令和 04 度卒業論文

# 自己注意による コンテクスト抽出を 用いた 画像ノイズ除去手法

千葉工業大学　先進工学部

19C3020　有働　和矢

指導教員　宮田　高道　教授

提出日　令和 2023 年 1 月 17 日

# 自己注意によるコンテクスト抽出を用いた画像ノイズ除去手法

19C03020　有働　和矢

**概要：**Transfomer に代表される自己注意機構を利用することが画像処理の分野で拡がっている. 画像ノイズ除去タスクにおいて長距離間の依存関係を認識できる自己注意機構を応用することにより, 画像のコンテクストを把握しながらノイズ除去を行うことが可能となるため, 高いノイズ除去性能が実現できることが明らかになっている. 一方でコンテクスト抽出と（抽出したコンテクストに基づく）ノイズ除去の二つのタスクを異なる二種類の CNN で行うことで効率的にノイズ除去を行える GTCNN と呼ばれる手法が提案されており, 画像ノイズ除去タスクにおいてコンテクストの把握とノイズ除去処理のすべてを自己注意機構のみで行うことは CNN と同様に効率的でない可能性がある. そこで本研究では GTCNN のコンテクストを抽出する CNN を自己注意に置き換えた SAGTCNN を提案する. 実験結果より, GTCNN と比較して高いノイズ除去性能またはより効率的なノイズ除去性能を得ることは叶わなかった. 本実験では画像サイズの小さい領域でしか自己注意を用いなかったり, ViT 等の画像処理に自己注意を活用する手法で行われている位置埋め込みを行なっていないなどの理由から自己注意が本来の性能を発揮できなかった可能性が考えられる.

# 目次

# 第 1 章

# 序論

## 1.1 背景

画像ノイズ除去はノイズが発生した画像から, ノイズを含まない原画像を推定することを目的としており, 画像処理分野における基礎的なタスクであることから多様な研究が行われている. 近年, 自動運転や監視カメラ, 医療などの分野において画像認識技術が幅広く活用されており, 今後もその対象は拡がっていくと予想されている前述の画像ノイズ除去は, このような画像認識の認識精度を向上させるための前処理としても重要な役割を担っている.

現在, 画像のノイズ除去においては高いノイズ除去性能を持つ深層学習を用いた手法が主流となっており, 特に深層畳み込みニューラルネットワーク（CNN）が高い性能を示すことが明らかになっている. CNN の欠点としては, 畳み込み層の受容野が比較的狭く, 長距離の画素間依存性を認識できないことが挙げられる. このため, 既存の CNN を用いた画像のノイズ除去手法では, 画像の持つコンテクスト（文脈情報）を判断することが困難であり, 除去すべきノイズと残すべき細部の特徴とを区別できないことが知られている.

このような CNN の問題を解決するための新たな手法として, Transfomer[1] に代表される自己注意機構を利用することが画像処理の分野で拡がっている [2, 3]. 自己注意機構は自然言語処理の分野で開発された手法で, 離れた位置に存在する単語間の関係を認識することで, 様々な自然言語処理を高精度で行えることが明らかになっている. 長距離間の依存関係を認識できる自己注意機構をノイズ除去に応用することにより, 画像のコンテクストを把握しながらノイズ除去を行うことが可能となり, 高いノイズ除去性能が実現できることが明らかになっている [3, 4]. 一方で, コンテクス

ト抽出と（抽出したコンテクストに基づく）ノイズ除去の二つのタスクを異なる二種類の CNN で行うことで効率的にノイズ除去を行える GTCNN と呼ばれる手法が提案されており [5], 画像ノイズ除去タスクにおいてコンテクストの把握とノイズ除去処理のすべてを自己注意機構のみで行うことは, CNN と同様に効率的でない可能性がある.

## 1.2　目的

前述のようにノイズ除去の二つのタスクを異なる二種類の CNN で行うことで効率的にノイズ除去を行うことができる. このうちコンテクスト抽出タスクを行う CNN を CNN よりコンテクスト把握に有効であるといわれている自己注意機構に置き換えることで, 既存手法と比較してより高いノイズ除去性能または効率的なノイズ除去性能を得ることを目的とする.

# 第 2 章

# 関連研究

## 2.1 Gated context CNN

Gated context CNN (以下 GTCNN) は, 図 2.1 に示すように入力層, 出力層, 中間層である $L$ 個の GCBR 層（Gated CBR Layer）からなる. GCBR 層はノイズ除去のみを行う CNN (CBR) とコンテクスト抽出のみを行う CNN (GTL) とに分離されている. GTL が抽出したコンテクストは gate 機構を介して CBR に渡されノイズ除去を制御する. このように機能分離を行うことで, 機能分離を行わない既存手法と比較して高いノイズ除去性能を獲得でき, またパラメータ数の削減に成功した.

### 2.1.1 GTL

GTL はマルチスケール構造の U-Net[6] をベースに設計されている. 一般的な U-Net のエンコーダは $S$ 段の層を持ち, その階層が下がるたびに特徴量のスケールを半減させ, チャンネル数を倍増させる. デコーダはエンコーダと同数の $S$ 段の層を持ち, 階層が上がるたびに特徴量のスケールを倍増させ, チャンネル数を半減させる. エンコーダ側では段階的にスケールのサイズが半減するため, CNN のカーネルのサイズが一定でも受容野の広さが倍に増大し, これによって広くコンテクストを把握できるようになる. 一方で, チャンネル数を倍増させることが計算コストの増大にもつながる. GTL では, 階層を下げるときにチャンネル数を増やさないように変更を加えることで計算効率を高めている. 一般的な U-Net を用いたノイズ除去とはことなり, GTCNN の GTL はコンテクストの把握のみに注力できる（ノイズ除去は CBR 層が行っている）ため, チャンネル数を増加させずとも良い性能が得られていると推測さ

れている.

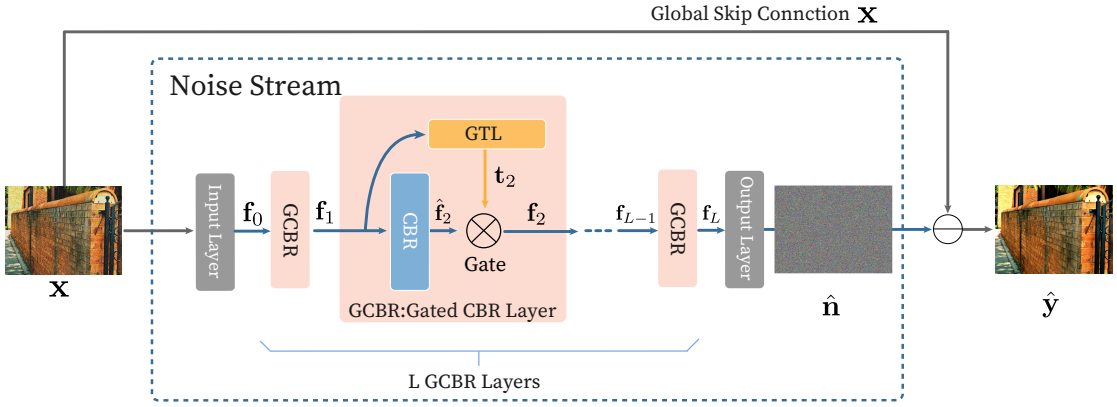


図 2.1　GTCNN のアーキテクチャ全体図. 文献 [5] より引用. (C) 2020 Springer

## 2.2　Vition Transformer

Vision Transformer (ViT) [2] は自然言語処理の分野で用いられていた Transformer を画像処理の分野で活用可能なものにし, また画像分類タスクの性能を向上をさせた手法である. それまで画像処理タスクは CNN に大きく依存しており, また CNN はその性質上広域なコンテクストの取得が難しかった.

ViT は画像を複数のパッチの集合とすることで自己注意主体のネットワーク構造で画像を処理することが可能となった. パッチをトークンに見立て, それに一般的な Transformer 同様に埋め込み処理を行いそれを Transformer の Encoder に入力する. そして出力されたデータを MLP を通しクラスデータを抽出する.

図 2.2 の Transformer Encoder 部を式で記述すると以下のようになる. Transformer Encoder にはパッチに埋め込み処理を施されたデータ $\boldsymbol{z}_0$ が初めに入力される.

$$\boldsymbol{z}'_l = MSA(LN(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1}, \qquad l = 1 \cdots L \tag{2.1}$$

$$\boldsymbol{z}_l = MLP(LN(\boldsymbol{z}'_l)) + \boldsymbol{z}'_l, \qquad l = 1 \cdots L \tag{2.2}$$

$LN$ はレイヤーノルム, $MLP$ は多層パーセプトロン, $MSA$ が自己注意を表している. 自己注意は式 (2.3) のように表される [1]. $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ は入力にそれぞれ異なる埋

め込み処理をしたものである.

$$Attention(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{2.3}$$

式 (2.3) では $\mathbf{Q}$ と $\mathbf{K}$ から得られる全データ間の類似度をもとに, $\mathbf{V}$ を処理することを表しており, このことから, 局所的な処理をする CNN とは異なり, 自己注意は大域的な処理を行っているということである. ViT ではこの特性のため広域なコンテクストの取得を可能とした. しかし自己注意機構には, 入力データ長の二乗に比例し計算量が増えてしまう課題が存在する.

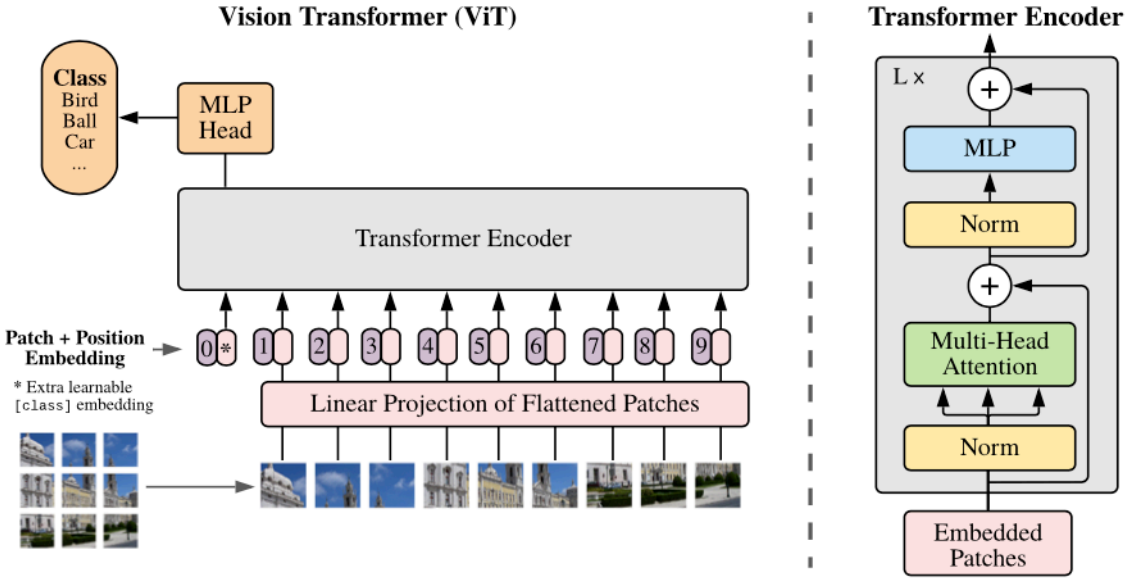


図 2.2　ViT のアーキテクチャ全体図. 文献 [2] より引用. (C) 2021 ICLR

## 2.3　Restormer

Restormer[4] は Transformer を活用した手法で, ノイズ除去をはじめとした画像復元タスクにおいて高い性能を示す. 図 2.3 の通り U-Net をベースに作られており, 各層には Transformer Block が配置されている. また Transfomer Block の数は階層が下がるほど増加する. Transformer Block には Multi-Dconv head transposed attention (MDTA) と Gated-Dconv feed-forward network (GDFN) が含まれており, このうち Transformer を持つ MDTA が画像の文脈情報を抽出している. Restormer は, 画像ノイズ除去を含む様々な画像復元タスクにおいて従来手法を上回る優れた性能を示す一方で, 複雑なネットワーク構造をしていることや Transformer を多用していることから学習および推論時の計算コストが大きいことが課題である. 手法の詳細は文献 [4] を参照されたい.

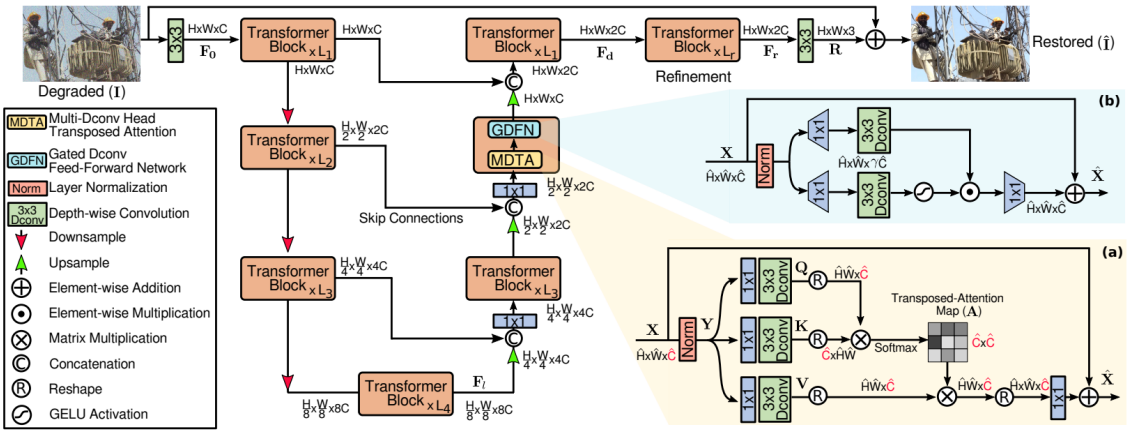


図 2.3　Restomer のアーキテクチャ全体図. 文献 [4] より引用. (C) 2022 IEEE/CVF

# 第 3 章

# 提案手法

既存手法である GTCNN において, コンテクスト抽出を担う GTL は U-Net をベースに設計されている．本論文では GTL の各層に CNN 以上にコンテクスト抽出性能の優れた Transformer を組み込んだ手法, SAGTCNN (Self-Attention GTCNN) を提案する．Transformer を挿入する場所を変更したいくつかのバージョンを提案し, それらの性能を比較する．実験を通して, Transformer のヘッドの数は 1 としている．提案手法は GTL を変更した以外は従来の GTCNN のアーキテクチャを踏襲している．またパラメータサイズの比較は表 3.1 にて示している．

## 3.1 SAGTCNN-S4

SAGTCNN-S4 は図 3.1 のような構造になっている．青い枠の部分が従来手法に変更を加えたところである．従来手法では全ての層が CNN Block で形成されていたが SAGTCNN-S4 では 4 段目を SA Block に変更しているところに特徴がある．さらに, プーリングよるダウンサンプリングを畳み込みで行うようにし, バイリニアでのアップサンプリングを畳み込みで行うように変更した．SA Block は畳み込みを 2 度行う D-conv 層と Attention 層の 2 層からなっている．Transformer は画像のスケールが大きくなるほど計算量が増大するため, スケールの小さい 4 段目で Transformer を用いることで計算量の増大を防いでいる．

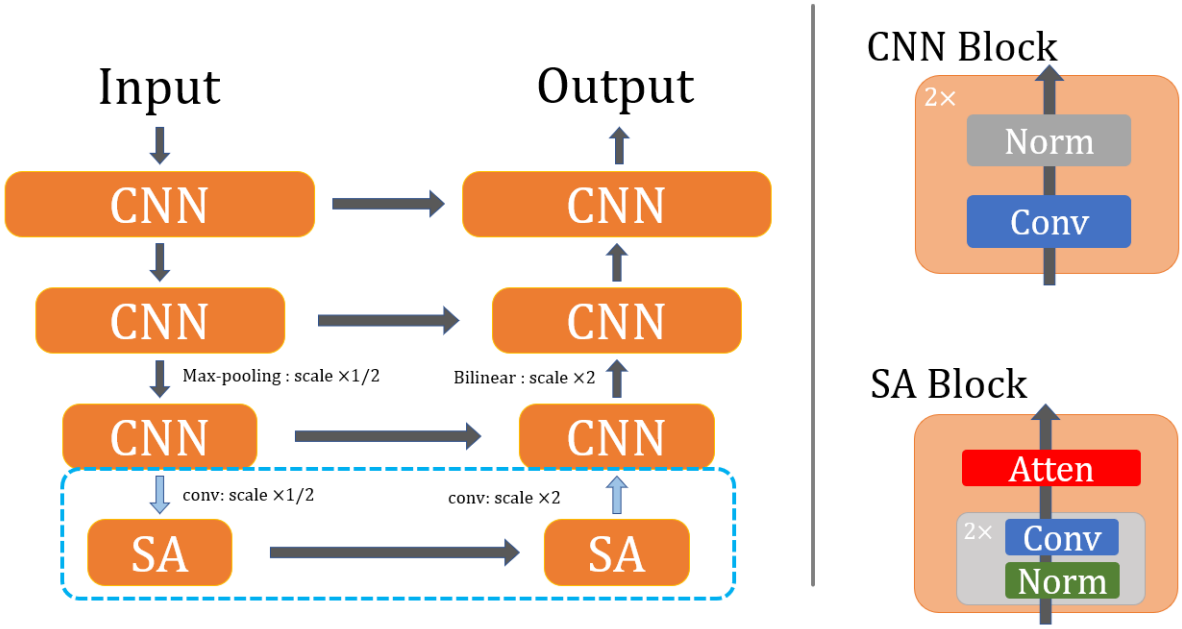


図 3.1　SAGTCNN-S4 の GTL のアーキテクチャ

## 3.2　SAGTCNN-M

SAGTCNN-M では本来なんの処理もなされていなかったミドル層に Transformer を追加した．アーキテクチャは図 3.2 の通りである．ミドル層は他の層と比べてスケールが最も小さく Transformer の処理も最も軽いため, 効率的にノイズ除去性能の向上を行える可能性があり実験を行った．

## 3.3　SAGTCNN-S4M

SAGTCNN-S4 および SAGTCNN-M を合わせた手法で, Transformer を 4 段目とミドル層に挿入した．計算量およびパラメータサイズは他の手法より大きくなる．

表 3.1　パラメータサイズの比較

| GTCNN | SAGTCNN-S4 | SAGTCNN-M | SAGTCNN-S4M |
|---|---|---|---|
| **851k** | 996k | 1,016k | 1,161k |

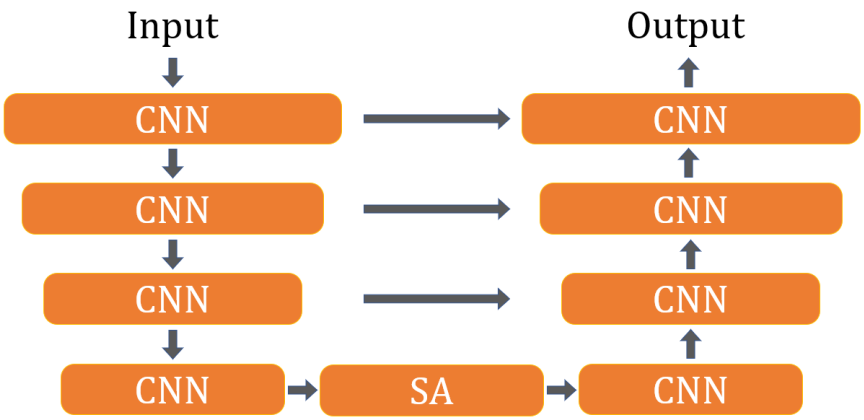


図 3.2　SAGTCNN-M の GTL のアーキテクチャ

# 第 4 章

# 実験

提案手法の有効性を調べるため, 既存手法と提案手法のノイズ除去性能を比べた. その際, 生成ノイズを加算したグレースケール画像に対するノイズ除去性能を比較した.

## 4.1 実験設定

概ね GTCNN の設定に従って実験を行う [5].

### 4.1.1 ネットワーク構成

既存手法 GTCNN の CBR 層の数を 1 つに固定した. 提案手法である自己注意機構を追加した GTCNN(Self-Attention GTCNN: 以後 SAGTCNN) は GTL の 4 層目の CNN を自己注意に置き換えた SAGTCNN-S4, ミドル層に自己注意を追加した SAGTCNN-M, そのどちらの変更を加えた SAGTCNN-S4M の 3 種類のパターンで比較を行った. GTCNN および SAGTCNN の GTL は共に同じ段数 ($S = 4$) で構成した.

### 4.1.2 学習設定

データセットには DIV2K を用いる. DIV2K の全画像から $192 \times 192$ 画素のパッチをストライド 192 画素で切り出したものを学習画像とする. 学習のエポック数は 600 とする.

### 4.1.3　評価手法

既存手法と同様に, ノイズ除去評価手法において一般的に使われるガウシアンノイズを原画像に加算した画像（生成ノイズ画像）から原画像を推定する. ノイズ除去結果の評価は PSNR を用いる. 原画像には Set12, BSD68, Urban100 のデータセットを使用した. 従来手法である GTCNN の評価値は, 論文に記載されている値を引用した.

## 4.2　実験結果

表 4.1 に示すように $\sigma = 50$ での性能比較の結果, 今回 SAGTCNN で試したどのパターンにおいても既存手法の性能を上回ることはなかった. SAGTCNN の中で最も性能の高い SAGTCNN-S4 と GTCNN を $\sigma$ の値を 30, 50, 70 と変更して比較しても, 表 4.2 の通り性能の向上は見られなかった. SAGTCNN-M は Urban100 のデータセットにて他の手法と比べて性能が特に悪い. 出力画像の一枚を比べると他の手法より画像が劣化していることが図 4.1 から分かる.

表 4.1　生成ノイズ $\sigma = 50$ における GTCNN と SAGTCNN の性能比較

| 手法 | Set12 | BSD68 | Urban100 |
|---|---|---|---|
| GTCNN | **27.56** | **26.46** | **26.97** |
| SAGTCNN-S4 | 27.55 | 26.43 | 26.95 |
| SAGTCNN-M | 27.55 | 26.29 | 23.48 |
| SAGTCNN-S4M | 27.47 | 26.38 | 26.84 |

表 4.2　生成ノイズの $\sigma$ の値を変更した時の GTCNN と SAGTCNN の性能比較

| 手法 | Set12 | | | BSD68 | | | Urban100 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 30 | 50 | 70 | 30 | 50 | 70 | 30 | 50 | 70 |
| GTCNN | 29.80 | **27.56** | **26.08** | **28.53** | **26.46** | **25.21** | 29.43 | **26.97** | **25.36** |
| SAGTCNN-S4 | **29.81** | 27.55 | 26.07 | 28.52 | 26.43 | 25.19 | **29.46** | 26.95 | 25.34 |

表 4.3　SAGTCNN の学習画像を増加させた際の変化:Normal は DIV2K のみ. Increase は DIV2K+SSID

| 手法 | Set12 | BSD68 | Urban100 |
|---|---|---|---|
| SAGTCNN-S4 (Normal) | 27.55 | 26.43 | 26.95 |
| SAGTCNN-S4 (Increase) | 27.52 | 26.43 | 26.90 |

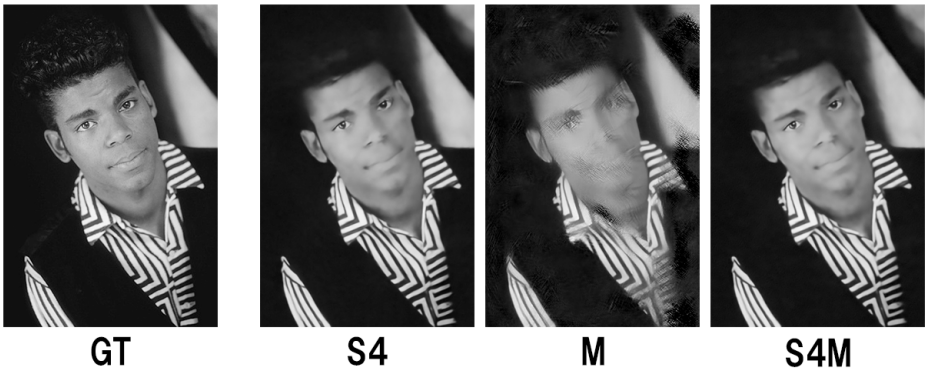


図 4.1　出力画像比較

## 4.3　考察

実験では GTCNN に対し SAGTCNN は生成ノイズ $\sigma = 30$ の場合においてノイズ除去性能をわずかに上回った．ものの, 明確な優位性を示すことはできなかった. 過学習の懸念から学習画像枚数を増やしてみたものの, 表 4.3 の通り, 性能が向上することはなかった．ノイズ除去性能の向上が見られなかった理由はいくつか考えられる．まず一般的に自己注意を用いる手法では大規模なデータセットを用いており, 今回の実験で使用した学習用のデータセットでは学習画像枚数が少ない可能性が考えられる．また自己注意は広域のコンテクストを捉えることを得意とするが, 今回 GTL の 4 段目という画像サイズの小さい領域でしか自己注意を用いておらず, 本来の性能が活かせなかったことが考えられる．最後に画像に自己注意を用いる際に ViT などで行われている位置埋め込みを本実験では行っていないため, コンテクストの抽出が上手く行えなかった可能性が考えられる．

# 第 5 章

# 結論

本論文では, GTCNN のコンテクスト抽出を行う GTL に自己注意機構を組み込んだ手法, SAGTCNN を提案した. 自己注意機構を挿入する場所を変更した SAGTCNN-S4, SAGTCNN-M, SAGTCNN-S4M の 3 種類のパターンで実験を行った. 実験では GTCNN に対し SAGTCNN はいくつかのケースでわずかに上回る性能を示した一方で, 明確な優位性を示すことはできなかった. ノイズ除去性能の向上があまり見られなかった理由はいくつか考察できる. まず一般的に自己注意を用いる手法では大規模なデータセットを用いており, 今回の実験で使用した学習用のデータセットでは学習画像枚数が少ない可能性が考えられる. また, 自己注意は広域のコンテクストを捉えることを得意とするが, 今回 GTL の 4 段目という画像サイズの小さい領域でしか自己注意を用いておらず, 本来の性能が活かせなかったことが考えられる. 最後に画像に自己注意を用いる際に ViT などで行われている位置埋め込みを本実験では行っていないため, コンテクストの抽出が上手く行えなかった可能性が考えられる.

今後の課題としては, 位置埋め込みの実装やリアルノイズでの性能比較, 自己注意を GTL の上層にて使用した際の有効性の検討などが挙げられる.

# 謝辞

本研究を進めるにあたり, 様々な助言および御指導を賜った宮田高道教授に心より感謝いたします. また GTCNN のコードをお借りした今井海斗氏にも深く感謝いたします. 研究室の同輩や先輩の方々にも助けて頂きました感謝いたします.

# 参考文献


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

[2] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit and Neil Houlsby Alexey Dosovitskiy, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[3] 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 *International Conference on Computer Vision*, 2021.

[4] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] Kaito Imai and Takamiti Miyata. Gated texture cnn for efficient and configurable image denoising. In *European Conference on Computer Vision AIM Workshop*, pp. 665–681, 2020.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.