のコンテンツをリピートするような生成になるので，山脈の画像など，扱えるシーンが限定されている．

本研究の目標．この研究の目標は，既存研究の限界でもある次の３つの改善である．

1. 生成しようとするピクセルの位置が入力画像のピクセルと離れていることが原因で畳み込みによる情報の伝達が十分にできていない課題に対処する．

2. 多様なシーンの画像で利用可能な拡張手法を実現する．

3. 生成されるコンテンツの形状をコントロールできる拡張を実現する．

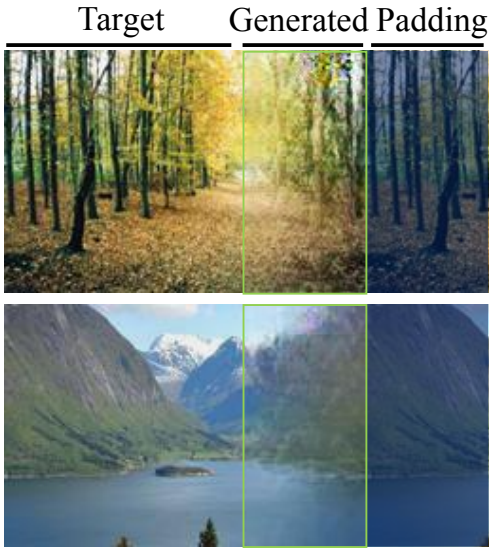
提案手法のキーアイデアは，拡張する領域を何らかの画像で挟み込むことである．こうすることで，生成するピクセルの近くに信頼できるピクセルを配置した状況になり，さらには外挿 (Extraploation) から内挿 (Interpolation) の問題になる．そして，挟み込んだ画像が生成の条件となるので，どの画像で挟み込むかによってユーザーによる生成コンテンツのコントロールが可能となる．

技術的課題．しかしながら，本研究の目的を満たす，補完領域の挟み込み方法は自明ではない．例えば，全く関係のない画像の挿入は，ターゲット画像とその画像を補完画像によって繋ぐことは大抵の場合で困難であるので，生成画像の質が下がる原因となる．図3.2の (b) のようにターゲット画像の一部を用いてパディングすると，多くの場合で下段の結果のように整合性のない結果が得られる．主な理由は，ターゲット画像の境界とパディング画像の境界がピクセルレベルで連続する保証が常にはないことである．ターゲット画像とパディング画像の両端同士が繋がり得るようにする方法は，単純には全てミラーする方法である．しかし，図3.2の (c) のように，この方法では似ているピクセルを水平方向に繰り返してしまい，既存手法と同じような結果を得てしまう．したがって，補完領域が挟み込まれている状態の入力画像の作成にはさらなる工夫が必要である．

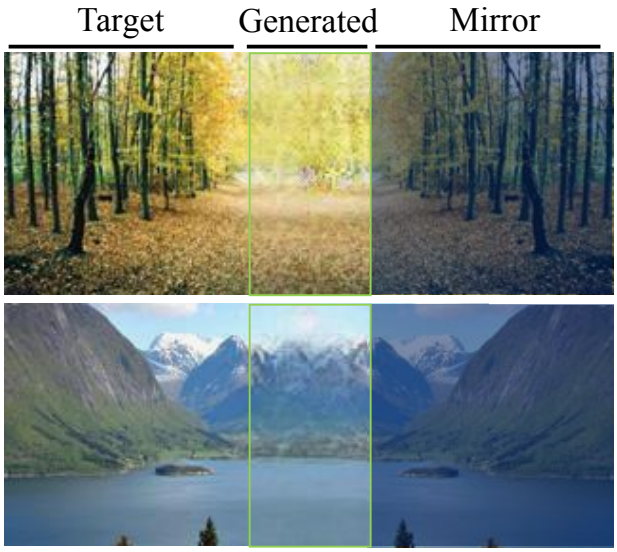
提案手法．提案する補完領域の挟み込み方法は，図3.2の (d) のように，ターゲット画像の一部分のみをミラーする方法である．この入力画像を Mirrored input と呼ぶ．一見，ターゲット画像とパディング画像の両端は連続しないように見えるが，実際はターゲット画像の右端を中心にミラーした場合が少なくとも一つの解として存在しているのでピクセルレベルで繋がる保証がある（詳細は3.2.1節）．セマンティックレベルで考えると，ミラーをすることに適さない物体が存在する画像も少なくないが，それらは今後の課題とした（詳細は3.4.2節）．この Mirrored input は推論時のみで使用し，学習時は同様の補完領域をもつ Inpainting 問題としてネットワークを学習する．

インパクトと利点．Mirrored input を使うことによって，補完領域のピクセルは

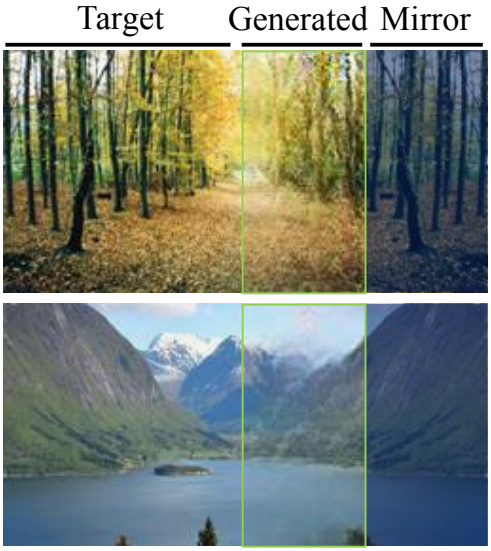
(a) Target image



(b) Left-half horizontally moved



(c) All mirrored



(d) Left-half mirrored

図 3.2 Mirrored input のアイデアと効果．(a) のように補完領域を挟み込むためにどのような画像を配置するかが重要な課題である．(b) のように入力画像の一部を水平方向にスライドさせると，しばしば不整合な結果となる．(c) のように入力画像を全てミラーしてしまうと，補完領域には水平方向に類似したピクセルが生成される傾向にある．一方で，(d) のように画像の左半分をミラーすることによって，ピクセルレベルの連続性を持ち，水平方向への繰り返しよりも複雑な形状のセマンティックス領域を形成する．本手法では (d) を Mirrored input と呼ぶ．

左右両方向への整合性を考慮できるだけでなく，本来は遠くに位置した補完ピクセルも近くの信頼できるピクセルから情報を得ることもできる．さらに，これはOutpainting 問題を Inpainting 問題に置き換える．したがって，多様なシーンクラスを扱える Inpainting 手法の様々なテクニックを利用可能という利点がある．それだけではなく，Inpainting の設定は，ターゲット画像とミラーした画像をピクセルレベルだけでなくセマンティックレベルでも整合性を持つように，補完領域のセマンティックを調整するための制約となる．例えば，図3.5のように，ターゲット画像の端にビルがあるとしても，ミラーされた画像にそのビルが存在しなければ，そのビルは補完領域の途中で途切れるべきである．このように Mirorred input は，水平方向に同一のセマンティックを拡張する既存の手法とは違って，より複雑なセマンティックスの形状を含んだ拡張画像を導くことが可能である．

本論文で示すこと．実験を通して，外挿の問題を解くよりも，Mirrored input を使った内挿の問題として解く方が，Inpainting や一般的な画像間変換 (Image-to-image translation) でそれぞれ定性的にも定量的にも向上することを示す．さらに，提案手法の Inpainting ネットワークと Mirrored input で，SoTA の生成的画像拡張を見た目も FID スコア [58] も両方超えることを示す．Ablation study を通して，そのネットワークの各コンポーネント（新たに提案する Bottleneck feature matching 損失関数を含む）の役割を検証する．

- Mirrored input を提案する．それは生成的な画像拡張を内挿の問題に置き換えることで，より質の高いピクセルの生成を実現するための手法である．これにより水平方向への繰り返しよりも複雑な形状でセマンティックスを拡張でき，コントロールの機会も与える．

- 提案する Inpainting ネットワークと Mirrored input は，多様なシーンを含むデータセット上で，Outpainting の SoTA よりも見た目も FID スコアも両方超える画像拡張を実現する．これにより，今後の生成による画像拡張の SoTA は Inpainting の SoTA 手法によって実現される可能性が高いことを示す．

3.2 提案手法

3.2.1 Mirrored input

3.1節で述べたように，Mirrored input は，Outpainting 設定の問題を Inpainting の問題に変換し，既存の生成による画像拡張の課題を解決する手法である．この節では，具体的な作成手順に関して述べる．提案する Mirrored input は，推論時のみ利用する．図3.3のように (i) ターゲット画像をある方向へ拡張するとき，(ii) その拡張される方向にターゲット画像の端を基点にミラーして 2 倍の画像サイズにする．(iii) そして拡張される領域にマスクを掛け，Generator G への入力画像とする．このような作成をすることによって，一見ターゲット画像の端とパディング画像の端のつながりは明らかではないが，ピクセルレベルの連続性を満たす解として，少なくとも 1 つは (ii) が解として存在する．このような保障の中で補完することで，決して繋がらない画像をパディングする場合と比べ，良いクオリティーの結果を期待できる．

本研究においては，正方形のターゲット画像に対して，その右方向へ 1.5 倍へ拡張するモデルを得ることを目標とする．そのため，Mirrored input を含めて，Generator への入力画像はターゲット画像の横方向への 2 倍とする．この方法で学習されたモデルを用いることで，1.5 倍よりも短い拡張を行いたい場合には，拡張したのちにクロップを行えばよく，一方で，1.5 倍以上に拡張を行いたい場合には，拡張結果を再び入力し目的の長さになるまで拡張を繰り返す．また，入力画像を左右反転することで，右方向への拡張モデルでも左方向へ拡張することが可能である．

ネットワークの学習時は，あらかじめ水平方向に 2 倍の画像サイズで Inpainting の学習させる．Mirrored input の補完領域に対応する領域にマスクをかけることで入力画像を作成し，Generator を学習する．Mirrored input は推論時のみで利用するので，このような学習方法を必ずしも行う必要はなく，SoTA の Inpainting 手法のような学習方法でも良い．しかし，実験では Mirrored input と同じ補完領域でネットワークを学習する方が良い結果を得られたためこの方法を実施する．

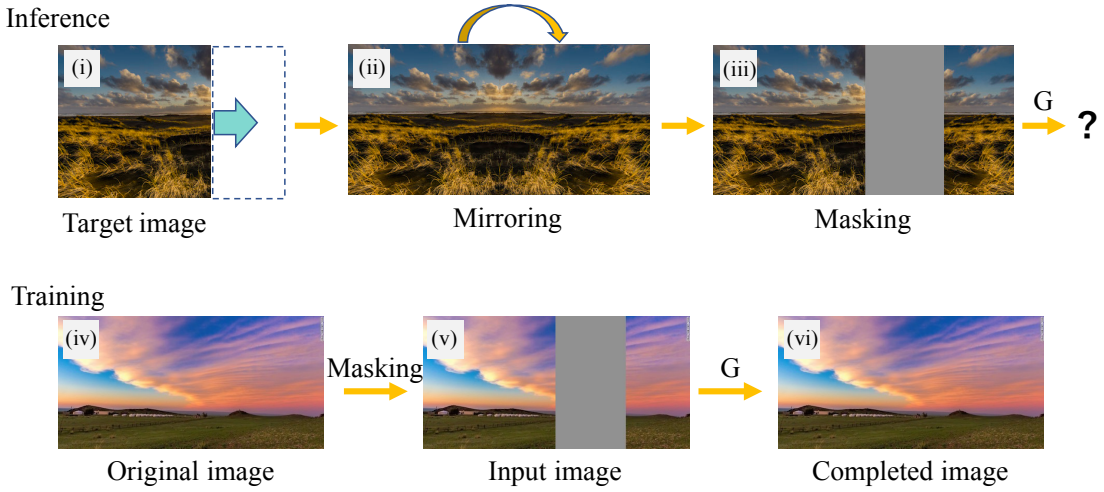


図 3.3 推論と学習の流れ．ターゲット画像の右端を基点にミラーして左右対称な画像を作成し，マスクをかけて入力画像を作成する．学習時は，Mirrored Input での推論を想定して，対応する領域にマスクをかけた入力画像を作成する．Mirrored Input は推論時のみ実施する．

3.2.2 提案ネットワーク

提案手法は pix2pixHD [16] をもとにした補完ネットワークを使用する．最新の Inpainting 手法の多くが自由形状 (Free-form) でかつ小さい穴の欠損領域に対応するために新しい畳み込み方法 (Gated convolution) を提案している．しかしながら，本研究が扱う補完の対象は長方形でかつ大きい穴で固定されているので，それらの最新手法は向いていない．GLCIC [32] が示したように，自然な画像補完には大域的かつ局所的な整合性が必要である．したがって，シーンコンテキストやセマンティックのような大域的な特徴を扱うことと同時に，テクスチャやストラクチャのような局所的な特徴を扱うことが必要である．提案手法のネットワークは，それらの性質を考慮する Resblock [3] を Generator に導入する．

Generator．Generator ネットワーク G は，図3.4のように，複数の Resblock を持つ U-net [15] である．それぞれの Resblock の内部は，複数の Dilated convolution が直列かつ並列に構成される．直列にすることで受容野 (Receptive field) を大きくし，並列にすることで大域的な情報と局所的な情報を同時に取り扱って補完を行う．テクスチャの生成の質を向上させるために Resblock の最後に Self-attention block [62]

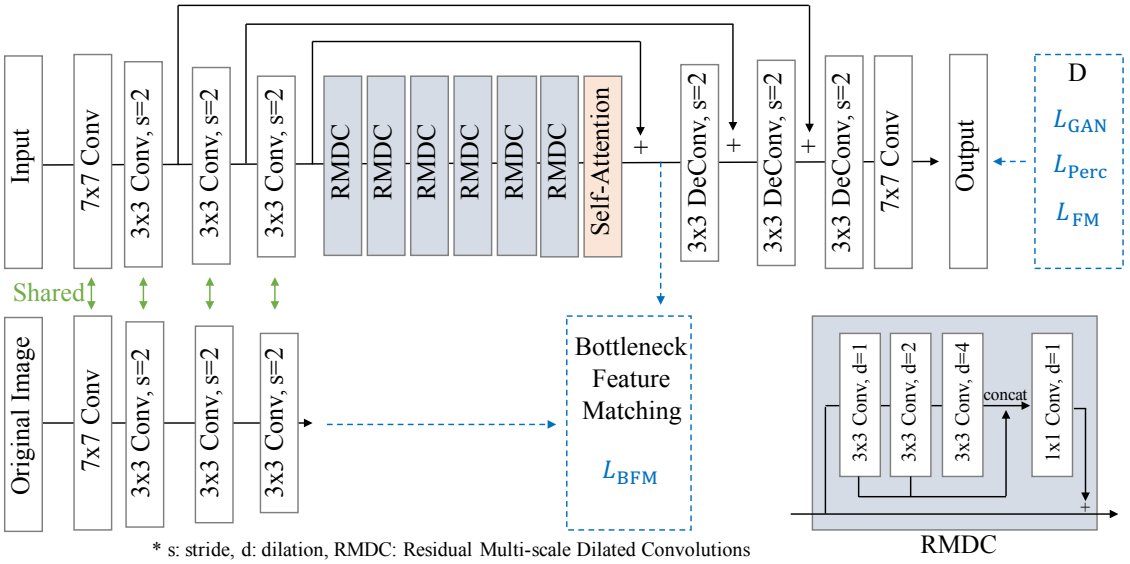


図 3.4 ネットワーク概要．ボトルネック部分に Residual blocks を持つ U-net を用いる．

を加える．この Resblock は [3] とおよそ同様のネットワーク構成だが，それと異なる点として，Dilation が 1 の Convolution 層を追加する．

3.2.3 損失関数

Discriminator．GLCIC などのように，大域的 (global) かつ局所的 (local) な整合性を考慮するために Discriminator D は，Multi-scale discriminator D_k [16] を導入する．Discriminator D_k の i 番目の層の特徴抽出を $D_k^{(i)}$ と記す．それぞれの Discriminator は 3 層の Convolution 層から成る Patch discriminator[16, 15] である．スケールダウンした画像をそれぞれ入力することで，異なるスケールでの real/fake の判定を行える．本実験では $k = 2$ として，目的関数は，

$$\min_{G} \max_{D_1, D_2} \sum_{k=1,2} \mathcal{L}_{GAN}(G, D_k) \tag{3.1}$$

と表される．

さらに，生成結果のテクスチャの質を向上させるために，VGG ネットワークを用

いた Perceptual loss [16, 63, 64],

$$\mathcal{L}_{Perc}(G) = \mathbb{E}_{(\mathbf{z},\mathbf{x})} \sum_{i=1}^{N} \frac{1}{M_i} [\| F^{(i)}_{VGG}(\mathbf{x}) - F^{(i)}_{VGG}(G(\mathbf{z})) \|_1] \tag{3.2}$$

を導入する．ここで，N とは指定された層の総数であり，$F^{(i)}_{VGG}$ とは M_i 個の要素数を持つ VGG network の i 番目の層の特徴量を表す．また，$\mathbf{z}$，$\mathbf{x}$，そして $G(\mathbf{z})$ は，それぞれ入力画像，オリジナル画像，出力画像を表す．ここでは L2 norm [64] ではなく，[16] と同様に L1 norm を使い使い距離を算出する．加えて，Discriminator の学習を安定させるために，生成画像とオリジナル画像を用いた Feature matching loss[16],

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(\mathbf{z},\mathbf{x})} \sum_{i=1}^{T} \frac{1}{N_i} [\| D^{(i)}_k(\mathbf{x}) - D^{(i)}_k(G(\mathbf{z})) \|_1] \tag{3.3}$$

を用いる．ここで，T は層の総数であり，$T = 3$ である．N_i はそれぞれの層の要素数を表す.

Bottleneck feature matching．入力画像と生成領域がセマンティックな連続性を持つようにするために，さらには，Generator の Bottleneck が補完の役割を持つように明示的に制約をかけるための Bottleneck feature matching loss を提案する．これは Bottleneck での higher level の特徴量は Decoder を通る前までにすでに補完されているべきという考えに基づく．図3.4のように，Original 画像を学習中の Encoder に通し Bottleneck の特徴量を計算する．これは欠損領域の無い特徴量である．一方で，欠損領域を持つ入力画像を Encoder に通すと欠損領域を持つ特徴量が得られる．Resblcok を通過した後の特徴量と Original 画像の特徴量の損失を計算することで，Resblcok でセマンティックレベルの特徴量の補完を行うことを強制することができる．この損失関数は，式3.3を参考にして，

$$\mathcal{L}_{BFM}(G) = \mathbb{E}_{(\mathbf{z},\mathbf{x})} \sum_{i=1}^{N'} \frac{1}{M'_i} [\| F^{(i)}_B(\mathbf{x}) - F^{(i)}_B(G(\mathbf{z})) \|_1] \tag{3.4}$$

と表せる．ここで，N' は層の総数であり，$F^{(i)}_B$ は M'_i 個の要素数を持つ i 番目の bot-

tleneck 部分の特徴量である．

この損失は，Encoder が更新されるにつれて抽出される Feature も変化するという点で D における Feature matching loss [16] に似ている．この点は，Encoder の代わりに学習済みモデルを使って抽出された Feature と比較する手法 [65] と異なる点である．Bottleneck feature matching loss は，Original 画像から得られる特徴量との比較を行うが，本研究での補完のゴールはオリジナル画像の生成ではないので，小さい係数を使って，この損失の影響が大きくなりすぎること避ける．トータルの損失関数は次のように表される．

$$\sum_{k=1,2} \mathcal{L}_{GAN}(G, D_k) + \lambda_1 \sum_{k=1,2} \mathcal{L}_{FM}(G, D_k) + \lambda_2 \mathcal{L}_{Perc}(G) + \lambda_3 \mathcal{L}_{BFM}(G). \tag{3.5}$$

すべての実験で，$\lambda_1 = 10$，$\lambda_2 = 10$，そして $\lambda_3 = 0.5$ を使用した．実装には PyTorch を使用した．Generator と Discriminator の学習には Adam を利用し，そのパラメータは，$\alpha = 10^{-4}$，$\beta_1 = 0.5$，そして $\beta_2 = 0.999$ である．ミニバッチサイズは 4 で，1 台の NVIDIA 2080Ti を用いてモデルの学習を行った．

3.3 実験

3.3.1 実験設定

データセット．データセットとして Places subset[60] と，VLNS [2] で提案された Scenery dataset を使う．Places subset は，Places365-Challenge dataset[60] の `trainset` の中から屋外シーンの 20 クラスを含み，それぞれ 20,000 枚の計 400,000 の訓練画像とし，それぞれ 1,000 枚の計 20,000 のテスト画像とする．その 20 クラスとは，`aquarium`, `athletic_field/outdoor`, `beach`, `cliff`, `coast`, `forest_path`, `golf_course`, `harbor`, `lake/natural`, `mountain`, `ocean`, `pier`, `pond`, `rainforest`, `river`, `skyscraper`, `swamp`, `underwater/ocean_deep`, `valley`, `vegetable_garden` である．Scenery dataset

は，VLNS [2] で実際に使われたデータセットの組み合わせとは異なり，5,000 枚の学習画像と 1,040 枚のテスト画像から構成される．Scenery dataset の多くが山脈の画像というのに対して，Places subset は多様性のある画像から構成されていて，それゆえ，GAN が学習するのは比較的難しい [12, 13, 14]．その一方で，多様なカテゴリや多様な画角を含んでいることにより，Places を用いて学習されたモデルは推論時により一層汎化することが期待できる．

3.3.2 評価方法

定量評価指標．画像補完を含め，生成による画像拡張の適切な定量評価は一般に知られていないが，PSNR と SSIM，FID を示す．PSNR と SSIM は拡張された領域のみから算出し，FID は拡張領域とターゲット画像の一部を含む正方形の領域から算出する．Boundless [1] では 25%分の拡張は生成部分が小さいため人によって判断することが難しく，75%の拡張は不自然な結果になると報告されている．したがって本研究はその中間の長さで結果を評価する．特に生成による画像拡張は本来の画像 (Original image) を再現することを目的としていないので，PSNR や SSIM は参考数値であることに注意されたい．一方で，[1, 2] と同様に，GAN の評価指標として使われる FID スコアを重視する．

3.3.3 結果

Outpainting 設定 vs. Inpainting 設定．この実験では，2 つの補完設定（Outpainting 設定と Mirrored input を用いた Inpainting 設定）でどちらが良い結果を導くか，3 つの手法を用いて定性的かつ定量的に評価する．図3.5(a) に示した Deep fill v2 [66] は SoTA の Inpainting 手法の一つである．図3.5(b) に示した pix2pix は汎用的な画像間変換の手法である．本実験では，Deep fill v2 として著者が配布している学習済みモデルを利用しているため，テスト画像が彼らの訓練画像にリークしていることが考え

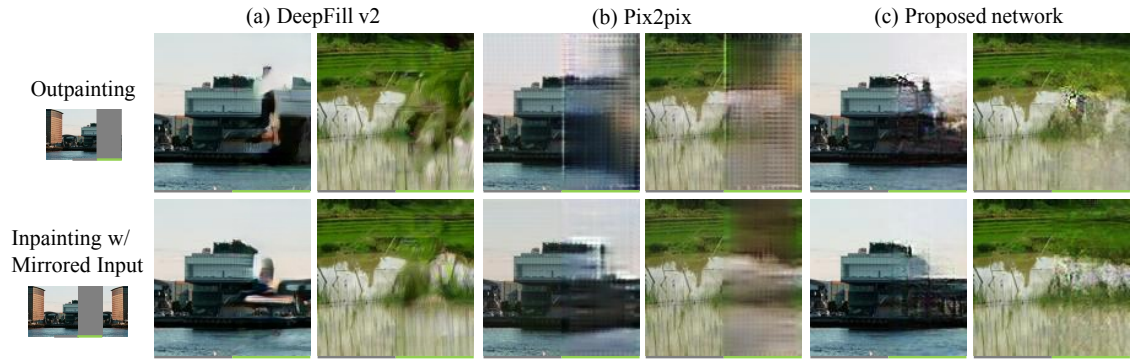


図 3.5 Outpainting 設定と Mirrored input を用いた Inpainting 設定の比較．各手法において，Outpainting と比べて Mirrored input を用いた Inpainting の方が視覚的に優れた結果を導いている．特に，画像の外側のテクスチャや構造 (建物など) である．注目すべき点は，Mirrored input に白い建物が含まれていないという条件からセマンティックのコントロールが行われていることである．つまり，白い建物は途中で拡張が止まっている一方で，その下の黒い土台ははっきりと延長されている．

られるので，手法間で比較することはできない．各手法において Outpainting 設定と Mirrored input を用いた Inpainting 設定での結果の比較を行うことができる．Pix2pix [15] と提案ネットワークの学習は Outpainting 設定と Inpainting 設定でそれぞれ個別の学習を行う．この実験では，Places subset を利用した．図3.5の拡張結果は，それぞれの手法において Outpainting 設定よりも Mirrored input を用いた Inpainting 設定で補完する方が質の高い結果を得られることを示している．具体的には，ターゲット画像から離れた位置のテクスチャや構造 (建物など) に改善が見られる．さらに注目すべき点は，Mirrored input に白い建物が含まれていないという条件に適応し，セマンティックのコントロールが行われている．具体的には白い建物は途中で拡張が止まっている一方で，その下の黒い土台ははっきりと延長されている． 追加の結果を図3.13に示す．表3.1は，定量評価の結果を示している．それぞれの手法において，Outpainting 設定よりも Mirrored input を用いた Inpainting 設定の方が良い FID スコアを導いている．画像拡張のために Inpainting の SoTA を利用する場合，Outpainting 設定では本来の性能が発揮されないので，Mirrored input を用いて Inpainting 設定を扱うようにすることでより本来の性能を発揮することが可能と考えられる．

他の生成による画像拡張手法との比較．この実験では，提案手法と既存の生成によ

表 3.1 Outpainting と，Mirrored input を用いた Inpainting の Places subset での定量比較．Inpaint. w/ M.I. は Inpainting with Mirrored Input を表す．

	DeepFill v2 [66] PSNR↑/SSIM↑/FID↓	Pix2pix [15] PSNR↑/SSIM↑/FID↓	Ours PSNR↑/SSIM↑/FID↓
Outpainting	**13.97**/0.2371/10.56	14.33/0.2144/88.27	**15.16**/**0.2732**/10.46
Inpaint. w/ M.I.	13.57/**0.2372**/**9.00**	**14.49**/**0.2603**/**68.22**	14.05/0.2508/**9.37**

表 3.2 Places subset と Scenery dataset での，他の生成による画像拡張手法との定量比較．

	Places subset PSNR↑/SSIM↑/FID↓	Scenery dataset PSNR↑/SSIM↑/FID↓
Boundless [1]	**15.19**/**0.2511**/17.30	**18.56**/0.4037/56.81
VLNS [2]	13.51/0.2035/24.16	17.74/0.3822/43.51
Ours	14.05/0.2508/**9.37**	17.35/**0.4157**/**31.48**

る画像拡張手法との比較を行う．Boundless [1] は，多様なシーンクラスで Outpainting を行えるように，シーンクラスの特徴量を Discriminator での識別に利用する手法である．公式実装は非公開であるため，この実験では再現実装を Places subset と Scenery dataset のそれぞれで学習した時の結果を示す．Boundless [1] は 64×128 の入力画像から 128×128 の拡張結果を得る．VLNS [2] は，水平方向へ特徴量を効率的に伝達するための Resblock を導入した手法である．公式実装を２つのデータセットで学習した結果を示す．VLNS [2] は，128×128 を入力として 128×256 に拡張する．図 3.6は，Places subset での結果を示している (追加の結果を図3.14に示す)．それぞれの手法では入出力の画像サイズは異なるが，同じ 128×128 の領域のみを示している．Boundless [1] は，ターゲット画像から離れた領域ほど生成の質が下がる．多くの結果では，ターゲット画像から離れた領域のテクスチャがブラーになっている．一部の結果では，画像の端に明らかなアーティファクトが存在する．その理由の一つは生成しようとするピクセルの位置が入力画像のピクセルと離れているため，畳み込

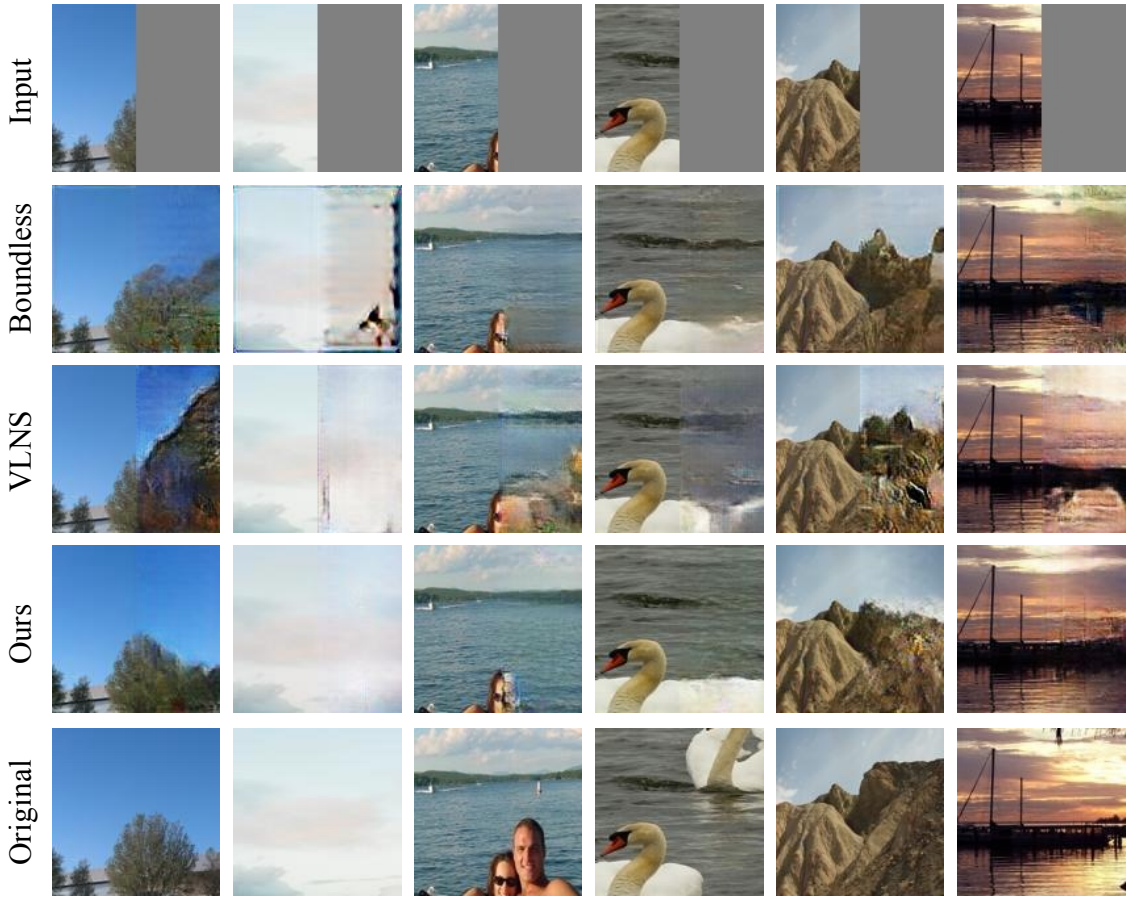


図 3.6 他の生成による画像拡張手法との Places での定性比較．Boundless [1] は，外側の生成ピクセルほど詳細なテクスチャが失われている．VLNS [2] は Places のような多様なシーンクラスのデータセットを学習するのは困難であると言える．一方，提案手法は多様なクラスのデータセットに対して，細部を維持した拡張を実現している．

み層による情報の伝達が十分にできていないからである．他の理由は，境界の条件は，拡張する領域の片側だけが利用可能であることである．つまり，補完領域の左側は入力領域に接しているので，そことの整合性は補完の良い制約となる．VLNS [2] は，本来山脈など，水平方向に類似セマンティックが続く画像の拡張を想定した手法であるため，Places subset に含まれる多様なシーンを学習するのは困難である．一方で，同一のシーンクラスで構成された Scenery dataset は，Places subset よりも比較的学習しやすい．そのため，図3.7のように，Scenery dataset での各手法の結果は視覚的には意味のある違いは見られない．表3.2には，Places subset と Scenery dataset での PSNR, SSIM, FID スコアを示している．本研究が最も重要視している FID スコ

アは，いずれも Ours が他の生成による画像拡張手法を超えている．まとめると，提案する Inpainting ネットワークと Mirrored input の組み合わせは，生成による画像拡張手法の SoTA よりも見た目も FID スコアも両方超える画像拡張を実現する．これは，今後の生成による画像拡張手法の SoTA は Inpainting の SoTA 手法によって実現される可能性が高いことを意味する．

Ablation study．提案ネットワークと損失関数 (Loss function) の効果を検証するために，それぞれ要素を一つ欠落させた次の条件でモデルを学習した：(a) Bottleneck feature matching loss を用いずに提案手法を学習した場合，(b) U-net から Skip connection を除いた場合，(c) Generator の Resblock を RMDC ではなく，pix2pixHD と同じ Resblock にする場合である．その結果を，図3.8に示す (追加の結果を図3.15に示す)．まず，(a) `No bottleneck feature matching loss` の結果は，生成されるコンテンツのストラクチャ・シルエットが曖昧である．一方，`Ours` の結果では建物の構造や木の形がはっきりと認識可能である．Bottleneck feature matching loss の働きは，Bottleneck が high-level 特徴の補完を実行することを強制することであり，後続の Decoder が高周波成分の補完とストラクチャの決定に専念できるようにすることができる．また，定量的には，`No bottle feature matching loss` の FID スコアは 9.62 で，`Ours` (9.37) よりも少し悪い．(b) `no skip connection` の結果では，Boundless [1] でも主張されていることだが，高周波成分の生成が難しいように見える．さらに，しばしばターゲット画像には含まれない色のアーティファクトがある．(c) `No RMDC` では，生成されるコンテンツのストラクチャなどがはっきりとしていない．この理由は，大きな受容野を獲得する機構がないので，画像中の離れた位置にある，参考にすべきコンテンツからその情報を獲得できていないと考えられる．各コンポーネントは重複する役割も持っていると考えられ，全てを含む `Ours` が視覚的に最も良い．

左右両方向への拡張．図3.9は，左右両方向へ拡張を行った結果画像を示している．本手法では，入力画像の右側を補完する学習を行うが，入力画像を反転させることにより左側の補完をすることが可能である．この方法を用いて，左右両方向への拡

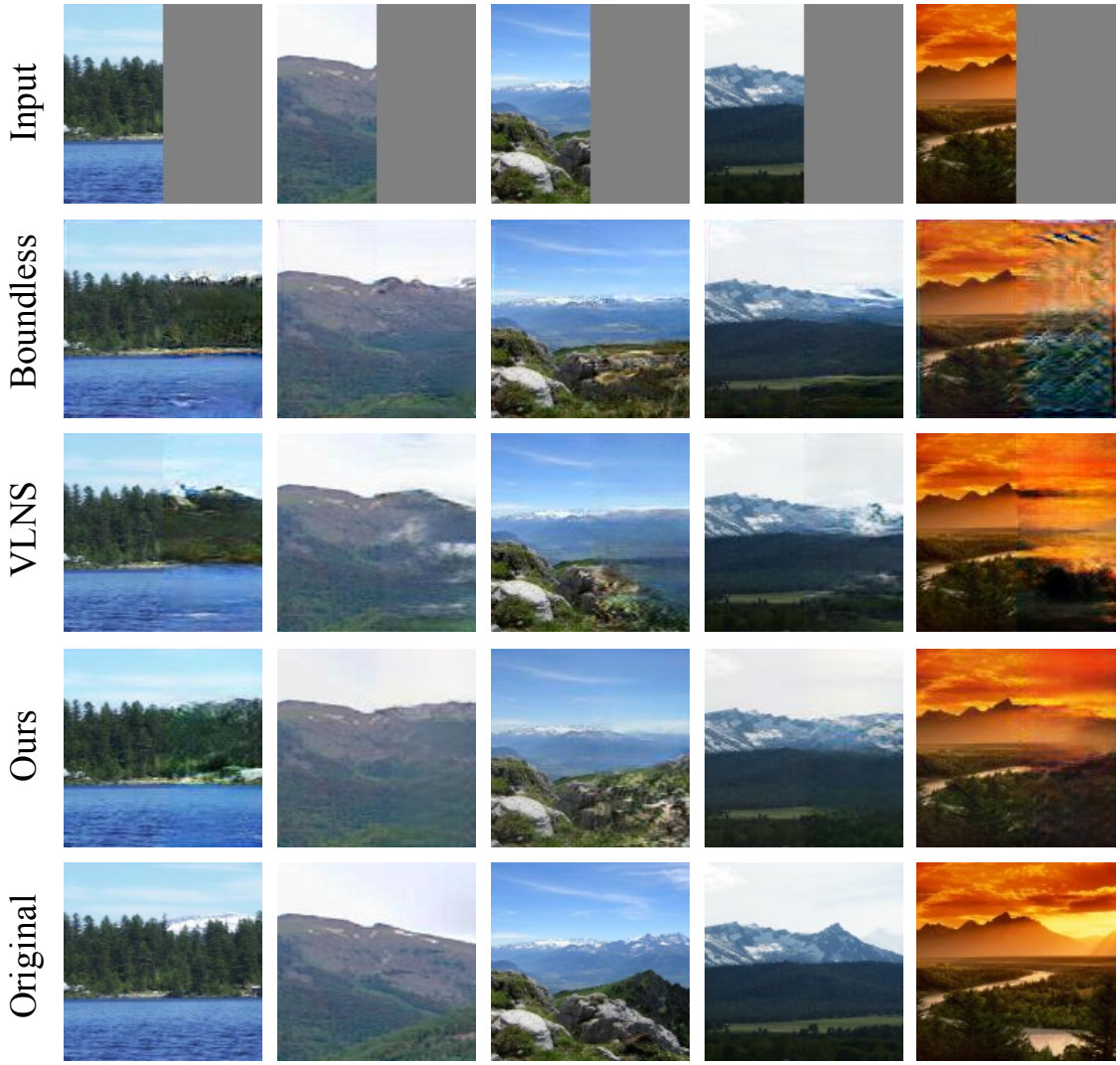


図 3.7 Scenery dataset での，他の生成による画像拡張手法と提案手法の定性比較．限定的なシーンクラスのみを含む Scenery dataset は Places subset よりも比較的簡単に学習することが可能である．そのため，それぞれの手法の結果には視覚的に意味のある違いが見られない．

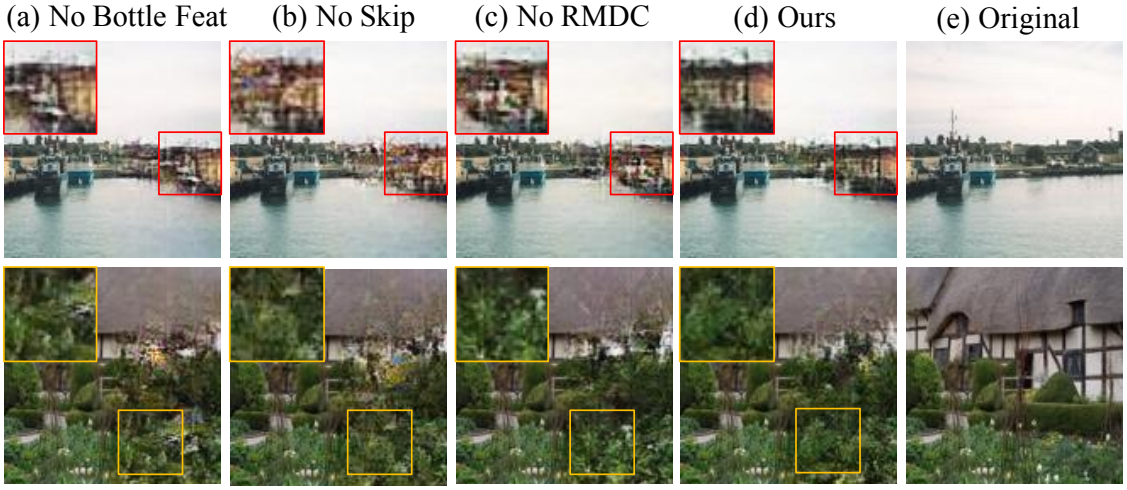


図 3.8 Ablation study．左から，bottleneck feature-matching loss，skip connection，そして RMDC を取り除いた場合の生成結果である．それぞれ画像の右半分が補完領域である．

張を本手法の応用的な使用方法として示す．

3.4 議論

3.4.1 ミラーに適した位置の調査

どの領域がミラーすることに適しているかを調べるために，図3.10に，ミラーする領域を変えた生成結果を示す．図3.10の (a) は図3.2(c) と同じで，図3.10(f) は図3.2(d) と同じものである．図3.10が示すように，ミラーする位置の数ピクセルの差が結果の大きな違いにはならないが，一方で，図3.2の (c) と (d) ような違いは明らかな結果の差になる．特に，図3.10の (a)，(b)，(c) のように，生成領域の両端が同じようなピクセルによって囲まれる場合，生成されるピクセルは横一列に同じようなパターンが繰り返されるので，不自然な結果になりやすい．3.10(d) のように左半分をミラー（ターゲットの右端を中心にミラー）は，3.2.1節で述べたように，少なくとも一つの最適な解が存在するので，自然な結果を得ることが可能となる．

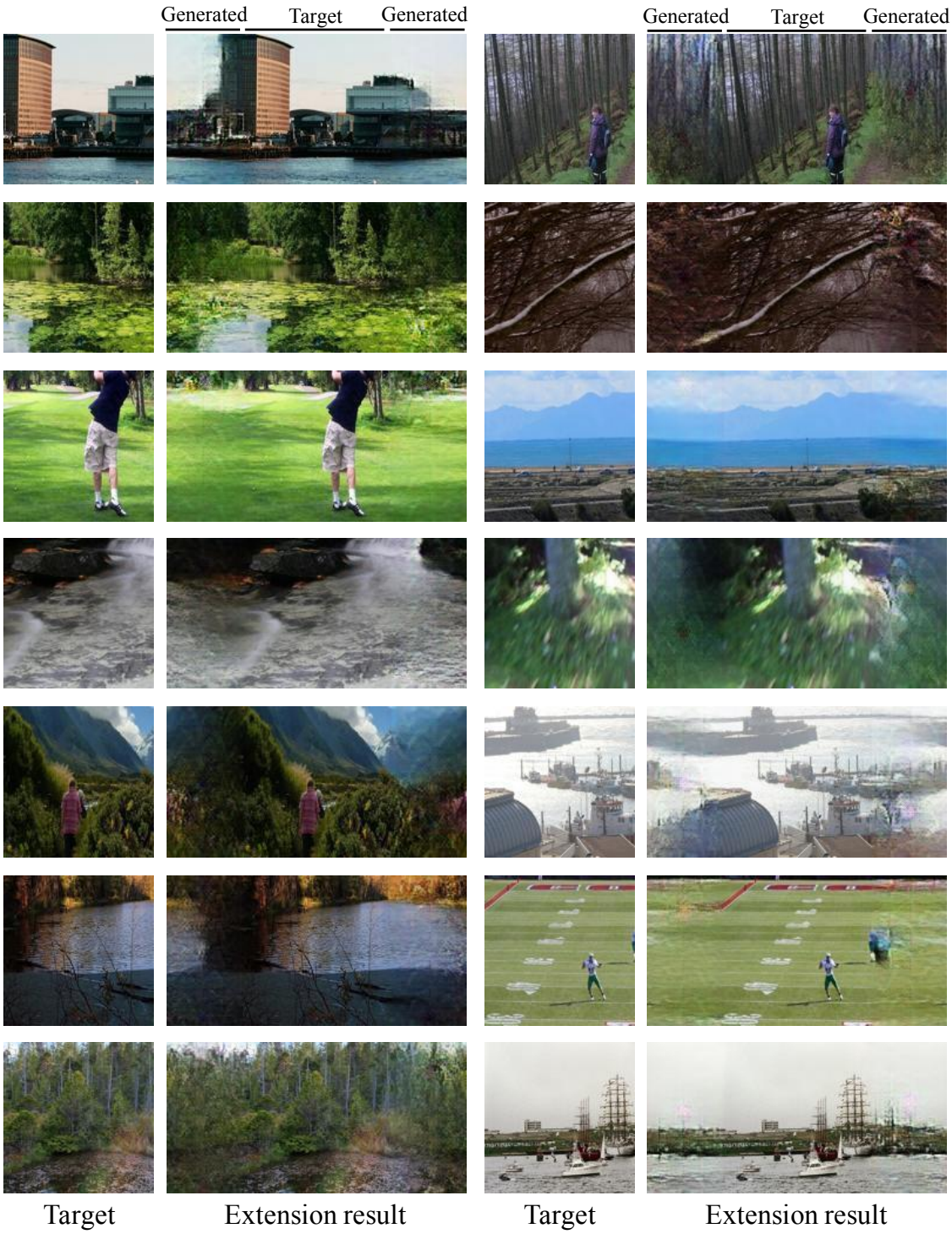


図 3.9 提案手法による左右両方向への拡張結果.

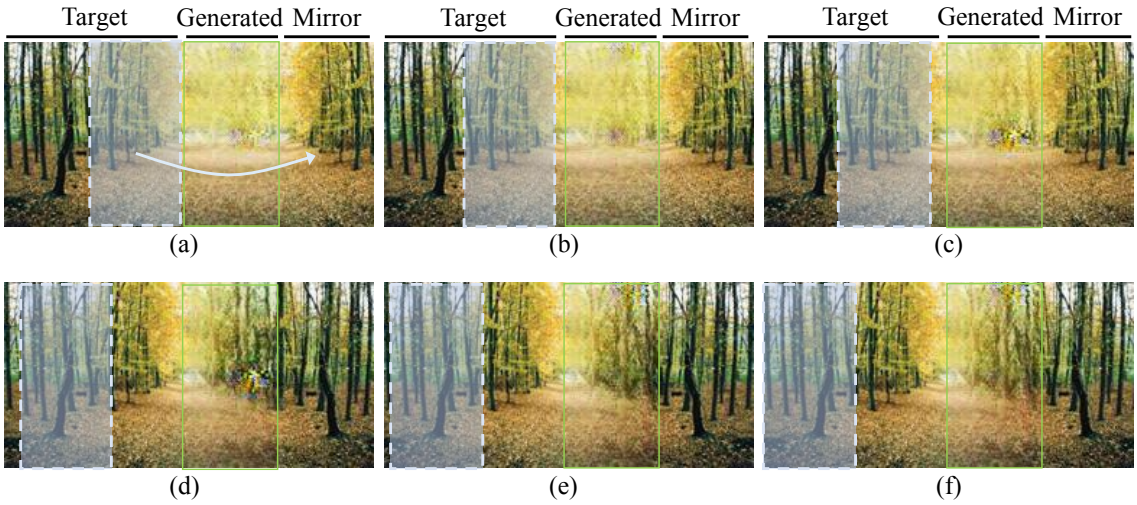


図 3.10 ミラーの位置による結果の違い．これらの画像は，ミラーされるターゲット画像の一部の領域の位置を変更した場合の結果を示している．画像 (a) は図3.2(c)，画像 (f) は図3.2(d) と同じ画像である．

3.4.2 提案手法の限界

既存の Inpainting 手法と同様に提案手法も複雑なオブジェクト（人物など）を補完するのは難しい (図3.11(a))．他にも，図3.11(b) のように，補完領域に接する境界に複雑なオブジェクトが配置されるように，そのオブジェクトの一部をミラーすることは拡張領域にアーティファクトが発生する原因となる．また，そもそもミラーには適さないコンテンツ（文字など）が含まれる画像は少なくない．このような画像の対処方法は，Adobe Photoshop に搭載されているシームカービングを拡張した `content-aware scale` のように，ミラーさせるべきでないコンテンツをあらかじめ選択するという発展方法が考えられる．図3.12のように，ミラー画像としてパディングする前に Inpainting によってそのようなコンテンツを取り除くことで，ミラーに適した背景のみをパディングすることも可能である．

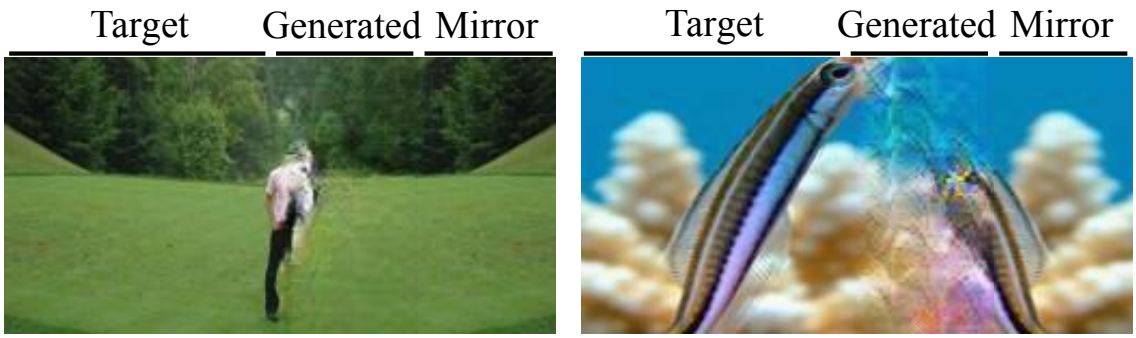


図 3.11 失敗例．提案手法は，例えば人のような複雑な形状を持つ物体の補完が困難である．したがって，補完領域の境界へ複雑な物体の一部分をミラーしてしまうとアーティファクトの原因となる．

(a) Mirrored input (b) Output (c) Modified mirrored input (d) Output

図 3.12 Mirrored input に対する修正での結果の操作．提案手法は，あらかじめ入力の一部を修正することによって，出力を変化させることが可能である．これによって，ユーザーはミラーに適さない物体を取り除くことができ，ユーザーが出力をコントロールする機会を与える．

3.4.3 今後の展望

提案手法は Inpainting の SoTA ではないが，生成による画像拡張においては，Inpainting の SoTA を用いた場合に匹敵する視覚的結果を得ることができている．近年の Inpainting は自由形状を補完できることが重視されている一方で，本研究では，補完領域がどの画像でも同じという前提を利用しているからだと考えている．今後出現する Inpainting の SoTA が自由形状と長方形でかつ大きな穴も自然に補完できるようになればその Inpainting 手法と Mirrored input の組み合わせによるアプローチが今後の画像拡張の手法としても有望である．

3.4.4 まとめ

本研究では Mirrored input を提案した．生成による画像拡張を内挿の問題に置き換えることで，既存 Outpainting 手法の，生成ピクセルがターゲット画像から離れるほど詳細なテクスチャが失われるという限界や水平方向に同一のセマンティックを繰り返すという制限に対処した (目標１と目標３)．提案した Inpainting ネットワークと Mirrored input は，多様なシーンを含むデータセット上で，Outpainting の SoTA よりも見た目も FID スコアも両方超える画像拡張を実現した (目標 2)．このことは，今後の Outpainting による画像拡張の SoTA は Inpainting の SoTA 手法によって実現される可能性が高いことを意味する．また，Mirrored input によって遠くのピクセルを近くに配置することで見た目の質が向上したということは，既存の Outpainting 手法はまだ遠くのピクセルから十分に情報を得られないネットワーク構造であるという課題が存在することを意味する．

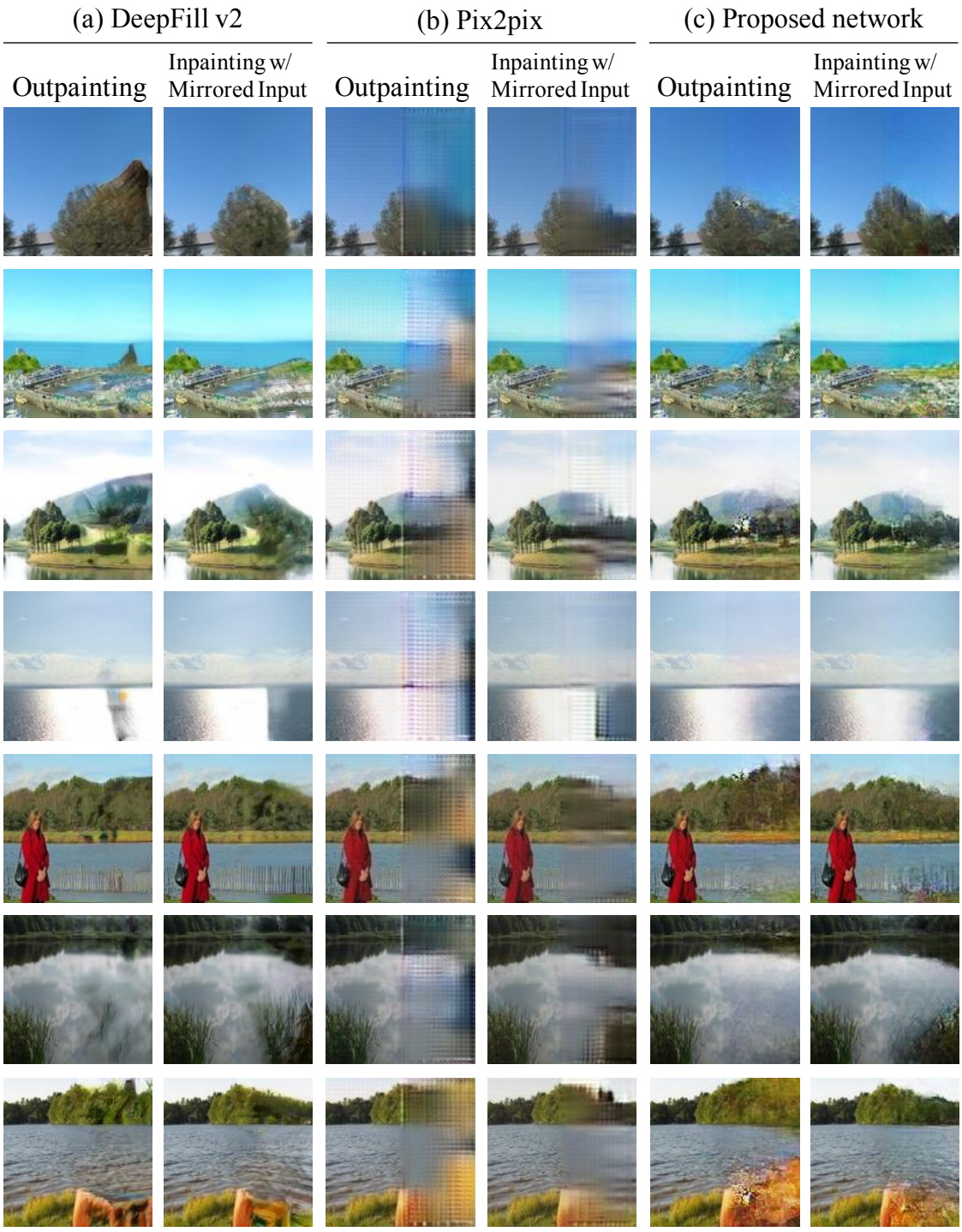


図 3.13 Outpainting 設定と Mirrored input を用いた Inpainting 設定の比較の追加結果. それぞれの手法において，Mirrored input を用いることによって Outpainting 設定と比べて視覚的により良い結果となることを示している.

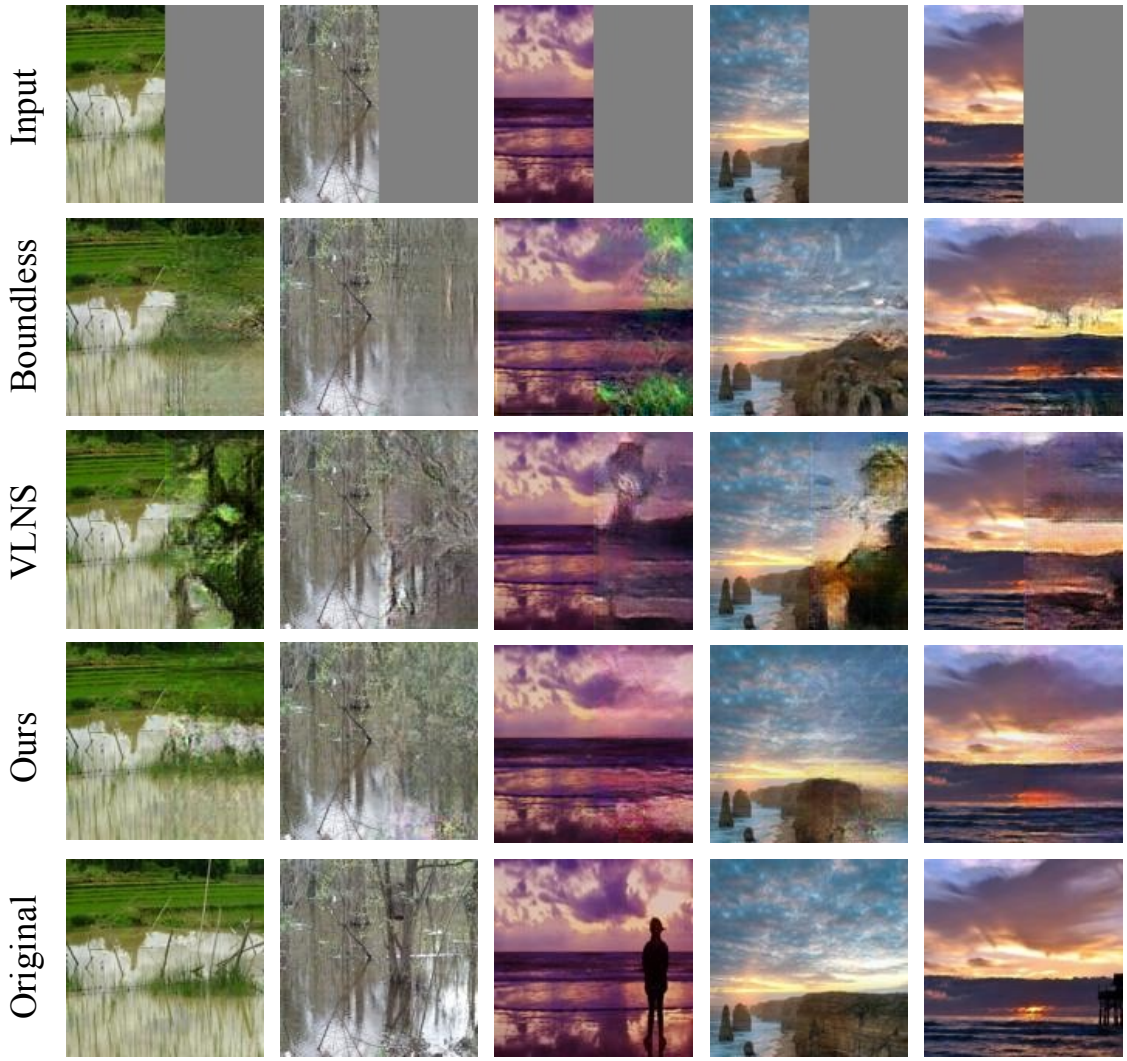


図 3.14 Places subset での他の生成による画像拡張手法との定性比較の追加結果．Boundless [1] は外側の生成されるピクセルに対して詳細なテクスチャが得られない結果となる．この実験では，VLNS [2] は Places subset のような多様なシーンクラスを含むデータセットでは学習が難しいことが観測される．一方，提案手法は，データセットに含まれる多様なシーンでより詳細を保ちつつ拡張を達成する．

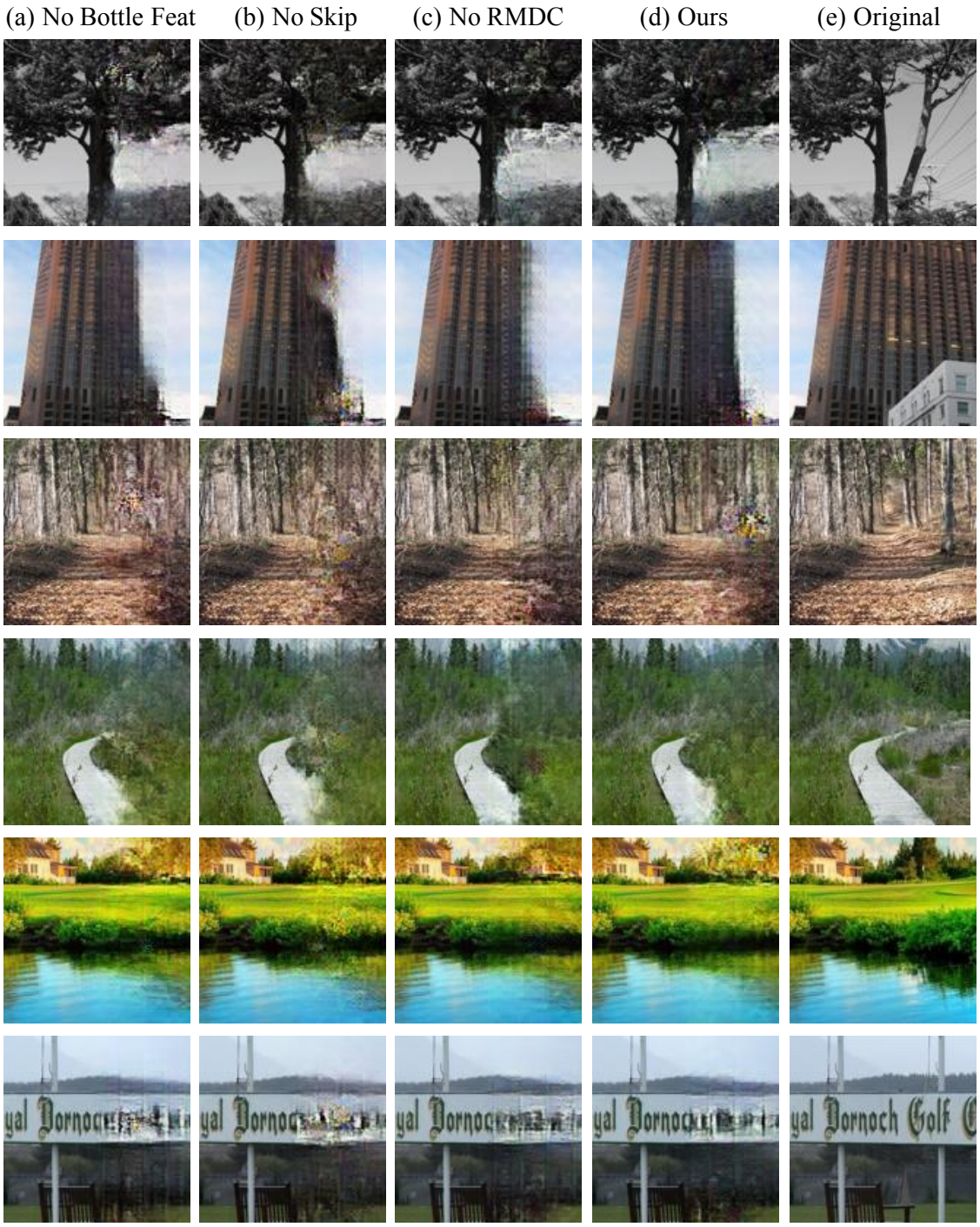


図 3.15 Ablation study の追加結果．左から，bottleneck feature-matching loss，skip connection，そして RMDC を取り除いた場合の生成結果である．それぞれ画像の右半分が補完領域である．

第4章　360度画像補完

4.1　概要

研究動機．360 度画像は近年の 3DCG 制作においてライティングや背景を効率的に作成する際に有用である．例えば，デザイナーたちは，近景の 3D 物体を時間をかけて作成し，背景は通常画角の 2D 画像 (Normal Field of View 画像) や 360 度画像を用いて短時間で作成するという方法を取ることがある．しかし，3D 物体の後方に 2D 画像を配置し背景を作るという制作方法 (図4.1 Background Creation Demo 参照) において，3D 物体の表面上に反射する景観を完全に表現することはできない．物体を 360 度囲うように画像が存在する場合にはこの問題は起きないが，一般に 360 度画像，特に High Dynamic Range Image (HDRI)，は Normal Field of View (NFoV) 画像に比べて用意するコストが高い．

本研究は，背景として用意している画像と整合性のある 360 度画像を獲得する方法として，NFoV の写真の周囲を補完し 360 度画像に変換するという問題に取り組む．この問題を解くことによって，ユーザーは NFoV の画像のみで，物体に周辺環境を反射したり [3, 4]，HDRI の場合では Image-based lighting によって自然な影や照明を実現したりすることが可能となる [4, 54].

既存研究の限界．デザイナーによる利用を想定すると，任意のサイズの NFoV 画像に対して推論が可能であり，整合性のある 360 度画像を多種多様に生成することで選択肢があることが望ましい．しかしながら既存の手法は，推定が決定的であり，また一度の学習で固定の解像度しか正しく推論できない．図4.2のように，360IC[3] は，512×256 で学習した場合，1024×512 では多くのアーティファクトが発生する．

図 4.1 研究概要．360 度画像の性質を考慮した Transformer ベースの Outpainting 手法を使うことにより，狭い画角の画像から尤もらしい環境を生成する．それは効率的な 3DCG の制作を実現することにつながる．

また，1 つ入力に対しては 1 つの出力のみである．TT[25] は，Transformer を用いた画像間変換 (Im2Im) 手法であり，Transformer によって学習された分布からサンプリングすることで多様なシーンを生成することができる．しかしながら，TT で提案された，学習時よりも大きな解像度を変換するための Sliding attention window を使用しても，360IC 同様にアーティファクトが発生する．360 度画像固有の歪みがその原因と考えられ，ネットワークの学習の際に，360 度画像固有の歪みの相対位置関係が記憶されるからである．また，両端の連続性のような 360 度画像の性質を満たさない．

本研究の目標．以上を踏まえ，本章の目標は，次の性質を持つ 360 度画像の Outpainting の実現である．

1. 1 つの入力に対して多様な出力をサンプリング可能であること．

2. 学習時とは異なる解像度においても推論可能であること．

3. 尤もらしい見た目の出力結果を得ること．

提案手法．提案手法のキーアイデアは，360 度画像の性質を考慮しつつ，Outpainting 手法に Transformer [18] を導入することである．そこで，多様な補完結果が得られる 2 段階生成フレームワークの提案と，360 度画像特有の性質を持たせるための 2 つのテクニックを提案する．

図 4.2 先行手法の限界．(a) CNN ベースの手法 [3] と，Transformer ベースの変換手法 [25] は学習時の解像度 (512×256) に過適合してしまう．さらに，(b) は左右両端のつながりがない．

具体的には，まず，CompletionNets と AdjustmentNet からなる 2 段階生成フレームワークを提案する．(1)CompletionNets は Encoder-Decoder である VQGAN [25] と，Transformer Encoder [18] を用いた補完モジュールである．固定サイズの画像を入力し，多様な補完結果をサンプリングすることが可能である．しかしながら，CompletionNets のみでは，図4.2の TT と同様に，条件として入力した画像との整合性が不十分であり，さらに固定サイズのみしか生成できない．この課題に対処するために，提案手法には U-net 構造の (2)AdjustmentNet を 2 ステージ目として導入する．CompletionNets の出力結果を入力画像と，色・つなぎ目・解像感の面で整合性を向上させる．入力画像のサイズに合わせて CompletionNets の解像感を調整するので，任意の画像サイズで補完結果を得ることが可能である．

次に，2 段階生成フレームワークの結果をさらに 360 度画像特有の性質を持つよう

にするためのテクニックを導入する．360 度画像の特徴の 1 つである画像両端の連続性を実現するために，Transformer の新たな自己回帰推定として，Circular inference を提案する．これは，画像を巡回するように推論を行うことで，画像の両端をピクセルレベルでもセマンティックレベルでもつながりを向上させる．さらに，知覚的な品質を向上させるために，VQGAN の学習のための WS-perceptual Loss を提案する．これは 360 度画像が緯度方向に情報量の差があることを反映し，情報量が多い領域で重点的に損失を計算する損失関数で，360 度画像をモデリングする性能を向上させる．

利点．実験では，提案手法が多様な補完が任意の解像度で実行可能なことだけでなく，提案手法が定性的にも定量的にも複数の SoTA 手法の性能を上回ることを示す．例えば，FID スコアにおいて提案手法は 360IC より 1.7 倍改善し，EnvMapNet [4] (256×128) に対して 16 倍のピクセル数の画像 (1024×512) でもっともらしい補完を実現する．

応用．さらに，1 枚の通常画角の画像から HDR 環境マップを作成し，3DCG 上の背景作成と照明を行うパイプラインを提案する．このデモを通して，提案手法が 3DCG に利用できる 360 度画像補完のクオリティに達していることと，効率的な背景制作の助けになることを示す．

提案手法は 1 枚の NFoV 画像のみから高品質な 360 度画像を経て，3DCG 上で自然な反射や照明の表現を可能にするだけでなく，多様な補完結果を提示することで，ユーザーに結果の中から選択の余地を与えることも可能とする．この性質を踏まえて，最後に潜在的な応用についての議論を行う．

4.2 提案手法

本研究は，NFoV の画像の周囲を補完して 360 度画像を生成する．ここでは 360 度画像として Equirectangular Projection image (ERP 画像) を利用する．図4.3は提案手法のフレームワーク概要を示している．フレームワーク全体の入力は，不完全な画像

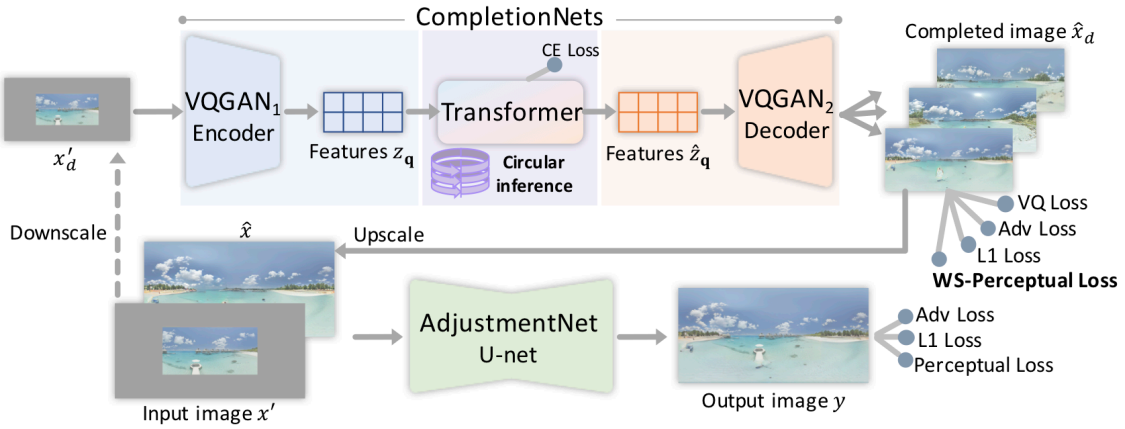


図 4.3 フレームワーク概要．提案手法は 2 つのモジュール，CompletionNets と AdjustmentNet から構成される．CompletionNets は，固定サイズの入力画像から多様な補完をサンプルすることができる．AdjustmentNet は CompletionNets の出力と入力画像との色味，つながり，解像度の整合性を向上させ，結果として任意の画像サイズの補完を行える．

$x' \in \mathbb{R}^{H\times W\times 3}$ である．学習中は，ERP 画像 $x \in \mathbb{R}^{H\times W\times 3}$ から切り出した一部の領域と残りをグレー値で埋めて作成する．出力は完全な 360 度シーンの復元画像 $y \in \mathbb{R}^{H\times W\times 3}$ である．

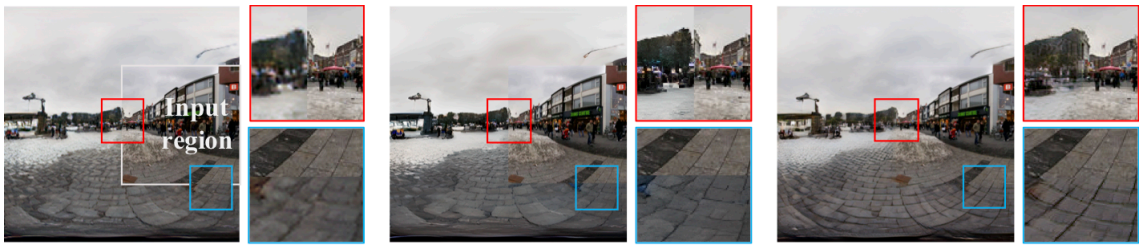
提案手法は，２段階ステージで構成され，CompletionNets と AdjustmentNet である．まず，不完全な入力画像 x' を固定サイズにダウンサイズして CompletionNets の入力とする．CompletionNets は，その不完全な画像 $x'_d \in \mathbb{R}^{h\times w\times 3}$ を Transformer を使って補完する．この補完画像 $\hat{x}_d \in \mathbb{R}^{h\times w\times 3}$ は固定サイズであるので，本来の入力画像のサイズに元に戻す．次に，学習時の解像度とは異なる任意の解像度で推論可能とするために，AdjustmentNet は補完画像 $\hat{x} \in \mathbb{R}^{H\times W\times 3}$ と入力画像 x' を使って，入力画像 x' に合わせて補完画像上の高周波のテクスチャを推定する．それに加えて，つなぎ目や色味も修正し，最終出力 y を得る．

以下では，多様な補完を任意の解像度で行うためのネットワーク構造の提案と，360 度画像の性質を向上させるための新たな損失関数 WS-perceptual loss と新たな Transformer の推論方法 Circular inference を提案する．

4.2.1 ネットワーク構造

CompletionNets. CompletionNets の目的は，Transformer を用いた多様なシーン補完である．TT はすでに Transformer を用いた Im2Im で補完が行えることを示している．しかしながら4.1節で考察したように，Transformer を使ったそのままの Im2Im は 360 度画像には適していない．それゆえ，TT を拡張し，360 度画像の性質を満たす補完を行うための機構を新たに導入する．CompletionNets の基本のネットワーク構造は TT と同様であり，2 つの VQGAN[25] と Transformer である．TT のアプローチは，Vector-quantized image modeling と呼ばれ，量子化された画像トークンのシーケンスをモデリングする方法である．Transformer がピクセル表現の代わりに量子化された特徴量の並びを推定することで計算量を抑えつつ画像全体のモデリングを行う．VQGAN は，Encoder-decoder CNN のボトルネック部分で特徴量を量子化する機構 [26] を使って，Transformer で扱うための量子化された特徴量を得る．VQGAN は end-to-end で学習可能である．提案するアプローチでは，CompletionNets は，不完全な画像をエンコードするための VQGAN_1 と，Transformer によって補完された特徴量をデコードするための VQGAN_2 を持つ．CompletionNets は Transformer を持つため計算量を抑える目的で，かつ，ERP 画像の固有の歪みの影響で VQGAN もサイズに過剰適合することを考慮して，提案手法は固定サイズの画像を入出力として扱う．提案手法の Transformer は，360 度画像のシーンを量子化特徴量の並びとしてモデリングし，学習された分布からサンプリングを行うことで多様な画像補完を行う．

AdjustmentNet. 任意の画像サイズでの補完を実現するために，AdjustmentNet を提案する．AdjustmentNet は，CompletionNets の出力と入力領域との整合性を向上させるためのネットワークである，2 段階ステージが採用されている高解像度画像を扱う画像補完手法 [67, 37] では，2 ステージ目の主な役割は，高周波成分を補い画像のリファインメントを行うことである．しかしながら，図4.4のように提案手法では画像の超解像だけでは不十分である．図4.4(a) は，アップスケールされた補完画像に入力領域を合成した画像である．Transformer によるサンプリングでは補完領域



(a) CompletionNets + Bicubic　(b) CompletionNets + Real-ESRGAN　(c) CompletionNets + AdjustmentNet

図 4.4 AdjustmentNet の効果．(a) と (b) は CompletionNets の出力に対して解像度だけではなく，色味と入力領域とのつながりも調整する必要があることを示す．

だけではなく，入力領域も再サンプリングを行う．そのため，得られた補完領域は再サンプリングされた入力領域に合うように予測されているため，本来の入力画像との整合性が取れていない．SoTA の超解像手法 [68] を行なった (b) でも，リファインメントを行うだけでは不十分であることを示している．一方で提案手法は，(c) のように色，つなぎ目，解像度という点で CompletionNets の出力を入力画像に合うように調整する．AdjustmentNet を使って補完領域と入力領域との整合性を向上することにより，結果として，提案フレームワークは任意の画像サイズでの補完を実現する．AdjustmentNet は，U-net 構造であり，VQ 機構 [26] を取り除いた VQGAN と同じ CNN 構造で実装する．

4.2.2 学習

WS-perceptual loss．VQGAN は，量子化された画像の特徴量を得るためのネットワークであり，CNN によって画像の局所領域をモデリングする．TT は Adversarial loss $\mathcal{L}_{\text{GAN}}$, L1 loss $\mathcal{L}_1$, Perceptual loss $\mathcal{L}_{\text{Perc}}$ と VQ loss $\mathcal{L}_{\text{VQ}}$ を使った自己教師学習を提案している．一方で，より一層 ERP 表現に適した局所領域のモデリングを行うために，新たなロス関数 WS-Perceptual loss を提案する．この損失関数は，緯度方向に沿って領域ごとの情報量の差があるという ERP 表現の性質を反映する．従来手法 [4, 54] では，L1 などのピクセルレベルの差分ロスに対して球への投影を考慮して重み付けをする．しかしながら，ERP 画像における小領域ごとの情報の密度は，中央

部分が大きいことを考慮すると，ピクセルレベルだけでなくセマンティックスなどの high-level 特徴量のモデリングも中央領域に重点的に行うべきであると考える．提案する Loss は，WS-Perceptual loss であり，Perceptual loss (LPIPS) [69] を単位球上での loss へ拡張した関数である．WS-PSNR[70] のように球への投影を考慮して次のような重みを用意する．

$$w'_l(u,v) = \cos((v - H_l/2 + 1/2) \cdot \pi / H_l), \tag{4.1}$$

ここで，u と v は特徴抽出器の l 番目の層での特徴量 (サイズは $H_l \times W_l$) の座標を表す．式4.1を用いて，Perceptual loss $\mathcal{L}_{\text{Perc}} = \sum_l \frac{1}{H_l W_l} \sum_{u,v} \|w_l \odot (y^l_{uv} - x^l_{uv})\|_2^2$ を各解像度で重み付けをする．

$$\mathcal{L}_{\text{WS-Perc}} = \sum_l \frac{1}{\sum_{u,v} w'_l} \sum_{u,v} w'_l \odot \|w_l \odot (y^l_{uv} - x^l_{uv})\|_2^2. \tag{4.2}$$

VQGAN．それぞれ Encoder と Decoder を持つ VQGAN_1, VQGAN_2 の両方を，どちらも次の式で学習する．

$$\mathcal{L}_{\text{VQGAN}} = \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_1 + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{WS-Perc}}\mathcal{L}_{\text{WS-Perc}}. \tag{4.3}$$

VQGAN_1 は，不完全な入力画像の量子化された特徴量 $z_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ を得るために，欠損領域のある 360 画像を使用して再構成学習する．一方で VQGAN_2 は，量子化された特徴量 $\hat{z}_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ から完全な 360 度画像を得るデコーダーを得るために，完全な 360 度画像を使用して再構成学習を行う．

Transformer．Transformer を，360 度シーンのモデリングを行い，補完を行うために学習する．学習済みの VQGAN_1 の量子化特徴量 $z_{\mathbf{q}}$ から学習済みの VQGAN_2 の量子化特徴量 $\hat{z}_{\mathbf{q}}$ への変換を教師として，Transformer はインデックスのシーケンス c を条件として連続するインデックスの次のインデックスの分布を予測するように次

の式で学習する．

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x\sim p(x)}\left[-\log p(s|c)\right], \tag{4.4}$$

ここで，$p(s|c) = \prod_i p(s_i|s_{<i}, c)$ であり，Transformer は量子化された特徴量 ($z_{\mathbf{q}}$ と $\hat{z}_{\mathbf{q}}$) を直接には扱わず，それらに割り当てられたインデックス (c と s) を扱う．

AdjustmentNet．AdjustmentNet が Completion stage の出力画像を入力画像と整合するように調整するためのネットワークになるように学習する．それゆえ，入力画像を次のように前処理し，それを AdjustmentNet で元の画像に復元するという自己教師方法で学習する．あらかじめ，上下部分の歪みを過学習することを避けるために，ERP 画像の一部のみを切り出して GT 画像とする．(1) 正解領域との境界部分の調整のために，学習済みの VQGAN_2 で GT 画像の再構成を行うことで，GT 画像との差分を持った再構成画像を獲得する．(2) 色の調整を学習させるために，VQGAN_2 での再構成の前に入力画像に Color jitter を追加することにより，GT 画像と色味が異なる再構成画像を獲得する．(3) 解像度の調整のために，VQGAN_2 で再構成する GT 画像を，前もってスケールダウンする．再構成後にバイキュービック法でオリジナルのスケールに戻す．以上の方法で，GT 画像と比べて，色味が異なり，高周波テクスチャが不足している再構成画像を獲得する．この画像よりもさらに小さい領域の GT 画像と結合して AdjustmentNet の入力画像とすることで，GT 領域をヒントしつつ，色とつなぎ目，解像度の調整を行う学習が可能である．GT 画像は ERP 画像を切り出しているものであるため，式4.3のように WS-perceptual loss ではなく本来の Perceptual loss を利用する．Adversarial loss, L2 loss, 本来の Perceptual loss を用いて学習する．以上をまとめると次の式にようになる．

$$\mathcal{L}_{\text{Adjust}} = \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_1 + \lambda_{\text{Perc}}\mathcal{L}_{\text{Perc}} \tag{4.5}$$

最後に学習手順について述べると，各ネットワークは個別に学習する．２つの VQGAN を式4.3で学習し，その後 Transformer を式4.4で，AdjustmentNet を式4.5で

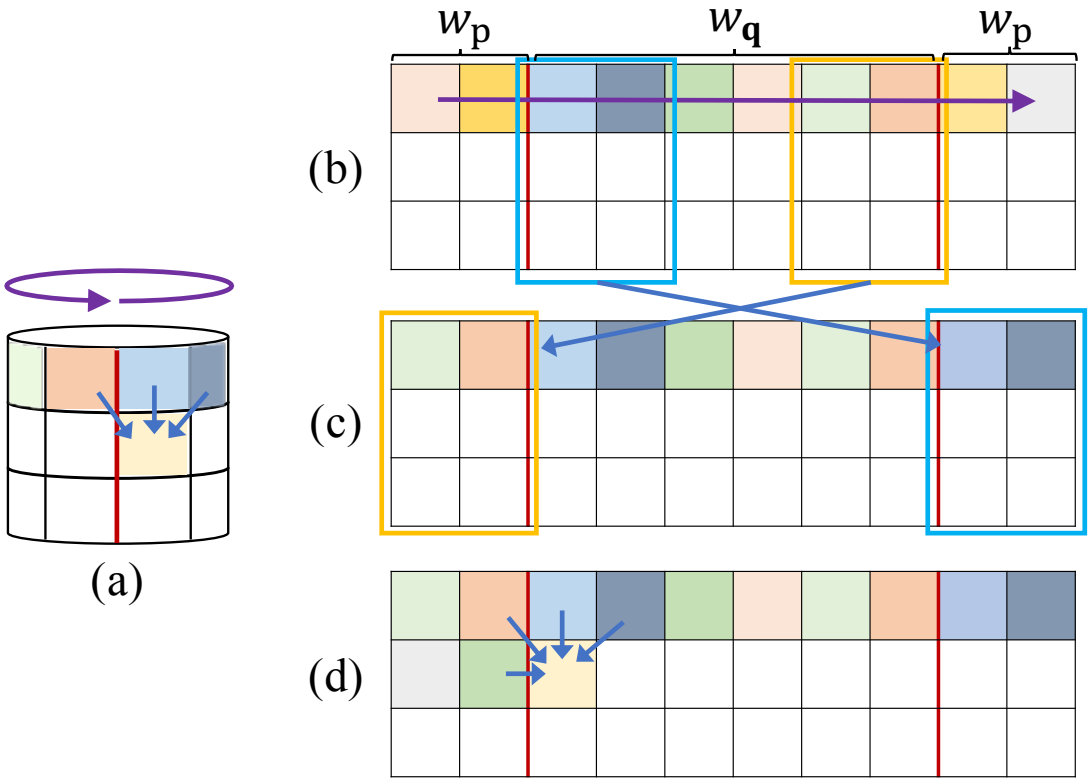


図 4.5 Circular inference. (a) が目標とする推論方法を示しており，赤線は画像の左右両端を表す．左右反対側からの情報も用いて推論を行う．(b) のようにパディングを行い，推論を行う．行の推論が終われば (c) のように左右に結果をコピー&ペーストする．このようにすることで，(d) のように左右反対側の予測結果を用いて推論を行うことが可能となる．

学習する．

4.2.3 推論

Circular inference. 360 度画像の両端は連続するという性質を反映する補完結果を得るために，Transformer の推論手順として，Circular inference を提案する．TT は，Transformer の自己回帰の推定順としてラスタオーダーの Sliding attention window を提案している．しかしながら図4.2(b) に示すように，この方法では本来繋がるべき 360 度画像の両端が不連続となる．SIG-SS では両端につながりを持たせるために Circular padding を提案している．しかしながらこれは畳み込み層に対する padding の一種であり，Transformer による 360 度画像の大域的なセマンティックスの推定の

段階では作用できない．そこで，Transformer による推定の段階で両端の連続性を考慮できるように Circular inference を提案する．図4.5(a) と図4.5(d) に示す．キーアイデアは，一部の領域は Transformer によって 2 度推定し，重なりを生じさせることである．実装として，量子化された特徴量 $z_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ の両端 (長さ w_{p}) をあらかじめ複製する ($h_{\mathbf{q}} \times w_{\mathbf{q}}$ から $h_{\mathbf{q}} \times (w_{\mathbf{q}} + 2w_{\mathrm{p}})$ になる)．そして，Transformer がラスタオーダーで推定を行う (図4.5(b))．1 つの行の推定の完了後に，量子化された特徴量 $\hat{z}_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ の両端の推定結果は反対側へコピーし置き換える (図4.5(c))．つまり，左端から長さ w_{p} は，本来の長さ $w_{\mathbf{q}}$ の右から w_{p} の推定結果によって置き換えられる．逆側も同様の処理を適用する．以上のように，Transformer の推定を画像を巡回するように行うことで，ハイレベルな特徴量での両端のつながりを持たせることができ，画像にデコードするとセマンティックレベルでのつながりが向上し，より 360 度画像としてより尤もらしい補完が可能となる．

全体の流れ．以上を踏まえて推論を式で表すと次のようになる．

$$y = G_{\mathrm{Adjust}}(G_{\mathrm{VQGAN}_2}(T(E_{\mathrm{VQGAN}_1}(x'))), x'), \tag{4.6}$$

ここで，G_{Adjust}，G_{VQGAN_2}，T，そして，E_{VQGAN_1} はそれぞれ，AdjustmentNet，VQGAN_2 の Decoder，Transformer，そして VQGAN_1 の Encoder を表す．

4.3 実験

本章では，提案手法と先行手法を定量的かつ定性的に比較し，提案手法の構成要素の効果を検証する．全周囲の見渡しや物体の挿入のアプリケーションで比較を行い，提案手法が高品質な挿入合成の結果を作ることができることを示す．

4.3.1 実験設定

実装詳細. オプティマイザーとして Adam[71] を利用し，Learning rate = 4.5e-06. $\lambda_1 = 1.0$，$\lambda_{\text{Perc}} = 1.0$，$\lambda_{\text{VQ}} = 1.0$，$\lambda_{\text{WS-Perc}} = 1.0$，そして λ_{GAN} は Esser ら [25] が提案した adaptive weight である．Transformer の学習は 20 epoch, それ以外のネットワークは 30 epoch 行う．提案手法は 1024×512 の画像を生成する．

データセット. Dataset として SUN360[61] と Laval Indoor Dataset[54] を利用する．SUN360 の `Outdoor` クラスを 47938 枚の学習用画像と 5000 枚の評価用画像に分割した．また，Laval Indoor dataset は提供された学習用と評価用の分割方法と同じであり，1837 枚の学習用画像と 289 枚の評価用画像である．Laval Indoor dataset はネットワークを学習するには枚数が少ないので，EnvMapNet が追加データセットを使用したのと同様の方針で，提案手法のモデルも SUN360 で学習したモデルをこのデータセットで追加学習 (Fine-tuning) をする．訓練時の ERP 画像のデータ数水増しとして，[3, 5] と同様に，視線方向 (View direction) を水平方向に沿ってランダムに変更する．これらのデータセットは，水平線が画像に対しても水平になるように調整されており，また，撮影時にカメラの地面からの高さもある程度統一されている．そのため，学習されたモデルもそのような入力画像に最適化されることが予想される．

評価. 定量評価指標は，生成画像の質と多様性を評価するために，FID スコア [58] を利用する．提案手法と他の手法を比較する際には，Transformer のシーケンス長は 512 である．提案手法の各要素を調査する際には，学習を効率化するために Transformer のシーケンス長を 256 とする．

ベースライン手法. 360IC と比較を行うために，このモデルを用意したデータセットで学習した．本来この手法は，600 枚の少ない訓練データセットでの過学習を避けるために 2 ステージアプローチを導入している．本実験では Hara ら [5] での 360IC の実装と同様に 1 ステージで実装し，十分な量の学習用データでそのネットワークを学習する．Encoder-Decoder network には VQGAN と同様の CNN 構造 (図4.2(a)) を利用する．そのボトルネックには VQ 機構の代わりに，提案された Parallel Dilated

図 4.6 提案手法の多様な出力．提案手法は入力画像 (左端) に対して多様で尤もらしい補完を達成する．図4.1の上段も同様の実験結果を示している．

Conv Blocks を 4 つ追加して実装した．ネットワーク構造と学習方法を提案手法と同じにすることによって，Transformer と CNN のシーンモデリングの違いを比較する．

SIG-SS と比較を行うために，著者らの学習済みモデルを使って，テスト用画像に対して推論を行なった．この著者らも SUN360 でモデルを学習している．しかしながらその学習用データと評価用データの分割方法は不明であるため，テスト画像でネットワークを学習している可能性，つまりテスト画像がリークしている可能性があることには注意されたい．Ours は 360IC で比較する際と同じ学習済みモデルである．

EnvMapNet は，コードを公開していないが，評価プロトコルとそのスクリプトのみ公開している．彼らが公開している手順に従い，Laval Indoor Dataset での同じ分割データにおいて彼らの評価スクリプトで評価を実施する．比較のために，EnvMapNet の結果の画像やスコアは彼らの論文から引用する．彼らの実験と本実験では，入力領域の位置，トーンマッピングの方法が全くは同じではない点に注意されたい．

4.3.2 多様な出力

図4.6は，提案アプローチが複数かつ多様に補完を行えることを示している．いずれの結果も 1024×512 の画像である．左の列は，入力画像である．上 2 段は SUN360 での生成結果であり，同じ学習モデルで異なる入力領域である．この上 2 段は経度

図 4.7 360IC との定性比較．入力領域は，図4.6の 1 行目と同じである．

方向に 180 度，緯度方向に 90 度分の領域を入力領域としている．次の上の 2 つは 90 度の画角に相当する領域を切り出した入力領域である．入力領域が大きい方が，生成されるテクスチャの質が良いことが分かる．最下段は Laval Indoor Dataset での実験結果である．多様な構造を持った室内シーンを推定することができている．

4.3.3 定性評価

図4.7は 360IC と提案手法の比較である．360IC は，360 度画像固有の歪みを捉えているが，中央付近の細かいテクスチャにアーティファクトが見られる．一方で提案手法は各物体のテクスチャや形状をより正確に生成できている．360IC と比べ，地面のタイルの歪みを学習できていることがわかる．360IC と CompNets only を比較することで，Convolution を使ったシーンモデリングと Transformer によるシーンモデリングの差がわかる．Dilated convolution を使うことで CNN の受容野を大きくすることができるが，これによって伝達する情報がスパースになり，中央領域のテクスチャの生成にアーティファクトが発生する原因と考えられる．対して，Transformer

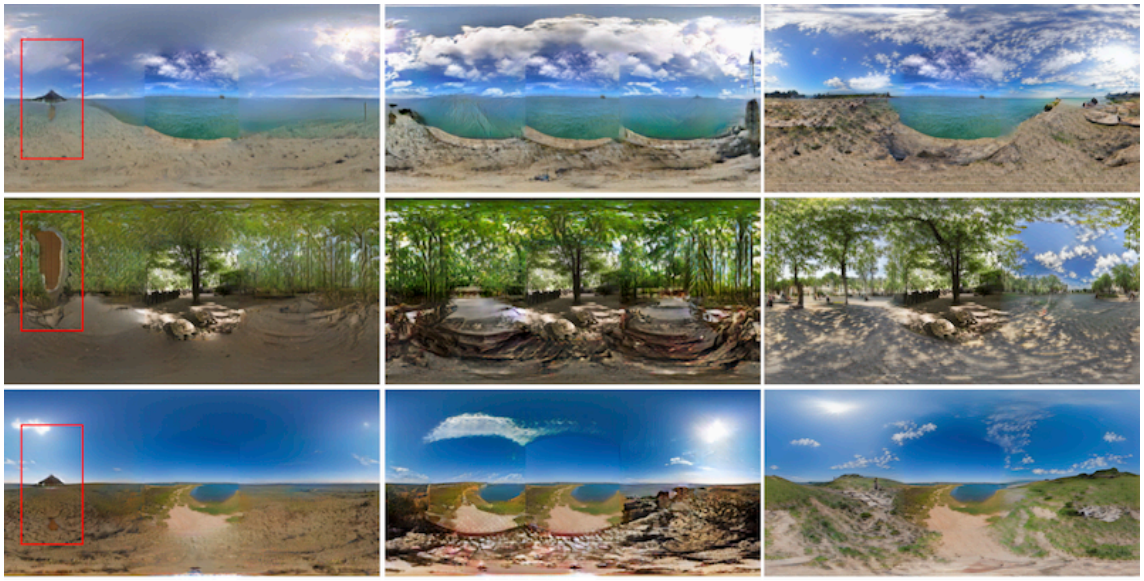


図 4.8 SIG-SS との定性比較．入力領域は，図4.6の 2 行目と同じである．

は中央領域においても詳細なテクスチャを生成することができ，さらに画像上下部分の ERP 画像特有の歪みが大きく発生する領域でもその歪みを表現することができている．したがって，本手法が球状のジオメトリの考慮をモデルに直接含めていないことを考慮すると，これらの歪みは学習データセットからその傾向が学習され，出力結果において歪みが表現されていると言える．正距円筒図法からパースペクティブ画像へ再投影した際に，例えば地面のタイルの歪みを見ると，不自然なタイル形状となっているため，正距円筒図法として正確な歪みではないことが見て取れ，改善の余地がある．

図4.8は Ours と SIG-SS の比較である．どちらも入力領域は画角 90 度に相当する領域である．SIG-SS は再構成 (rec) とサンプリング (gen) の結果がある．解像度は 512×256 である．再構成結果は同じような物体（右端の電柱のようなものや左側の茶色の山のようなもの）が出現する過学習が起きている．また，サンプリング結果は町並みの入力領域に対して木が生成されるなど生成されるコンテキストが一部に偏っている．一方で提案手法は入力領域のコンテキストに合わせた結果を得ることができている．

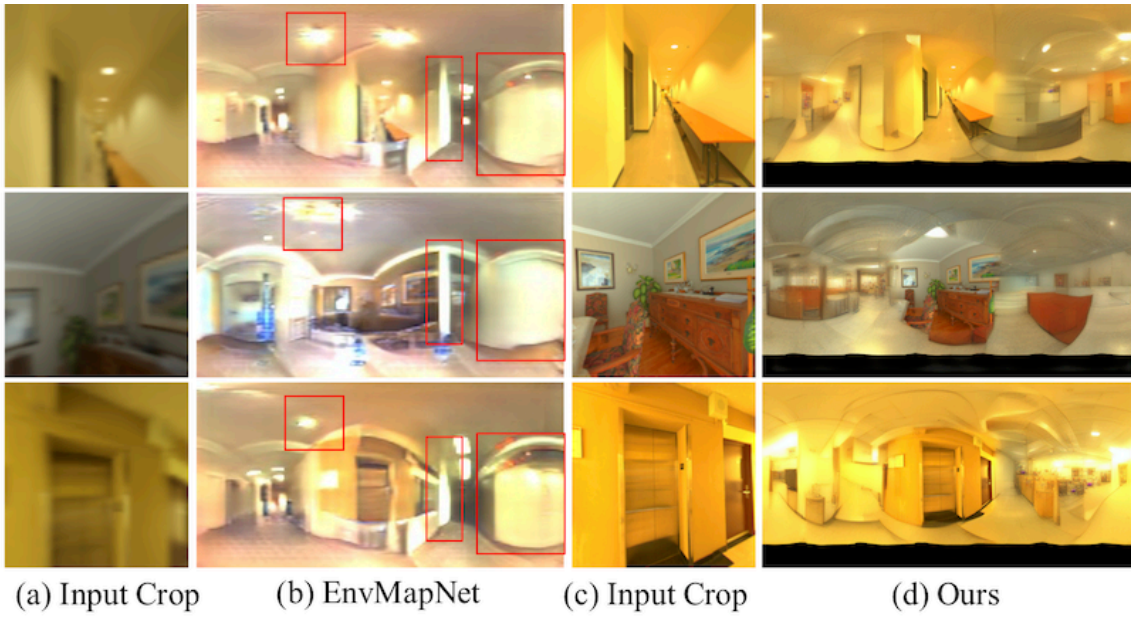


図 4.9 EnvMapNet との定性比較．入力形状は，図4.6の 3 行目と同じである．EnvMapNet(a) と Ours(c) の Input Crop は正確には同じではないことに注意されたい．

図4.9は Ours と EnvMapNet の比較である．EnvMapNet は 256×128 の補完画像で，Ours は 1024×512 である．提案手法は EnvMapNet に比べて 16 倍のピクセル数の解像度であるにも関わらず補完が行えている．EnvMapNet は敵対的学習を安定化させるために，事前処理としてデータセットをクラスタリングし，補助情報として入力するが，それらを必要とせずに提案手法は学習できる．一方で，SUN360 で学習した場合に比べて，詳細なテクスチャは生成することができていないように見える．これの原因は少ないデータセットで学習していることであると考えており，多くのデータを必要とすることが提案手法の限界の 1 つである．

以上の SoTA 手法との比較は，提案手法の補完結果が，定性的に，360 度画像として大きな差をつけてより尤もらしい画像を補完することができることを示している．

4.3.4 定量評価

提案手法の生成の質と多様性を評価するために，Fréchet Inception Distance (FID) スコアを使用する．この評価に使用した結果画像は4.3.2節と同じ入出力で得た画像

表 4.1 FID スコア．SUN360 で，180° × 90° の入力を用いる．

	360IC[3]	CompletionNets Only	Ours
FID↓	16.44	14.96	**9.52**

表 4.2 FID スコア．SUN360 で，90° の入力を用いる．

	SIG-SS(rec)[5]	SIG-SS(gen)[5]	Ours
FID↓	31.91	26.81	**23.13**

である．表4.1は，データセットとして SUN360 Outdoor を使い，FID の算出スクリプトとして clean_FID[72] を用いる．提案手法は 1024×512 で生成する．結果は，提案手法が先行手法である 360IC を上回る．さらに，この比較結果は Dilated convolution を何層も重ねた CNN でシーンをモデリングするよりも，Transformer でモデリングする方がより学習データの分布に似た分布を獲得できることを示している．同様の評価方法で，表4.2では SIG-SS との比較結果を示している．CVAE を用いた彼らの手法よりも，Transformer を用いた提案手法の方がより，本来のデータセットに近い結果を得ることができる．また，表4.3は提案手法と EnvMapNet と Gardner ら [54] との比較を，Laval Indoor Dataset で行なった際の結果である．入力画像は 90 度の画角に相当する領域である．FID を算出するために，彼らの評価プロトコルに従い，Cubemap に変換し，情報がほとんどない `Top` と `Bottom` の面は評価に含めない．この FID の算出には Tensorflow を用いた．この FID の比較結果は，提案手法が生成結果の質や生成結果の多様性という点で優位であることを示している．

4.3.5 分析

Circular inference の効果の検証．図4.10と表4.4は，提案する Circular inference が，ERP 画像として整合性を持つ推定が行えることを示す．図4.10は，CompletionNet の出力 (512×256) のうち，補完領域が画像の中央に来るように画像両端を合わせた後，一部を切り出した画像 (256×128) である．それぞれは同じ学習済みモデルであり，推

表 4.3 Laval Indoor dataset での FID スコア.

	Gardner *et al.* [54]	EnvMapNet[4]	Ours
FID↓	197.4	52.7	**46.15**

表 4.4 180° × 90° の入力の SUN360 での，Circular inference の効果の評価.

	Raster order	Circular padding	Circular inference
FID↓	30.03	**26.33**	26.96

論時の一部のみが異なる．図4.10(a) が示すように，TT で用いられる Transformer の自己回帰推定の順序であるラスタオーダーは，左右の端での繋がりが得られない．EPR 画像の特徴であるこの両端のつながりを持たせる 1 つの方法が，図4.10(b) の Circular padding[5] であり，Convolution を行う際に ERP 画像の反対側の端のピクセルをパディングするテクニックである．図4.10(b) では，Transformer の推定結果によって得られた特徴量を画像へデコードする際にこれを利用する．この方法はピクセルレベルでの連続性を向上するのに役立つのに対して，Circular inference は，Transformer による推定時にセマンティックレベルの連続性を向上するのに役立つことが期待される．表4.4が示すように Circular padding と Circular inference の方の FID スコアは同程度でありつつも，定性的には Circular inference は，より ERP 画像の特徴を持った画像の生成に寄与することが分かる．

(a) Raster order (b) Circular padding (c) Circular inference

図 4.10 Circular inference の効果．Circular inference は，360 度画像の両端をピクセルレベルとセマンティックレベルでつなげる．

表 4.5 提案ネットワークでの WS-perceptual loss の効果．180° × 90° の入力の SUN360 を用いる．

	Perceptual loss	WS-L1 loss	WS-perceptual loss
FID↓	29.00	35.00	**26.96**

表 4.6 360IC のネットワークでの WS-perceptual loss の効果．180° × 90° の入力の SUN360 を用いる．

	Perceptual loss	WS-perceptual loss
FID↓	67.47	**50.87**

WS-perceptual loss の効果の検証．この損失関数は，low-level の差分だけでなく，high-level の差分でも ERP 画像の緯度方向の情報量の差を考慮する．表4.5と表4.6では，より直接効果を検証するために，CompletionNet の出力を使って評価する．出力画像 (512×256) のうち，入力画像の領域を含まない，緯度 90 度，経度 180 度に相当する画像中央の生成領域 (256×128) を切り出し，FID を算出する．表4.5は提案手法ネットワークを用いて，WS-perceptual loss を使わない場合 (Perceptual loss) と low-level のみを考慮する WS-L1 loss で学習した場合と比較する．表4.6は 360IC のネットワークの学習においても，WS-perceptual loss の使用が FID スコアの向上に寄与することを示す．この WS-perceptual loss は，high-level な特徴量でも球を考慮した重みつけをすることで，より ERP 画像の特徴を持った画像の生成に寄与することが分かる．

4.4 応用

4.4.1 背景作成と照明のデモ

補完された 360 度画像を背景として利用しつつ，ライティングにも使用する応用例を実装する．図4.11が示すように，3DCG ソフトウェアである Unreal Engine 4 (UE4) 上で `Image Based Lighting` を行うためのパイプラインを提案する．UE4 のプラグ

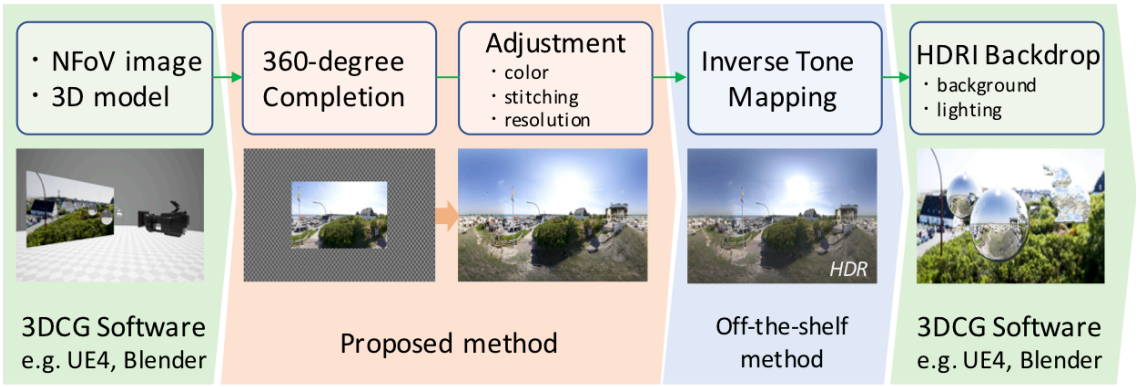


図 4.11 3DCG シーンの照明と背景のために補完結果を用いる際のパイプライン.

インである HDRI backdrop を使用し，360 度画像を背景かつ照明として使用するために，提案手法の補完結果をあらかじめ HDR 画像に変換する必要がある．ここでは，既存の手法による Inverse tone mapping [73] によって，Low Dynamic Range (LDR) から High Dynamic Range (HDR) へと変換する．このプロセスの導入により，提案手法は LDR の NFoV を補完して，HDR environment map を生成することができる．

このパイプラインを使って，3DCG シーン内の鏡面物体に対して，補完された 360 度画像で背景作成と照明を行うアプリケーションを示す[1] [2]．図4.12がこのアプリケーションのデモ動画のスクリーンショットである．通常 3DCG モデルの後方に NFoV 画像をコンポジットして，カメラに投影する場合，金属など反射する物体がある際にその物体の反射面が表現できない．先行手法 [4, 53] では，室内に限定する手法が多く，鏡面反射を表現できるように Environment map を生成するが，提案手法に比べて低解像度である．それゆえ，提案手法の結果のように，鏡面物体をカメラの近くに挿入できるのは他にはない．さらに，デモ動画では，このアプリケーションを使って，先行研究の結果画像との比較を行っている．提案手法が解像度高くかつ尤もらしい 360 度画像補完ができることを利用して，鏡面物体の挿入だけでなく，カメラを動かし周りを見渡すデモも行う．知る限りこのようなデモを行った研究は他にはない．他の手法と異なり，このように高解像でバリエーション豊かな生成がで

[1] https://www.youtube.com/watch?v=FxfudEt_Fds
[2] https://akmtn.github.io/omni-dreamer/

きる提案手法は，全周囲の背景を効率的に作成することで，デザイナーに対して新たな 3DCG 制作のワークフローを提供する可能性がある．

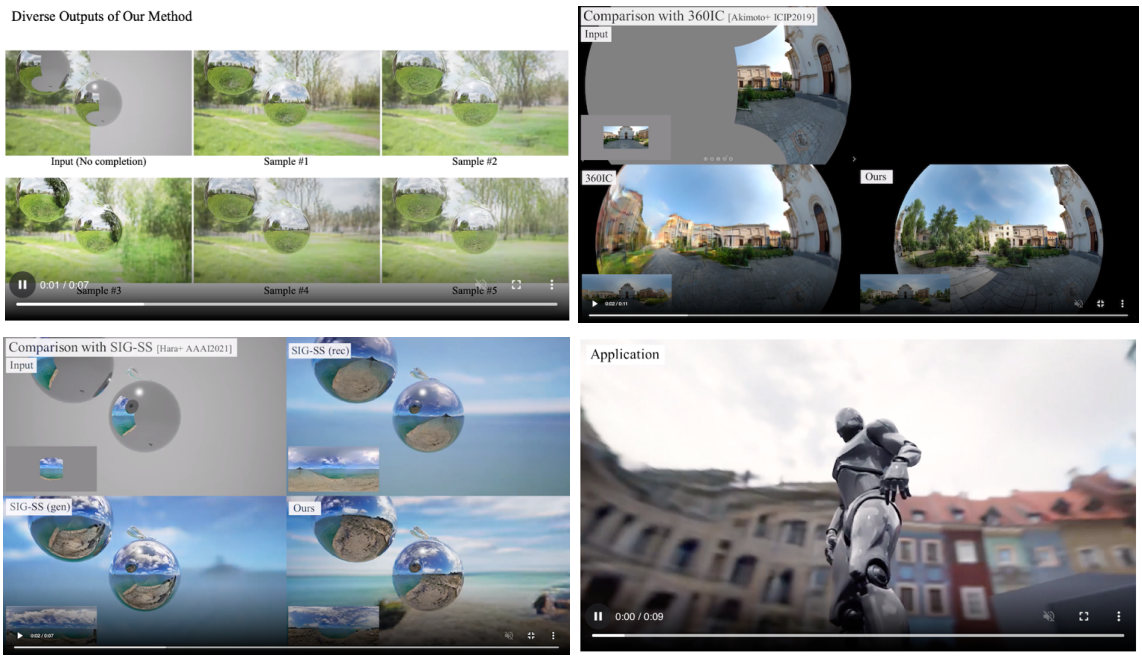


図 4.12 デモ動画のスクリーンショット．

4.4.2 スマートフォン写真に対する推論

実際にカメラで撮影した写真に対して，提案手法を適用し 360 度画像を作成する例を示す．これまでの実験では 360 度画像から切り出したものを入力画像として扱っていたが，ここでは通常画角のカメラ (スマートフォン) で撮影した写真を扱う．

本研究の目的が，背景画像を作成し効率的な 3DCG 制作の支援であることを考慮すると，元画像をストックフォトサイトのような Web 上から得て利用する場合が想定される．そのような画像は 360 度画像に補完されることを想定されず作成されたものである．一方で，あらかじめ 360 度画像に補完されることを想定し，複数枚の画像もしくは動画を撮影し，カメラ内部パラメータや外部パラメータを利用し，360 度画像にマッピングする応用も存在する [4]．本研究はそれとは異なり，360 度画像に補完されることを想定していない画像を対象とする．したがって，元画像が 360

度画像のどこを占めるかを特定できないまま補完する必要がある．本実験では，元画像を 3 つのバリエーションで 360 度画像内に配置して，その結果を確認する．

より詳細な実験設定を述べる．実験には iPhone XR によって撮影された 2 つの写真を利用する．撮影後に画像をクロップしているため，iPhone のカメラ画角と画像が写す画角は異なる．また，360 度のどこを写し出しているか不明な状態での補完を想定するため，iPhone のカメラ内部パラメータや，外部パラメータは利用しない．元画像を横方向に 180 度分，120 度分，90 度分あると仮定した 3 つのパターンで 1024×512 の入力画像を作成する．

結果を図4.13に示す．元画像が 360 度画像中で適切と思われる領域をカバーする場合には，その生成結果の見た目の質が高まることが分かる．明らかに元画像の占める範囲が適切でない場合には，結果も不自然となることが分かる．例えば，あまりに入力領域が小さい場合には，生成領域には入力を無視したようなテクスチャが得られている．学習時には入力領域は横方向に 180 度分で学習するが，90 度分の入力でも生成できているため横方向へロバストであると言える．しかしながら，元画像が縦方向に大きすぎると自然なつながりを得ることができていない．これは，画像の上下の領域は 360 度画像固有の歪みを持っているべきであるのに対して，元画像がそのような歪みを持たないため，モデルが適切に対処できないためと考えられる．提案手法のモデルは，360 度画像から切り出した画像によって学習しているため，360 度画像として正しいデータしか学習中に扱っておらず，推論時にそのようなデータから大きく逸脱するものに対しては高い品質で補完することが難しいと考えられる．

4.5 議論

近年，コンピュータビジョンや機械学習のコミュニティでは，モデルやデータセットに関する透明性や倫理を考慮しながら研究発表をする意識が高まっており，それに習い，この研究が社会に及ぼしうる影響について議論する．

図 4.13 スマートフォン写真に対する推論.

4.5.1 ネガティブな影響

提案手法による出力結果は未だ，カメラによって撮影される実際の 360 度画像に比べると改善の余地があるが，仮に生成結果がより良くなれば Deepfake の一種となりうる．存在しないシーンの生成や，オブジェクトの挿入合成が自然な見た目で行うことが可能となれば，人々に対して生成画像に写ったシーンが現実のシーンであると勘違いを引き起こす可能性がある．

4.5.2 潜在的な応用

- **バーチャルプロダクション**．この手法は，バーチャルプロダクションにおいてフォトベースの背景表現で役に立つ．例えば，平面の LED wall に投影するために，デザイナーは，しばしば，3D モデルで作られた 3D 環境を使う代わりに，ストックフォトから得た画像をコンポジットすることによって背景を製作することがある．その際であっても写っていない領域の状況を推定することができ，挿入される物体に対して反射を表現することもできる．

- **HDRI ストックフォト**．HDRI ストックフォトサイトには限られた枚数しかないため，クリエーター同士で同じ画像を使い，オリジナリティが損なわれるという課題がある．例えば，よく知られたサイト [74] でさえ約 490 枚の画像しか提供されておらず，テスト画像 5000 枚と比べて非常に少ない．提案手法は，新たな 360 度の HDRI を多様に生成することでこの問題を解決する可能性を持つ．

- **メタバース**．メタバースのような VR 空間の映像配信などでユーザーの視線方向にある領域の映像のみを配信することによって，通信容量を削減することができると考えられる．しかし，例えばエンドユーザーのアバターが手に鏡を持つ場合には何も反射を写すことはできない．この場合に，提案手法は推定されたシーンを表示することで反射の問題を解決できる．

4.5.3 本手法の限界

推論スピードと計算メモリの限界．EnvMapNet は，モバイルデバイス上で動くアプリケーションを想定した手法になっている．それに比べると提案手法は，推論に画像 1 枚あたり，NVIDIA 2080Ti で約 30 秒かかってしまうため，モバイルデバイス上で使用するためには速度面で改善が必要である．これらの原因は主に Transformer を使用しているためであり，時間と計算メモリを消費する．一方で，本研究の対象ユーザーである 3DCG のデザイナーは多くの場合でハイエンドの GPU を搭載したコンピューターを使っており，提案手法の性能は実用としては十分な可能性もある．また，一般的な Transformer の改善はソフトウェア・ハードウェアの両面から現在活発に取り組まれていることであり，提案手法はその恩恵を直接受けることができる可能性がある．

コントローラビリティ．提案手法は補完領域に何が生成されるかをコントロールしない．1 つの解法として，補完領域に出現する物体を直接貼り付け，それと滑らかにつながるような補完を行うようにするということが考えられる．

新たな深層生成モデルとの比較．本手法では，画像のハイレベルな特徴量を Transformer で補完することにより，CNN で補完する場合よりも良い結果になることを示している．つまり，360 度画像のシーンのモデリングをどの深層生成モデルで行うべきかということを検証している．現在新たな深層生成モデルとして拡散モデルが注目されており，このモデルでの検証も今後行う必要があると考える．

データセットの規模．通常画角の風景画像データセットに比べると，360 度画像の公開データセットの規模は小さい．通常画角の画像は数億枚規模のデータセットも一般に公開されている．一方で，SUN360 のような 360 度画像のデータセットの規模は，数万枚の規模である．近年の深層生成モデルの傾向では，データが多いほど生成結果の品質が上がると言えるため，360 度画像の補完においても，データセットの規模の拡大が品質の向上に貢献すると考えられる．

4.6 まとめ

この論文では，まず先行手法には学習時の解像度に過適合することと出力が決定的であるという課題があることを明らかにし，次にその課題を解決するためのフレームワークを提案した (目標 1 と目標 2)．さらに 360 度画像の性質が向上した補完結果を得るための 2 つの新規テクニックを提案した．比較実験で，提案手法は他の手法より尤もらしい見た目の出力結果が得られることを示した (目標 3)．最後に，デモでは，提案手法が全周囲の背景を効率的に提供することで，デザイナーに 3DCG 制作の新しいワークフローを提供する可能性があることを示した．

第5章　結論

本論文では，まず，入力された画像の端に関連する尤もらしいピクセルを生成することによって画像を拡張するという Outpainting による画像拡張に取り組んだ．次に，通常画角相当の画像の周辺ピクセルを生成することによって 360 度画像を生成するという 360 度画像の Outpainting に取り組んだ．画像両端の拡張を繰り返すと最終的には 360 度の景観を映すことになるという意味で，Outpainting による画像拡張の延長上に 360 度画像の Outpainting の問題が存在する．

Outpainting による画像拡張の研究で得られた知見の 1 つに，Mirrored input を用いて遠くのピクセルを近くに配置することにより補完結果の視覚的品質を向上することが可能であること，つまり，CNN による既存の Outpainting は未だ遠くのピクセルから十分に情報を取得していなかったことが挙げられる．この知見を生かして，360 度画像の Outpainting では，Transformer を導入し，補完領域と入力領域に距離があっても Attention 機構によって十分に考慮するアプローチを採用した．結果として先行手法を大きく上回る自然で尤もらしい見た目の補完画像を得ることができた．以下では，各問題での総括と今後の展望について述べる．

Outpainting による画像拡張．Outpainting による先行手法の生成ピクセルがターゲット画像から離れるほどブラーになるという限界や水平方向に同一のセマンティックを繰り返すという課題に対して，Mirrored input を提案した．これによって Outpainting の問題を Inpainting の問題に置き換えた．提案する Inpainting ネットワークと Mirrored input は，多様なシーンを含むデータセット上で，Outpainting の SoTA よりも見た目も FID スコアも両方超える画像拡張を実現した．さらに Mirrored input によって生成されるコンテンツをコントロールする方法を示した．

360 度画像の Outpainting．この研究では，まず先行手法には学習時の解像度に過適合することと出力が決定的であるという課題があることを明らかにし，次にその課題を解決するためのフレームワークを提案した．さらに 360 度画像の性質が向上した補完結果を得るための 2 つの新規テクニックを提案した．先行手法と比べて圧倒的に尤もらしい見た目を持つ 360 度画像の生成を達成した．デモは，提案手法は全周囲の背景を効率的に提供することで，デザイナーに 3DCG 制作の新しいワークフローを提供する可能性があることを示した．

今後の展望．本研究で取り扱った研究の最終目標は，その技術が実利用されることである．実利用されるとは，Inpainting が Adobe Photoshop に搭載され一般的に利用されているように，Outpainting が商用ソフトウェアに搭載される状態のことである．従来の Outpainting 手法から本研究までにおいても未だ，搭載に至るまでには課題が存在する．既存研究に比べて，提案手法での結果は定性的にも定量的にも良好であることが多かったが，現状では背景や遠景を作るという目的で補完画像を使用するという程度に留まる．その原因の一つは，扱えている画像解像度である．つまり，4K や 8K のコンテンツに対応するにはさらなる向上が不可欠である．また他の課題として，近景として利用されるには，3D としての整合性を考慮した生成も必要であろう．特定のシーンに特化したモデルにユースケースが存在する場合はあるものの，扱えるシーンのドメインを増やすことも重要である．コントロール性の向上も必要となる．例えばマルチモーダル化による方法が考えられ，テキスト入力によってどのような補完結果を出力したいかを指示することでコントロールするという方法があり得る．

おわりに．デジタル社会の現在は，テキストの共有から写真の共有，動画の共有へと進み，3D 情報の共有へと変遷つつあり，メタバースや VR というトピックがバズワードとして認知され，大衆化を迎えようとしている．一方で，3DCG 制作やコンテンツ制作には労力や時間が必要なものであり，大きな需要に応えてコンテンツを供給するには，制作の効率化を行うことが必須である．効率化の一つの方法が，

既存のコンテンツの活用である．本研究では，Outpainting による画像拡張も，360 度画像の Outpainting も，既存の画像コンテンツの使い道を拡張させる役割を担えることを示した．本研究のアイデアが，次世代デジタル社会を支える技術のひとつとなっていくことを願い，本論文を締めくくる．

謝辞

本研究は，慶應義塾大学大学院理工学研究科後期博士課程在学中に行われました．研究を遂行するなかで，多くの方にお世話になりました．この場を借りて感謝の意を述べさせていただきます．

まず，本論文の主査であり，指導教員である慶應義塾大学理工学部青木義満教授に深く感謝いたします．本研究の立ち上げから本論文の執筆に至るまで，多くの指導をしていただきました．加えて，研究者としての研究に対する向き合い方についても学ばせていただきました．研究室での生活を通して様々な体験を共にした青木研究室の皆様にも深く感謝申し上げます．

本論文の審査にあたり，副査を快くお引き受けいただいた慶應義塾大学理工学部池原雅章教授，斎藤英雄教授，杉本麻樹教授に深く感謝申し上げます．本論文をまとめるにあたり，多くのご助言をいただきました．厚く御礼申し上げます．

最後に，常に支えてくれた家族に深く感謝いたします．

参考文献

[1] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10521–10530, 2019.

[2] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10561–10570, 2019.

[3] Naofumi Akimoto, Seito Kasai, Masaki Hayashi, and Yoshimitsu Aoki. 360-degree image completion by two-stage conditional gans. In *IEEE International Conference on Image Processing*, pp. 4704–4708. IEEE, 2019.

[4] Gowri Somanath and Daniel Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11298–11306, 2021.

[5] Takayuki Hara, Yusuke Mukuta, and Tatsuya Harada. Spherical image generation from a single image by considering scene symmetry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[6] Naofumi Akimoto, Daiki Ito, and Yoshimitsu Aoki. Scenery image extension via inpainting with a mirrored input. *IEEE Access*, Vol. 9, pp. 59286–59300, 2021.

[7] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11441–11450, 2022.

[8] Naofumi Akimoto and Yoshimitsu Aoki. Image completion of 360-degree images by cgan with residual multi-scale dilated convolution. *IIEEJ Transactions on Image Electronics and Visual Computing*, Vol. 8, No. 1, pp. 35–43, 2020.

[9] 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*, pp. 2672–2680, 2014.

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[12] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2642–2651. JMLR. org, 2017.

[13] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[14] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2223–2232, 2017.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020.

[25] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, June 2021.

[26] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

[27] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Vol. 2, pp. 1033–1038. IEEE, 1999.

[28] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 341–346. ACM, 2001.

[29] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, Vol. 28, p. 24. ACM, 2009.

[30] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH 2007)*, Vol. 26, No. 3, 2007.

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[32] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, Vol. 36, No. 4, p. 107, 2017.

[33] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3911–3919, 2017.

[34] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1438–1447, 2019.

[35] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, Vol. 28, pp. 3483–3491, 2015.

[36] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020.

[37] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4692–4701, October 2021.

[38] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. *arXiv preprint arXiv:2104.07652*, 2021.

[39] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1399–1408, 2019.

[40] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7467–7477, 2020.

[41] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14458–14467, 2021.

[42] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4512–4521, 2019.

[43] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14144–14153, October 2021.

[44] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *ACM SIGGRAPH 2007 papers*, pp. 10–es. ACM New York, NY, USA, 2007.

[45] Josef Sivic, Biliana Kaneva, Antonio Torralba, Shai Avidan, and William T Freeman. Creating and exploring a large photorealistic virtual space. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8. IEEE, 2008.

[46] Johannes Kopf, Wolf Kienzle, Steven Drucker, and Sing Bing Kang. Quality prediction for image completion. *ACM Transactions on Graphics (TOG)*, Vol. 31, No. 6, pp. 1–8, 2012.

[47] Yinda Zhang, Jianxiong Xiao, James Hays, and Ping Tan. Framebreak: Dramatic image extrapolation by guided shift-maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1171–1178, 2013.

[48] Miao Wang, Yukun Lai, Yuan Liang, Ralph Robert Martin, and Shi-Min Hu. Biggerpicture: data-driven image extrapolation using graph matching. *ACM Transactions on Graphics*, Vol. 33, No. 6, 2014.

[49] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4492–4501, 2019.

[50] Jiayuan Mao, Xiuming Zhang, Yikai Li, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Program-guided image manipulators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4030–4039, 2019.

[51] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4570–4580, 2019.

[52] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[53] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[54] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2017.

[55] Marc-Andre Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagne, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[56] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[57] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5918–5928, 2019.

[58] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, Vol. 30, , 2017.

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[60] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[61] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2695–2702, 2012.

[62] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pp. 7354–7363, 2019.

[63] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pp. 658–666, 2016.

[64] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

[65] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4170–4179, 2019.

[66] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4471–4480, 2019.

[67] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7508–7517, 2020.

[68] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[69] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[70] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, Vol. 24, No. 9, pp. 1408–1412, 2017.

[71] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[72] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021.

[73] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1651–1660, 2020.

[74] Greg Zaal. *Poly Haven*. `https://polyhaven.com/` (Accessed: 11/16/2021).