# 修 士 学 位 論 文

# 審美性を考慮した
# 画像の凸四角形領域切り出し

# Aesthetic-aware image cropping
# with convex quadrilaterals

令和 元年度　　1月30日提出


情報理工学系研究科

電子情報学専攻

48-186431　　西保 匠

指導教員　　佐藤 洋一 教授

# 論文要旨

写真としての美しさを考慮して画像のトリミングを行うタスクである，審美性を考慮した画像切り出しはコンピュータビジョン分野における重要な一分野として従来から取り組まれてきた．審美性を考慮した画像切り出し技術は，写真撮影時の自動フレーミングや別媒体へ画像を表示する際の自動のトリミングなど，主にアマチュア写真家の写真撮影や画像編集の支援への応用が期待されている．

従来までの審美性を考慮した画像切り出しは，矩形領域を切り出す手法のみが取り扱われてきた．取り扱う画像は，矩形領域を切り出すことで審美性が向上する画像であった．しかし多くの場合，撮影した画像には予期せぬ回転や視点のずれにより生じる，形状の不自然さが含まれる．そのような画像から，コンピュータが審美性を考慮して画像領域を切り出す場合，矩形領域を切り出す手法では対応できない場面が多い．そこで本研究では，審美性を考慮した凸四辺形領域を切り出す手法に取り組む．

従来の審美性を考慮した矩形切り出しでは，審美性評価器を用いた切り出し領域の選択手法や物体認識ベースの矩形切り出し手法が提案されてきた．しかしながら，従来の審美性評価器では，被写体とカメラが正対していない状態で撮影された画像に対する審美性評価ができるのかが未知である．また，従来の物体認識ベースの手法の拡張では，任意の凸四辺形を表現することが難しい．

そこで，本論文では，回帰による凸四辺形領域切り出し手法と形状の不自然さを考慮に入れた審美性評価器を提案した．そして，ランダムに変形させた評価用データセットを作成し，切り出しの精度評価の比較を行った結果，既存手法による切り出し精度を上回ったが，凸四辺形領域切り出し手法を実現するには至らなかった．

# 謝辞

本研究を行うにあたり多くの方々にお世話になりました．指導教員である佐藤洋一先生には，素晴らしい研究環境を用意していただきました．また，研究の方針や物事の考え方など様々な指導をいただきました．准教授の菅野先生には，研究テーマ設定や提案手法の議論でお世話になり，そのたびに的確なアドバイスをいただきました．元特任講師の樋口啓太さん（現株式会社 Preferred Networks）には，研究の話題を含め様々な相談に乗っていただきました．行き詰まっているときにも暖かく励ましていただきました．助教の松井勇佑さんには，研究や論文執筆の進め方を指導をしていただきました．特にエンジニアリングの基本的なことである，最もシンプルで確実に動くものから作って積み上げることの大切さを教わりました．また，技術的な面でのアドバイスを何度もいただきました．佐藤研究室秘書の鈴木咲恵さんと今川洋子さんには，研究生活における様々な支援をしていただきました．元特任研究員の神窪利絵さん（現 Carnegie Mellon University）と博士課程の八木拓真さんには，何度も研究テーマの相談に乗っていただきました．また，研究室の黄逸飛さん，李振強さん，Kaipeng ZHANG さん，Onur GULER くん，Donghao WU くん，Chuyi WANG さん，Lijin YANG さん，實平暁海くん，王誉錫くん，王純一さん，Zhehao ZHU くん，佐藤禎哉くん，２年間ありがとうございました．最後に，大学院生活を見守り，支援してくれた家族に感謝します．

西保　匠

2020.1.30

# 目次

**第 1 章　序論** **1**

1.1　背景 . . . . . 1

1.2　本研究における課題とアプローチ . . . . . 3

1.3　本論文の構成 . . . . . 4

**第 2 章　関連研究** **5**

2.1　正解領域が明示的に与えられていないデータで学習する手法 . . . . . 7

2.2　正解領域情報が明示的に与えられているデータで学習する手法 . . . . 12

2.3　関連研究のまとめと本研究の立ち位置 . . . . . 14

**第 3 章　凸四辺形領域切り出し手法** **16**

3.1　定式化 . . . . . 16

3.2　凸四辺形領域の頂点座標を出力する手法 (DRN) . . . . . 17

3.3　射影変換による矩形への変換 . . . . . 20

3.4　ホモグラフィー行列を出力する手法 (HRN) . . . . . 21

**第 4 章　データセット** **23**

4.1　概要 . . . . . 23

4.2　データセットの作成方法 (Random FCDB) . . . . . 23

**第 5 章　実験** **26**

5.1　実験 1 ：提案手法の切り出し精度評価 . . . . . 26

5.2　実装 . . . . . 28

5.3　結果と考察 . . . . . 29

5.4　実験 2：回転や変形を考慮した審美性評価器の評価 . . . . . 35

5.5　実装 . . . . . 38

5.6　結果と考察 . . . . . 39

**第 6 章　結論** **44**

6.1　結論 . . . . . 44

6.2　今後の方針 . . . . . 44

**参考文献** **46**

# 図一覧

1.1 切り出し手法の違いによる切り出し画像の比較 . . . . . 2

2.1 ABP と AA ネットワークの概観 . . . . . 6

2.2 ABP ネットワークの学習プロセス . . . . . 7

2.3 A2RL の概観 . . . . . 8

2.4 A2RL の挙動の例 . . . . . 9

2.5 E2E の概観 . . . . . 10

2.6 VPN と VEN の概観 . . . . . 11

2.7 Grid Anchor based Approach の概観 . . . . . 13

3.1 提案手法の概念図 . . . . . 17

3.2 凸四辺形領域の頂点座標を出力する手法 (DRN) . . . . . 18

3.3 2 step の学習による DRN の学習の流れ . . . . . 19

3.4 ホモグラフィー行列を出力する手法 (HRN) の概観 . . . . . 22

4.1 評価用データセットの作成方法 . . . . . 24

4.2 作成したデータセットの一例 . . . . . 25

5.1 Random FCDB における提案手法の結果 . . . . . 31
5.2 FCDB における HRN のバッチサイズごとの評価精度 . . . . . 34
5.3 RandomFCDB における HRN のバッチサイズごとの評価精度 . . . . 34
5.4 審美性評価器 (TVEN) の概観 . . . . . 35
5.5 審美性評価器による凸四辺形切り出し手法の概観 . . . . . 37
5.6 RandomFCDB における審美性評価器 (TVEN) の出力画像 . . . . . 41
5.7 RandomFCDB における審美性評価器 (TVEN) 出力画像を射影変換した画像 . . . . . 42

# 表一覧

2.1 関連するデータセットの一覧 . . . . . 15

2.2 関連手法の一覧 . . . . . 15

5.1 FCDB における各手法の精度比較 . . . . . 29

5.2 Random FCDB における各手法の精度比較 . . . . . 30

5.3 DRN の Ablation Study . . . . . 32

5.4 学習データ別の性能比較 . . . . . 32

5.5 FCDB における審美性評価器の性能比較 . . . . . 39

5.6 Random FCDB における審美性評価器の性能比較 . . . . . 39

5.7 Random FCDB における VFN の性能比較 . . . . . 43

# 第 1 章

# 序論

## 1.1 背景

ソーシャルメディアの普及とスマートフォンのカメラ機能の発展に伴い，自分で写真を撮影し，撮影した画像を公開する機会が増大した．しかし，多くの初心者のアマチュア写真家にとって，撮影時の適切なフレーミング技術や撮影した画像の加工の技術は，習得が難しく習得までに時間がかかるものである．そうした背景から，コンピュータビジョン分野において，自動で写真の品質を評価する手法 [26, 8] や画像中の幾何的な関係を保ったまま画像中の主役を移動させる手法 [9, 11] など画像の質を向上させるための様々な研究がなされてきた [12]．その中の一分野として，写真としての美しさを考慮して画像のトリミングを行う手法（審美性を考慮した画像切り出し）が研究されてきた．

審美性を考慮した画像切り出し技術は，写真撮影時の自動フレーミングや別媒体へ画像を表示する際の自動のトリミングなど，主にアマチュア写真家の写真撮影や画像編集の支援への応用が期待されている．近年では，使用者の撮影の仕方に合わせて切り出しの仕方を学習するアプリケーション [23] や集合写真撮影時の動画から最適な切り出しを出力するアプリケーション [29] など，さまざまな応用事例研究が登場している．

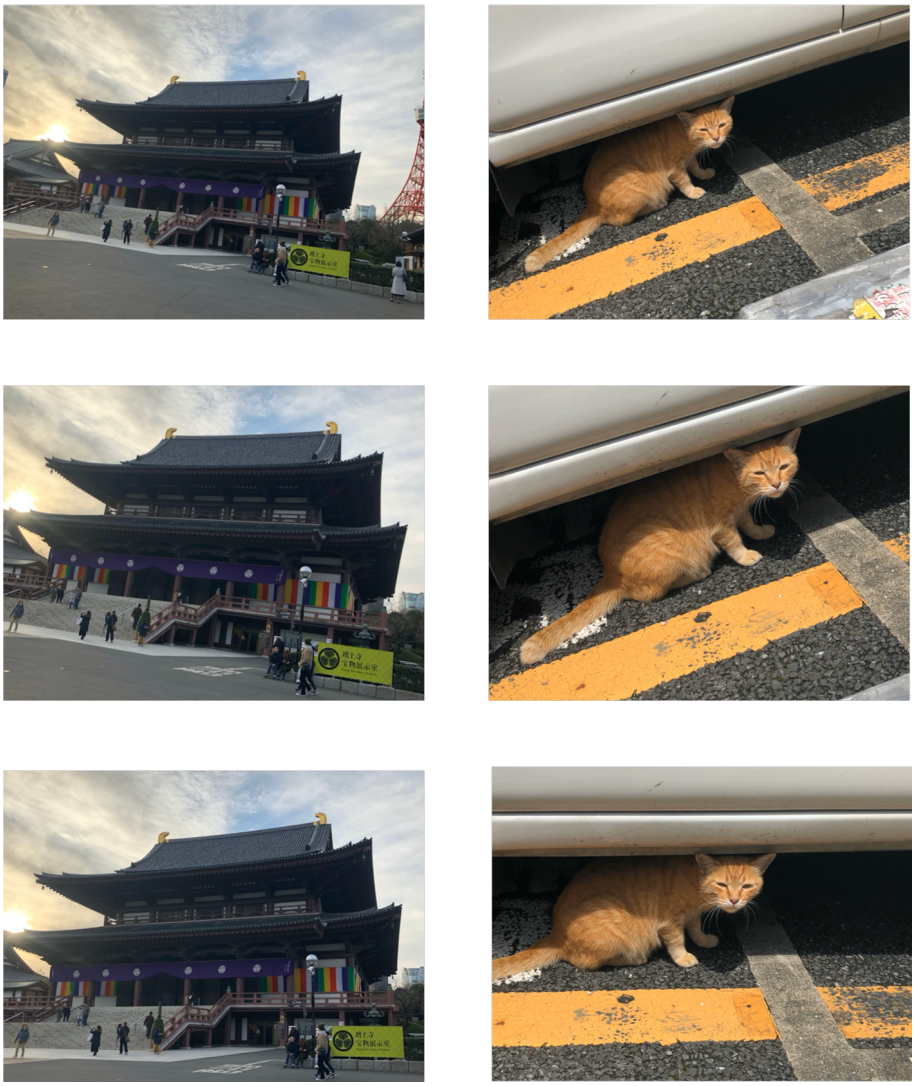

図 1.1: 切り出し手法の違いによる切り出し画像の比較．上段：手持ちカメラで撮影した画像．中段：従来手法によって切り出され得る画像の例．下段：本研究で目標とする切り出し画像の例．本研究では，入力画像から余分な領域を取り除き，変形を加えることで審美性の高い画像を切り出すことを目指す．

審美性を考慮した画像切り出しは，画像中のコンテンツを保持しながら画像から重要な領域を切り出す手法である．近年の審美性を考慮した画像切り出しは，データドリブンな手法が多く提案されている．その中の代表的な手法は，学習に使用す

るデータのタイプによって 2 つに分類できる．1 つ目は，正解領域情報がついていない画像データセットで顕著性領域検出器や審美性評価器を学習させ，活用するアプローチである．2 つ目は，正解領域情報が複数ついた画像データセットを用いた物体認識ベースのアプローチである．いずれの手法も，既存のベンチマークでは高い切り出し精度を記録しており，少ない計算資源で実行可能になりつつある．

従来の研究で取り扱っていた画像は，矩形領域を切り出すことで審美性が向上する画像であった．しかし多くの場合，撮影した画像には予期せぬ回転や視点のずれによって生じる，形状の不自然さが含まれる．例えば，図 1.1 の上段のような水平方向に回転してしまっている画像や視点のずれのある画像が撮影されたりする．これらの画像からは従来手法では，図 1.1 の中段のような切り出ししか得ることができない．

図 1.1 の上段のような画像から，コンピュータが審美性を考慮して画像領域を切り出す場合，矩形領域を切り出す手法では対応できない場面が多い．そこで本研究では，画像中から凸四辺形領域を切り出し，切り出した凸四辺形領域を透視投影変換によって矩形に変換した画像を審美的な画像にする手法に取り組む．つまり，審美的な画像の凸四辺形領域切り出しによって，図 1.1 の下段のような画像を取得することを目標とする．

## 1.2　本研究における課題とアプローチ

従来の審美性を考慮した矩形切り出しでは，審美性評価器を用いた切り出し領域の選択や物体認識ベースの矩形切り出し手法が提案されてきた．しかしながら，従来の審美性評価器は，形状の不自然さを考慮に入れずに設計されているため，形状が不自然な領域を切り出してしまう．また，従来の物体認識ベースの手法の拡張では，任意の凸四辺形を表現することが難しい．

そこで，本論文では，回帰による凸四辺形領域切り出し手法と形状の不自然さを考慮に入れた審美性評価器を提案する．回帰による凸四辺形領域切り出し手法では，凸四辺形の 4 点を直接出力する手法と，ホモグラフィー行列を出力し入力画像を変形する手法の 2 つを提案し，比較する．また，既存のデータセット中の画像に変形

を施した画像を用いて，形状の不自然さを考慮に入れた審美性評価器を作成し，その評価性能を調査する．

## 1.3　本論文の構成

本論文は全 6 章で構成され，審美性を考慮した画像の凸四辺形領域切り出し手法を提案し，その有用性を検証するものである．

1 章では，序論として本研究の背景とアプローチについて述べた．2 章で審美性を考慮した画像切り出しに関する関連研究について，研究の変遷をまとめ，代表的な手法について説明する．続く 3 章では，回帰による凸四辺形領域切り出し手法を提案する．4 章では，本研究では使用するデータセットの作成方法を記す．5 章では提案手法の性能評価実験と審美性評価器の性能評価実験を行い，それぞれについて議論する．最後に 6 章で，本稿の結論と今後の方針を記す．

# 第 2 章

# 関連研究

審美的画像切り出し手法 (Aesthetic-aware Image Cropping) の研究における最終的な目標は，画像中の不必要な領域を削除し，切り出す前の画像よりも審美性の高い画像を出力することである．個別の関連研究について言及する前に，以下では審美的画像切り出し手法の発展の大まかな流れを述べる．

審美的画像切り出しの手法の初期の研究では，画像の審美性の評価のために重要な要素だと考えれる特徴量をモデル化し，最も審美性スコアが高い切り出しを選択するものであった [27, 10]．特徴量には色やエッジ，コントラストなどの低レベルの特徴の他，顕著性マップ [31] や審美性マップ [16] などの高レベルの特徴が用いられることもある．しかし，切り出し候補の選択手法にスライディングウィンドウ方式が用いられていたために処理時間が大きいことが課題であった．

その後の研究では，処理時間の削減のため，候補領域を絞り込む手法について研究された．当時，画像に対する審美性を考慮した切り出しの正解データを集めるにはコストがかかるため，正解領域と全体画像のペアを用いて直接学習するための十分なデータが揃っていなかった．そのため，画像投稿サイトから取得できる審美性スコア付き画像大規模データセット [25] を用いて学習を行う手法を用いて計算時間の削減を提案を図っていた [6, 33, 34, 17, 18, 38]．

2018 年からの 2 年ほどの間に，一枚の画像に対して複数の切り出しの情報がついた大規模データセットが作成された．これにより，一枚の画像中の複数切り出し画像のスコアを学習することが可能になり，より速い処理時間で高い精度を出せるようになってきた [21, 35, 37]．

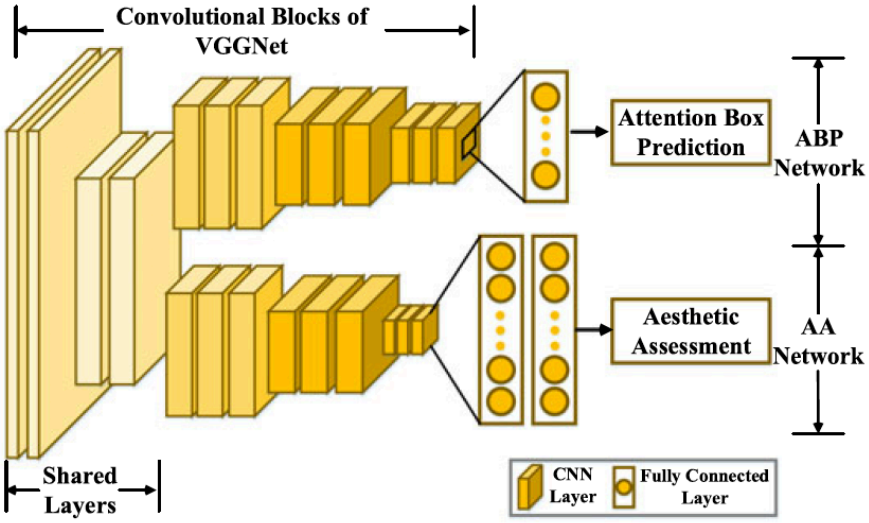


図 2.1: ABP と AA のネットワーク．下層のネットワークを共有している．上段の ABP ネットワークでは，顕著な領域を囲むバウンディングボックスを出力し，下段の AA ネットワークでは，候補画像の審美性を評価する．画像は [34] からの引用．

以上のように，審美的画像の切り出しの大まかな流れを整理し，関連研究を分類すると，精度のみのを考慮に入れた研究と切り出しの精度と計算の速度の両方を考慮した研究に分けることが出来る．本研究では，切り出しの精度と計算の速度の両方を考慮した研究について取り扱う．以降では，切り出しの精度と計算の速度の両方を考慮した研究における手法を，正解領域が明示的に与えられていないデータで学習する手法と正解領域情報が明示的に与えられているデータで学習する手法の 2 つに分類し，それぞれの代表的な研究を説明する．そして本研究との関連性を述べ，立ち位置を明確にする．

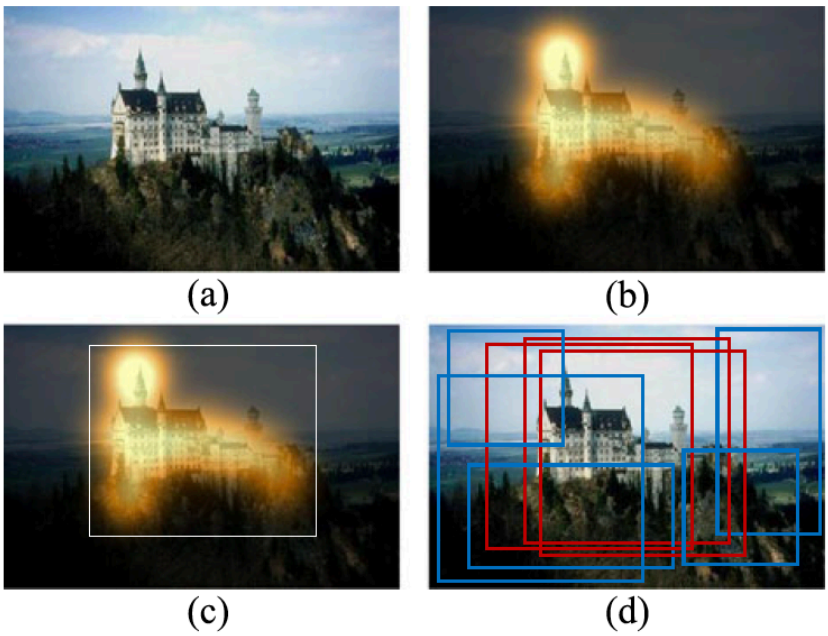


図 2.2: ABP の学習プロセス．(a) 入力画像．(b) [14] データセットに含まれる正解の attention map．(c) [4] によって生成された正解領域．(d) 赤色の矩形が正解と判定されるデフォルトボックスであり，青色の矩形が不正解と判定されるデフォルトボックス．画像は [34] からの引用．

## 2.1 正解領域が明示的に与えられていないデータで学習する手法

### ABP-AA

ABP-AA [34] は，画像中で顕著な領域を検出しその領域を囲む候補領域を複数作成し，それぞれの候補領域を審美的評価モジュールで選別することによって，計算時間の削減を試みた．

ネットワークの構造を図 2.1 に示す．ネットワークの下層は共通のレイヤーで構成されており，ネットワークの上層は上段の顕著性の高い領域を検出する Attention Box Prediction Network (ABP) と下段の画像の審美性を評価する Aesthetic Assessment Network (AA) で構成されている．

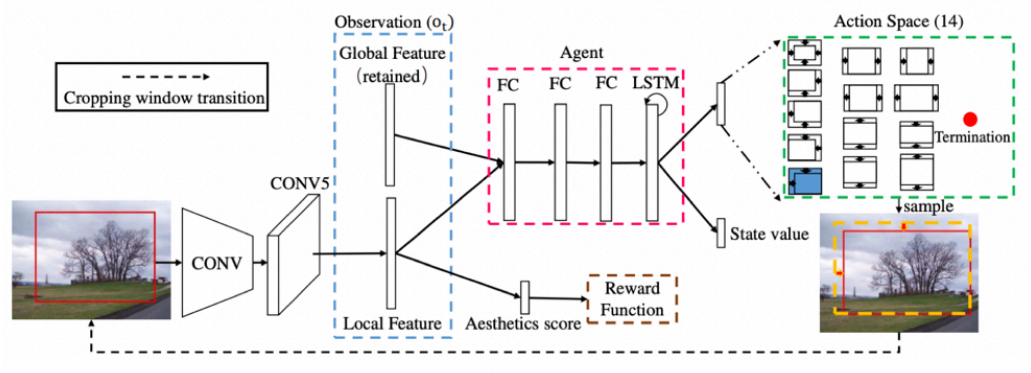


図 2.3: A2RL [17] のネットワークの概観．最初に入力画像を 5 層の畳み込み層を持つ特徴抽出器に入力し，Global Feature を得る．次に最初の切り出し画像を入力し同様に Local Feature を得る．その後は 2 つのブランチに分かれている．下段では，既存の審美性評価器による審美性評価を行う．上段の Agent では行動の選択と推定価値を出力する．上段の Agent への入力には先に取得した Global Feature と現在取得した Local Feature をあわせて渡す．Agent では方策関数に従い，14 種類の行動から一つを確率的に選択し，画像切り出す．また，現状態の推定価値を出力する．これを Agent が終了状態 (Terminal) を選択するまで行う．画像は [17] からの引用．

ネットワークの学習は上段の ABP と下段の AA を交互に異なるデータセットを用いて学習させている．上段の ABP の学習には，注視情報がついた画像データセットである SALICON [14] に対して， [4] の手法を用いて正解のバウンディングボックスを作成することで物体認識の手法 [28, 19] と同じ要領で学習させている（図 2.2 参照）．下段の AA の学習には，審美性評価のための大規模データセットである AVA [25] を用いている．これらの学習方法によって，正解領域のないデータセットを用いてネットワークを学習している．

### Reinforcement Learning Approach (A2RL)

A2RL [17] は審美的画像の切り出しに強化学習を取り入れた手法である．ABP+AA の手法では，はじめに ABP によって評価する候補画像領域を削減していたが，それでも $6^4 = 1296$ 通りの画像を評価するため，処理に時間がかかるという問題があっ

図 2.4: A2RL [17] の挙動の例．各ステップでサンプルされた行動に従って画像領域が切り出されている．赤い点は終了 (Terminal) を示しており，この時点での切り出し領域が出力される．画像は [17] からの引用．

た．A2RL は，強化学習アプローチで候補画像領域の絞り込み方を学習させることで，20 回程度の絞り込みで審美性の高い画像を切り出すことができる．この手法では，膨大な数の画像を比較することがないため，高速に審美性の高い切り出しを出力することができる．

A2RL では，強化学習の Actor-Critic 手法である A3C アルゴリズム [24] を採用している．以下図 2.3 を用いて学習のプロセスを説明する．図 2.3 に示すように，エージェント (Agent) は入力画像と切り出した領域から観測特徴を得る．その観測特徴とそれまでの経験に従い，行動空間から行動をサンプルする．エージェントはサンプルされた行動に従い画像領域を切り出す．それぞれの行動のあと，エージェントは切り出した画像に対する審美的スコアに従った報酬を得る．エージェントは累積報酬を最大化することによって最適な切り出しを得ることを目指してネットワークを学習する．

報酬を与えるモジュールには，既存の審美性評価モジュールである [6] の学習済みモデルを用いている．A2RL は特徴量抽出層を既存の審美性評価モジュールと共有しているため，行動によって得られた画像領域特徴を直接審美性評価に用いることができる．そのため切り出しで得られた報酬を即時に与えることができる．前状態の審美性スコアと現状態の審美性スコアの差を符号関数に与え，前状態より高い審美性スコアが得られれば $+1$ を，前状態より低い審美性スコアが得られれば $-1$ を報酬として出力する．

この手法のメリットは，1. 教師なしで学習させた点，2. 任意のアスペクト比の画

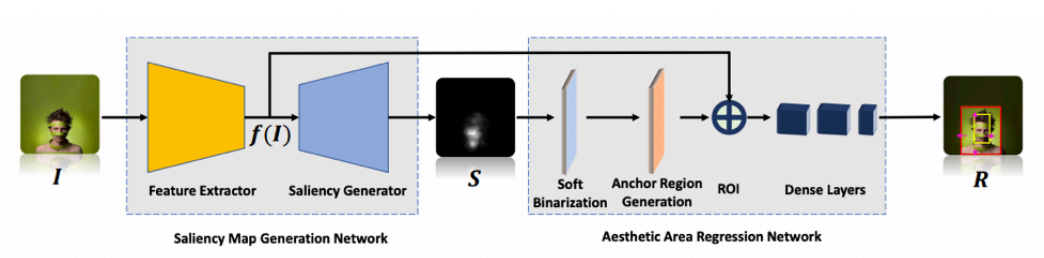


図 2.5: End-to-End Approach の概観．はじめに入力画像から Feature Extractor で特徴量を抽出する．その後顕著なオブジェクト領域を生成し，それをもとに選ばれた Anchor Region で，最初に抽出した特徴量から RoI pooling を行う．最後に Dense Layers で最終的な切り出しへのパラメータを補正して画像を出力する．画像は [20] からの引用．

像を切り出せる点，3. 高速に画像の切り出しを行える点である．この手法のデメリットは，学習時とテスト時の出力が安定しないことである．方策関数は確率的に次の行動を選択するため，同じ画像に対してクロップする場合でも，毎回異なる出力が現れる．そのため，再現性を確かめることが難しい．

### Saliency + Regression (E2E)

Lu らは正解領域の情報が存在しないデータを用いた End-to-End(E2E) なネットワークを提案した [20]．彼らは正解領域の情報が存在しない大規模なデータセットである AVA の画像を正解の切り出しとして設定した．この研究では，顕著な物体を囲むバウンディングボックスから正解の領域を囲むバウンディングボックスまでの差分をパラメータにし，適切な差分パラメータを回帰のアプローチによって学習させている．

図 2.5 に提案されたネットワークを示す．下層のネットワークは，Saliency Map Generation Module で特徴量を抽出し顕著性マップ（Saliency Map）を生成する．抽出された特徴量は，Aesthetic Area Regression Network に送られる．Saliency Map Generation Module は，生成された Saliency Map と SALICON [14] で作成された正解の Saliency Map との誤差を損失として学習する．次に，2 値化された Saliency

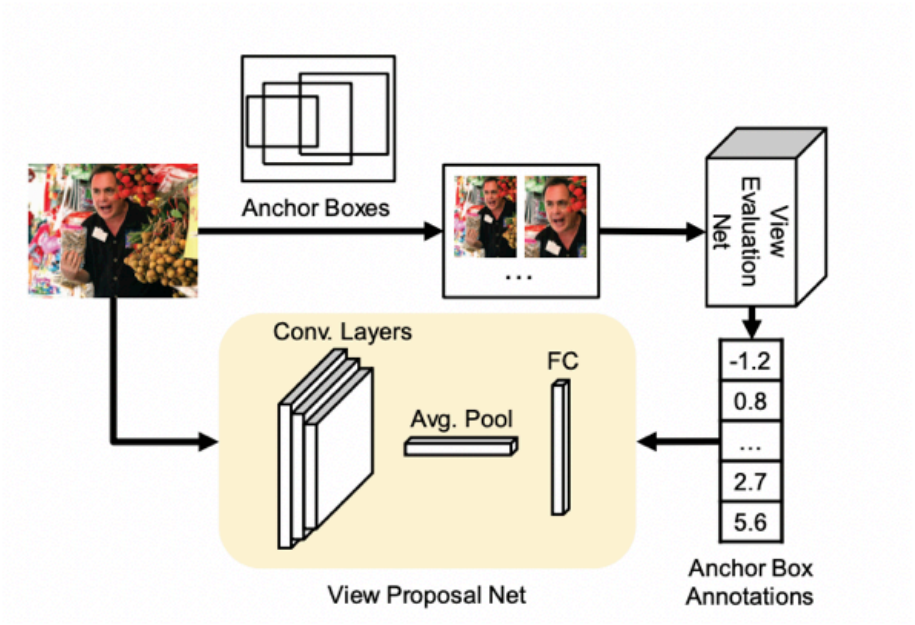


図 2.6: VPN と VEN の学習プロセスの概観．View Proposal Net(VPN) の学習の際には，Anchor Boxes によって切り出された画像のペアを学習済みの View Evaluation Net(VEN) でスコアを記録し Anchor Box Annotations を作成する．各画像の各 Anchor Boxe に対して VEN で正解スコアをつけておくことで，VPN の学習の際に，教師あり学習として物体認識アプローチと同じ要領で学習させることができる．画像は [35] からの引用．

Map の重心と分散を元に，Anchor Region が生成される．最後に，RoI Pooling したあと，Dense Layer で 2 点のパラメータを調節し出力する．デフォルトボックスを使うことなく，Saliency Map の重心と分散を用いて Anchor Region の生成しているため，ネットワークの下層まで損失を逆伝播できるようにし，End-To-End なネットワークを構築した．

## 2.2 正解領域情報が明示的に与えられているデータで学習する手法

### Good View Hunting (VPN,VEN)

Wei らは物体認識と同じ要領でネットワークを学習させるために，一枚の画像の複数の領域に対して正解スコアが付与された，大規模なデータセットを作成した [35]．作成したデータセットを使用し, View Evaluation Net(VEN) と View Proposal Net(VPN) の 2 つのネットワークを学習させた．以下この学習のプロセスを説明する．

まず VFN [6] と同様に，Siamese Architecture [7] で VEN を学習させた．Siamese Architecture は 2 つの重みパラメータが共通の VEN で構成され，それぞれからスコアを出力する．２つのスコアのうち，高い方が望ましいスコアと低い方が望ましいスコアの差が大きくなるように損失関数を与える．このようにしてネットワークを学習させる手法が Siamese Architecture である．

上で学習した VEN を使って，各 Anchor Box で切り出した領域画像に対しスコアを付けていき，VPN で学習する際の正解アノテーションを作成する．このようにして正解アノテーションを与えることで，VPN を SSD [19] のような物体認識アプローチで学習させることができる．2.1 節の ABP+AA も物体認識アプローチを取っているが，ABP+AA では初期切り出しの正解領域を IoU(Intersection over Union) を基準に与えているのに対し，この研究では各切り出しがどれほど良い切り出しであるかを指標にしている点で異なる．画像切り出しの正解の与え方としては IoU を基準にしたものよりも適していると考えられる．

公開されている他のデータセットを用いて VEN を学習させた結果，異なる画像どうしを比較して学習した画像評価モジュールよりも，同一の画像中から切り出した画像どうしを比較して学習した画像評価モジュールの方が，画像切り出しにおいては高い性能を出すことが確認された．また，同一の画像から取得する正解つき候補領域の数が多いほど，画像評価モジュールの性能が高くなることが報告されている．

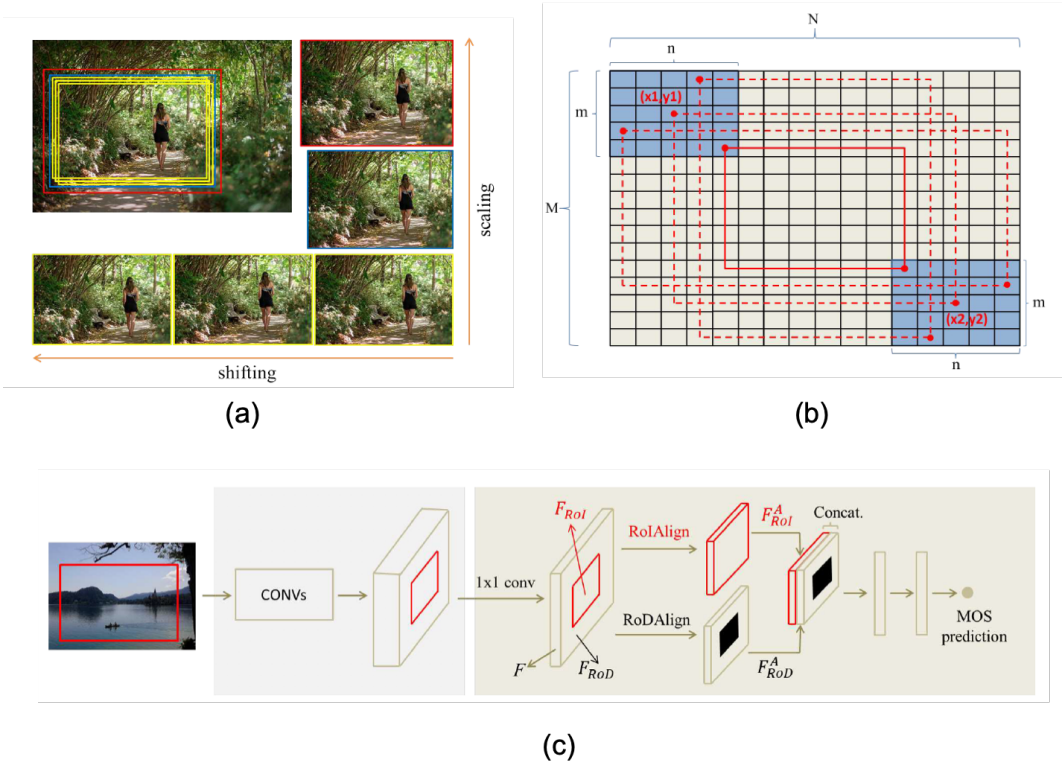


図 2.7: Grid Anchor based Approach の考察とネットワークの構造．(a) 切り出し位置の違いによる構図の違いを無視している．(b) 選択する領域の数をグリッドに区切ることで候補領域の数を制限している．(c) 各候補領域の相対順位を出力している．画像は [37] より引用．

### Grid Anchor based Approach

2.2 節の Good View Hunting で同一画像中の複数領域に対してスコアが付いた大規模データセットが作成され，ネットワークの学習に使用された．しかし，評価指標に対する見直しはなされていなかった．正解の領域が一つに限られないような画像の場合には，IoU ベースでの比較は適していないと Zeng らは主張している [37]．その主張のもと，新たな評価指標として，各候補領域のランキングの相関係数である $SRCC$ と順位の高い領域候補に含まれる正解の領域の数を表す $Acc_{K/M}$ を導入した．これらの指標を導入するためには，これらの指標を適用できるデータセットが必要になる．そこで彼らは，候補領域を限定したデータセットを作成した．

データセットの設計の際，どの程度まで候補領域を絞ることができるかを検討した．彼らは図 2.7(a) に示すように，領域の位置の小さな違いが構図の変化を生じさせないことに注目し，画像の各領域を図 2.7(b) のように小さなグリッドに分けることにした．これにより，一枚あたり約 90 個の領域に制限できるため，一人あたり 35000 枚をアノテーションすることでデータセットを構築することができた．

彼らの提案したネットワークは図 2.7(c) である．特徴は，Region of Interest(RoI) に加えて Region of Discard(RoD) を導入している点である．これは取り除いた領域情報が重要であるという考えから得たものである [36]．例えば，画像内の重要でない領域を取り除けば画像の審美性は向上するが，画像内の重要な領域を取り除いてしまうと切り出した画像の審美性は劇的に減少してしまう．物体認識の研究では，物体が存在している領域のみを考慮すればよいのに対して，画像切り出しの研究では，取り除かれた画像領域の情報も考慮に入れることが有効とされている．最後にこのネットワークは，画像の審美性のスコアである mean opinion score(MOS) の予測を出力する．

## 2.3 関連研究のまとめと本研究の立ち位置

これまでに提案された代表的なデータセットと手法をそれぞれ表 2.1 と表 2.2 にまとめた．

AVA データセットは多くの手法で学習に使用されている．これまで，正解領域のついたデータセットを学習に使った研究は，データセットの規模が小さかったため，あまりなされていなかった．近年作成された CPC や GAICD は，同一画像の複数領域に対して審美性スコアが付与されている．そのため，同一画像中の複数切り出し間の小さな違いも学習できるようになった．画像切り出しに用いられるデータセットの詳しい特徴は [2] にまとめられている．

これまで述べてきた先行研究は，与えられた画像の中から矩形領域を，どれだけ正確にどれだけ高速に切り出せるかを目的としていた．正確さにおいては，VEN のように，同一画像内の複数領域をペアにして審美性評価を学習させたものをあらゆる候補画像に対して評価するものが最も高いが，その分だけ計算コストがかかってしまう．また，検出速度に関しては，VPN や Grid Anchor Approach のような物体

認識手法由来の手法が最もよいが，学習に同一画像中の複数領域の評価スコアが必要なため，データセット作成の難しさの観点から，本研究では踏襲できないアプローチである．

本研究では，上述の知見を取り入れつつ，画像中から凸四辺形領域を切り出す手法について検討する．最初に，回帰ネットワークにより，直接凸四辺形領域を出力することを目指す．その後，回転や変形に反応する VEN のような審美性評価器を作成できるのかを調査し，ABP+AA や A2RL のような審美性評価器を用いた手法での応用可能性について考察する．

表 2.1: 関連するデータセットの一覧

| Dataset | Images | Views per Images | Evaluation | Ranking |
|---|---|---|---|---|
| AVA [25] | 250,000 | 0 | No | No |
| ICDB [36] | 950 | 3 | Experts | No |
| FLMS [10] | 500 | 10 | Experts | No |
| FCDB [5] | 1608 | 1 | AMT workers | No |
| CPC [35] | 10,797 | 24 | AMT workers | Yes |
| XPView [35] | 992 | 24 | Experts | Yes |
| GAICD [37] | 1236 | 90 | Experts | Yes |

表 2.2: 関連手法の一覧

| Method | IoU (FCDB) | Steps | Datasets for Learning Process |
|---|---|---|---|
| ABP-AA [34] | 0.650 | 1296 | AVA, SALICON |
| A2RL [17] | 0.663 | 20 | AVA |
| Saliency + Regression [20] | 0.673 | 1 | AVA |
| VPN [35] | 0.711 | 1 | CPC |
| VEN [35] | 0.735 | SW | CPC |
| Anchor Grid Approach [37] | 0.674 | 1 | GAICD |

# 第 3 章

# 凸四辺形領域切り出し手法

本研究では，審美性を考慮した画像の凸四辺形領域切り出しを，回帰問題に帰着させ取り組んだ．図 3.1 に提案手法の概観を示す．入力の画像が与えられたときに，Regression Network が凸四辺形の頂点座標を出力する．Regression Network からの出力を用いて，凸四辺形の内部領域を入力画像中から切り出す．切り出した領域を射影変換によって入力画像のアスペクト比の画像に変形する．

本研究では，凸四辺形領域の頂点座標を求める手法として 2 つの手法を提案する．1 つは，画像中から矩形領域の座標と変形のための調整パラメータを出力し，それらを足し合わせることで凸四辺形領域の頂点座標を表現する手法 (DRN) である．もう 1 つは，正解領域を矩形に射影変換する際のホモグラフィー行列を出力し，元画像の頂点座標を射影変換して凸四辺形領域の頂点座標を表現する手法 (HRN) である．以下では，2 つの提案手法における共通の定式化を行い，その後各手法について説明する．

## 3.1 定式化

$\{\boldsymbol{I}_i, \boldsymbol{r}_i\}_{i\in[1,n]}$ が与えられたとする．ここで，$\boldsymbol{I}_i$ は入力画像，$\boldsymbol{r}_i$ は入力画像に対する正解領域（凸四辺形）の 4 点であり，$\boldsymbol{r}_i = (x_1, y_2, x_2, y_2, x_3, y_3, x_4, y_4)^\mathsf{T}$ である．本研究における 2 つの手法での損失関数は以下で表される．

$$L = \frac{1}{n}\sum_{i=1}^{n} |\hat{\boldsymbol{r}}_i - \boldsymbol{r}_i|^2 \tag{3.1}$$

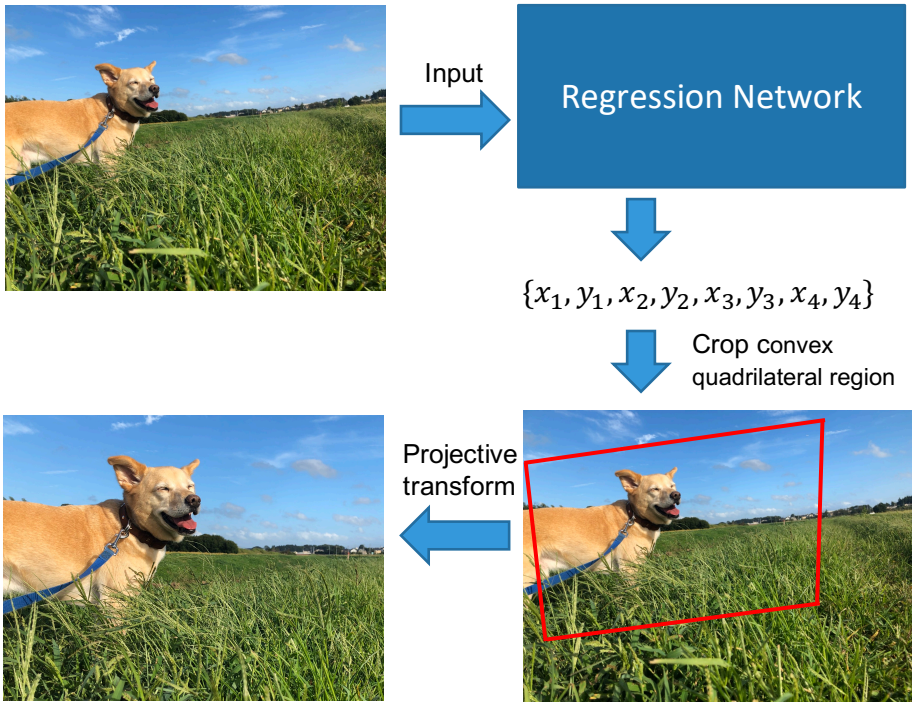


図 3.1: 提案手法の概念図．Regression Network は画像を受け取り，対応する頂点座標を出力する．その後，Regression Network から出力された頂点座標を用いて，入力画像中から凸四辺形領域を切り出す．最後に，切り出された凸四辺形領域を，射影変換によって矩形画像に変形する．

ここで，$\hat{\boldsymbol{r}}_i$ は入力画像 $\boldsymbol{I}_i$ をネットワークに入力した際の出力である．この損失関数を最小化することでネットワークを学習させる．

## 3.2 凸四辺形領域の頂点座標を出力する手法 (DRN)

本節では，入力画像が与えられたときに対応する頂点座標を出力する手法 (DRN) について説明する．本手法の概観を図 3.2 に示す．提案手法のネットワークは VGG16 をベースとしている．ネットワークは画像を入力として特徴量を抽出する．得られた特徴量を 2 つのブランチに送る．上段では矩形領域の座標表す $\hat{\boldsymbol{r}}_i^{\mathrm{rec}} = \left(\hat{x}, \hat{y}, \hat{w}, \hat{h}\right)^{\mathsf{T}}$ を出力し，下段では微調整パラメータを表す $\hat{\boldsymbol{r}}_i^{\mathrm{refine}} = \left(\hat{dx}_1, \hat{dy}_1, \hat{dx}_2, \hat{dy}_2, \hat{dx}_3, \hat{dy}_3, \hat{dx}_4, \hat{dy}_4\right)^{\mathsf{T}}$

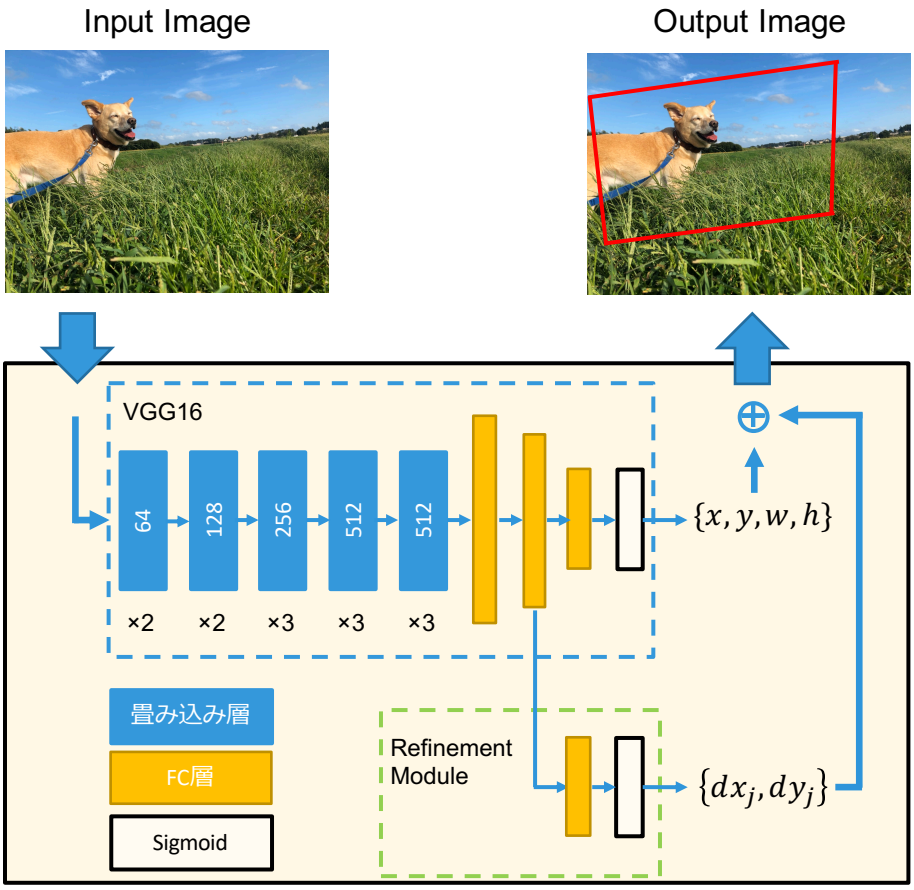


図 3.2: 凸四辺形領域の頂点座標を出力する手法 (DRN)．上段で矩形領域の座標表すパラメータを出力し，切り出し範囲を大まかに決定する．また，下段で微調整パラメータを出力し，最終的な凸四辺形領域の切り出し位置の微調整を行う．この 2 つのプロセスにより，凸四辺形領域切り出しを行う．

を出力する．まず，上段で出力される，矩形領域の座標表すパラメータで切り出し範囲を大まかに決定し，下段で出力される微調整パラメータで切り出し位置の微調整を行い，最終的に切り出す凸四辺形領域を決定する．

形状の変形を制限するために，式 3.2 を以下のように分解する．

$$L = \frac{1}{n}\sum_{i=1}^{n} |\hat{\boldsymbol{r}}_i - \boldsymbol{r}_i|^2 + \lambda \frac{1}{n}\sum_{i=1}^{n} \left|\hat{\boldsymbol{r}}_i^{\text{refine}}\right|^2 \tag{3.2}$$

ここで $\lambda$ は正則化パラメータを表す．

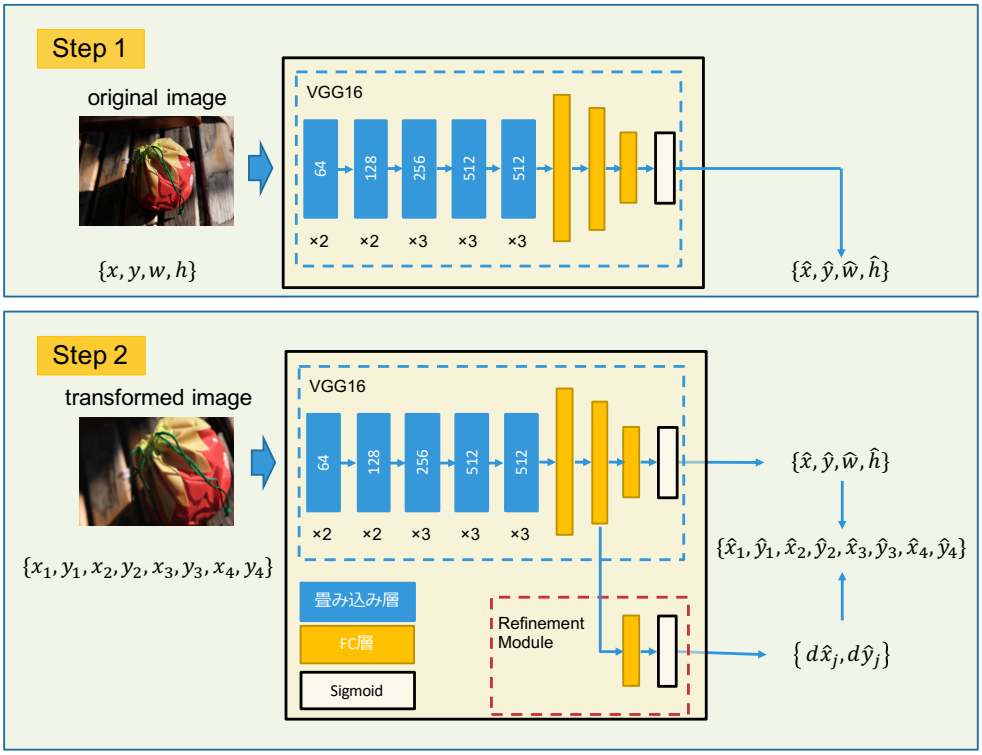


図 3.3: 2 step の学習による DRN の学習の流れ．Step 1 では矩形領域が正解領域を入力し，推定した矩形領域のパラメータの差を損失として学習する．その後 Step 2 で，凸四辺形領域が正解領域である画像を入力し，上段と下段で推定したパラメータを足し合わせて凸四辺形領域の各頂点推定し，正解情報との差を損失として学習させる．

本手法では 2 ステップの学習によってネットワークを学習する．図 3.3 に 2 step の学習の流れを示す．まず，Step 1 では，VGG16 をベースとしたネットワークに矩形領域が正解領域である画像を入力し，正解の矩形領域を出力するように学習する．その後，Step 2 では Refinemet Module を追加し，最終的なネットワーク構造を構築する．そのネットワークに，凸四辺形領域が正解領域である画像を入力し，凸四辺形領域を出力するように学習する．このように 2 step の学習を行うことで，出力する切り出しを凸四辺形に限定することができる．

### 3.3 射影変換による矩形への変換

回帰ネットワークから出力された 8 パラメータは入力画像中の 4 点に対応する．出力された 4 点と入力画像の 4 隅の点を対応させ，4 組の点のペアを得る．次に，この 4 組の点のペアを用いて，入力画像と凸四辺形切り出し領域のホモグラフィー行列を求める．求めたホモグラフィー行列を凸四辺形切り出し領域の各点に適用し，射影変換することで切り出した凸四辺形領域を入力画像のアスペクト比の矩形に変換する．以下では，ホモグラフィー行列の求め方を説明する．

ホモグラフィー行列は以下のような任意の 3 × 3 の行列である．

$$\boldsymbol{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \tag{3.3}$$

ある対応する点の組が

$$\boldsymbol{p} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \boldsymbol{P} = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} \tag{3.4}$$

のように与えられたとき，

$$\boldsymbol{P} = \boldsymbol{H}\boldsymbol{p} \tag{3.5}$$

の関係が成り立つ．式変形によって，

$$\begin{pmatrix} x & y & 1 & 0 & 0 & 0 & -xx' & -yx' \\ 0 & 0 & 0 & x & y & 1 & -xy' & -yy' \end{pmatrix} \begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{pmatrix} = \begin{pmatrix} x' \\ y' \end{pmatrix} \tag{3.6}$$

のように変形できる．上式において，未知の変数は 8 つである．1 組の点と点の関係がわかると，2 本の連立方程式が立てられるため，対応する 4 組の点がわかれば，連立方程式を解いてホモグラフィー行列を求めることが出来る．

得られたホモグラフィー行列を用いて画像の各ピクセルを射影変換し，矩形画像を得る．

## 3.4 ホモグラフィー行列を出力する手法 (HRN)

本節では，入力画像が与えられたときに，対応するホモグラフィー行列を直接出力する手法 (HRN) について説明する．本手法では，元画像の各頂点とネットワークによって出力されたホモグラフィー行列によって，凸四辺形の各頂点座標を推定する．推定した各頂点座標と正解領域の各頂点座標の差を損失として学習する．

本手法の概観を図 3.4 に示す．ネットワークの上段では画像を上段の出力と，下段の VGG16 をベースとした Localizer Network に入力し，特徴量を抽出する．得られた特徴量を用いてホモグラフィー行列を推定する．推定されたホモグラフィー行列を用いて上段の画像を射影変換することで凸四辺形領域切り出しを行う．

この提案手法は，spatial transformer networks [13] と似た構造をとっている．spatial transformer networks は物体の姿勢推定 [32] や顔識別 [39]，医用画像における細胞の分類 [1] などのタスクに用いられている．spatial transformer networks は，目的のタスクのためのネットワークの途中に挿入することができ，その際に新たに損失関数を定義せずに特徴量の変形を学習することできる．本研究では損失を定義しているため，spatial transformer networks の本来の用途ではないが，実装上は spatial transformer networks と同様の構造を採用し，特徴量から画像に変換する前段階で射影変換によって凸四辺形領域を出力している．

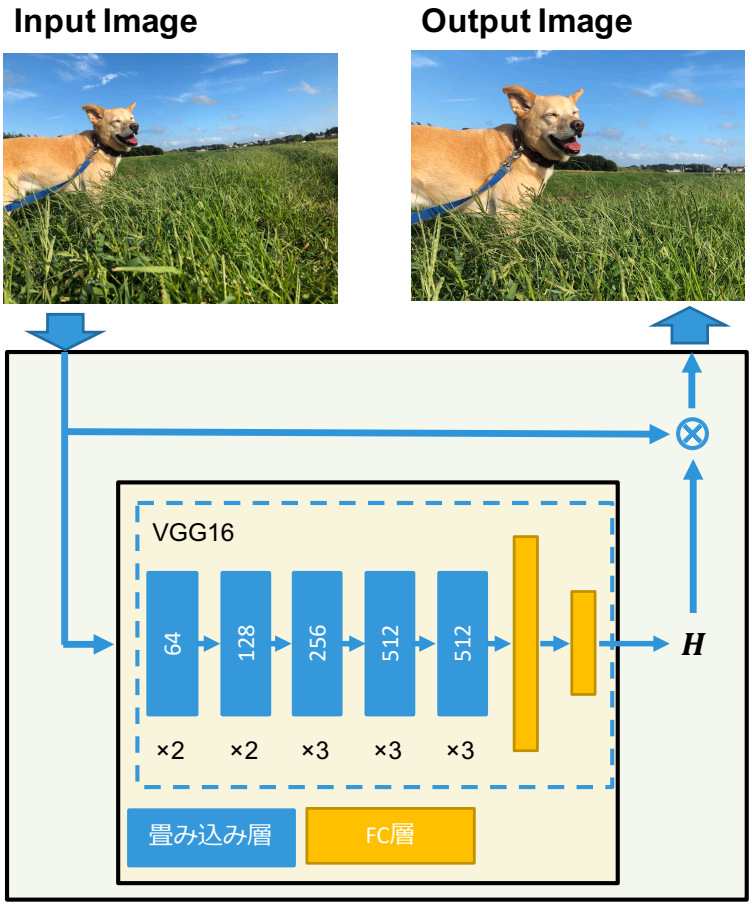


図 3.4: ホモグラフィー行列を出力する手法 (HRN) の概観．VGG16 をベースとした Localizer に画像を入力し，ホモグラフィー行列 $\boldsymbol{H}$ の各パラメータを出力する．出力したパラメータを用いてネットワークの内部で射影変換を行い，元画像中の凸四辺形領域を矩形に変換した画像を出力する．

# 第 4 章

# データセット

## 4.1 概要

本研究では，画像の回転や変形を考慮した凸四辺形領域の切り出し手法を提案する．矩形の正解領域が含まれる画像データセットは第 2 章に挙げたとおり，いくつか存在しているが，凸四辺形の正解領域が含まれる画像データセットは存在しない．

審美性を考慮した切り出しのためのデータセットを作成するためには，プロの写真家や芸術大学の卒業生らの協力のもとで作成されることが望ましいため，大規模な矩形の正解領域のアノテーションデータセットを構築するのは難しく，専門家によるアノテーションの付いた大規模なデータセットは存在していない．また，本研究で想定している，凸四辺形領域が正解領域としてアノテーションされているデータセットは存在していない．

そのため，本研究では凸四辺形領域が正解領域であるような状況を擬似的に作るために，関連研究で作成された，矩形領域が正解領域である画像を変形し歪ませることで，凸四辺形領域が正解領域であるデータセットを擬似的に作成した．

## 4.2 データセットの作成方法 (Random FCDB)

2 点位置がアノテーションされているデータセットの代表的なものに，FCDB [5] がある．FCDB は約 1600 枚の画像に対して正解領域がアノテーションされている．内訳はトレーニング用途の画像が約 1300 枚，テスト用途の画像が約 300 枚となって

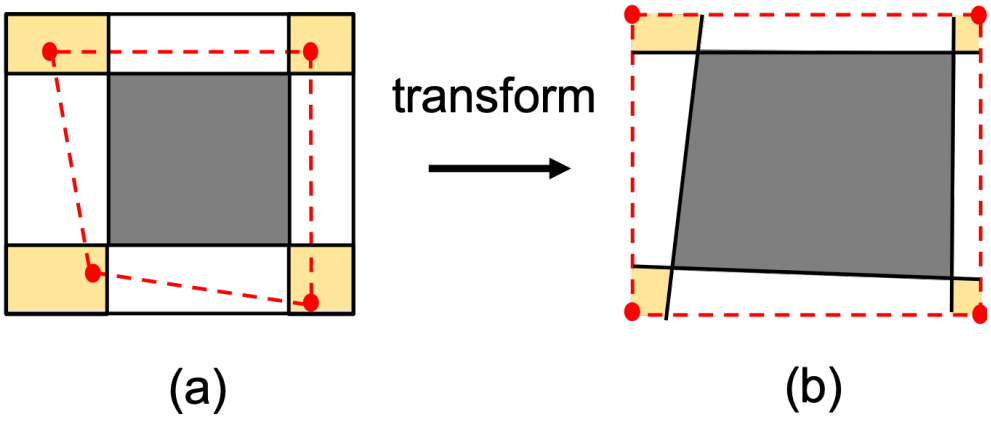


図 4.1: 評価用データセットの作成方法．(a)FCDB [5] の画像に対してグレー色の正解領域がアノテーションされている．グレー色の正解領域の外側にある 4 つの黄色の領域から点の座標をランダムに選択し，赤色の点線領域を作成する．(b) 選択した赤色の点線領域を入力画像の 4 隅に移動させるように射影変換を行う．

いる．本研究では，これらの画像をランダムに形状変形し，凸四辺形領域が正解領域となるデータセットの作成を行った．以降，作成したデータセットを Random FCDB と呼ぶこととする．

図 4.1 に評価用のデータセットの作成方法を示す．まず，画像にアノテーションされた正解領域の外側領域からランダムに点の座標を選択する．その後，選択した点の座標を結んだ四角形を元の画像のアスペクト比になるように射影投影変換を行うことでアノテーション付きのデータセットを作成した．作成したデータセットの一例を図 4.2 に示す．緑の線が正解領域がアノテーションされたものである．

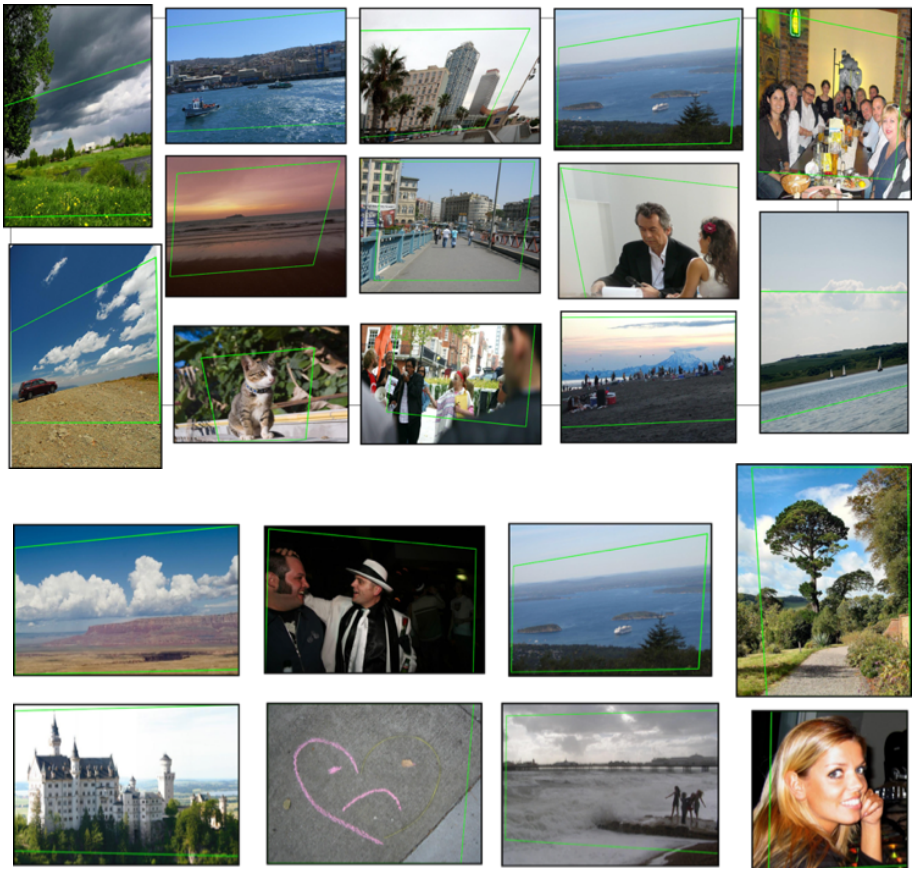

図 4.2: 作成したデータセットの一例．正解領域が大きいものは歪みが小さくなっているが大きく歪んだ画像も存在している．

# 第 5 章

# 実験

## 5.1 実験 1 ：提案手法の切り出し精度評価

本研究では提案手法の評価のために第 4 章で作成したデータセットを用いて切り出し精度の評価を行った．本研究で提案する 2 つの手法は，入力画像が与えられたときに，矩形を表現するパラメータと歪ませるパラメータを組み合わせて，凸四辺形領域の切り出しを行う．本章では，矩形領域が正解領域となるデータセットでの切り出し精度と凸四辺形領域が正解領域となるデータセットの両方を切り出し精度を評価し，異なるデータセット間での精度の比較と学習手法による精度の比較を行い，各手法の性能を評価した．

### 比較手法

比較する各手法は以下の 7 つである．

- DRN(4 points)：4 点の座標を出力し，入力画像から 4 点を結んだ凸四辺形領域を切り出す手法（提案手法 1）
- DRN(2 points)：矩形の左上の座標と右下の座標を出力し，入力画像から 2 点を結んだ矩形領域を切り出す手法（提案手法 1 の Ablation Study）
- HRN：ホモグラフィー行列を出力し，入力画像を変形することで凸四辺形領域を切り出す手法（提案手法 2）

- VPN [35]：物体検出をベースとした矩形領域を出力する既存手法（ベースライン 1）
- E2E [20]：Saliency をベースとした矩形領域を出力する既存手法（ベースライン 2）
- Baseline_N：入力画像を切り出ししない手法（ベースライン 3）
- Baseline_C：入力画像の高さと幅を 0.9 倍した領域を画像の中心矩形領域を切り出す手法（ベースライン 4）

既存研究の矩形領域を切り出す手法である，VPN [35] と E2E [20] をベースラインとして用意した．また，各手法が評価用のデータセット特有の正解領域分布に過度に適合していないかを確認するために，ベースラインとして Baseline_N と Basekine_C を用意した．さらに，提案手法 1 の Refinement Module の効果による，凸四辺形領域の切り出し精度の向上性能評価のために，DRN(2 points) を比較手法として用いた．

#### 評価指標

評価指標として，IoU (Intersection over Union) と Disp (Points Displacement Error) を以下のように定義した．

$$IoU = \frac{\mathrm{Area}^{\mathrm{gt}} \cap \mathrm{Area}^{\mathrm{pred}}}{\mathrm{Area}^{\mathrm{gt}} \cup \mathrm{Area}^{\mathrm{pred}}} \tag{5.1}$$

$$Disp = \sum_{j=1}^{4} \frac{|\hat{B}^j - B^j|}{4} \tag{5.2}$$

IoU は 2 つの領域の重なる具合を示す指標であり，物体認識や画像切り出しの研究で広く用いられている評価指標である．式 5.1 において，$\mathrm{Area}^{\mathrm{gt}}$ と $\mathrm{Area}^{\mathrm{pred}}$ は，それぞれ正解の領域と推定した領域を表す．一方，Disp は 2 つの領域の端点の距離の差を表す指標である．式 5.2 において，$\hat{B}^j$ と $B^j$ は，それぞれ正解領域の各頂点の座標と推定した領域の各頂点の座標を表している．本研究では，これらの指標を用いて定量評価を行った．

## 5.2 実装

### DRN の学習用データセット

DRN の最初のステップの学習のためのデータセットを作成した．データの数が限られているため，データ拡張によって擬似的に FCDB のデータ数を増やした．まず，入力画像に対してランダムに数ピクセル変更し（×5）水平反転（×2）する．次に，入力画像のアスペクト比を保ったまま，長辺を 244 に変更する．その画像を，$C \in 50\%, 60\%, 80\%, 90\%$ のダウンスケールをする （×4）．その後，ダウンスケールした画像の {左上，右上，右下，左下} のいずれかの領域にパッディングする（×4）．最後にアスペクト比を考慮しないで画像サイズを 224 まで下げたものを用意する．このような方法で，一枚の画像から $5 \times 2 \times (4 \times 4 + 1) = 170$ 枚の画像を作成した．

### DRN の実装

DRN のベースは VGG16 で構築した．VGG16 の最後の pooling layer より上層を取り除き，新たな全結合層を 2 層重ね，全結合層のチャネル数をそれぞれ 1024 と 512 に変更した．DRN の重みパラメータは，ImageNet で事前学習済み VGG16 モデルの重みパラメータで初期化した．DRN の学習は，図 3.3 のように 2 つの段階に分かれている．最初に 5.2 節と同様の手順で作成した矩形領域が正解領域であるデータセットを用いて，上段の VGG ベースのネットワークを学習させ，矩形領域を精度高く出力できるように学習させる．その後，4 章で作成する凸四辺形領域が正解領域であるデータセットを用いて，ネットワーク全体を学習させる．

どちらの段階においても，オプティマイザーには確率的勾配降下法 (SDG) を用い，学習率を 0.01 に設定し，モーメンタムを 0.9 に設定した．最初の 1 ステップ目の学習は 5 イテレーション行った．2 ステップ目の学習は 45 イテレーション行った．2 ステップ目の学習では 20 エポックごとに学習率を 0.1 倍した．バッチサイズは 32 とした．5 イテレーションごとに評価データセットでスコアを比較し，最も評価スコアが高いものを選択した．

**HRN の実装**

HRN のベースは VGG16 で構築した．VGG16 の最後の pooling layer より上層を取り除き，新たな全結合層を 2 層重ね，全結合層のチャネル数をそれぞれ 1024 と 512 に変更した．HRN の重みパラメータは，ImageNet で事前学習済み VGG16 モデルの重みパラメータで初期化した．5.2 節で作成する凸四辺形領域が正解領域であるデータセットを用いて，ネットワーク全体を学習させる．

オプティマイザーには確率的勾配降下法 (SDG) を用い，学習率を 0.01 に設定し，モーメンタムを 0.9 に設定した．学習は 45 イテレーション行い，20 イテレーションごとに学習率を 0.1 倍した．バッチサイズは 32 とした．5 イテレーションごとに評価データセットでスコアを比較し，最も評価スコアが高いものを選択した．

## 5.3 結果と考察

表 5.1 と表 5.2 に各手法の精度比較を示す．本実験では，どちらの評価データセットにおいても DRN と HRN はともに，いずれのベースラインの手法よりも精度が高かった．以下では，本研究で提案する回帰による凸四辺形領域切り出し手法の各機能の調査とその結果から導かれる考察を述べる．

表 5.1: FCDB における各手法の精度比較

| Method | Avg IoU | Avg Disp |
|---|---|---|
| DRN | 0.685 | 0.073 |
| HRN | 0.689 | 0.073 |
| VPN [35] | 0.670 | 0.084 |
| E2E [20] | 0.670 | 0.086 |
| Baseline_N | 0.645 | 0.087 |
| Baseline_C | 0.670 | 0.085 |

表 5.2: Random FCDB における各手法の精度比較

| Method | Avg IoU | Avg Disp |
|---|---|---|
| DRN | 0.781 | 0.048 |
| HRN | 0.786 | 0.052 |
| VPN [35] | 0.721 | 0.073 |
| E2E [20] | 0.750 | 0.066 |
| Baseline_N | 0.766 | 0.053 |
| Baseline_C | 0.774 | 0.058 |

#### 定性的な考察

ここでは定性的な考察を述べる．図 5.1 に DRN と HRN による出力画像を示す．図 5.1(a) は DRN による出力を表す．赤色の矩形が正解領域を表し，緑色の矩形が各手法が出力した凸四辺形領域である．図 5.1(b) は (a) の緑色の領域を射影変換した画像である．図 5.1(c) は HRN による出力を表し，図 5.1(d) は (c) の緑色の領域を射影変換した画像である．

どちらの手法の結果も正解領域の大半をカバーしている．また，どちらの手法による切り出しにおいても，矩形が出力されており，出力は矩形に近い形状をしている．図 5.1 の上段 3 つの画像のように水平線が傾いた画像や下段のような形状が歪んだ画像から，変形を考慮した画像の切り出しはできていないことが見て取れる．

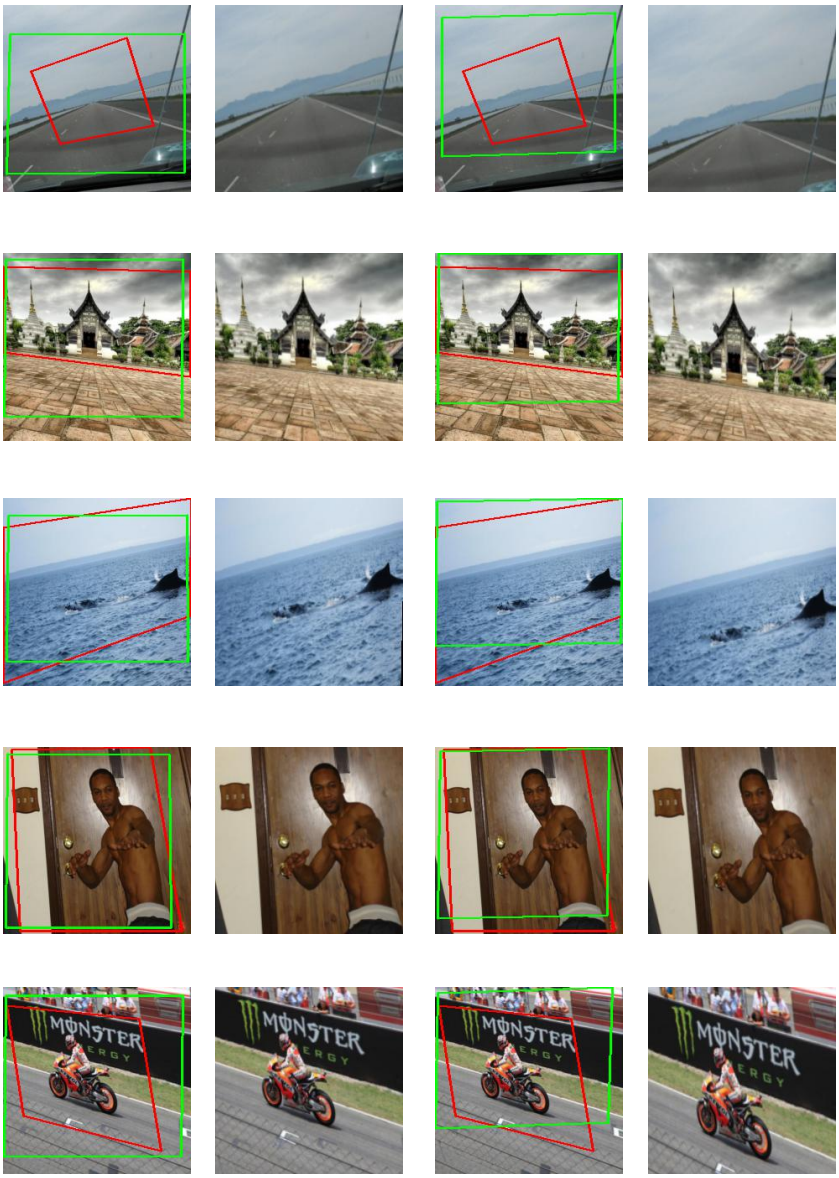


図 5.1: Random FCDB における提案手法の結果．(a)(c) が DRN と HRN による結果．赤色が正解領域であり，緑色が手法によって推定された領域である．(b)(d) が DRN と HRN による結果を射影変換した画像である．どちらの手法の結果も正解領域の画像の大半を囲んでいる．

表 5.3: DRN の Ablation Study．Refinement Module(RM) によって推定精度が向上している．

| Method | Avg IoU | Avg Disp |
|---|---|---|
| DRN (FCDB) | 0.685 | 0.073 |
| DRN w/o RM (FCDB) | 0.679 | 0.073 |
| DRN(Random FCDB) | 0.781 | 0.048 |
| DRN w/o RM (Random FCDB) | 0.772 | 0.049 |

表 5.4: 学習データ別の性能比較．評価データには Random FCDB を用いている．学習で使用したデータに含まれる正解領域の大きさの違いによりテストスコアに差が生じている．

| Method | Avg IoU | Avg Disp |
|---|---|---|
| DRN(aug) | 0.676 | 0.068 |
| DRN | 0.781 | 0.048 |

**DRN における Refinement Module の効果**

DRN における Refinement Module の有無による評価精度の違いを考察する．表 5.3 の結果では，両方の評価データセットにおいて，DRN モデルのほうが Refinement Module を持たない DRN よりも切り出し精度が大きい．このことから，Refinement Module が凸四辺形領域の切り出し精度向上に影響することが確認できる．

**DRN の学習に使用するデータセットの効果**

DRN の学習に使用するデータセットごとの評価精度の違いをもとに，学習に使用するデータセットが DRN の性能に与える効果を考察する．表 5.4 は，DRN の学習に用いたデータセット別の評価精度の表である．DRN(aug) は 5.2 節で作成した画像を用いたものである．通常の DRN は 5.2 節で作成したデータセットでの学習後に，4 章で作成したデータセットを用いて学習する．表 5.4 からは，4 章で作成したデータセットを用いて学習した DRN の方が精度が高いことが確認できる．2 つの学習用データセットは同じデータセットを元にして作成されている．つまり，画像全体に

対する正解領域の割合の分布は等しい．しかし，5.2 節で作成した画像は元画像を縮小して，元画像のサイズ画像に配置し，空白部分をパディングすることでデータ拡張を行っている．そのため，学習用のデータにおける DRN(aug) 出力される凸四辺形領域は元画像全体に対する正解領域の割合の分布は等しいとみなせるが，空白領域を含めた全体画像に対する正解領域の割合は，4 章で作成したデータセットよりも平均して小さくなっている．このデータセットの画像全体に占める正解領域の割合の違いが DRN の性能に影響していると考えられる．

**HRN のバッチサイズごとの精度比較**

バッチサイズは，ネットワークが学習するまでに使用するデータの数である．バッチサイズが大きい場合，ネットワークはバッチサイズ分だけのデータの平均を学習することになる．バッチサイズを大きく設定すると学習にかかる時間を削減することができる．反対に，バッチサイズが小さい場合には，データの一つ一つに敏感に反応するように学習するが，学習にかかる時間は大きくなる．

図 5.2 に FCDB における HRN のバッチサイズごとの評価精度を示す．FCDB データセットにおいては，バッチサイズの大きさによる評価精度の違いに明確な傾向は見られない．図 5.3 に RandomFCDB の HRN のバッチサイズごとの評価精度を示す．RandomFCDB データセットにおいては，バッチサイズに比例して HRN の評価精度が小さくなる傾向が見られる．これらの結果から，HRN でのホモグラフィー行列を推定することで凸四辺形領域を切り出す際には，バッチサイズを比較的小さく設定することが望ましいと考えられる．

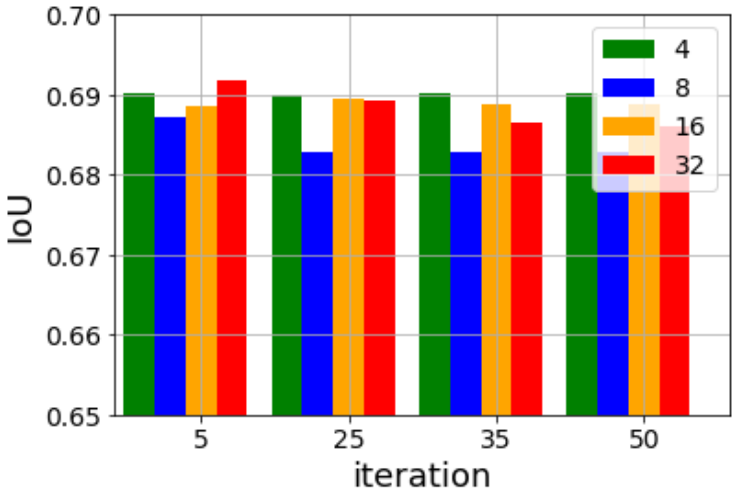


図 5.2: FCDB における HRN のバッチサイズごとの評価精度．バッチサイズによる精度の変化は小さい．

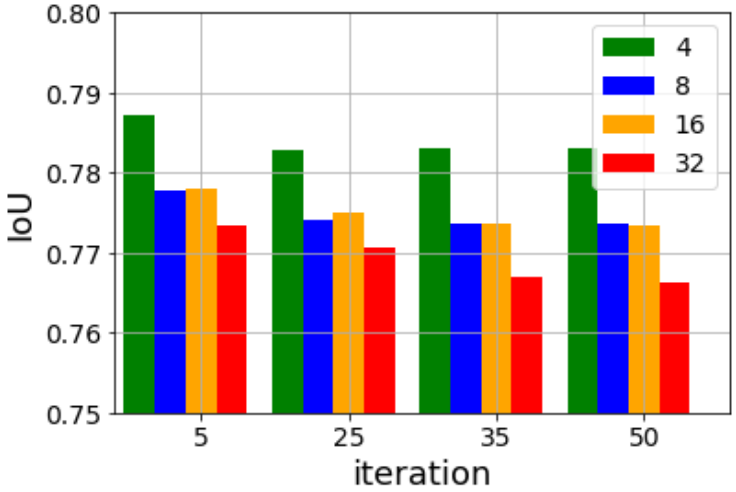


図 5.3: RandomFCDB における HRN のバッチサイズごとの評価精度．バッチサイズに比例して精度が小さくなっている．

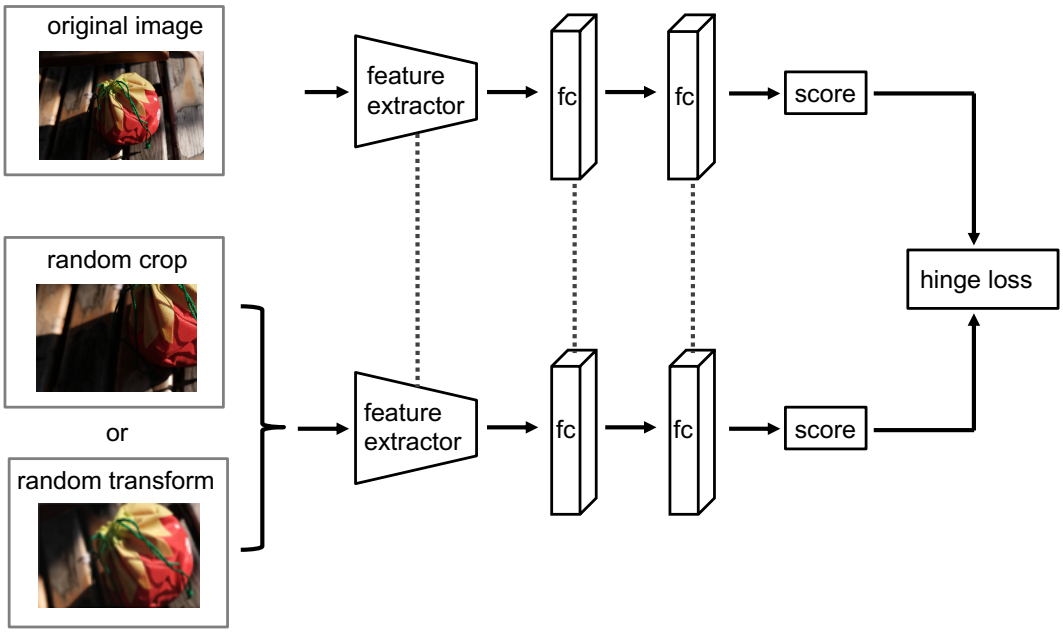


図 5.4: 審美性評価器 (TVEN) の概観．VFN [6] や VEN[35] と同様の構造である．本研究ではこのネットワークが回転・変形を考慮した評価ができるように学習させる．

## 5.4 実験 2：回転や変形を考慮した審美性評価器の評価

本研究では，画像の回転や変形を考慮した審美的画像の切り出しに取り組んでいる．既存研究では，審美性評価器を用いて，総当たり法によって画像切り出しを行う手法や，探索する範囲を限定して限られた画像候補に対して審美性評価を行う手法が提案されてきた．それらの手法では審美性評価器の性能が手法の性能を決定することになる．これまで提案されてきた VFN [6] や VEN [35] などの審美性評価器は，画像の回転や変形は考慮されてこなかったため，回転や変形による審美性の違いをうまく評価できないことが考えられる．ただ，同じ画像中の複数の切り出しどうしの審美性のスコアを学習させた手法 [35] や似たようなコンテンツの画像どうしから審美性の高い画像を選ぶ手法 [3] などに見られるように，学習に用いるデータセットの工夫によって回転や変形による審美性の違いを評価できることが期待される．そこで本研究では，回転や変形を考慮した審美性評価器 (Transformed Evaluation Net) を作成し，審美性に対する精度を既存手法と比較した．

### 回転や変形を考慮した審美性評価器の作成

図 5.4 に本研究で作成する，Transformed Evaluation Net(TVEN) のネットワーク構造と学習のプロセスの概観を示す．上段のネットワークには，ランダムに切り出した画像またはランダムに変形させた画像を入力する．下段のネットワークにはオリジナルの画像を入力する．上段と下段のネットワークはそれぞれ審美性スコアを出力し，出力されたスコアの差を損失として審美性評価器の機能を学習する．注意されたいのは，上段と下段のネットワークの重みパラメータは共有されていることである．このような共通のネットワーク重みを共有した 2 つのネットワークは Siamese Architecture [7] と呼ばれており，距離学習の 1 手法として広く使われている．審美性評価器の目標は，以下のマッピング関数 $f$ を学習することである．

$$f\left(I_i\right) > f\left(C_{ij}\right) \tag{5.3}$$

$I_i$ は全体画像を表し，$C_{ij}$ はと全体画像からランダムに切り出された $j$ 個目の画像を表す．ここで，$I_i$ は常に $C_{ij}$ よりも審美性スコアが高いことを想定している．画像のペアの損失 $l\left(I_i, C_{ij}\right)$ を以下のヒンジ損失で定義することで，マッピング関数 $f$ を学習する．

$$l\left(I_i, C_{ij}\right) = \max\left\{0, g + f\left(C_{ij}\right) - f\left(I_i\right)\right\} \tag{5.4}$$

ここで $g$ はギャップを表すパラメータで，$I_i$ と $C_{ij}$ の間の最小のマージンとしての働きがある．本研究では [6] に従い，$g = 1$ に設定する．

### 審美性評価器の探索範囲選択

審美的画像切り出し手法の最もシンプルな方法として，審美性評価器を総当たり (brute-force) で行い，スコアの最も高いものを選ぶ方法がある．しかし，凸四辺形選択における審美性評価器を総当たり法では，パラメータの数が従来研究の 2 倍になるため，計算量は従来研究の計算量が 2 乗倍になり Sliding Window 方式による総当りの評価ができない．そこで本実験では， [37] にならい，点の選択範囲をグリッ

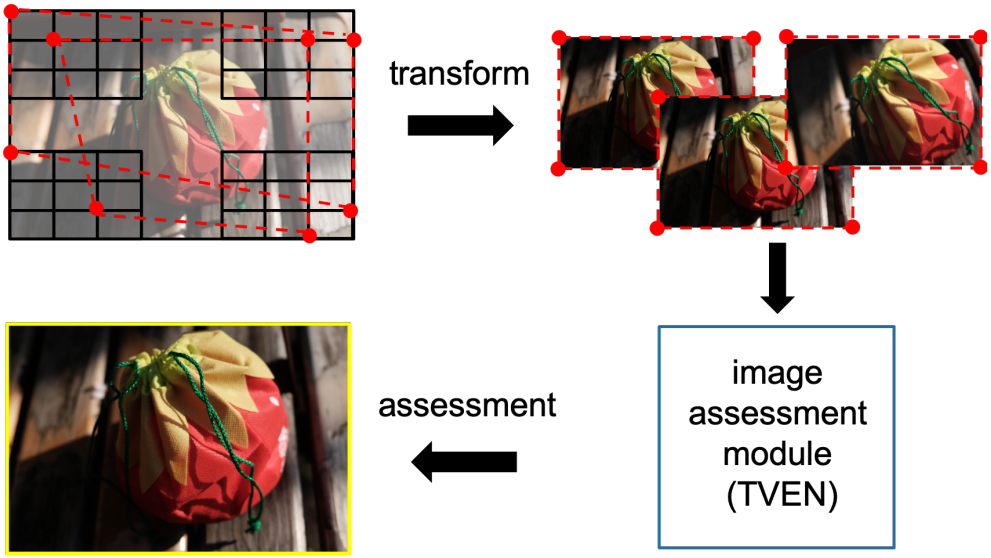


図 5.5: 審美性評価器による凸四辺形切り出し手法． [37] でのように選択する範囲をグリッドに限定して，各グリッドから点を選択し凸四辺形を一つ取り出す．取り出した凸四辺形を矩形に射影変換したものを審美性評価器で評価する．

ドに限定した総当り手法での評価を行う．本実験では，TVEN と VFN [6] の審美性評価器の性能の比較を行い，歪みを考慮した画像を用いた学習の有効性を検証する．

図 5.5 に審美性評価器による凸四辺形切り出し手法の手順を示す．画像の縦横それぞれを 11 分割し，144 個の格子点を作成する．$m$ を自然数としたとき，画像の 4 隅の点から中心に向かう $(m, m)$ の範囲の格子点から 1 点を選ぶ．例えば，$m = 3$ の場合には，凸四辺形の各頂点を，それぞれ 9 個の格子点の中から選ぶ．選んだ格子点を結んだ凸四辺形領域を射影変換によって矩形に変形し，審美性評価器に入力する．審美性評価器は各切り出しに対して審美性スコアを出力し，最も高いスコアのものを最適な切り出し結果とする．このような簡易的な切り出し手法を用いて，審美性評価器の性能を評価する．

## 5.5 実装

### 審美性評価器の学習データセット

作成する審美性評価器は，歪みがある不自然な画像からは低いスコアを出し，歪みのない自然な画像からは高いスコアを出すように学習させたい．そこで，審美性評価の高い画像を変形させて擬似的に形状が不自然な画像を大量に作成し，元の画像とのペアを作成する．元画像には，[6] で構築された画像を用いた．[6] には，Flicker において高評価が付けられた画像 21045 枚と，それらの画像に対して 14 個のランダムな切り出し領域情報が含まれている．

Flicker において高評価が付けられた画像は審美性の高い画像とみなせる．審美性の高い画像は，少しの回転・変形の操作を加えることで十分に画像の品質を損ねることができるため，[6] に含まれる画像をランダムに変形することで，歪みのある画像を作成した．審美性評価器の学習では，[6] のランダムな切り出し画像の 4 割を元画像を歪ませた画像に置き換えて学習させる．

歪ませた画像は，第 4 章の図 4.1 の評価用データセットの作成方法と同様の過程で作成した．その際，Flicker において高評価が付けられた画像には正解領域が含まれていないため，正解領域を自ら設定した．ここでは，審美性評価器が少しの回転・変形に対して敏感に反応して審美性を評価するように学習させるために，元画像の 0.85 倍の領域を元画像の中心に配置し，正解領域とした．これらの過程によって学習用のデータセットを作成した．

### 審美性評価器の実装

TVEN は VGG16 [30] を基本構造とした．VGG16 の最後の pooling layer より上層を取り除き，新たな全結合層を 2 層重ねた．全結合層のチャネル数をそれぞれ，1024 と 512 に変更した．TVEN の訓練では，重みパラメータを ImageNet で事前学習済み VGG16 モデルで初期化し，4.2 節で作成したデータセットの画像ペアを用いて学習する．最初の 60 エポックでは学習率を 0.001 に設定し，20 エポックごとに 0.1 倍した．オプティマイザーには確率的勾配降下法 (SDG) を用い，モーメンタムを 0.9

表 5.5: FCDB における審美性評価器の性能比較．回転を考慮した画像を含んだ TVEN の精度は既存手法の VFN と比較して，グリッドの数が多い場合での評価精度は高いが，グリッドの数が少ない場合での評価精度は低い．

| Method | Avg IoU | Avg Disp |
|---|---|---|
| VFN (2 points)($4 \times 4$) | 0.6198 | 0.1052 |
| VFN (4 points)($2 \times 2$) | 0.6610 | 0.0941 |
| TVEN (2 points)($4 \times 4$) | 0.6326 | 0.1017 |
| TVEN (4 points)($2 \times 2$) | 0.6576 | 0.0946 |

表 5.6: Random FCDB における審美性評価器の性能比較．回転を考慮した画像を含んだ TVEN の精度は既存手法の VFN と比較して，グリッドの数が多い場合での評価精度は高く，グリッドの数が少ない場合での評価精度も高い．

| Method | Avg IoU | Avg Disp |
|---|---|---|
| VFN (2 points)($4 \times 4$) | 0.7012 | 0.0828 |
| VFN (4 points)($2 \times 2$) | 0.7334 | 0.0753 |
| TVEN (2 points)($4 \times 4$) | 0.7576 | 0.0659 |
| TVEN (4 points)($2 \times 2$) | 0.7677 | 0.0639 |

に設定した．1 エポックごとに validation を行い，最も validaton score の高いモデルを用いた．

## 5.6 結果と考察

表 5.5 と表 5.6 に FCDB と RandomFCDB における各審美性評価器の性能比較の結果を示す．

### 評価精度の比較

Random FCDB における結果から，回転や変形を考慮したデータセットで学習した評価器 TVEN は，既存手法の VFN と比較して精度が高いことが確認された．この結果から，学習データセットに回転や変形を施した画像を含めることで，歪みの含まれた画像に対しては既存の手法よりも性能が向上することが示された．しかし，FCDB に対する評価において，VFN と比較すると，わずかではあるが評価精度が低い．そのため，回転や歪みに対する審美性評価器の性能と回転や歪みを考慮しない審美性評価器の性能はトレードオフの関係にあることが考えられる．

また，どちらのデータセットに対しても，TVEN と VFN の両方とも $(4 \times 4)$ のグリッドから点を選択する手法よりも，$(2 \times 2)$ のグリッドから点を選択する出力する手法のほうが精度が低い．このことから審美性評価器は比較的小さい面積の領域を審美性が高いと評価していることが考察される．

図 5.6 と図 5.7 に TVEN を用いた簡易的な切り出し手法による出力結果と，出力した画像を射影変換した画像を示す．図 5.6 の 1 行 2 列の画像や 4 行 2 列の画像などの歪んだ画像であっても画像全体を切りだしており，歪みを考慮した切り出しになっていない画像も見受けられる．しかし，1 行 3 列の画像や 4 行 4 列の画像のような被写体が明確な物体である場合には，歪みを改善するような切り出しができている．これらをまとめると，TVEN は形状を考慮したのではなく，物体などのシンプルなコンテンツが画像に占める割合が，大きくなるように切り出していると考察できる．

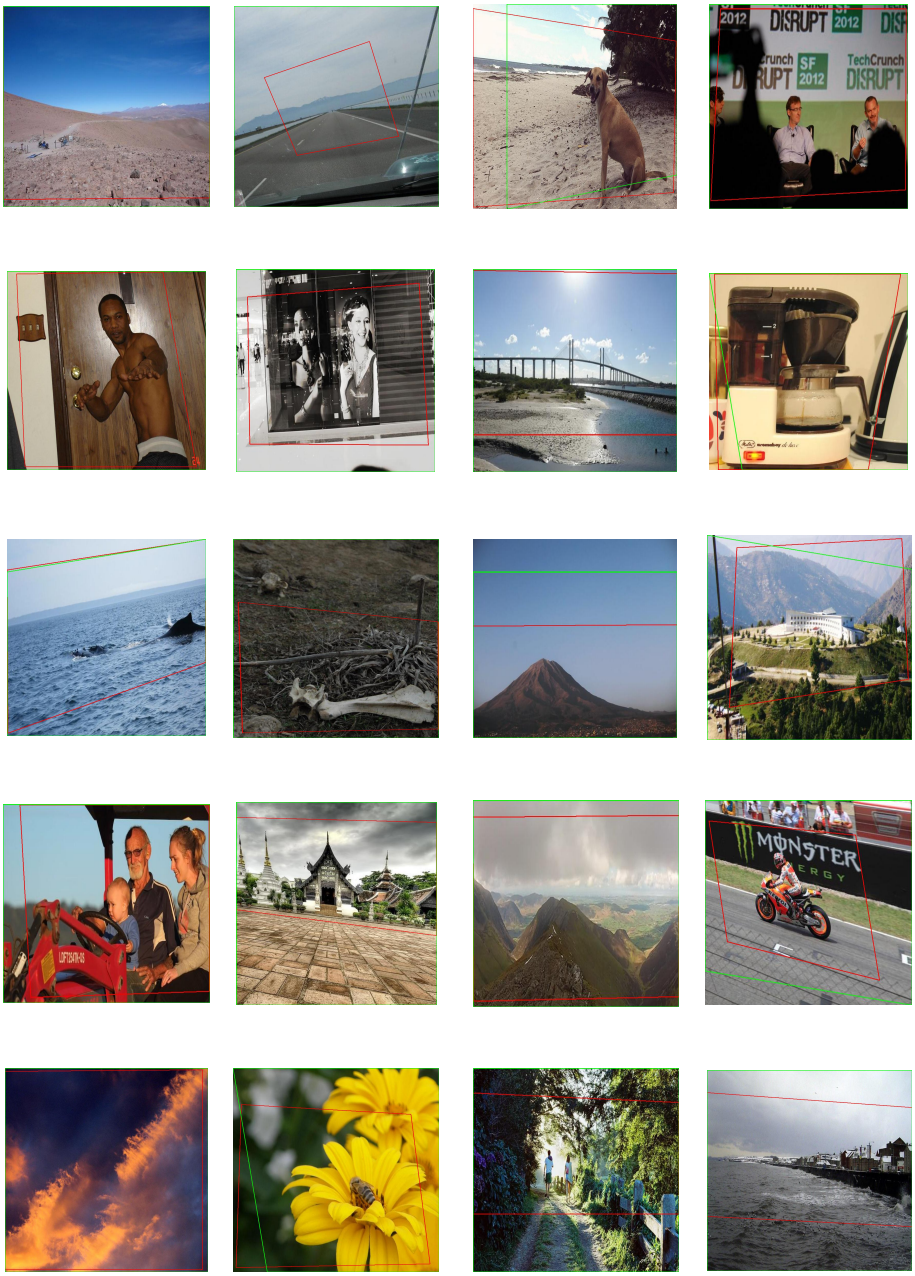


図 5.6: RandomFCDB における審美性評価器 (TVEN) の出力画像．赤色の凸四辺形が正解領域を示し，緑色の凸四辺形が TVEN の出力する審美性スコアが最も高い凸四辺形である．

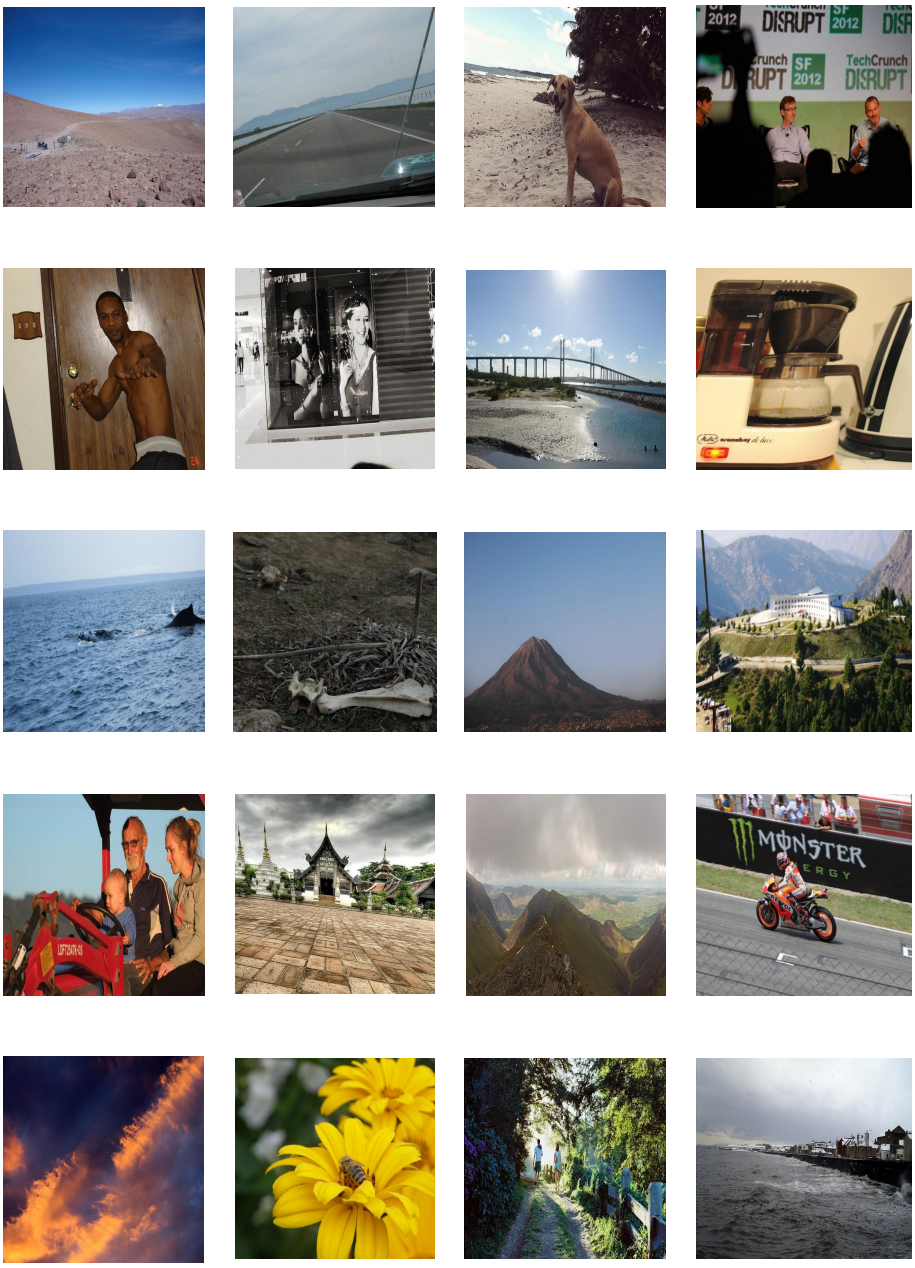


図 5.7: RandomFCDB における審美性評価器 (TVEN) の出力画像を射影変換した画像一例.

表 5.7: Random FCDB における VFN の性能比較

| Method | Avg IoU | Avg Disp | Avg Steps | Avg Times(s) |
|---|---|---|---|---|
| VFN (4 points)($2 \times 2$) | 0.7334 | 0.0753 | 256 | 6.53 |
| VFN (4 points)($3 \times 3$) | 0.6879 | 0.0896 | 6561 | 390 |

#### 計算時間の考察

最後に本手法を審美性を考慮した凸四辺形画像切り出し手法としての使用可能性について，出力にかかる計算時間の観点から考察する．表 5.7 に Random FCDB における VFN の性能比較を示す．グリッド数が $(2 \times 2)$ の場合，1 枚の画像から切り出す候補画像の数は 256 枚になり，その処理に 7 秒程度かかっている．さらに，グリッド数が $(3 \times 3)$ の場合には，1 枚の画像から切り出す候補画像の数は 6561 枚になり，その処理に 390 秒かかっている．この結果から，凸四辺形画像切り出しにおいては審美性評価器を用いた網羅的な切り出しは現実的でなく，対象領域を絞った手法を取る必要があることが確認された．

# 第 6 章

# 結論

## 6.1 結論

本論文では，審美性を考慮した凸四辺形領域切り出し手法における課題を示し，その解決策として，回帰による凸四辺形領域切り出し手法を提案した．本研究では，凸四辺形領域の 4 隅の座標を推定する手法とホモグラフィー行列パラメータを推定する手法を提案し，凸四辺形領域が正解領域となる画像の擬似的なデータセットを用いて，各手法の性能評価を行った．評価実験の結果，どちらの凸四辺形領域の切り出し手法も，既存手法よりも高い精度を示した．しかし，切り出された画像を観察すると，画像の歪みや被写体の形状を考慮したモデルでないことがわかった．また，回転や変形を考慮した審美性評価器を作成し評価した．提案した審美性評価器は歪みの含まれる画像データセットにおいては高い性能を示したが，歪みのない画像データセットにおいては精度が向上しなかった．

## 6.2 今後の方針

ここでは，審美性を考慮した凸四辺形切り出し手法における今後の方針を取り上げる．

### データセットの構築

本研究における提案手法では，既存のデータセットの画像を歪ませて正解領域が凸四辺形領域となるを擬似的なデータセットを作成し，学習と評価を行った．しかし，既存のデータセットの画像は，正解領域が画像に占める割合が大きいため，歪ませたとしても大きな歪みを生じさせることができなかった．そのため，今後は正解領域が小さい画像データセットを構築することが求められる．データセットの構築の一例として，広角の一人称視点カメラとスマートフォンのカメラを用いて同時に撮影し，後処理で画像間の対応点マッチングを行うことで広角の一人称視点カメラで撮影された画像に凸四辺形領域の情報をアノテーションする方法が挙げられる．

### 強化学習アプローチ

強化学習アプローチは，報酬の設定次第で少ない試行回数で目的の切り出し行える手法である．既存の強化学習ベースの手法である [17] では報酬を与える審美性評価モジュールとして VFN を用いている．新たなデータセットによって，審美性評価器を学習させることができれば，回転や変形を考慮した審美性評価を用いた強化学習手法による凸四辺形切り出しに拡張できる可能性がある．また，エージェントにあたえる報酬に IoU を用いた手法 [38] のように，報酬に審美性スコア以外のものを設定することも挙げられる．

### Rotation Region Proposal Network の適用

回転を考慮した文字認識手法である [22] [15] を任意の凸四辺形を考慮する手法に拡張することも考えられる．これらの手法はアフィン変換までを対象にしているが，これらの研究で用いられているアプローチを射影変換まで拡張することで，審美性を考慮した凸四辺形領域切り出し手法に取り組むことが考えられる．その際に考慮に入れる必要があるのがデフォルトボックスの選定であり，一般的な物体認識に必要な小さいサイズのデフォルトボックスが必要ない代わりに，様々な形状のデフォルトボックスを作成する必要があるため，精度と計算量のトレードオフを考慮する必要がある．

# 参考文献


[1] Marc Aubreville, Maximilian Krappmann, Christof Bertram, Robert Klopfleisch, and Andreas Maier. A guided spatial transformer network for histology cell differentiation. *arXiv preprint arXiv:1707.08525*, 2017.

[2] Luigi Celona, Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Autocropping: A closer look at benchmark datasets. In *Proc. IAPR International Conference on Image Analysis and Processing (ICIAP)*, pages 315–325, 2019.

[3] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)*, 35(4):148, 2016.

[4] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping : A computational complexity study. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 507–515, 2016.

[5] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 226–234, 2017.

[6] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *Proc. ACM Conference on Multimedia (ACMMM)*, pages 37–45, 2017.

[7] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005.

[8] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine (SPM)*, 34(4):80–106, 2017.

[9] Seyed A Esmaeili, Bharat Singh, and Larry S Davis. Fast-at: Fast automatic thumbnail generation using deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4622–4630, 2017.

[10] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proc. ACM international conference on Multimedia (ACMMM)*, pages 1105–1108, 2014.

[11] Jingwei Huang, Huarong Chen, Bin Wang, and Stephen Lin. Automatic thumbnail generation based on visual representativeness and foreground recognizability. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 253–261, 2015.

[12] Md Baharul Islam, Wong Lai-Kuan, and Wong Chee-Onn. A survey of aesthetics-driven image recomposition. *Multimedia Tools and Applications (MTA)*, 76(7):9517–9542, 2017.

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2017–2025, 2015.

[14] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, 2015.

[15] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

[16] Yueying Kao, Ran He, and Kaiqi Huang. Automatic image cropping with aesthetic map and gradient energy map. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1982–1986, 2017.

[17] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8193–8201, 2018.

[18] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Transactions on Image Processing (TIP)*, 28(10):5105–5120, 2019.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. IEEE European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[20] Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint arXiv:1907.01432*, 2019.

[21] Weirui Lu, Xiaofen Xing, Bolun Cai, and Xiangmin Xu. Listwise view ranking for image cropping. *arXiv preprint arXiv:1905.05352*, 2019.

[22] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.

[23] Shuai Ma, Zijun Wei, Feng Tian, Xiangmin Fan, Jianming Zhang, Xiaohui Shen, Zhe Lin, Jin Huang, Radomír Měch, Dimitris Samaras, et al. Smarteye: Assisting instant photo taking via integrating user preference with deep view

proposal network. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, pages 471:1–471:12, 2019.

[24] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 1928–1937, 2016.

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2408–2415, 2012.

[26] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. Aesthetic quality classification of photographs based on color harmony. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 33–40, 2011.

[27] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato. Sensation-based photo cropping. In *Proc. ACM international conference on Multimedia (ACMMM)*, pages 669–672, 2009.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[29] Tomoya Sawada, Masahiro Toyoura, and Xiaoyang Mao. Auto-framing based on user camera movement. In *Proc. ACM Computer Graphics International Conference (CGI)*, page 18, 2017.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. *arXiv preprint arXiv:1911.10492*, 2019.

[32] K. Ueno, G. Irie, M. Nishiyama, and Y. Iwai. Weakly supervised triplet learning of canonical plane transformation for joint object recognition and pose estimation. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 2476–2480, 2019.

[33] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2186–2194, 2017.

[34] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(7):1531–1544, 2019.

[35] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5437–5446, 2018.

[36] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 971–978, 2013.

[37] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5949–5957, 2019.

[38] Yaqing Zhang, Xueming Li, and Xuewei Li. Photo cropping via deep reinforcement learning. In *Proc. IEEE International Conference on Agents (ICA)*, pages 86–90, 2019.

[39] Yuanyi Zhong, Jiansheng Chen, and Bo Huang. Toward end-to-end face recognition through alignment learning. *IEEE Signal Processing Letters (SPL)*, 24(8):1213–1217, 2017.