氏　　名	Minh-Duc VO
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2195 号
学位授与の日付	2020 年 9 月 28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Learning-based Image Synthesis by Utilizing Disentangled Feature Representations
論 文 審 査 委 員	主　査　教授　杉本　晃宏
	教授　佐藤　いまり
	助教　池畑　諭
	教授　宮尾　祐介 東京大学　大学院情報理工学系研究科
	教授　石川　博 早稲田大学　理工学術院

(Form 3)

Summary of Doctoral Thesis

Name in full: Minh-Duc VO

Title: Learning-based Image Synthesis by Utilizing Disentangled Feature Representations

Image synthesis is to render a novel image from given inputs. Besides having a wide range of applications, it plays an important role to further step in computer vision research in which many work effort to move from visual-level understanding to reasoning-level understanding. Thanks to deep learning, learning-based image synthesis has been established where a deep network is used to first learn feature space from given inputs and then map the learned feature space to the image domain. Though learning-based models obtain remarkable results, they still limit in generating faithful and realistic images. Their shortcomings come from the fact that the inputs of image synthesis are themselves diverse and multi-meaning, leading multiple descriptive feature representations can be obtained in the feature space along with the depth of the network. Consequently, simply mapping all the feature representations (at the same layer in the network) is unable to elaborate the contribution of each feature representation in the generated image as our expectation. Based on the fact that the more helpful feature space is attained, the more chance to generate better images, faithfully understanding and effectively utilizing the feature representations thus is crucial in robustly elevating the performance of image synthesis.

Needless to say, the feature representations are disentangled and have different roles in generating image. This dissertation, therefore, addresses learning-based image synthesis by an introduction to a novel approach that fully takes into account the feature space. Our approach first selects disentangled feature representations depending on the role to obtain descriptive information. It then combines the disentangled feature representations using an appropriate mapping process to generate images faithfully and realistically. Generally, our approach is potential to deal with a wide range of image synthesis tasks. We therefore apply our proposed way on three interestingly challenging tasks including (i) rendering image contents in different styles (i.e., style transfer), (ii) image manipulation with text and (iii) text-to-image synthesis tasks. The comprehensive experiments manifest that our proposed approach is sufficient and flexible to handle many tasks in image synthesis.

The first task, rendering image contents in different styles, is to render given image contents in given styles. This task requires to preserve contents and to faithfully render of styles. We thus propose a feed-forward network having two distinct streams to learn disentangled content and style features. These features are then combined using our

proposed adaptive feature injection and concatenation which fully take into account contribution of the content and the style features in stylized images. In order to train our proposed network, we employ a loss network, the pre-trained VGG-16, to compute content loss and style loss, both of which are efficiently used for the feature injection as well as the feature concatenation.

The second task, image manipulation with text, on the other hand, is to render foreground (object) given as a text description into a given source image. It requires to disentangle the semantics contained in (source) image and text information and then combine the disentangled semantics to synthesize realistic images. We propose a generative adversarial network (GAN) where the network possesses one generator and a pair of discriminators with different architecture, call *Paired-D GAN*. The generator has encoder-decoder architecture with skip-connections and synthesizes an image matching the given text description while preserving other parts of the source image. Two discriminators, on the other hand, judge foreground and background of the generated image separately to meet an input text description and a source image. We also propose a three-player adversarial learning process to simultaneously and effectively train our *Paired-D GAN*.

The third task, text-to-image synthesis, is to render a novel image that is consistent with a given text description. This task requires not only entity information (i.e., object type, attribute, shape...) but also relation among entities (i.e., position, interaction...). Since the gap between text description and image is large, we thus follow two-step approach where inference of the scene layout as an intermediate representation between text and image is followed by using the layout to generate images. The layout in previous work is constructed through only the comprehensive usage of relation among entities for bounding-boxes' localization, resulting generated images may have poor scene structure as a whole even if each entity is realistically rendered. We step further in predicting visual-relation layout by employing not only all available relations together but also individual relation separately. More precise, we first comprehensively use all available relations together to localize initial bounding-boxes of all the entities. Next, we use individual relation separately to predict from the initial bounding-boxes relation-units for all the relations in the input text. Since each entity may involve in multiple relations, we then unify all the relation-units to produce the visual-relation layout. Finally, our visual-relation layout is conditioned on a stack of three GAN, namely stacking-GAN, to generate images that consistently capture the scene structure.

We evaluate our approach on publicly available dataset. More precise, for rendering image contents in different styles, we use images in the MS-COCO 2014 dataset as our content images, and six famous paintings widely used in style transfer as our style images. For image manipulation with text, we conduct experiments on the Caltech-200 bird dataset and the Oxford-102 flower dataset. For text-to-image synthesis, we verify our method on challenging 2017 COCO-stuff dataset and Visual GENOME dataset. The

intensive and extensive experiments show outperformances of our approach against state-of-the-arts.

Form8・Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　Minh-Duc VO

Title
論文題目　Learning-based Image Synthesis by Utilizing Disentangled Feature Representations

博士論文は、「Learning-based Image Synthesis by Utilizing Disentangled Feature Representations（役割に応じた特徴表現を利用した学習にもとづく画像生成）」と題し、英文で書かれている。画像やテキストを用いて画像を編集・生成するという技術は、知的画像生成をはじめ多くの応用があり脚光を浴びている。本博士論文は、深層学習を使った画像編集・生成をテーマとし、画風変換、テキストによる画像編集、テキストからの画像生成という３つのタスクに取り組んでいる。タスクにおける入力の役割に応じて特徴量を使い分けるという視点に立脚し、３つの異なるタスクに共通のアプローチで手法を開発し、それぞれのタスクにおいて、従来手法に対する優位性を示している。

博士論文は６章で構成されている。第１章では、画像やテキストを用いた画像編集・画像生成の重要性を論じ、博士研究で取り組む問題の意義を議論している。そして、従来法でのアプローチの問題点をあげ、画像編集・画像生成における入力の役割に応じた特徴量使い分けの必要性を論じて博士研究の位置づけを明確にしている。第２章では、それぞれの立場での従来研究の動向と問題点を示している。引き続く３つの章が本論文の主要部分となっている。第３章では、コンテンツ画像を与えられたスタイル画像の画風に変換する問題に対して、コンテンツ画像の特徴、スタイル画像の特徴をその役割に応じて使い分け、コンテンツとスタイルを切り分けて評価することで、バランスの取れた画風変換を実現する手法を提案している。そして、提案手法による画風変換の質を最先端の関連手法と比較し、その優位性を示すとともに、提案手法を構成する各要素が最終結果にどのように貢献しているかを実験的に検証している。第４章では、画像の前景物体を与えられたテキストによって編集する問題（セマンティック画像生成）に対して、前景の特徴量、背景の特徴量を区別して用いるとともに、前景、背景それぞれに特化した識別器を有するモデルを提案している。そして、１つの生成器、２つの識別器の三者による敵対的学習法を提案している。ここでも、提案手法が最先端の関連手法に対して優位であることや導入した学習法の有効性などを実験的に検証している。第５章では、テキストから画像を生成する問題に対して、テキストで与えられた物体間の主語 - 述語 - 目的語という関係を、多者間の包括的な関係と二者間の個別の関係とに区別して用いることで、生成画像中の物体配置がテキストの記述と整合するレイアウト生成手法を提案している。これにより、複数の物体が含まれる複雑なテキスト入力に対しても、そのテキストの記述と整合する物体のレイアウトを有する画像生成に成功している。最先端の関連手法に対する優位性を示すとともに、生成された画像の質、入力テキストとの整合性など提案手法の有効性を多角的に検証している。第６章では、まとめと今後の課題を示している。

出願者による約４５分の発表もこの順で説明が行われ、その後、１５分程度の質疑応答があった。審査委員からは、特徴量の使い分けと提案モデルの関係やセマンティック画像生成における入力テキストの多様性の限界などに質問とコメントが寄せられ、それらに対し出願者は適切に回答した。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究が深層学習に基づく画像編集・生成における異なるタスクに対して共通の方法論で取り組むアプローチは独創的であることが評価されるとともに、研究成果として、査読付き国際学術雑誌論文１件が採択され、また、トップカンファレンスを含む国際会議において査読付き論文４編が採択されていることが確認された。以上の理由により、審査委員会全員一致で、博士論文として十分な水準にある研究であると認め、本論文が博士の学位請求論文として合格であり、学位の授与に値すると結論づけた。