Self-Attention 機構に着目した顕著領域の抽出と Neural Style Transfer への応用

山内　良介†　　青野　雅樹††

† 豊橋技術科学大学 博士前期課程情報・知能工学専攻　〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町字雲雀ヶ丘 1-1
E-mail: †yamauchi@kde.cs.tut.ac.jp, ††aono@tut.jp

あらまし　近年，Prisma や DeepArt などスタイル変換ができるアプリケーションが注目されており，これらのアプリケーションの背景技術である Neural Style Transfer は盛んに研究されている．Neural Style Transfer は，多様なスタイルへの対応やリアルタイム処理など様々な改善がされている中，画風のみにとらわれたスタイル変換が多く，空間情報をうまく反映していない問題点が存在する．一方で Attention 機構は深層学習の研究においてさまざまなタスクに効果的であることが証明されている．画像に対する Attention 機構では，空間情報の重みづけを行うことで，顕著領域を強調することができる．本研究では，Neural Style Transfer と Attention 機構を組み合わせた手法を検討し，空間情報を保持し，画風のみにとらわれないスタイル変換手法を提案する．

キーワード　深層学習，スタイル変換，Attention 機構

1　はじめに

Neural Style Transfer は，図 1 のようにコンテンツ画像とスタイル画像を入力として，コンテンツ画像の物体構造とスタイル画像の画風を併せ持った合成画像を出力するアルゴリズムである．元来，画素値やテクスチャーの類似度計算によって画像合成を行った Image Analogies [1] やテクスチャー合成によって画像合成を行った Style-Transfer via Texture-Synthesis [2] などのアルゴリズムによって画像のスタイル変換が行われていたが，近年では畳み込みニューラルネットワーク (CNN) を利用した画像合成手法によって，クオリティーの高い画像のスタイル変換が可能となった．

現在の Neural Style Transfer は，2016 年の Gatys らの手法 [3] から大きな進歩を遂げ，未学習のスタイル画像に対してのスタイル変換や処理速度が向上してリアルタイムでスタイル変換ができるようになった．しかし，Yao ら [4] はこれまでの Neural Style Transfer 手法の問題点を挙げ，Neural Style Transfer と Self-Attention 機構を組み合わせ，コンテンツ画像の Attention マップの重要度によって，スタイル変換の粒度を変更することで，スタイル画像の画風のみに捉われない滑らかなスタイル変換を可能とした．

Neural Style Transfer の一般的な用途は，ユーザーが撮った写真と絵画を入力することで人工的なアートワークを作成するエンターテインメント性のあるアプリケーションがあげられる．写真編集アプリケーションの Prisma や独自の芸術的な画像を作成できる Web サイトの DeepArt には Neural Style Transfer の技術が用いられている．更に応用先として考えられるのは，クロマキー合成の様な画像合成が考えられる．クロマキー合成は，特定の色成分から画像の一部を透明にし，その部分に別の画像を合成する技術である．この技術は，背景が単色でないときちんと合成できないが，Attention マップの重要でない部分を背景に置き換えることができれば，マスク処理によって背景が単色でなくてもクロマキー合成と同じ効果が得られると考えられる．

本研究では，Yao らの手法の Self-Attention 機構を改良し，高精度な顕著領域の抽出手法を検討し，Neural Style Transfer への応用を目的とする．最終的には，コンテンツ画像の Attention マップから物体部分と背景部分を自動で判別し，物体部分のみスタイル変換をすることを目標とする．

(a) コンテンツ画像　**(b)** スタイル画像　**(c)** 合成画像

図 1: Neural Style Transfer

2　関連研究

2.1　Neural Style Transfer

近年のスタイル変換は CNN を用いた手法が普及しており，Gatys らの手法 [3] は，我々が知る限りスタイル変換に CNN を取り入れた最初の研究である．この手法では，CNN に VGG19 から全結合層を除いたネットワークを用いており，これ以降の手法も VGG19 に似たネットワークを用いることが多くなっている．この手法は，ホワイトノイズをコンテンツ画像とスタイル画像に近づくように最適化するもので，スタイル変換するたびに誤差逆伝播を行うため計算コストがかなり大きくなるとい

う問題がある．

Huang らの手法 [5] は，VGG19 のオートエンコーダモデルであり，デコーダを訓練することで学習に使用していないスタイル画像に対してもスタイル変換することができ，Gatys らの手法 [3] に比べて，遥かに計算コストが小さい．この手法は AdaIn と呼ばれ，コンテンツ画像とスタイル画像を用いてモデルの学習を行い，エンコーダから得られたスタイル特徴の平均と標準偏差を用いて Instance Normalization したものから，スタイル変換を行うようにデコーダを訓練するものである．

Li らの手法 [6] は，スタイル画像の学習が不要になったことが特徴である．この手法も，VGG19 のオートエンコーダのモデルであるが，エンコーダから得られた特徴を WCT(Whitening and Coloring Transforms) レイヤーを通して，デコーダに入力している．WCT レイヤーを通すことで，コンテンツ画像の物体構造の情報は保持したまま，画風の情報を削減することが可能となり，その特徴をスタイル画像のパラメータで変換することで，シンプルな手法にも関わらずスタイル変換することが可能となった．

Chen らの手法 [7] は Style Swap と呼ばれ，コンテンツ画像の中間特徴量量をスタイル画像の中間特徴量で置換したことが特徴である．エンコーダから得られたコンテンツ画像およびスタイル画像の中間特徴を，パッチに切り出してパッチ間のコサイン類似度が最大のもので置換し，置換された特徴をデコーダに入力することでスタイル変換を行う．

Sheng らの手法 [8] は Avatar-Net と呼ばれ，Li らの手法 [6] と Chen らの手法 [7] を組み合わせた手法であり，WCT レイヤーの途中で Style Swap を入れることでスタイル変換を行うものである．

2.2 Attention 機構

深層学習の研究において Attention 機構が注目されている．Attention 機構は様々なニューラルネットワークに取り入れることができ，画像分類，VQA，機械翻訳などさまざまなタスクに効果的であることが証明されている．

画像に対しての Attention は 2 種類あり，Channel Attention と Spatial Attention の 2 つに分けられる．Channel Attention はチャンネル方向の Attention であり，意味があるものに焦点を当てる．Hu らの手法である Squeeze-and-Excitation [9] や Wang らの手法である ECA-Net [10] などが Channel Attention を適用している．これらの手法の多くには，Global Average Pooling や Global Max Pooling などが用いられる．

Spatial Attention は空間方向の Attention であり，どこに物体があるかについて焦点を当てている．Zhang らの手法である Self-Attention [11] や Choe らの手法である ADL [12] などが Spatial Attention を適用している．これらの手法の多くには，Channelwise Average Pooling や Pointwise Convolution どが用いられる．

また，Channel Attention と Spatial Attention を組み合わせた Attention 機構も存在する．Woo らの手法である CBAM [13] は，Channel Attention 後に Spatial Attention をシリアルに接続しているのに対して，Hu らの手法である CSAR block [14] は，Channel Attention と Spatial Attention をパラレルに接続している．

2.3 AAMS

Yao らの手法 [4] である AAMS(Attention-aware Multi-stroke Style Transfer) は，Neural Style Transfer と Self-Attention 機構を組み合わせたものである．Yao らは，Style Swap [7] では，スタイル画像の低レベルのスタイルパターンのみが反映されることや，AdaIn [5] では，スタイル依存の訓練によるテクスチャーパターンの偏りが発生してしまうことや，Avatar-Net [8] では，コンテンツ画像の物体形状の歪みをうまく処理できないなこれまでの Neural Style Tranfser の様々な問題点を挙げた．それらの問題点を解決するために Yao らの手法 [4] では，Self-Attention 機構を導入することで，コンテンツ画像の Attention マップの重要度によって，スタイル変換の画風の粒度を調整して，空間情報を壊さずスタイル画像の画風にとらわれないスタイル変換を行うことを可能とした．

AAMS は VGG19 のオートエンコーダ間に 3 つの機構を持っている．1 つ目は，Self-Attention 機構である．この機構ではコンテンツ画像の特徴量 f_c の Self-Attention マップを出力している．Self-Attention 機構は SA-GAN [11] と同じものを使用している．

2 つ目は，Multi-scale Style Swap 機構である．この機構では，コンテンツ画像の空間情報情報に対応して，粗いタッチから滑らかなタッチといった様々な画風を持つ特徴量を作成している．

3 つ目は，Multi-stroke Fusion 機構である．この機構では，コンテンツ画像の Attention マップの重要度によって画風の粒度を決定している．

3 提 案 手 法

AAMS [4] で用いられる Self-Attention 機構では画像中の小さな物体が顕著物体である場合，その物体全体に Attention がかかり，画像中の大きな物体が顕著物体である場合は，その物体中の細かな領域に Attention がかかるという特徴がある．この手法では顕著物体が小さい場合，顕著物体の物体構造が保持されるが，顕著物体が大きい場合，細かな領域のみの物体構造が保持され，物体全体の構造が保持されないという問題点がある．また，Attention の強度の問題点もある．既存手法の Self-Attention では，コンテンツ画像の特徴マップの特に大きな値のみを強調するため，顕著物体全体が強調されず，全体的に弱い Attention がかかってしまう．そのため，顕著物体と背景部分の分離ができず，空間情報が考慮されない．

そこで，本研究では Attention 機構の改善を行うことで，顕著物体全体に強い Attention をかけて，顕著物体と背景部分の分離を行う．

一般的に畳み込み層が浅ければ，物体の色やテクスチャが抽出されるのに対して，畳み込み層が深ければ，物体の形状が抽

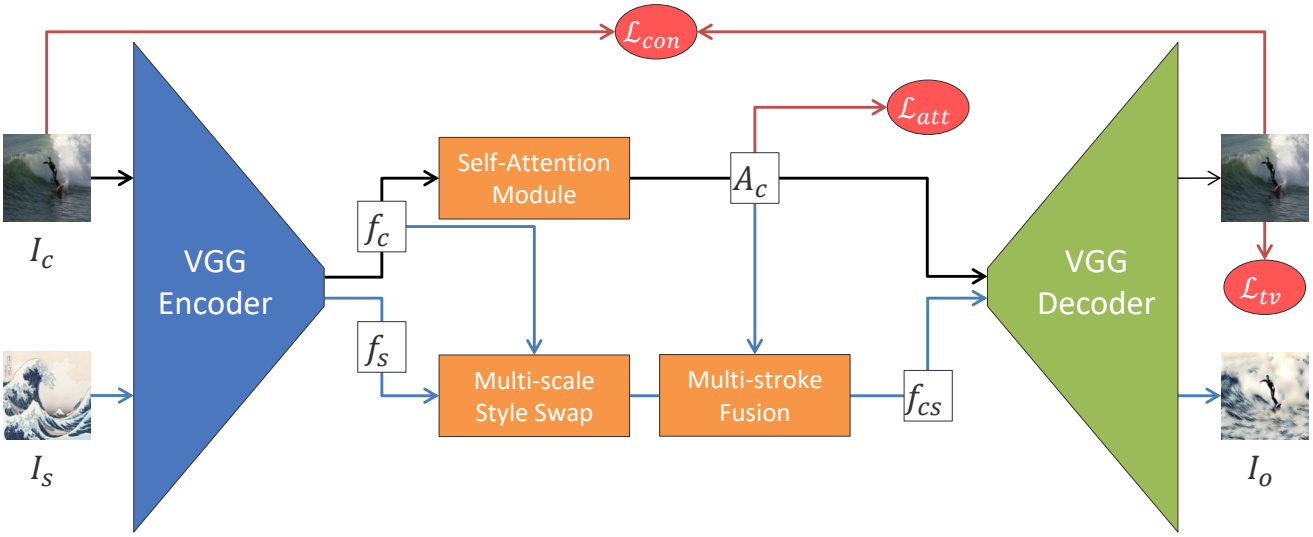

図 2: 提案手法ネットワーク図

出される．既存手法では，エンコーダの畳み込み層は VGG19 の比較的真ん中の層である 9 層目を使用しているが，畳み込み層の層数によって，スタイル変換後の画像の差異を調査するために，オートエンコーダの層数の変更を行った．

具体的には Self-Attention 部分を CBAM [13] に変更し，畳み込みの層数も 5 層 (conv3_1 まで)，9 層 (conv4_1 まで)，13 層 (conv5_1 まで) の 3 種類のパターンを用意した．本提案のネットワーク図を図 2 に示す．

3.1 Self-Attention 機構

本研究では Self-Attention 機構に CBAM(Convolutional Block Attention Module) を採用した．CBAM の概要図を図 3 に示す．CBAM は Channel Attention 機構と Spatial Attention 機構をシリアルに接続した Attention 機構である．

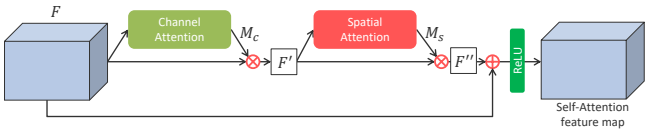

図 3: CBAM の概要図

Channel Attention 機構では特徴マップの意味があるものに焦点を当てる働きがある．Channel Attention の概要図を図 4 に示す．Channel Attention では特徴マップ F を Global Average Pooling と Global Max Pooling を行い，1 つの隠れ層を有する MLP に入力して，出力された特徴ベクトルを足し合わせ，Sigmoid を通してチャンネル方向の特徴マップ M_c を作成する．

Spatial Attention では，物体がどの位置にあるかについて焦点を当てる働きがある．Spatial Attention の概要図を図 5 に示す．Spatial Attention では，特徴マップ F' を Channelwise

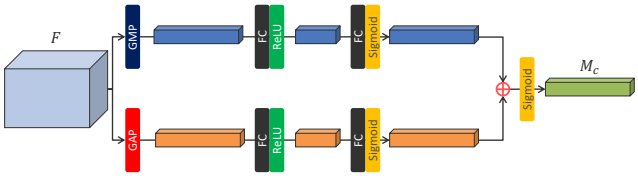

図 4: Channel Attention

Average Pooling と Channelwise Max Pooling を行い，フィルタサイズ 7×7 の Pointwise Convolution の後に Sigmoid を通して空間方向の Attention マップ M_s を作成する．

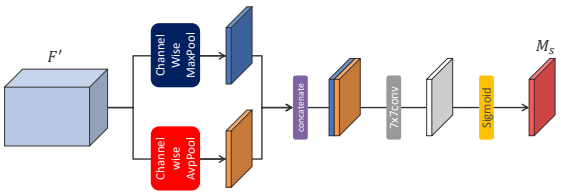

図 5: Spatial Attention

したがってコンテンツ画像の特徴マップ f_c の Self-Attention マップは式 (1)〜(3) のように計算される．ここで，式 (1) の M_c は Channel Attention 機構，式 (2) の M_s は Spatial Attention 機構である．

$$F' = M_c(f_c) \otimes f_c \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

$$A_c = f_c \oplus F'' \tag{3}$$

3.2 損失関数

損失関数は，既存手法 [4] と同様であり，式 (4) に示すような 3 つの損失関数の合計である．ここで，λ_{con}，λ_{att}，λ_{tv} は損失のバランスをとるための重みである．

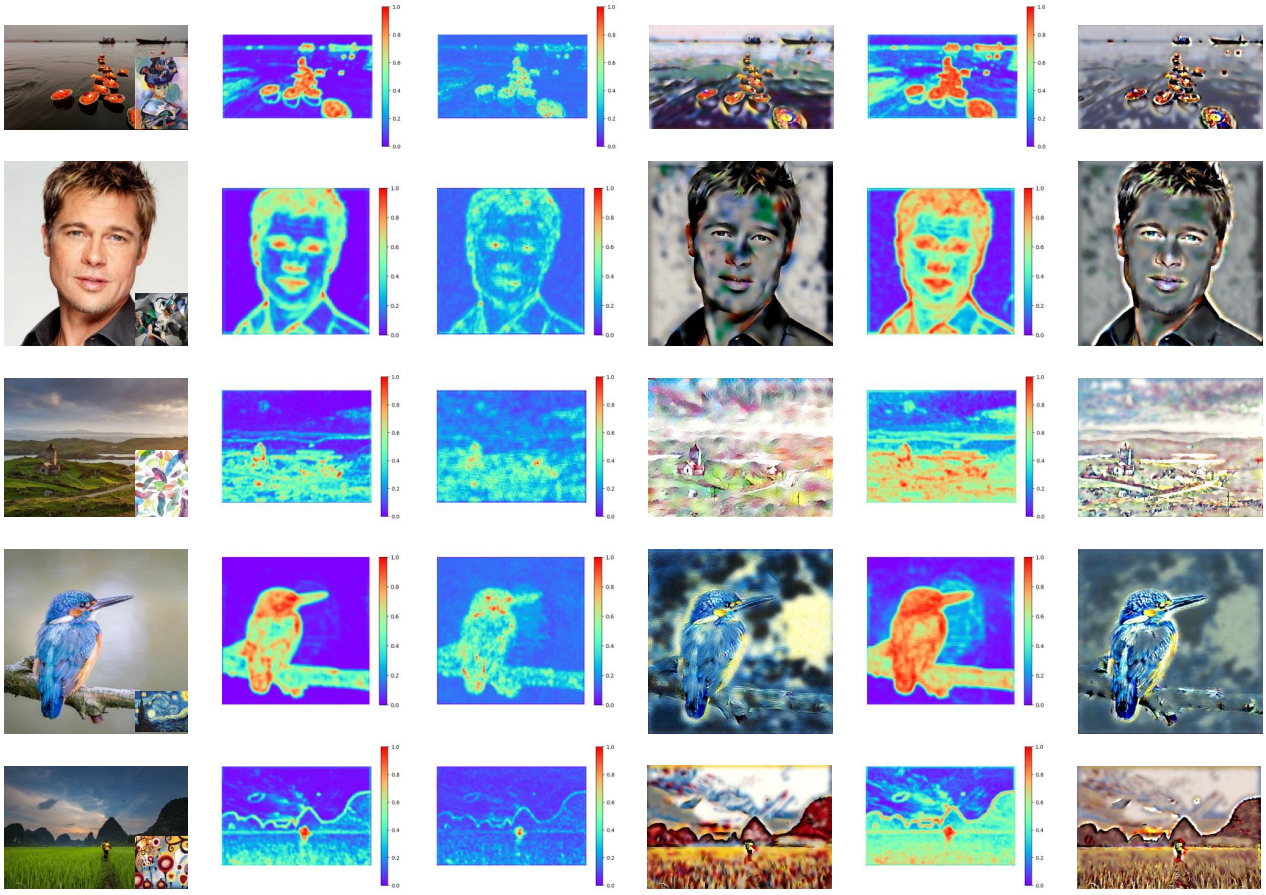

図 6: 実験結果 (conv3_1 までのモデル)

$$\mathcal{L} = \lambda_{con}\mathcal{L}_{con} + \lambda_{att}\mathcal{L}_{att} + \lambda_{tv}\mathcal{L}_{tv} \tag{4}$$

式 (4) 右辺の 1 項目は，コンテンツの損失である．これは，式 (5) で表され，コンテンツ画像と再構成された画像の二乗誤差と VGG19 のオートエンコーダのエンコーダとデコーダの l 番目の層の活性化関数の二乗誤差の和である．ここで，λ_p は 2 つの損失のバランスをとるための重みである．

$$\mathcal{L}_{con} = \sum_{l \in l_c} ||\phi_l(\hat{x}) - \phi_l(x)||_2^2 + \lambda_p ||\hat{x} - x||_2^2 \tag{5}$$

式 (4) 右辺の 2 項目は，Attention の損失である．これは式 (6) で表され，Self-Attention マップの絶対値である．この損失を導入することで，画像全体ではなく顕著な領域により Attention がかかるようにしている．

$$\mathcal{L}_{att} = ||A_x||_1 \tag{6}$$

式 (4) 右辺の 3 項目は，全変動損失である [15]．これは，再構成された画像のピクセルを操作して，再構成された画像を空間的に連続させることで，過度の画素化を回避する．つまり，再構成された画像を滑らかにする正則化の役割を果たしている．

4 実　　験

約 80,000 枚の画像で構成される MS-COCO データセット [16] を使用して，VGG19 オートエンコーダの訓練を行った．エンコーダには，ImageNet [17] で事前訓練された VGG19 のレイヤーが含まれており，デコーダはエンコーダに対して対称の構造である．VGG エンコーダの層数を変更したモデルを 3 種類用意した．最適化には Adam を用いて，イテレーション数は 10,000，バッチサイズは 8 とした．その他のパラメータも Yao ら [4] と同様のもので実験を行った．

4.1 実 験 結 果

図 6，7，8 はそれぞれ VGG19 エンコーダの畳み込み層が上から 5 層 (conv3_1)，9 層 (conv4_1)，13 層 (conv5_1) までのモデルでの実験結果である．

図 6 の ReLU3_1 の特徴マップを見てみると，畳み込み層が比較的浅い層であるため，灯篭 (1 行目の画像) や風景画の建物 (3 行目の画像) そして農家 (5 行目の画像) の様な，小さい顕著物体の値が大きい．一方で，顔 (2 行目の画像) や鳥 (4 行目の画像) の様に大きな顕著物体に対しては目，鼻，口，羽など顕著物体中の細かい領域の値が大きい．図 8 の ReLU5_1 の特徴

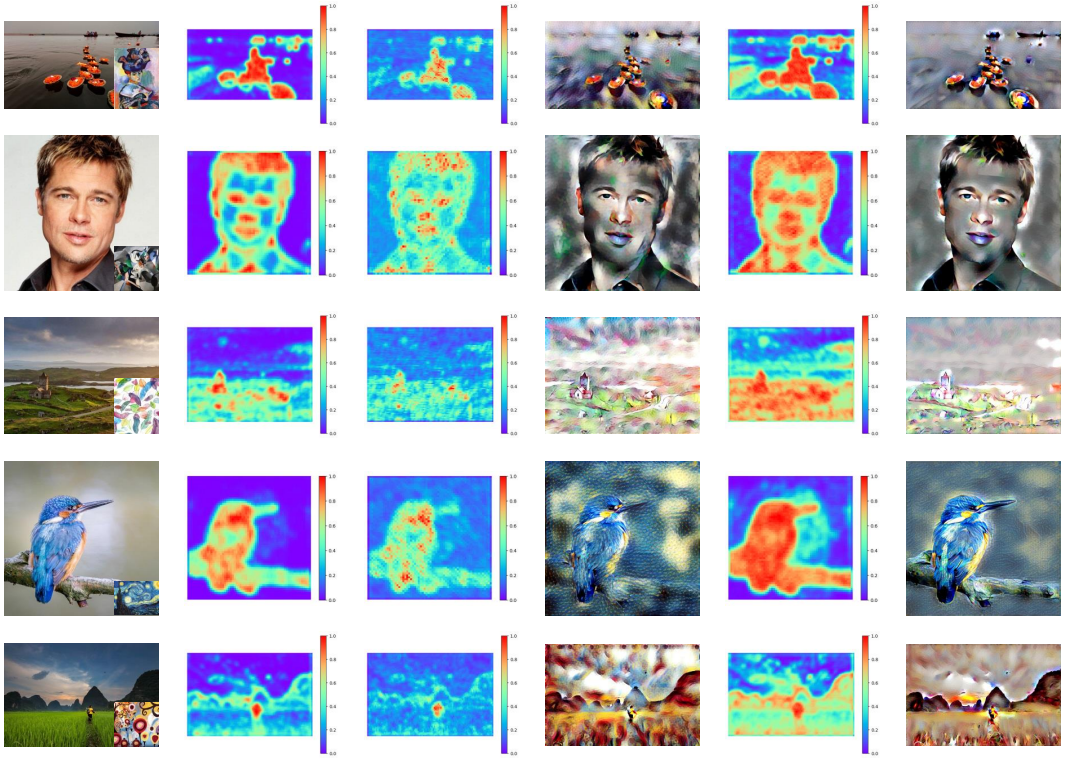

図 **7:** 実験結果 (conv4_1 までのモデル)

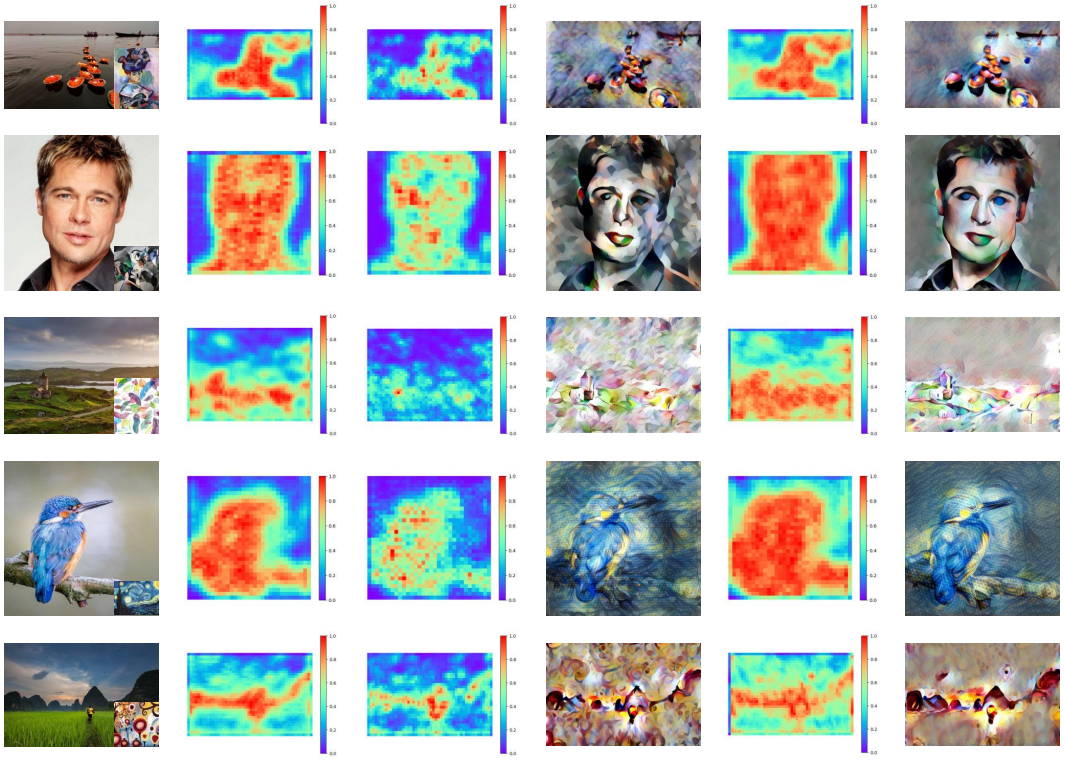

図 **8:** 実験結果 (conv5_1 までのモデル)

マップを見てみると，畳み込み層が比較的深い層であるため，顕著領域全体の値が大きい．

表 1 はそれぞれのモデルのパラメータ数である．既存手法の Self-Attention [11] は，空間 2 次元とチャンネル 1 次元の合計 3 次元をまとめて Attention しているためパラメータ数が多くなっており，計算コストもかかってしまう．一方で，CBAM は空間とチャンネルを分離して Attention を行うため，パラメータ数が少ないことがわかる．

表 1: パラメータ数

	AAMS	Ours
conv3_1	145,579,187	145,432,793
conv4_1	155,167,859	154,579,865
conv5_1	183,485,555	182,897,561

4.2 考　　察

図 6 の風景画 (3 行目の画像) のスタイル変換後の画像を比較すると，既存手法に比べて提案手法では小屋や道路など細部まで物体構造が保持できていることが分かる．これは，Attention によって細部の情報が強調されているためであると考えられる．さらに，提案手法では奥の山の輪郭部分にも Attention がかかっているため，提案手法では山が連なっている様子が反映されており，奥行の空間情報も保持していることが分かる．また，鳥の画像 (4 行目の画像) のスタイル変換後の画像を比較すると，既存手法に比べて提案手法では鳥と背景部分でタッチが異なっていることが分かる．これは，鳥と木の枝に強い Attention がかかり，背景部分には弱い Attention がかかっているため．顕著領域と背景部分の分離がうまくできており，空間情報を保持していることが分かる．

図 7 の灯篭 (1 行目の画像) のスタイル変換後の画像を比較すると，既存手法では小さい船の物体構造が歪んでいるのに対して，提案手法では小さい船の物体構造が保持されていることが分かる．また，鳥の画像 (4 行目の画像) のスタイル変換後の画像を比較すると，既存手法に比べて木の枝の構造情報が保持されていることが分かる．これらも，提案手法の Attention によって物体構造を強調する Attention がかかっているため構造情報が保持されていると考えられる．

図 8 の顔の画像 (2 行目の画像) のスタイル変換後の画像を比較すると，既存手法では右目や左頬の物体構造が歪んでいるのに対して，提案手法では物体構造が保持されていることが分かる．さらに，conv3_1 および conv4_1 の Attention マップと比べて肌全体にの Attention がかかっていることが分かる．そのため，既存手法では肌の部分に背景と同じようなタッチがされているが，提案手法では肌と背景のタッチが異なっていることが分かる．これらも，提案手法の Attention によって物体構造が強調されたためと考えられる．

結果全体を通して，既存手法の Self-Attention では，入力の特徴マップの特に大きい値に対して Attention がかかっている．一方で，提案手法は入力の特徴マップの大きな値付近に Attention がかかっている．畳み込み層が比較的浅い場合は，物体の細部であるエッジ部分の情報が Attention として強調されるためスタイル変換後の画像に対して，輪郭部分や鳥の羽といった細かい部分は反映されるがスタイル画像の画風は色だけの簡単なものが反映されるトレードオフの関係があると考えれらる．畳み込み層が比較的深い場合は，物体構造が Attention によって更に強調されるため，顔や鳥といった画像中の大きな物体と背景部分がうまく分離でき，空間情報に対応しスタイル変換ができたと考えられる．

5 おわりに

本研究では，Neural Style Transfer と Attention 機構を組み合わせ，空間情報を保持し画風のみに捉われないスタイル手法を提案した．Attention 機構には CBAM を用いてコンテンツ画像の顕著な物体全体に Attention をかけて空間情報の重みづけを行うことで，物体構造の保持と顕著な物体と背景の分離ができ，空間情報を考慮したスタイル変換が可能となり，提案手法が有効であることが示された．

今後の課題としては，顕著な物体はコンテンツ画像のスタイルそのままで，背景部分のみスタイル変換を行うようなクロマキー合成の提案や，Self-Attention ではなく，物体検出ネットワークのように，顕著領域の正解データを用意した訓練手法のアルゴリズムや，新たな Attention 機構の考案など考えられる．

謝　　辞

本研究の一部は，科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した．

文　　献

[1] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 327–340, New York, NY, USA, 2001. ACM.

[2] Michael Elad and Peyman Milanfar. Style transfer via texture synthesis. *Trans. Img. Proc.*, 26(5):2338–2351, May 2017.

[3] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017.

[7] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *CoRR*, abs/1612.04337, 2016.

[8] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. *CoRR*, abs/1910.03151, 2019.

[11] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[12] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[14] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *CoRR*, abs/1809.11130, 2018.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.