複数の天気状況下におけるセマンティックセグメンテーションのための GAN を用いたデータ拡張*

中 嶋 航 大^{**,***} 佐 藤 雄 隆^{***,**} 片 岡 裕 雄^{***}

GAN-based Data Augmentation for Semantic Segmentation in Multiple Weathers

Kodai NAKASHIMA, Yutaka SATOH and Hirokatsu KATAOKA

Datasets play an important role in determining the features that deep neural networks can acquire, but they can also contain unintended biases when constructing datasets. The BDD100K dataset, famous for its semantic segmentation task, was collected to include traffic scenes for multiple weather conditions. However, due to differences in frequency of occurrence, there is a bias in the number of data for each weather condition. Therefore, the segmentation network trained by BDD100K has poor recognition performance in some weather conditions. Semantic segmentation is an urgent issue because it is expected to be applied to traffic scene recognition systems. In this paper, we aim to improve the performance of semantic segmentation by designing a method that generates images of desired weather conditions and uses them for data augmentation. In our experiments, we first show that the image generation method we have developed produces images of a quality that can be used for data augmentation. Next, we examine the effect of data augmentation on the semantic segmentation task. As a result, compared to baseline, the mean intersection over union (mIoU) improved by about 15% in wet weather, about 9% at night, and about 7% overall.

Key words: data augmentation, semantic segmentation, GAN, multiple weathers, traffic scene

1. 緒 言

近年、大規模なデータセットで学習されたディープニューラルネットワークが様々な画像認識タスクで高い性能を示している。データセットはネットワークが獲得可能な特徴を決定づける重要な役割を担っているが、データセット構築の際に意図しない偏りを含んでしまうことがある。例えば一般物体認識タスクで有名なデータセットである ImageNet¹⁾ は、クラスごとのデータ数が一定でないことや²⁾、画像の撮影視点に偏りを持っていること³⁾ が明らかになっており、認識性能が低下する原因となっている。セマンティックセグメンテーションタスクも例外ではなく、BDD100K データセット⁴⁾ は複数種類の天気状況の交通シーンで収集されたが、発生頻度の違いから各天気状況のデータ数が偏っており、2,417 個の晴れのシーンがあるのに対して霧のシーンは 8 個しかない。そのため、BDD100K で学習されたセグメンテーションネットワークは、いくつかの天気状況では認識性能が低くなってしまふ。セマンティックセグメンテーションは、交通シーン認識システムへの応用が期待されているためこれは喫緊の課題である。理想的な解決策として新しいデータの収集がある。しかし、発生頻度の低い天気状況における多様なシーンの収集は年単位の時間を要し、さらにはセグメントラベルのアノテーションは人間が行っても 1 枚あたり数時間必要であるなど莫大な労力がかかるため現実的ではない。

本研究の目的は、各天気状況のデータ数の偏りを改善し、セマンティックセグメンテーションの性能を向上することである。

概念的には図 1 のように、3.2 節で提案する画像変換手法で生成された画像を用いて、実画像データセット (a) から全体的なデータ数を底上げして (b) に拡張する。その後 (a) および (b) でそれぞれセグメンテーションネットワークを学習し、図 1 のセグメンテーション結果のような認識性能の向上を確認する。ここで、従来の画像変換手法では、エンコーダが抽出した特徴に天気状況に関わる情報は含まれてはいるものの、明示的に分離されておらず、生成する画像の「天気状況のカテゴリ」のみを自由に切り替えることは実現できていなかった。これに対し我々の画像変換手法では、天気状況を分類するタスクで事前にネットワークをトレーニングすることにより、天気状況に関わる情報を抽出するエンコーダと、その他の情報を抽出するエンコーダを明示的に分離することに成功し、「天気状況のカテゴリ」を自由に切り替えることを可能にした。また、画像の見た目の天気状況が良好にコントロールできたとしても、各画素の意味（所属クラス）が大きく変化してしまつては、データ拡張に用いるには不適となる。そこで、入力画像と生成画像の各画素の「所属クラスが変化すること」に対してペナルティを与える項を目的関数に追加した。

我々の画像変換手法は、(b) 以外にも様々なタイプのデータセットを簡単に構築できる（図 2 を参照されたい）。一方で、学習のために複数の交通環境で収集された画像とそのセグメントラベル、および天気ラベルのセットが必要であるため、我々は BDD100K を用いることにした。BDD100K は全画像（訓練 7,000 枚とテスト 1,000 枚）に対してセグメントラベルがアノテーション済みであるが、天気ラベルは訓練画像 2,464 枚とテスト画像 0 枚にしかアノテーションされていないため、我々は残りの画像全てに対して追加アノテーションを行った。

実験セクションでは、まず我々の画像変換手法は、与えられ

* 原稿受付 令和 2 年 5 月 22 日

掲載決定 令和 2 年 9 月 30 日

** 筑波大学大学院（茨城県つくば市天王台 1-1-1）

*** 産業技術総合研究所（茨城県つくば市梅園 1-1-1）

Fig.1 Overview of this study

たセグメントラベルと入力ベクトルに対応した画像を生成可能であることを示す。また、生成画像を用いて3タイプのデータ拡張を行いセマンティックセグメンテーションタスクでの認識性能の影響を確認する。その結果、mean intersection over union (mIoU) がベースラインから、例えば雨のシーンでは15%、雪のシーンでは10%、全体では9%程度向上することが明らかになった。さらには、新たに追加する意味空間における距離を用いた制約の効果を明らかにするための実験も行う。

2. 関連研究

2.1 複数天気状況でのセマンティックセグメンテーション

セマンティックセグメンテーションは、入力画像の全ピクセルに事前に定義されたクラスを割り当てるタスクである。初期段階の手法では、Conditional Random Field (CRF)⁵⁾を最適化していた。例えばTextonBoost⁶⁾は、入力画像に畳み込んだ17種類のフィルタの応答を基に得られる粗いセグメンテーション結果にCRFを適用することで、テクスチャやエッジなどの情報の考慮を可能とした。近年では、Fully Convolutional Network (FCN)⁷⁾やSegNet⁸⁾、U-Net⁹⁾、DeepLab^{10)~12)}などのディープニューラルネットワークを活用した手法が高い性能を示している。しかし、これらの手法は有名なベンチマークで高い性能を達成することに焦点を当てているため、データセットの偏りへの対応は考慮していない。そのため、例えばBDD100Kを用いて学習されたネットワークは、いくつかの天気状況において認識性能が低くなってしまう。この問題に対処するアプローチの1つとして、ネットワークで抽出された特徴量が各天気状況で見分けがつかないようにするドメイン適応手法が挙げられる。このアプローチは特徴量分類器を追加すれば良いため、実装が簡単であるといった利点がある一方で、性能の上限が決まっている¹³⁾という問題がある。他のアプローチとしては、実画像だけでなく画像変換手法により生成された各天気状況の画

像も用いて、セグメンテーションネットワークを学習するものがある。Wenzelら¹⁴⁾は天気状況の変化に頑健なステアリング操作を目的に、セグメントラベルから生成した複数種類の天気状況の画像を訓練データに追加し、セグメンテーションネットワークを学習したときの効果を検証した。しかしこの研究は画像変換手法にCycleGAN¹⁵⁾を用いているため、天気状況の数だけGeneratorを学習する必要があり効率的とはいえない。またステアリング操作が目的であるため、セグメンテーションネットワークの評価がされていない。そのため、複数種類の天気状況下でのセマンティックセグメンテーションに対する画像変換アプローチの有用性は示されていない。一方本研究では、1回の学習で複数種類の天気状況の画像を生成可能な画像生成手法を設計し、またセグメンテーションネットワークの評価も行う。

2.2 画像変換

セグメントラベルから写実的な画像を生成する問題は、画像変換タスクのフレームワークで解決できる。既存の教師あり手法の例として、pix2pix¹⁶⁾は高い変換性能を示す手法であり、実画像または生成画像とセグメントラベルのペアをDiscriminatorに入力し、そのペアが真であるかどうかを区別する。pix2pixはGeneratorの入力層からセグメントラベルを与えるため、度重なる畳み込み層や正規化層での処理過程でセグメントラベルが持つクラス情報が失われる傾向がある。GauGAN¹⁷⁾はこの問題を解決することを目的として、SPADE (Spatially-Adaptive Denormalization)と呼ばれる正規化手法を提案した。これは、セグメントラベルに畳み込み操作を施すことで変換パラメータ γ と β を求めるものであり、Generatorの各正規化層に適用することでクラス情報を保持可能にした。一方教師なし手法には、CycleGANやMUNIT¹⁸⁾などがある。CycleGANは、セグメントラベル→画像→セグメントラベル(あるいはこの逆)のフローで変換したときの最終的な結果は元と同じになるというcycle consistency lossを導入した。これにより、ドメイン不変

な情報を保持しながらドメイン固有の情報を変換可能とした。MUNIT は画像表現をドメイン不変な情報とドメイン固有の情報に明示的に分割することで、1 入力から他出力が得られる手法である。一般的に教師なし手法は入力と教師データが一对一に対応する必要がないという利点があるが、教師あり手法よりも変換性能が低いという欠点がある。本研究では教師データ（セグメントラベルに対応した実画像と天気ラベル）が用意できるため、画像生成手法を GauGAN をベースに設計する。

3. 手 法

3.1 BDD100K データセットへの追加アノテーション

セグメントラベルから複数種類の天気状況の画像を生成するために、BDD100K の各画像に対して天気ラベルを追加でアノテーションする。BDD100K は複数の天気ラベル {clear, partly cloudy, overcast, rainy, snowy, foggy} が含まれている。このデータセットは全画像（訓練 7,000 枚とテスト 1,000 枚）に対してセグメントラベルがアノテーション済みであるが、天気ラベルは訓練画像 2,464 枚とテスト画像 0 枚にのみアノテーションされている。本研究では、実画像とそのセグメントラベルに加え天気ラベルを必要とするため、手動で残りの画像全てに対してアノテーションを行う^{*1}。アノテーション戦略および結果として得られた天気状況のデータ分布は次の通りである。

アノテーション戦略: アノテーション作業は 2 人のアノテータによって行われ、主に 2 つのステップに分けられる。はじめのステップでは、BDD100K ですでに天気ラベルが付与されている画像を参考に残りの画像にラベル付けを行った。例えば、対象の画像に雪や霧の領域が含まれる場合、それぞれ *snowy/foggy* のラベルを付与する。次はクロスチェックのステップであり、2 人のアノテータの意見を比較する。2 人の意見が一致すると、アノテーションが確定される。意見が異なる場合は、2 人のアノテータが慎重な議論を通してラベルを確定する。なお、アノテータはコンピュータビジョンに関する十分な知識を持っていることに注意されたい。

天気状況のデータ分布: 図 2 (a) は BDD100K の天気状況のデータ分布であり、データ数の偏りが確認できる（例えば *foggy* のデータ数は 8 と極端に少ない）。次節では不足する画像を生成し補充することで、データセットのデータ数および天気状況のバランスを改善することを検討する。

3.2 画像変換手法

本研究における画像変換手法の目標は、与えられたセグメントラベルから複数種類の天気状況の画像を生成することである。生成画像の物体構成はセグメントラベルにより決定できるため、我々は残りの画像構成要素である天気情報とシーン情報を入力ベクトルとして与える。なお、以降 GAN の研究においてよく用いられるやや独特な形式の図および式によって詳細な説明を行うので、馴染みが薄い場合には関連する論文¹⁵⁾等を併せて参照されたい。これらを踏まえ画像変換ネットワークを図 3 のように 2 つの Encoder (E_1, E_2) と 1 つの Generator (G) で構成する。天気情報とシーン情報はそれぞれ E_1 と E_2 により抽出される。ここで E_1 と E_2 は、それぞれ 5 層の CNN で構成されている。 G は与えられた天気情報とシーン情報、セグメントラベルから

画像を再構成する。ここで、セグメントラベルは G の各正規化層である SPADE から入力する。 G の構造は GauGAN と類似しているが、天気情報とシーン情報を 2 つの E を用いて明示的に分割する点で異なっている。これにより、生成画像の天気状況の直接的な制御が可能になるだけでなく、セグメントラベルや天気情報を固定してシーン情報のみの変更が可能となり、より多様なシーンが生成できる。

画像変換ネットワークは 2 段階の学習を行う。はじめに実画像と天気ラベルを用いた天気状況の分類タスクで E_1 を学習する。これにより、 E_1 では天気情報を抽出する能力を獲得することが期待される。次に、 E_2 と G を入力画像を再構成するタスクにより学習する。ここで E_1 の重みを固定することにより、 E_2 ではシーン情報を抽出する能力を獲得することが期待される。これらの学習により G は、天気情報とシーン情報、セグメントラベルから画像を生成可能となる。

学習後には次の 4 ステップで画像を生成する。(1) 生成したい画像の物体構成をもつセグメントラベルを選択もしくは手動で作成する。(2) 生成したい天気状況の実画像を選択する。ここで選択する実画像は天気情報の抽出にのみ利用されるため、(1) で選択したセグメントラベルと同じ物体構成である必要はない。(3) シーン情報を正規分布からサンプリングする。(4) (1) から (3) の情報を G に入力し、画像を生成する。

3.3 目的関数

前述のように画像変換ネットワークの目標は、天気情報 c_1 とシーン情報 c_2 、セグメントラベル x から Generator (G) により画像 $\hat{y}$ を生成することである。ここで、 c_1 と c_2 はそれぞれ実画像 y から E_1 と E_2 により抽出される。我々は生成画像をデータ拡張に利用することを前提とし、敵対的損失や特徴マッチング損失、潜在回帰損失、KL 損失といった広く利用されている損失に加え、意味空間における距離を制約とした意味損失を追加する。

敵対的損失: 実画像と生成画像の従う生成分布を同一にするために次式で表される敵対的損失 $\mathcal{L}_{GAN}$ を用いる。

$$\mathcal{L}_{GAN}(G, E_2, D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,c_1,c_2}[\log(1 - D(x, G(x, c_1, c_2)))] \quad (1)$$

ここで、 G は敵対的損失を最小化するのに対して、Discriminator (D) は最大化するように学習を行う。この損失により、我々の画像変換ネットワークは実画像と区別できないほど写実的な画像を生成できるようになる。

特徴マッチング損失: 近年の研究により敵対的損失に加え、L1 損失や L2 損失、特徴マッチング損失を用いることでより写実的な画像を得られることが分かってきた。これらの損失は生成画像と正解となる実画像の類似度を測るものであり、損失を最小化する学習を行うことで実画像に近い外観の画像を生成することが可能となる。L1 損失はセグメントラベルを無視するように G の学習が進む可能性があったり、L2 損失は生成画像がボケる特徴があることから、本研究では ImageNet で学習済みの VGG19¹⁹⁾ による特徴マッチング損失を用いる。特徴マッチング損失 $\mathcal{L}_{FM}$ は次式のように表される。

$$\mathcal{L}_{FM}(G, E_2) = \mathbb{E}_x \sum_{i=0}^T \frac{1}{N_i} \left[\|F^i(y) - F^i(G(x, c_1, c_2))\|_1 \right] \quad (2)$$

^{*1} 追加した天気ラベルは https://drive.google.com/drive/folders/1yiF7Kn_1JeGS16VO1gaPGt1B1NJijym?usp=sharing からダウンロードできる。

Fig.2 Data distribution of weather conditions for each dataset

Fig.3 Overview of our image-to-image transformation method

Fig.4 Samples of generated images

ここで、 F は VGG19、 T は VGG19 の層の数、 N_i は層ごとの重みを表している。また y はセグメントラベル x に対応する実画像である。

潜在回帰損失: $\mathcal{L}_{GAN}$ と $\mathcal{L}_{FM}$ は写実的な画像を生成するためのものであるため、 G が生成画像に入力ベクトル c_1 と c_2 を反映することは保証されていない。本研究では、様々な天気状況の画像を生成することが重要であるため、次式で表される潜在回帰損失を用いる。

$$\mathcal{L}_{latent}^{c_1}(G, E_2) = \mathbb{E}_{c_1} [\|c_1 - E_1(G(x, c_1, c_2))\|_1] \quad (3)$$

$$\mathcal{L}_{latent}^{c_2}(G, E_2) = \mathbb{E}_{c_2} [\|c_2 - E_2(G(x, c_1, c_2))\|_1] \quad (4)$$

KL 損失: より多様な画像を生成する方法の 1 つとして、ネットワークの出力を確率分布に従うようにする方法がある。我々は

E_2 の出力が正規分布に従うように KL 損失 $\mathcal{L}_{KL}$ を用いる。

$$\mathcal{L}_{KL}(E_2) = \mathbb{E}_y [\mathcal{D}_{KL}(E_2(y) \|\mathcal{N}(0, I))]] \quad (5)$$

ここで、 $\mathcal{D}_{KL}(p\|q) = -\int p(z) \log \frac{p(z)}{q(z)} dz$ である。この損失により学習後の画像生成時に、正規分布からのサンプリング結果を用いることが可能となるため、同一のセグメントラベルおよび天気情報を与えても多様なシーンを生成することができる。

意味的損失: 従来の画像変換手法と異なり、本研究ではデータ拡張への利用を前提としているため、生成画像は与えられたセグメントラベル x の情報を正確に反映している必要がある。そこで、生成画像 $\hat{y}$ と入力画像 y との意味空間における距離を損失として設定し、クラス情報が変化しないように制約をかける。

$$\mathcal{L}_{SEM}(G) = \mathbb{E}_y [\|S(y) - S(G(x, c_1, c_2))\|_1] \quad (6)$$

Table1 Results of evaluation 1 with R50 as the backbone. The Dataset column corresponds to Fig. 2.

Model	Dataset	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Motorbike	Bicycle	mIoU
FCN	a	91.2	48.0	76.7	16.1	20.0	25.3	23.9	19.5	78.7	38.1	93.0	35.5	12.2	81.4	35.4	44.9	21.7	10.4	40.6
	b	92.5	55.1	79.5	20.4	32.4	30.0	37.1	32.7	81.3	41.8	93.5	44.5	28.4	84.7	42.8	61.6	32.7	30.2	48.5
	c	92.1	51.7	78.8	16.1	26.0	29.8	37.2	32.0	80.2	42.0	92.8	43.8	15.2	83.4	39.7	50.2	30.1	28.8	45.8
	d	92.4	53.9	79.7	17.2	32.2	30.4	38.6	32.7	81.2	43.7	93.4	44.4	14.2	83.9	40.2	55.2	32.9	27.7	47.0
DeepLab	a	91.5	50.6	77.2	22.0	20.0	25.9	16.3	19.1	78.8	41.3	92.5	33.7	1.99	81.1	37.2	43.6	17.1	3.28	39.6
	b	92.3	54.0	79.5	20.7	34.0	30.0	36.4	31.8	80.8	42.3	93.3	44.9	20.3	84.3	46.3	66.2	29.6	19.4	47.7
	c	92.0	53.3	78.7	18.9	26.8	30.4	35.5	29.7	80.3	41.2	93.3	41.2	17.2	83.0	39.8	52.4	30.4	16.2	45.3
	d	91.3	51.7	79.2	22.7	32.1	29.1	34.8	31.4	80.2	39.2	93.3	43.2	8.60	82.8	41.6	56.3	25.3	17.7	45.3

Table2 Results of evaluation 1 with R101 as the backbone

Model	Dataset	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Motorbike	Bicycle	mIoU
FCN	a	91.1	47.2	75.2	15.1	17.0	23.6	19.1	15.5	77.5	38.1	92.5	30.7	3.92	80.7	33.0	42.5	23.1	11.6	38.8
	b	90.8	53.9	78.9	19.6	31.3	29.7	33.3	31.2	79.1	40.3	93.1	47.5	26.8	83.8	43.7	58.6	35.5	14.3	46.9
	c	91.9	51.2	78.6	22.4	25.2	29.4	38.8	29.9	80.3	40.2	93.2	45.8	23.6	83.7	39.1	45.6	30.1	22.7	45.9
	d	92.4	54.3	79.1	20.4	30.7	29.4	34.5	31.0	81.3	43.7	93.6	42.7	19.3	84.5	38.1	58.6	32.8	14.2	46.3
DeepLab	a	91.3	49.6	76.1	14.8	19.9	24.7	15.0	16.3	77.7	40.0	92.6	31.2	1.17	80.8	36.8	45.2	20.9	6.25	39.0
	b	92.1	54.4	79.2	20.1	34.0	28.6	33.9	30.3	81.0	43.0	93.5	44.1	11.1	83.4	45.7	60.2	25.2	14.8	46.0
	c	92.2	53.3	79.1	20.6	27.6	29.2	36.2	30.8	79.9	41.8	93.3	45.3	12.7	83.6	44.4	54.5	32.9	13.6	45.8
	d	91.6	53.8	78.9	22.8	31.3	28.6	35.2	30.0	80.7	40.0	93.2	45.9	17.3	82.8	47.4	60.3	28.0	10.6	46.2

ここで S は BDD100K を用いて事前に学習されたセグメンテーションネットワークである。

最終的な目的関数は次の通りである。

$$G^*, E_2^* = \arg \min_{G, E_2} \max_D \mathcal{L}_{GAN}(G, E_2, D) + \lambda_{FM} \mathcal{L}_{FM}(G) + \lambda_{latent}^{c1} \mathcal{L}_{latent}^{c1}(G, E_2) + \lambda_{latent}^{c2} \mathcal{L}_{latent}^{c2}(G, E_2) + \lambda_{KL} \mathcal{L}_{KL}(E_2) + \lambda_{SEM} \mathcal{L}_{SEM}(G) \quad (7)$$

ここで、 λ_{FM} や λ_{latent}^{c1} , λ_{latent}^{c2} , λ_{KL} , λ_{SEM} は各損失の重要度を表すハイパーパラメータである。

4. 実験と結果

本章では、まず提案した画像生成手法が、データ拡張に効果的であることを示す。また、セマンティックセグメンテーションの性能とデータ拡張方法の関係を調査する。さらには、意味的損失の効果をも明らかにするための実験も行う。

4.1 画像変換手法の定性評価

図 4 に 3.2 節で述べた方法により画像を生成した例を示す。ここで、セグメントラベルと天気状況の元になる実画像は BDD100K の訓練データセットから選択した。また、各セグメントラベルのシーン情報は固定されている。生成結果から我々の画像変換手法は、各天気状況の特徴を捉えた画像を生成できていることがわかる。例えば *rainy* では地面が濡れていたり、*snowy* では雪が積もり木が枯れていることが確認できる。他には *clear* と *partly cloudy*, *overcast* では雲の量に差があることも確認できる。*foggy* のような訓練データの数が少ない天気状況においても特徴を反映できている。全体を通して、空の明るさと地面の様子が自然であることや各物体の色が適切であることなどから、実画像が従う生成分布を近似できていると考えられる。各行に着目すると、車や建物、看板などの色が概ね同じ

であることから、与えられたシーン情報も反映できている。電柱のような入力セグメントラベルの細かな情報を失うことなく生成画像に反映することもできており、我々の画像変換手法はデータ拡張に利用可能な画像を生成できている。

4.2 データ拡張方法の定量評価

ここでは、BDD100K の訓練データセットに生成画像を 3 つの方法で追加し、セマンティックセグメンテーションタスクに対するそれぞれのデータ拡張の効果を検討する。3 タイプのデータ拡張方法は図 2 (b) から (d) のように、(b) BDD100K の訓練データセットの全セグメントラベル (7000 枚) から天気状況ごとに 1 枚ずつ画像を生成し追加する。ここで、天気情報とシーン情報はランダムに選択する。(c) 天気状況のデータ数バイアスを消去するため、各天気状況のデータ数が *clear* と同じになるように (b) で生成した画像セットからランダムに選択し追加する。(d) (b) で生成した画像セットと天気情報とシーン情報を再度ランダムに選択し生成した画像セットを追加し、データ数を 9,417 に統一する。

実験で使用するセグメンテーションネットワークは FCN と DeepLabv3 であり、バックボーンには ResNet50 (R50) と ResNet101 (R101)²⁰⁾ を使用する。データ拡張方法は BDD100K のテストデータセットを用いて、次の 2 つで評価する。(1) 全データでの mIoU。(2) 天気状況ごとの mIoU。

表 1 と表 2 は、BDD100K および 3 タイプのデータ拡張方法を用いて学習したセグメンテーションネットワークの評価 (1) の結果であり、それぞれバックボーンに R50 と R101 を用いている。また両表の Dataset は図 2 と対応している。はじめにデータ拡張を施していない場合 (a) と施した場合 (b, c, d) を比較すると、明らかに mIoU が向上している (例えば FCN w/ R50 にて、a 40.6% vs. b 48.5% と +7.9 ポイント向上)。またク

Table3 Results of evaluation 2 with R50 as the backbone

Model	Dataset	mIoU for each weather condition					
		clear	partly cloudy	overcast	rainy	snowy	foggy
FCN	a	38.0	42.7	44.4	36.6	38.9	38.4
	b	43.1	48.8	54.4	48.7	47.3	43.1
	c	42.1	49.3	50.1	44.1	42.3	48.9
	d	43.6	48.0	48.3	44.2	41.9	40.9
DeepLab	a	36.8	41.8	43.8	37.2	39.2	44.1
	b	44.8	46.9	49.7	49.4	42.2	41.6
	c	41.5	48.0	47.0	43.5	40.4	45.6
	d	41.1	47.4	48.1	47.4	41.5	41.9

Table4 Results of evaluation 2 with R101 as the backbone

Model	Dataset	mIoU for each weather condition					
		clear	partly cloudy	overcast	rainy	snowy	foggy
FCN	a	35.4	39.9	42.2	34.7	35.3	39.7
	b	42.7	48.2	51.6	46.7	42.3	38.4
	c	42.9	48.3	50.6	46.1	40.1	50.3
	d	42.8	49.1	50.5	49.5	45.0	38.4
DeepLab	a	36.3	40.9	43.5	34.0	36.9	42.9
	b	42.2	45.5	48.4	49.7	43.4	45.1
	c	40.6	49.2	48.2	49.2	41.8	38.3
	d	42.6	49.3	47.6	49.6	43.9	51.8

Table5 Comparison of mIoU with and without semantic loss

Model	Semantic loss	mIoU for each weather condition					
		clear	partly cloudy	overcast	rainy	snowy	foggy
FCN	✓	22.0	25.8	25.4	20.4	20.5	15.7
		28.4	37.1	35.5	31.0	33.6	34.4
DeepLab	✓	21.3	26.1	25.0	18.7	18.6	14.7
		31.6	36.9	36.2	32.3	34.1	37.5

ラスごとの認識性能に着目すると、*Traffic sign* や *Traffic light*, *Person*, *Rider*, *Motorbike*, *Bicycle* が特に向上している（例えば DeepLab w/ R101 にて *Traffic light* の認識率は、a 15.0% vs. c 36.2% と +21.2 ポイント向上）。これらは認識を誤ると重大事故に繋がりがねない重要なクラスであるため、好ましい傾向であるといえる。次に 3 タイプのデータ拡張方法 (b, c, d) を比較する。全体を通して (b), (d), (c) の順に mIoU が高い傾向があることから、データセットのサイズが重要であると考えられる。また (b) が (d) よりも mIoU が高い原因は、実画像と生成画像が従う生成分布が完全には一致しないことにあると考えられる。

表 3 と表 4 は、BDD100K および 3 タイプのデータ拡張方法を用いて学習したセグメンテーションネットワークの評価 (2) の結果である。ここにおいても両表の Dataset は図 2 と対応している。はじめにデータ拡張を施していない場合 (a) と施した場合 (b, c, d) を比較すると、いくつかの *foggy* を除いて mIoU が向上している。とりわけ (a) で mIoU の低い傾向にある *rainy* が向上している（例えば FCN w/ R50 にて、a 36.6% vs. b 48.7% と +12.1 ポイント向上）。*rainy* は年間を通して発生する天気状況であるため、これは良い傾向であるといえる。次に 3 タイプのデータ拡張方法 (b, c, d) を確認すると、学習するモデルにも依存するが拡張方法ごとに mIoU が向上しやすい天気状況があることが確認できる（例えば *overcast* や *rainy*, *foggy*）。また興味深いことに (c) の方法では、*clear* のデータを増やしていないにもかかわらず認識性能が向上している。これは、セグメンテーションネットワークがより汎用性のある特徴を獲得できたためであると考えられる。

以上のことから本研究の設定では、セマンティックセグメンテーションへのデータ拡張は効果があるといえる。

4.3 意味的損失の効果に関する実験

ここでは、意味的損失の効果を明らかにする。目的関数に意味的損失を加えずに学習、もしくは加えて学習した画像変換ネットワークによる生成画像のみを用いセグメンテーションネットワークの学習を行う。性能評価は BDD100K のテストデータセットを用いて、天気状況ごとの mIoU により行う。使用するセグメンテーションネットワークは FCN と DeepLabv3 で、バックボーンは ResNet50 とした。表 5 から意味的損失を加えることで、有用な画像を生成できることがわかる。これは、意味的損失を加えることで、 G が従来手法以上にセグメントラベルの情報を生成画像に反映したと考えられる。

5. 今後の課題

本論文では BDD100K が含む天気状況に注目し、あらかじめ用意された不完全な天気ラベルに対して手作業でアノテーションを追加し、画像変換ネットワークの学習を行った。BDD100K には不完全ながら時間帯のラベルも付与されているので、同様のスキームにより天気状況だけでなく時間帯状況も自由にコントロールできる可能性があり、今後の課題として検討する予定である。

6. 結 言

本研究は、データ拡張により各天気状況のデータ数の偏りを改善し、セマンティックセグメンテーションの性能を向上することを目的とした。そのために我々は、データ拡張に用いる画像を生成するための手法の提案も行った。我々の手法は、生成器への入力ベクトルを天気情報とシーン情報に明示的に分割したことや、データ拡張に用いることを前提として実画像と生成画像間の意味空間における距離を制約に追加した点で従来と異なっている。実験では、はじめに我々の画像変換手法が各天気状況の特徴を反映し、データ拡張に利用可能な画像を生成可能であることを確認した。次に BDD100K の訓練データセットに生成画像を用いた 3 タイプのデータ拡張を施し、セグメンテーションネットワークの性能を検討した。その結果、拡張方法に関わらず mIoU の向上が確認されたが、特に全天気状況で同数の生成画像を追加する方法 (b) が高くなる傾向があった。また、mIoU が全天気状況で向上する傾向があることから、本研究のアプローチによりセグメンテーションネットワークはより良い特徴を獲得できるようになったといえる。

参 考 文 献

- 1) J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei : ImageNet: A large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition, 2009.
- 2) S. Rahman, S. Khan, and F. Porikli : Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts, Asian Conference on Computer Vision, 2018.
- 3) 邱珉, 佐藤 雄隆, 鈴木亮太, 片岡 裕雄, 岩田 健司 : 一般物体認識用大規模画像データセットにおける視点依存性に関する考察及び定量化, 精密工学会論文誌, **85**, 12 (2019) 1087.
- 4) F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell : BDD100K: A diverse driving dataset for heterogeneous multitask learning, IEEE Conference on Computer Vision and Pattern Recognition, 2020.
- 5) J. Lafferty, A. McCallum, and F. Pereira : Conditional random fields: Probabilistic models for segmenting and labeling sequence data, International Conference on Machine Learning, 2001.
- 6) J. Shotton, J. Winn, C. Rother, and A. Criminisi : Textonboost for image understanding: Multi-class object recognition and segmentation by

- jointly modeling texture, layout, and context, *International Journal of Computer Vision*, **81**, 1 (2009) 2.
- 7) J. Long, E. Shelhamer, and T. Darrell : Fully convolutional networks for semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- 8) V. Badrinarayanan, A. Kendall, and R. Cipolla : Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *arXiv preprint arXiv:1511.00561*, 2015.
- 9) O. Ronneberger, P. Fischer, and T. Brox : U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.
- 10) L. -C. Chen, G. Papandreou, F. Schroff, and H. Adam : Rethinking atrous convolution for semantic image segmentation, *arXiv preprint arXiv:1706.05587*, 2017.
- 11) L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille : DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 4 (2017) 834.
- 12) L. -C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam : Encoder-decoder with atrous separable convolution for semantic image segmentation, *European Conference on Computer Vision*, 2018.
- 13) I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani : A survey on domain adaptation theory: learning bounds and theoretical guarantees, *arXiv preprint arXiv:2004.11829*, 2020.
- 14) P. Wenzel, Q. Khan, D. Cremers, and L. Leal-Taixé : Modular vehicle control for transferring semantic information between weather conditions using gans, *Conference on Robot Learning*, 2018.
- 15) J. Zhu, T. Park, P. Isola, and A. A. Efros : Unpaired image-to-image translation using cycle-consistent adversarial networks, *International Conference on Computer Vision*, 2017.
- 16) P. Isola, J. -Y. Zhu, T. Zhou, and A. A. Efros : Image-to-image translation with conditional adversarial networks, *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- 17) T. Karras, S. Laine, and T. Aila : A style-based generator architecture for generative adversarial networks, *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- 18) J. -Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman : Toward multimodal image-to-image translation, *Advances in Neural Information Processing Systems*, 2017.
- 19) K. Simonyan and A. Zisserman : Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*, 2014.
- 20) K. He, X. Zhang, S. Ren, and J. Sun : Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.