Vol.2022-CG-185 No.11
2022/3/12

絵画に対する感情表現を考慮したスタイル変換

刀根 千尋[1,a)]　髙橋 遼[1,b)]　栗山 繁[1,2,c)]

概要：近年，深層学習によるアート画像を対象としたスタイル変換手法が盛んに研究されており，自然画像の高度な意味情報を取得しながらその固有表現を保ちつつ任意の画風にスタイル変換する手法や，自然言語を入力としてアート画像を生成・変換する手法等が提案されている．アート作品はしばしば鑑賞者が引き起こす感情的効果を意図して制作されるため，自然画像と比較して感情表現を多く含む．したがって，絵画を対象とした既存のスタイル変換手法では画風の変換を主な目的としているが，感情を入力としてアート画像をスタイル変換する手法の開拓が望まれる．また機械学習において，画像から得られる視覚的作用とその感情的効果への理解を深めることは，生成モデルにおける表現力を高めることにも大きく貢献することが期待される．そこで本研究では，絵画の感情表現に着目し，感情識別モデルを学習して VQGAN による画像の変換を最適化するスタイル変換手法を提案する．ArtEmis データセットを用いた変換結果により，色彩だけでなく質感による感情表現を考慮した変換画像が得られたことを確認し，ユーザ評価によりその有効性と今後の課題を調査した．

キーワード：画像スタイル変換，感情表現と推定モデル，深層学習

Style Transfer of Painting by Emotional Expressions

Abstract: Many style transfer methods have recently been proposed using deep learning to transform natural scene images into arbitrary styles while maintaining intrinsic expressions. In addition, some methods estimate high-level semantic information for transferring art images via natural language. Since artworks are often created to produce emotional effects, they contain more emotional expressions than natural images. However, existing style transformations mainly focus on only painting styles, and it is desirable to develop a method using emotions as input. In this study, we propose a style transfer method using VQGAN by learning an emotion discrimination model, focusing on the emotional expression of paintings. Experiments with the ArtEmis dataset show that the images can be stylized, reflecting the emotional expression based on colors and textures.

Keywords: Image style transfer, Emotional expression and estimation model, Deep learning

1. はじめに

近年，アート画像を生成する深層学習によるスタイル変換手法が盛んに研究されている．Neural Style Transfer [1] では，CNN（Convolutional Neural Network）を用いて，入力自然画像の高度な意味情報を取得し，画像の固有表現を保ちつつ任意の画風にスタイル変換する．また，VQGAN-CLIP[*1]や StyleGAN-NADA [2] では，自然言語を入力としてアート画像を生成・変換する．

アート作品は，しばしば鑑賞者が引き起こす感情的効果を意図して制作されるため，自然画像と比較して感情表現を多く含む．しかし，既存のスタイル変換手法では画風の変換を主な目的としており，感情を入力としてアート画像のスタイルを変換する研究は少ない．機械学習において，画像から得られる視覚的作用とその感情的効果への理解を深めることは，生成モデルにおける表現力を高めることに大きく貢献すると考えられる．そこで，本研究では絵画の感情表現に着目した，感情を入力とするスタイル変換を提

1 豊橋技術科学大学
TUT, Toyohashi, Aichi 441–8580, Japan
2 株式会社サイバーエージェント
CyberAgent inc., Shibuya, Tokyo 150–0042, Japan
a) tone.chihiro.il@tut.jp
b) haruka@val.cs.tut.ac.jp
c) sk@tut.jp

*1 https://github.com/nerdyrodent/VQGAN-CLIP

© 2022 Information Processing Society of Japan

案する．提案手法では，ArtEmis [3] データセットを用いて感情識別モデルを学習し，VQGAN の生成する画像を最適化する．

2. 関連研究

2.1 Neural Style Transfer

Neural Style Transfer [1] は，スタイル画像の画風をコンテンツ画像に転送するスタイル変換手法である．この手法では，コンテンツ画像の固有表現を失うことなく画風を転送するため，事前学習済みの VGG19 [4] にコンテンツ画像を入力し，その固有表現となる特徴量を中間層から取得する．そして，元画像と変換画像を比較するコンテンツ損失を用いることにより，浅い層では形状や輪郭などの大域的な特徴を取得し，深い層では局所的な特徴を取得できることを示した．

2.2 VQGAN-CLIP

VQGAN-CLIP は，生成器として VQGAN [5] と，画像と文章の大規模分類モデルである CLIP [6] を組み合わせ，文章から画像を生成する．CLIP は画像と文章間の類似度の計算が可能であり，入力文と生成画像の類似度を用いて損失関数を定義する．入力文は CLIP の文章エンコーダで特徴 Z_t に変換され，VQGAN によって生成された画像は CLIP の画像エンコーダで特徴 Z_i に変換される．そして，Z_t と Z_i の類似度により生成器 VQGAN を最適化することで，より文意に適合した画像を生成する．

この手法では，VQGAN のエンコーダで生成した入力画像とその変換後の特徴量を比較し変換を制限することにより，スタイル変換が実現できる．例えば，”ピカソ”というテキスト表現を入力とし，生成画像をピカソの画風に変換できる．しかし，入力文章に感情表現を入力すると指示した感情として適切な表現に変換されず，固有表現を著しく損なうために良い生成結果が得られない問題点がある．

2.3 ArtEmis

一般的にアート画像は，客観的な内容と，鑑賞者にどのような感情的効果を与えることを意図しているかという間接的な意味情報を持つ．そのため，客観的な内容のみでラベリングが可能な自然画像とは異なり，画像の内容を単純に説明することが難しい．絵画に着目することで画像が持つ感情的な意味への理解を深め，自然画像に対するより豊かな理解にも応用することを目的とした，ArtEmis データセットが公開されている．このデータセットは，WikiArt[*2]に収録された約８万点の絵画作品に対して複数の注釈者によって８つの感情ラベルが付与されている：

正の感情: Amusement, Awe, Contentment, Excitement

[*2] https://www.wikiart.org

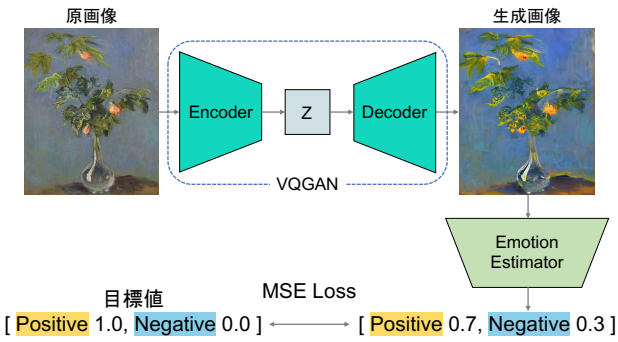

図 1 提案手法によるスタイル変換の構成．VQGAN で生成された画像の感情を Emotion Estimator で推定した値を用いてデコーダを更新し，スタイル変換を最適化する．

Fig. 1 The architecture of proposed method. The decoder is updated to optimize the style transformation using the estimation of an Emotion Estimator for the images generated by VQGAN.

負の感情: Fear, Sadness, Anger

どちらにも属さない感情: Something else

さらに，絵画から与えられた感情的効果についての説明を含むキャプションも付与されている．

絵画から与えられる感情的効果は注釈者間で個人差があるので，絵画によっては感情ラベルが一意に定まらない場合がある．しかしながら，ArtEmis では感情的な画像理解は，画像の内容だけでなく鑑賞者に与える影響を統合した分析が必要だとされている．したがって，鑑賞者の主観性を重視し，データセットは１つの絵画が感情ラベルを複数持つ形式となっている．

ArtEmis ではさらにキャプションに感情表現が多用され，自由や愛など抽象的な概念への言及や，多種多様な比喩表現などの画像内には含まれていない情報を含んでいる．そのため，主観的内容と豊富な表現を含む点において，一般的に広く利用されている COCO [7] データセットとは区別される．

3. 提案手法

本研究では，絵画における感情表現は大きく分けて正と負に分類できると仮定し，入力画像の感情表現のスタイルを正負変換するモデルを提案する．提案手法によるスタイル変換の構造は**図 1** に示す様に，VQGAN-CLIP の構造を基にしており，ArtEmis データセットを使用して絵画の感情表現を学習した感情推定モデルを適用することで絵画の感情を変換する．また，VGG19 を使用したコンテンツ損失により，元画像の固有表現の欠損を防ぐ機能を導入する．

3.1 感情推定モデルの学習

生成画像の感情を評価するために，画像の正負感情の分布を推定するモデル（**図 2** 参照）を学習する．

© 2022 Information Processing Society of Japan

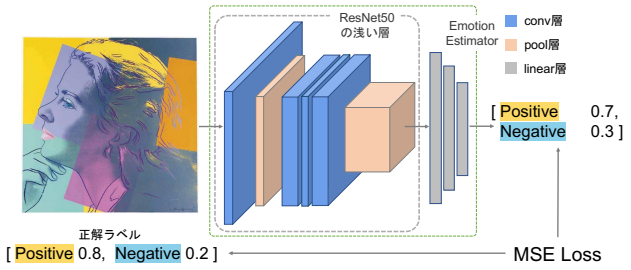

図 2　感情推定モデルの構成．感情ラベルの正負の比率と推定値間での MSE を損失とする．

Fig. 2　Architecture of the emotion estimation model. The loss is given by the MSE between The positive/negative ratio of all emotion labels and its estimation.

本研究では，キャプションは使用せず感情ラベルのみを使用し，ArtEmis で定義されている感情ラベルの正負の分類に基づいて整形したデータを学習した．ArtEmis では，1 つの画像に対して 5~10 人の注釈者が異なる感情ラベルを与えているので，1 つの絵画に対して与えられた全ての感情ラベルの正負の比率を正解ラベルに設定した．ネットワークは ResNet50 [8] の浅い層を使用し，推定値と正解ラベルの MSE をスタイル（感情）の損失関数として学習した．

3.2　スタイル変換の最適化

VQGAN を生成器として用いて得られた画像を感情推定モデルに入力し，感情の正負の分布を推定する．ここでは，感情スタイル損失 $\mathcal{L}_{\text{emotion}}$ とコンテンツ損失 $\mathcal{L}_{\text{content}}$ の重み付き和

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{emotion}} + \beta\mathcal{L}_{\text{content}} \tag{1}$$

を最小化するように VQGAN のデコーダを更新することで，画像の固有表現を保持しながら目標の感情に近づくように画像の変換を最適化する．ただし，α と β は各損失に対する重み係数であり，本実験では $\alpha = 1.00$，$\beta = 0.01$ を用いた．また，画像は 512×512 で生成され，最終的な出力は元画像の縦横比を維持するように生成画像を拡大した．

感情スタイル損失 $\mathcal{L}_{\text{emotion}}$ は，前節で述べた感情の分布を推定するモデルを用いて VQGAN で出力された画像の感情の推定値 $E^{'}$ と，その目標値 E との間の MSE として以下の式で与える．

$$\mathcal{L}_{\text{emotion}} = \parallel E - E^{'} \parallel_2 \tag{2}$$

また，スタイル変換では物体の固有表現を保持する必要がある．そこで，コンテンツ損失 $\mathcal{L}_{\text{content}}$ は，入力画像と変換画像を事前学習済みの VGG19 に入力し，特定の畳み込み層から取得した特徴間の MSE として以下の式で与える．

$$\mathcal{L}_{\text{content}} = \parallel F - P \parallel_2 \tag{3}$$

ただし，F と P は各々，VGG19 の第 5 ブロック 1 層目における生成画像と元画像の特徴量を示す．

4.　実験結果と考察

4.1　VQGAN-CLIP による感情の変換

VQGAN-CLIP は画像と文章を入力することによってスタイル変換を行う．そこで，画像としては絵画を，文章としては目標の感情表現を入力して，感情に基づく絵画のスタイル変換を試みた．しかし，指示した感情としての適切な表現には変換されず，固有表現を著しく失ったために良好な生成結果が得られなかった．その原因としては，CLIP が絵画における感情表現に着目したモデルではないことが考えられる．また，CLIP は単純に描かれている客観的な内容と入力文章を比較すると推察されるが，VQGAN-CLIP で定義されているスタイル変換の際の固有表現を保つための損失関数だけでは不十分なことが推察される．

4.2　感情推定モデルのネットワークの考察

感情推定モデルのネットワークには ResNet50 の浅い層のみを部分的に使用した．これは，いくつかのサイズの ResNet で全体を使用して学習した場合と比較して，正負の感情の分類精度が高くなった結果に依拠する．その原因としては，CNN は浅い層では形状や輪郭などの大域的な特徴，深い層ではより局所的な特徴を取得する傾向があり，絵画のような画像では質感など局所的な特徴が自然画像で事前学習された特徴と異なる場合が多いためと推察される．

4.3　提案手法によるスタイル変換結果と考察

絵画の感情表現を変換した結果を図 **3** に示す．全体的に正感情に変換した画像では輝度，彩度が高く，負感情に変換した画像では輝度，彩度が低くなった．また，図 **4** に示されているように，正の感情に変換した画像では明瞭な輪郭で塗りつぶした様に，負の感情に変換した画像では歪んだ輪郭で細かいタッチとなった．これらのことから，感情推定モデルは絵画の感情表現に対し，その色彩だけでなく輪郭やタッチを考慮して学習されていると考られる．しかしながら，複雑な輪郭を含んだり人物の顔が写っている絵画では，変換の際に輪郭や表情が大きく崩れる結果となった．これは，コンテンツ損失を計算する VGG19 から絵画の固有表現を正確に取得できないのが原因と考えられる．

4.4　ユーザ評価

提案手法による感情表現の正負へのスタイル変換を評価するために，24 人の被験者に対してアンケートを実施した．アンケートには，20 枚の元画像と変換画像の計 60 枚の各々から感じられる正負の感情の 5 段階評価と，変換の品質に関する記述式の質問を設けた．なお，感情の 5 段階

© 2022 Information Processing Society of Japan

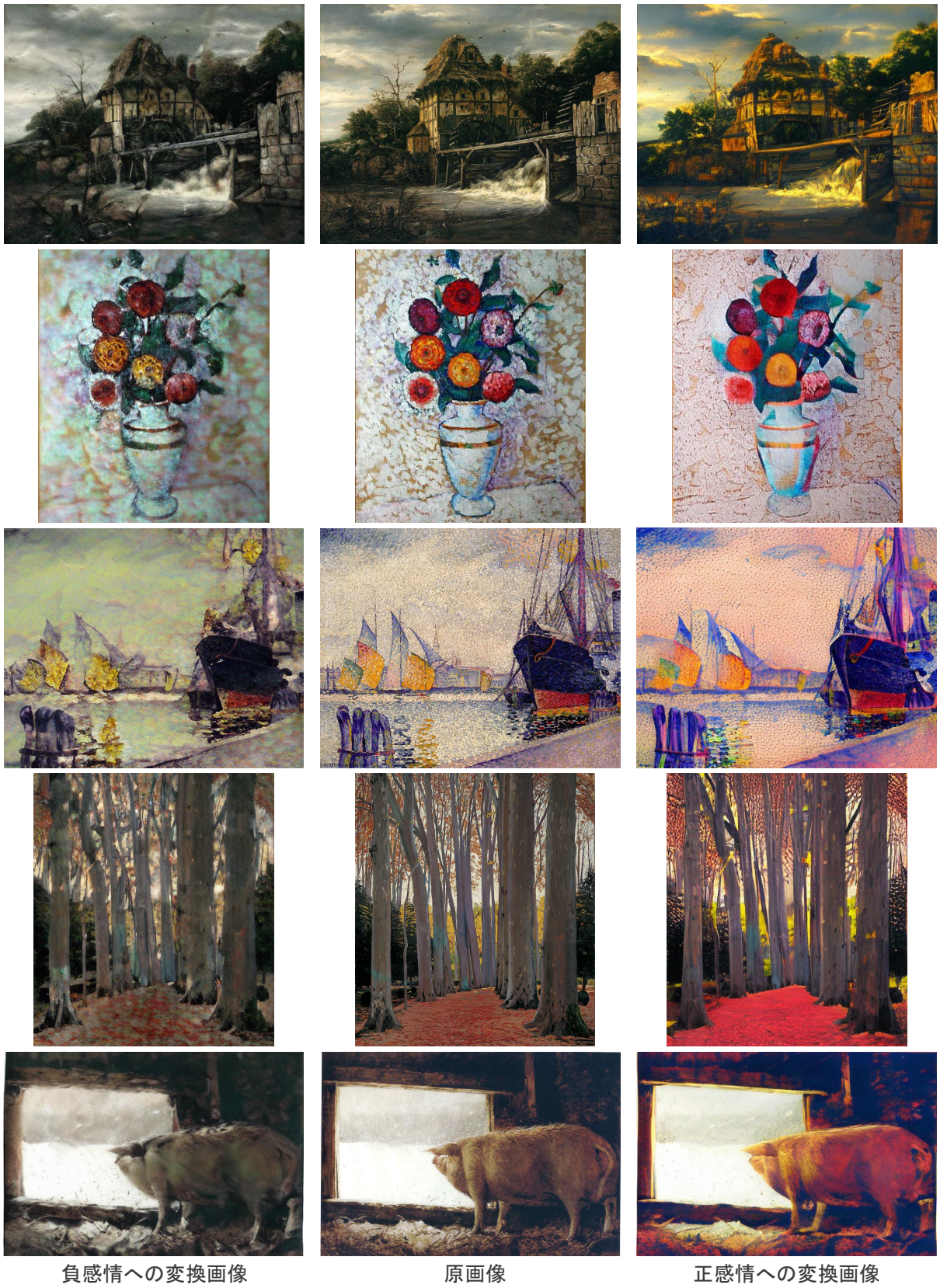

図 3 提案手法によるスタイル変換の結果．左から順に，負の感情への変換，元画像，正の感情への変換を示す．

Fig. 3 Generated results of the style transformation using out method. Images are arranged from the left: transformed to negative emotion, original image, transformed to positive emotion.

© 2022 Information Processing Society of Japan

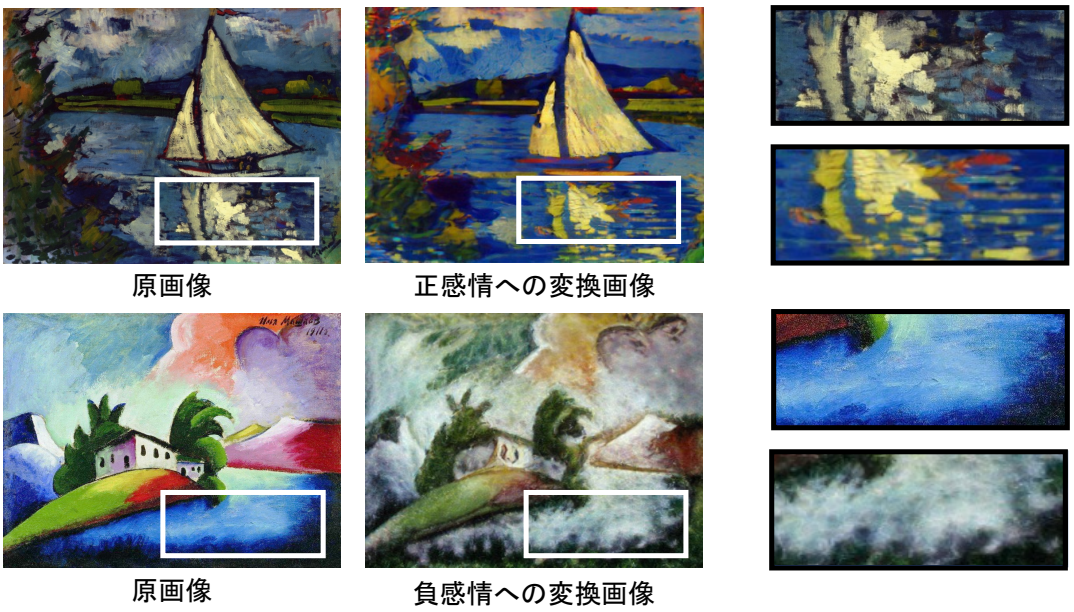

図 4　提案手法による変換前と変換後での輪郭とタッチの変化．白線で囲まれた矩形領域の拡大図を右に示す．

Fig. 4　Changes in outlines and touches before and after the transformation by the proposed method. The enlarged rectangular regions enclosed by white lines are shown on the right.

評価に用いた画像は，元画像と変換画像は並べて表示するのではなく，すべて個別に無作為順に並べて提示した．

元画像と正と負の各感情に変換した画像での感情の 5 段階評価の結果を図 **5** に示す．負感情への変換では，元画像と比較して変換画像が負感情が強いと評価されたが，一方で正感情への変換では，元画像と比較して変化が無いという回答や，負感情が強くなったと回答されたものが一部確認された．また，スタイル変換の品質に関する記述式の質問では，輪郭や表情の歪みが負感情を連想させるという回答が得られた．したがって，図 **6** に示す様な無彩色の絵画で大きな変化を出せなかったという点や，複雑なシーンや人物が写った絵画での歪みが原因だと考えられる．

5.　おわりに

5.1　まとめ

本研究では，ArtEmis データセットを用いて画像の感情表現を推定するモデルを作成し，VQGAN-CLIP の構造を基に，絵画の感情表現を変換する手法を提案した．また，提案手法を用いることで，絵画の色彩を変化させるだけでなく，輪郭やタッチを考慮したスタイル変換が可能となった．スタイル変換の品質を評価するためにアンケートを実施し，負の感情への変換では，元画像と比較して変換画像の方が負感情が強くなったという評価を得た．しかし，複雑なシーンや人物の写った絵画で大きく歪んだことにより，正の感情への変換をしたにもかかわらず負感情を連想させてしまうという問題が発見された．

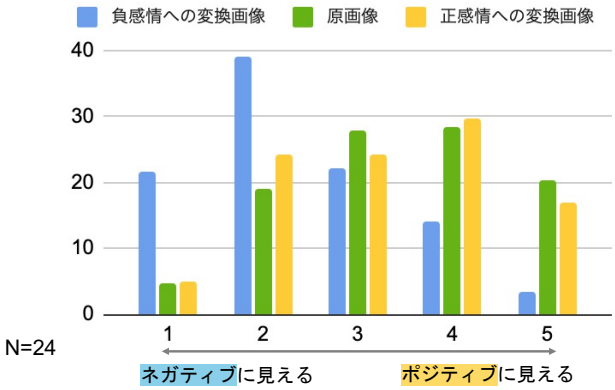

図 5　原画像，変換画像群に対して感じた正負の 5 段階評価の回答の分布．横軸は感情評価の段階，縦軸は選択された割合を表す．

Fig. 5　Distribution of responses to the five positive/negative evaluations for the original and transformed images. The horizontal axis shows the stage of the emotional evaluation and the vertical axis shows the percentage of the selected images.

5.2　今後の展望と課題

本研究では感情推定モデルの学習を正負の 2 感情で行ない，それを使用したスタイル変換の可能性を示せたが，ArtEmis で定義されたすべての感情ラベルを用いて感情推定モデルを学習することで，正負だけでなく，具体的な感情へのスタイル変換も可能になると考えられる．さらには，絵画に対するスタイル変換のみでなく，自然画像を感情表現を介して絵画風に変換する手法や，人物画のみに着

© 2022 Information Processing Society of Japan

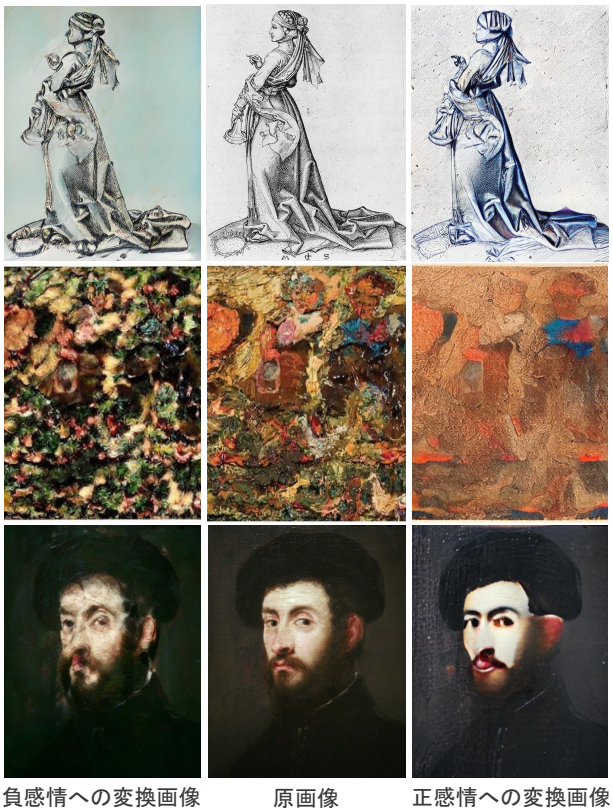

図 6 アンケート画像での失敗例．正の感情へ変換した際に，変化がない，負感情が強くなったと評価された例．

Fig. 6 Examples of failures in the questionnaire images. These examples transformed to positive emotions, but evaluated as having no change, or more negative emotions.

目し顔の特徴点等を併せて使用して人物の表情と感情の対応関係を学習し，表情を考慮してスタイルを変換する手法等が，今後に開発すべき課題として考えられる．

参考文献

[1] Gatys, L. A., Ecker, A. S. and Bethge, M.: Image Style Transfer Using Convolutional Neural Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[2] Gal, R., Patashnik, O., Maron, H., Chechik, G. and Cohen-Or, D.: StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators, *ArXiv*, Vol. abs/2108.00946 (2021).

[3] Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M. and Guibas, L. J.: ArtEmis: Affective Language for Visual Art, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11569–11579 (2021).

[4] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556 (2015).

[5] Esser, P., Rombach, R. and Ommer, B.: Taming Transformers for High-Resolution Image Synthesis, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883 (2021).

[6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *ICML* (2021).

[7] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *ECCV* (2014).

[8] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016).

© 2022 Information Processing Society of Japan