PRMU2022

StyleGANによるCLIP-Guidedな画像形状特徴編集

電気通信大学　大学院　情報学専攻

銭　雨晨　柳井啓司

研究背景

- 深層学習に基づいた画像変換や画像特徴編集の研究が盛んに行われている
- 自然言語が、人と機械のインタフェースになっている

⇒マルチモーダルモデルを利用し、自然言語を用いて画像特徴を編集

Open-Edit (ECCV 2020)

StyleCLIP (ICCV 2021)

Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang and Hongsheng Li. Open-Edit: Open-Domain Image Manipulation with Open-Vocabulary Instructions. ECCV, 2020から図を引用

Or Patashniky, Zongze Wu, Eli Shechtmanx, Daniel Cohen-Ory and Dani Lischinskiz. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arxiv: 2103.17249から図を引用

研究背景: 問題点

- 今までの研究には、画像内のオブジェクトの外観特徴（色やテクスチャなど）に対する編集が多い
- それに対して、オブジェクトの形状特徴（一部のサイズなど）に対する編集の研究が少ない

ManiGAN (CVPR 2020)

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz and Philip HS Torr. Manigan: Text-guided image manipulation. CVPR, 2020

研究目的

- 本研究の目的:
  事前学習済みGANモデルとマルチモーダルモデルを利用し、入力テキストに基づいた画像の形状特徴の編集を実現

A small-wheel car

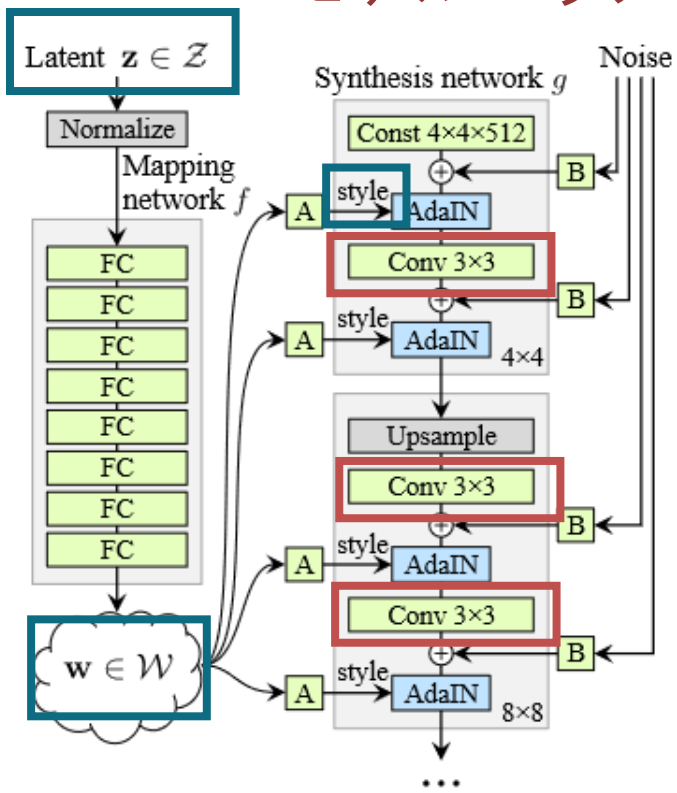
関連研究

事前学習済みGANモデルを利用した画像編集タスク

- モデルパラメータを操作する方法、と潜在空間を操作する方法
- モデルパラメータを操作する方法：NaviGAN
- 潜在空間を操作する方法：Paint by WordやStyleCLIP

潜在空間

モデルパラメータ



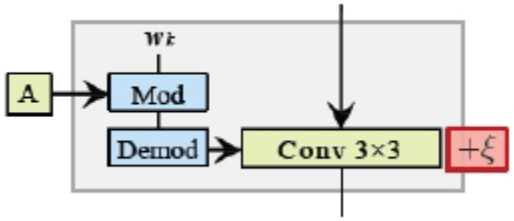
(b) Style-based generator

Anton Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. CVPR, 2021
Karras, Tero, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. CVPR, 2019
より引用

関連研究

事前学習済みGANモデルを利用した画像編集タスク

- NaviGAN:　生成器のパラメータを調整する
- 変換が起きそうな変換方向を探し、それらの方向を人工的にチェックする
- 探した変換の種類には形状特徴の編集に関する変換が多い



Anton Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. CVPR, 2021から図を引用

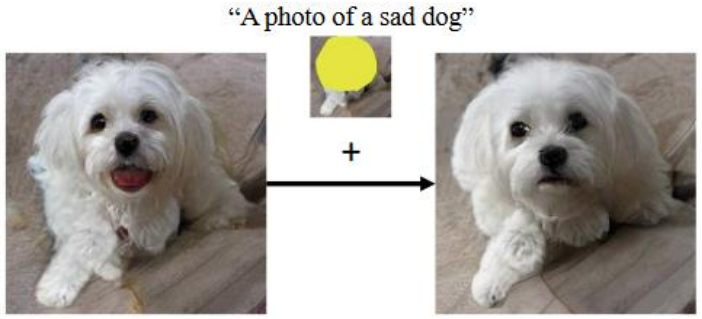
関連研究

テキストを用いた画像編集に関する研究

- 事前学習済みマルチモーダルモデルで画像編集のガイドを提供する方法
- StyleCLIPやPaint by Word: 潜在空間の潜在コードの調整
- 他の部分に影響を及ぼす形状編集はうまく処理できない

- 本研究は、画像形状特徴の編集に重点を置く

Paint by Word



StyleCLIP (ICCV 2021)

Bau David, Andonian Alex, Cui Audrey, Park YeonHwan, Jahanian Ali, Oliva Aude, and Torralba Antonio. Paint by Word. arXiv preprint arXiv:2103.10951
Or Patashniky, Zongze Wu, Eli Shechtmanx, Daniel Cohen-Ory and Dani Lischinskiz. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arxiv: 2103.17249から図を引用

提案手法

- NaviGANの「生成器のパラメータを調整する」アイデアに基づいて、事前学習済みStyleGAN2生成器を利用し、モデルを構築
- CLIPにより、生成画像と入力テキストの類似程度を測定し、最適化のためのロスを提供

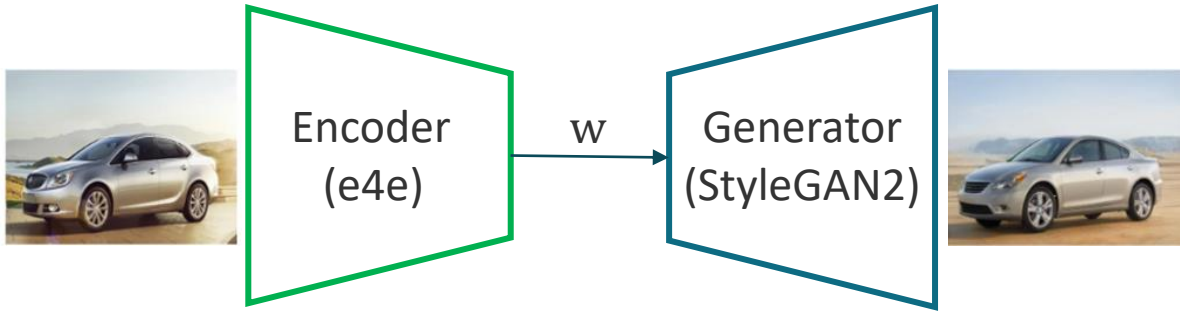
A small-wheel car

手法概要：事前学習済み画像生成モデル

StyleGANとその拡張版StyleGAN2

- 性能が最も良いな画像生成モデル
- StyleGAN2を使用して画像編集をするため、入力画像をStyleGAN2の潜在空間にマッピングできるモジュールencoder4editing (e4e)

Encoder4editing (SIGGRAPH 2021)



Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. CVPR, 2019
Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. CVPR, 2020
Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik and Daniel Cohen-Or. ACM Transactions on Graphics (TOG), 2021から図を引用、変更

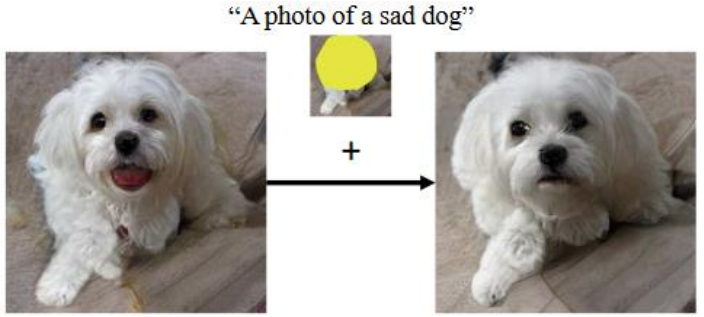
手法概要：マルチモーダルモデル

CLIP (Contrastive Language-Image Pre-Training)

- テキストと画像の類似程度を測定
- 最先端なimage-textマッチング性能で、テキストによる画像変換モデルで利用されている
- 400 millionの画像とテキストのペアデータで学習
- OpenAIで公開された学習済みモデルを利用

StyleCLIP

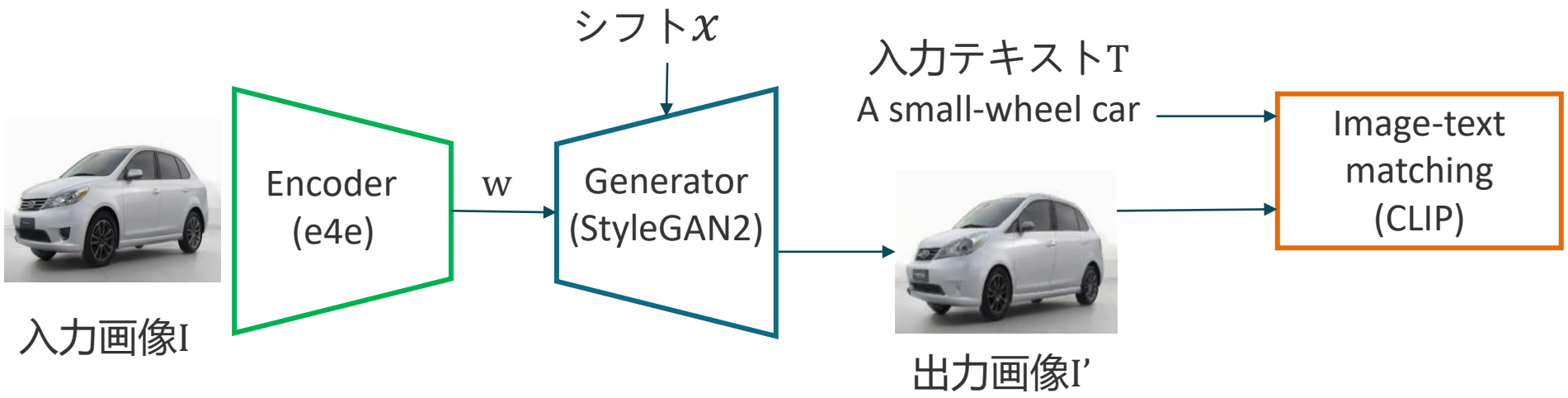
Paint by Word



Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever, Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020
Bau David, Andonian Alex, Cui Audrey, Park YeonHwan, Jahanian Ali, Oliva Aude, and Torralba Antonio. Paint by Word. arXiv preprint arXiv:2103.10951

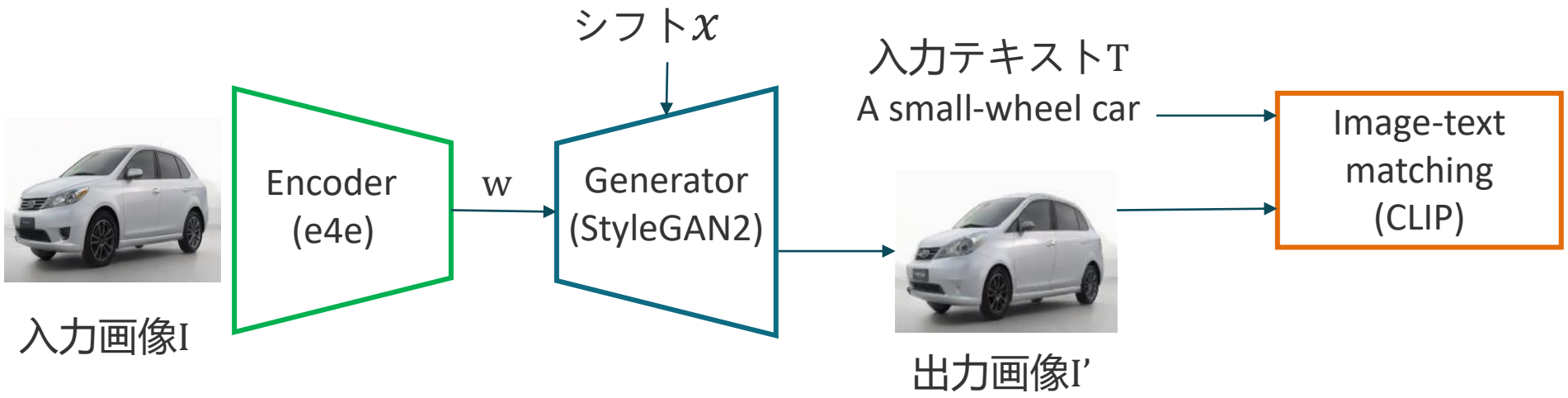
手法詳細

- 入力は入力画像I(または潜在コードw)と入力テキストT、出力画像はI’
- 入力テキストTのセマンティック意味に合うようにIを編集する
- 生成器は事前学習済みのStyleGAN2の生成器を使う



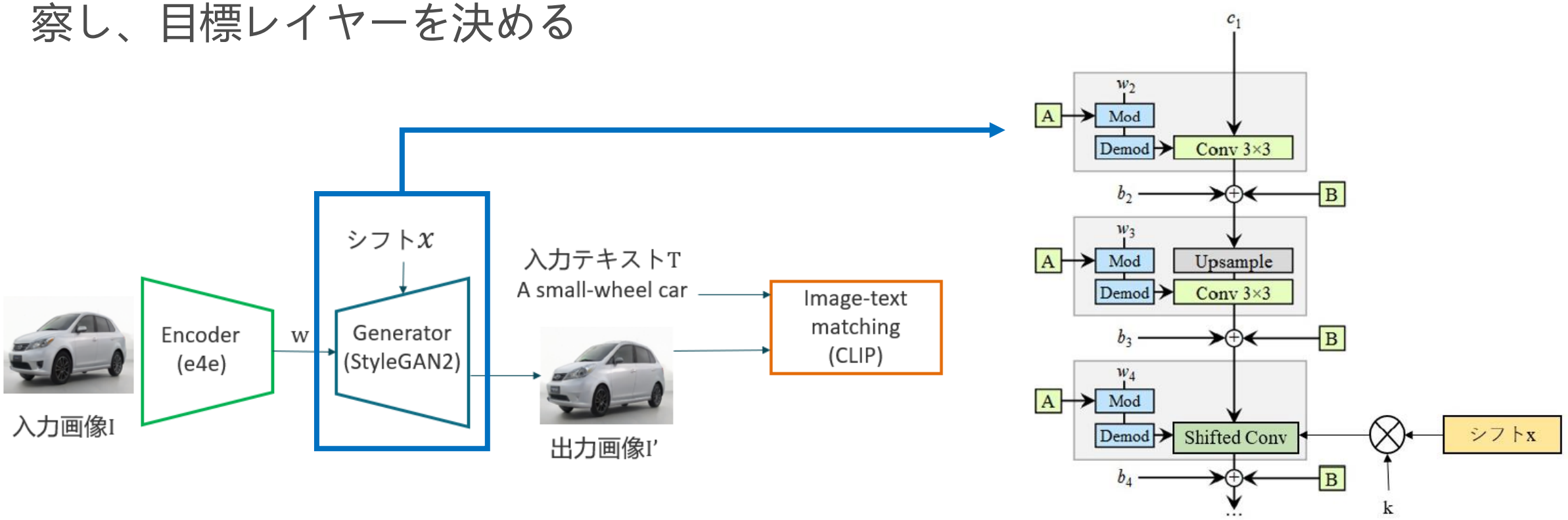
手法詳細

- シフトxで生成器のパラメータを調整し、出力画像I'に影響を与える
- 入力テキストTと出力画像I'を一致させるため、シフトxを最適化によって求める
- 一回の最適化に、一つの入力画像と入力テキストのペアデータを使用し、一つのシフトxを得る



手法詳細

- 一つのシフトxは一つの畳み込みレイヤーのパラメータを調整する
- レイヤーの選定：一定範囲のレイヤーを選び、それらのレイヤーの生成効果を観察し、目標レイヤーを決める



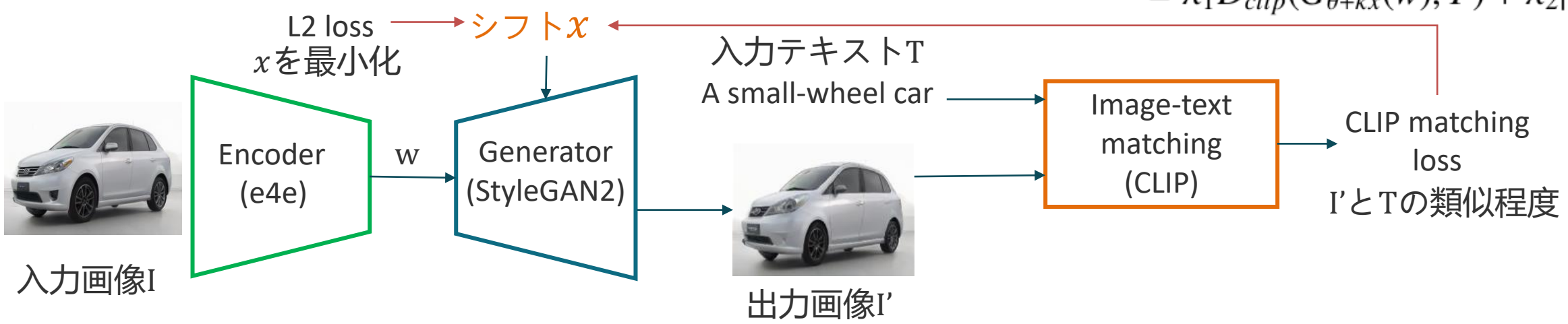
Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. CVPR, 2020
から図を引用、変更

手法詳細

ロスについて

- CLIP matching loss + L2 loss
- xまで逆伝播することでxを最適化する
- xだけに対して最適化し、ほかの部分のパラメータを固定する
- 一回の最適化におよそ1分30秒かかる

$$
\begin{aligned}
L &= \lambda_1 L_{clip} + \lambda_2 L_2 \\
&= \lambda_1 D_{clip}(I', T) + \lambda_2 \|x\|_2 \\
&= \lambda_1 D_{clip}(G_{\theta+kx}(w), T) + \lambda_2 \|x\|_2
\end{aligned}
$$

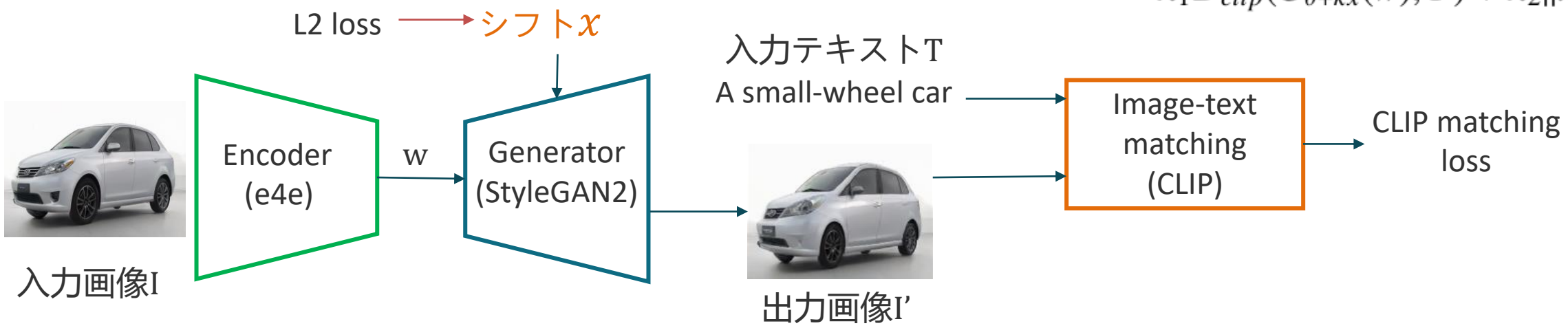


手法詳細

ロスについて

- ハイパーパラメータの選定:
  $\lambda_1 = 10$ で，λ_2の値は変換種類により0.01-0.1 の範囲に最適な数値を選択する

$$
\begin{aligned}
L &= \lambda_1 L_{clip} + \lambda_2 L_2 \\
&= \lambda_1 D_{clip}(I', T) + \lambda_2 \|x\|_2 \\
&= \lambda_1 D_{clip}(G_{\theta+kx}(w), T) + \lambda_2 \|x\|_2
\end{aligned}
$$



実験

StyleGAN2の事前学習に用いたデータセット

- FFHQ 1024x1024
- LSUN Car 512x384
- LSUN Horse 256x256

LSUN Horseで学習したモデル

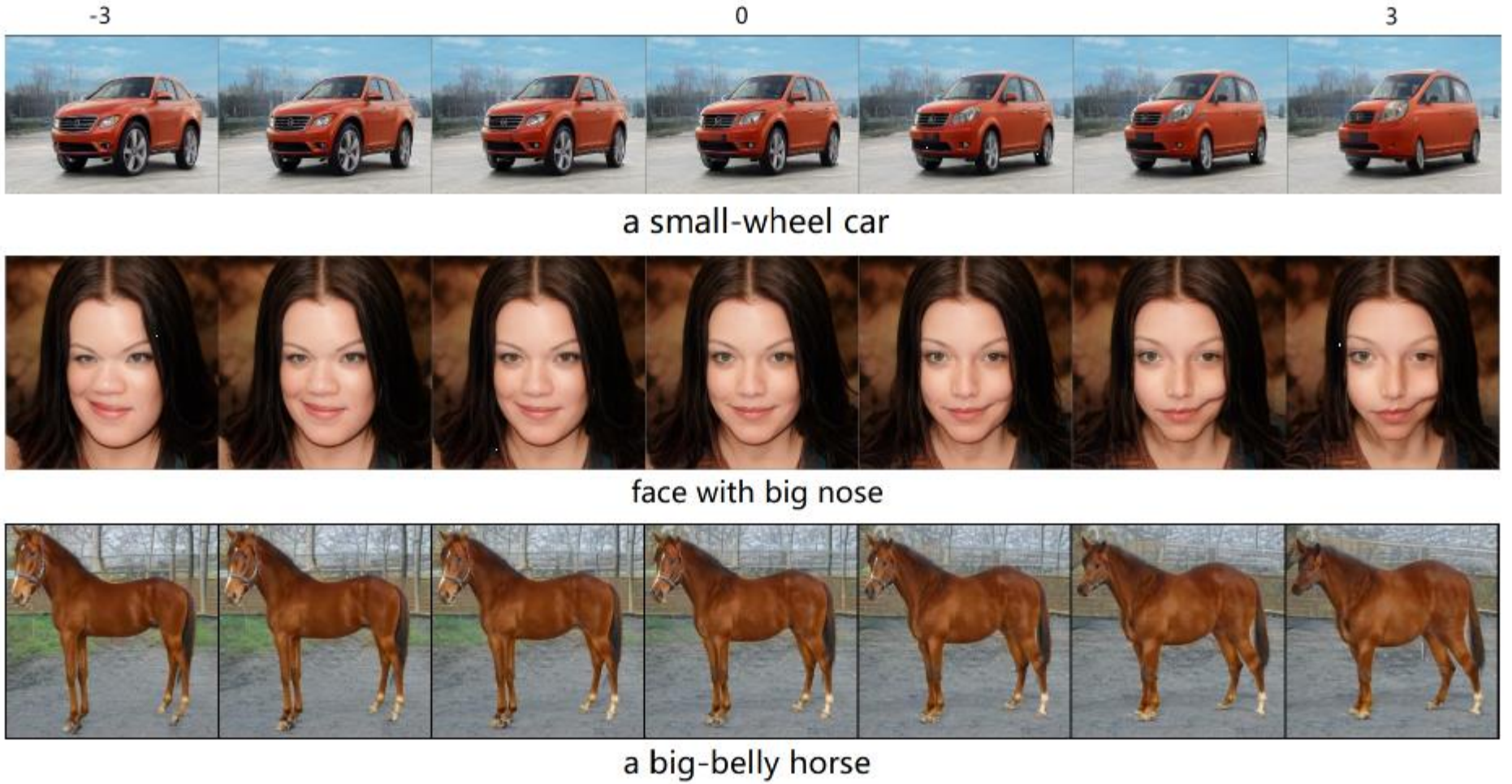
FFHQで学習したモデル

LSUN Carで学習したモデル

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. CVPR, 2020から図を引用

実験結果

- 最適化で得たシフトxに-3~3の係数をかけ、編集された画像を生成



実験結果：比較実験

比較モデル：StyleCLIP

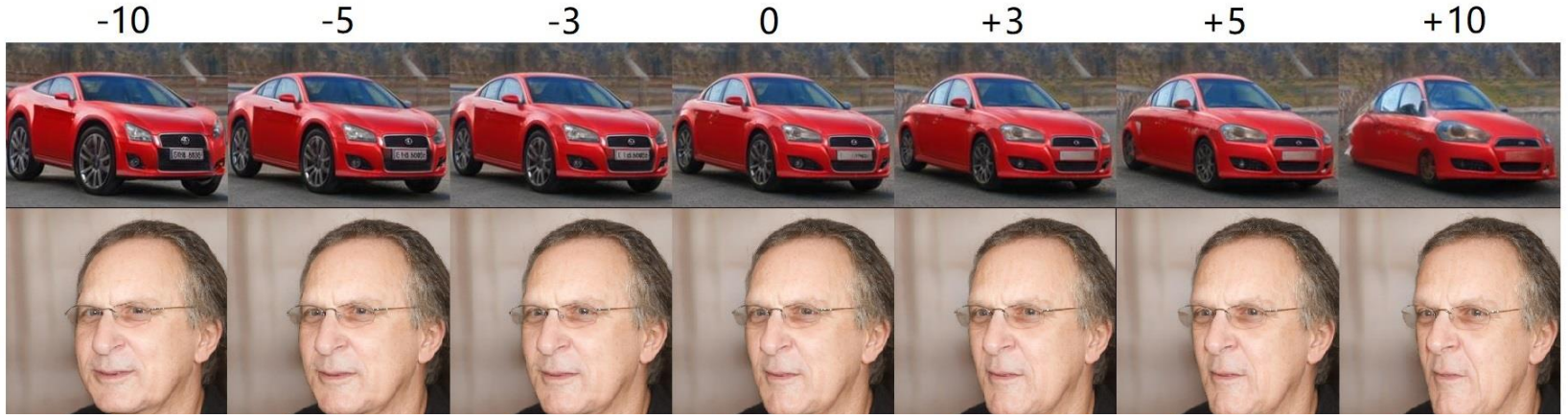
- StyleGAN2 + CLIP　三つの手法を提案
- 潜在空間での潜在コードを調整する
- 手法①②は潜在空間Wを利用し、手法③は潜在空間Sを利用する

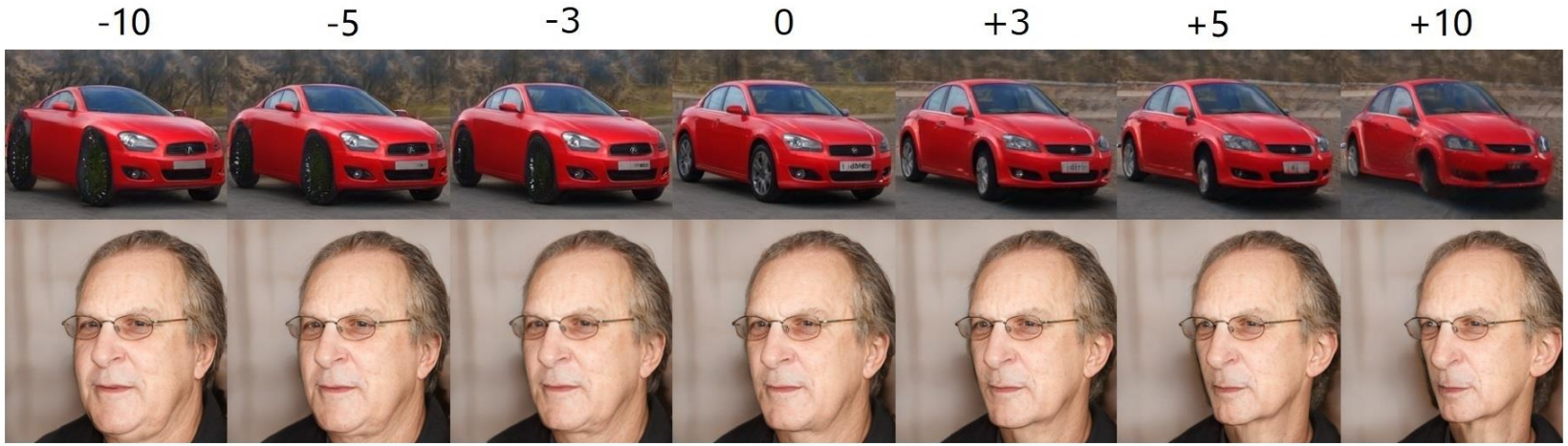
実験結果：比較実験


入力画像
StyleCLIP latent optimization
StyleCLIP latent mapper
StyleCLIP global directions
提案手法
鼻を短くする
face with short nose
顔の幅を大きくする
a wide face
お腹を大きくする
a big-belly horse
細くする
a thin horse

実験結果：定量評価

- 提案手法とStyleCLIPのglobal directionsの方法の比較
- 同じ強さの編集効果において、二つの手法で得られた画像の品質を比較する





実験結果：定量評価

- FIDという指標で評価を行った
- 3000枚の画像を前と同じように変換して、リアル画像とのFID数値を計算

- 多くの場合には、提案手法がより良いFID値を取得した
- StyleCLIPより、提案モデルの変換結果が画像品質を保つことができると証明

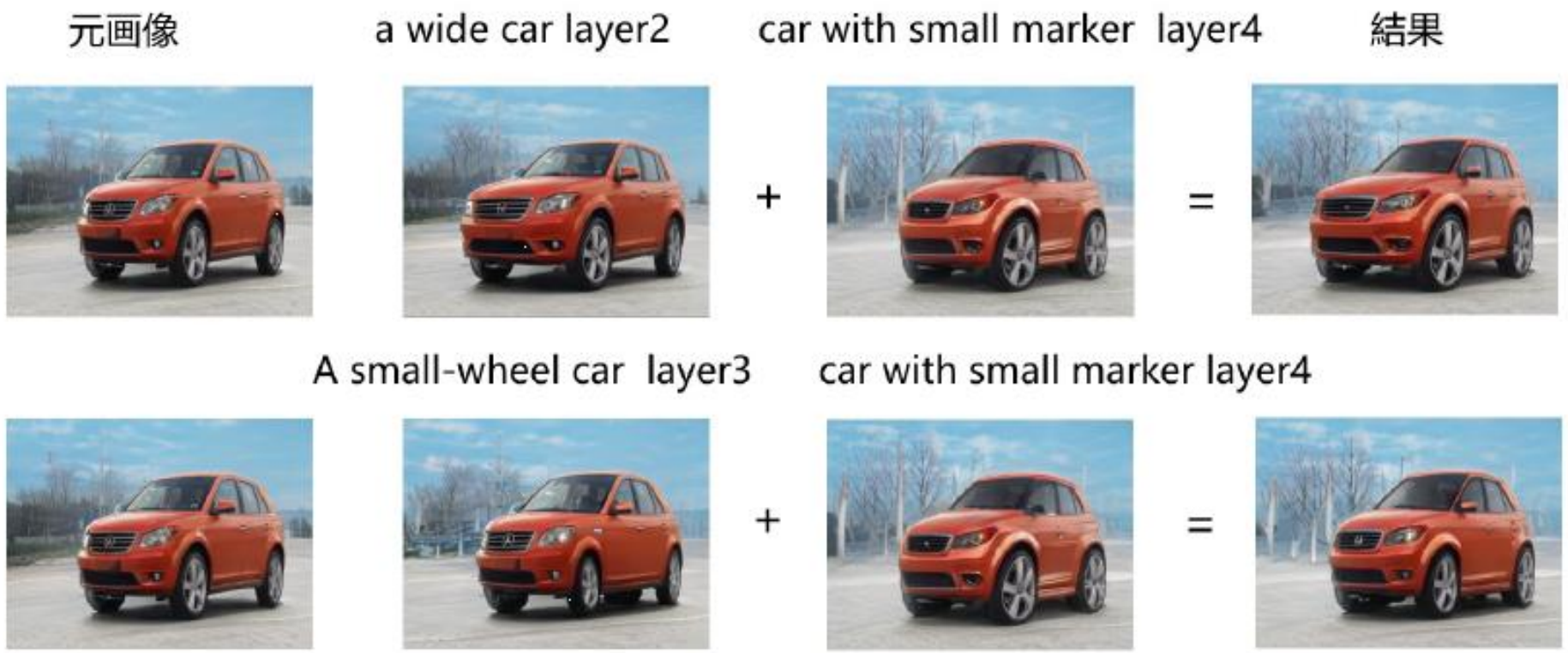
		幅	FID	幅	FID
Wheel Size	StyleCLIP	-3	33.36	+3	18.16
		-5	42.35	+5	23.33
		-10	54.34	+10	67.57
	提案手法	-3	**15.34**	+3	**15.22**
		-5	**17.96**	+5	**21.30**
		-10	**26.27**	+10	**62.39**
	逆マッピング	0	12.54		
Cheek Size	StyleCLIP	-3	28.90	+3	28.40
		-5	**29.16**	+5	29.20
		-10	30.55	+10	30.12
	提案手法	-3	**28.89**	+3	**27.96**
		-5	29.34	+5	**28.41**
		-10	**30.21**	+10	**29.86**
	逆マッピング	0	25.6		

実験結果：複数変換

- 複数の最適化されたシフトを同時にモデルに適用することにより，複数の変換をする画像を得られる
- 同じレイヤーの場合：それらを線形的に加えてモデルに適用する

実験結果：複数変換

- 違うレイヤーの場合：それぞれのシフトを対応するレイヤーに加える
- それぞれのシフトによる変換効果が互いに干渉することがある



失敗の例

- 一部の変換には、画像に不自然な部分が現れることがある
- 複数のレイヤーを用いて変換をすると、このような現象を抑えられる可能性がある

まとめと今後の課題

- 提案手法で、入力テキストに基づいた画像の形状特徴の編集に成功
- 従来の潜在空間を調整をする方法に比べて、形状特徴の編集において、提案手法の定性評価と定量評価での表現が上回る

- 一回の最適化に一つだけの目標特徴を編集できる
  - 複数の特徴を編集できる学習や最適化の方法
- 目標特徴以外の特徴の変化が大きい
  - なるべく他の特徴への影響を抑える