An Enhanced GAN Model for Automatic Satellite-to-Map Image Conversion

YING ZHANG	extsuperscript{1} (Member, IEEE), YIFANG YIN	extsuperscript{2}, ROGER ZIMMERMANN	extsuperscript{1} (Senior Member, IEEE), GUANFENG WANG	extsuperscript{3}, JAGANNADAN VARADARAJAN	extsuperscript{3}, AND SEE-KIONG NG	extsuperscript{2}

	extsuperscript{1}School of Computing, National University of Singapore, Singapore 119077

	extsuperscript{2}Institute of Data Science, National University of Singapore, Singapore 119077

	extsuperscript{3}R&D Geo, GrabTaxi Holdings, Singapore 018937

Corresponding author: Ying Zhang (dcszyi@nus.edu.sg)

This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2018-T2-1-103.

ABSTRACT Location-based service significantly relies on accurate and up-to-date maps. The conventional map generation involves labor-intensive and time-consuming manual efforts, which restricts the map-update frequency to a few years or even longer. In recent years, satellite images have become ubiquitous, and converting them to map-style images has attracted attention due to its frequent-updating and cost-effective nature. Generative adversarial network (GAN) is a promising approach for automatic satellite-to-map image conversion. However, it is still challenging to process satellite images when the underlying road structure is complex and irregular, or when some objects are visually indistinguishable due to obstruction or bad weather. To address these issues, we propose an enhanced GAN model to generate improved quality map images by bringing in the external geographic data as implicit guidance. The textual geographic data is converted to an image so that it can work collaboratively and seamlessly with the satellite image during the conversion. A high-level semantic regulation is also introduced to further reduce the noisy patterns generated during the translation, which occur frequently for the regions with sparse geographic data. The proposed method is versatile to various backbone GAN structures with a 20% performance improvement on three popular metrics (Inception Score, Frechet Inception Distance Score and SSIM score). Our proposed Semantic-regulated Geographic GAN (SG-GAN) is anticipated to reduce the manual identification efforts in broad geospatial applications.

INDEX TERMS Geographic information, location-based service, image to image translation, GAN.

I. INTRODUCTION

Accurate and up-to-date map images are fundamentally important for many location-based services such as map inference, road attribute detection [19]. The conventional map generation involves labor-intensive and time-consuming manual efforts [27], which restricts the map-update frequency to a few years or even longer for less populated and isolated places [8]. In recent years, satellite images have been quickly improved in terms of quantity, timeliness, quality and contents diversity. Automatic conversion from satellite images to map-style images draws increasing interest in the spatial information industry due to its cost-effective and frequent-updating nature.

Generative Adversarial Network (GAN) is a successful model for image translation and has been applied in satellite-to-map image conversion [17], [21]. However, it is still challenging to process satellite images when the underlying road structure is complex and irregular, or when some objects are visually indistinguishable due to obstruction or bad weather (shadow, cloudy, etc.). Fig. 1 shows an example when a GAN approach is applied to New York and a typical southeast Asian city. As shown in the satellite and map images, most of the streets in New York are straight and the road network follows regular and repeated patterns. Whereas in the southeast Asian city, the roads have many irregular shapes and small subbranches. We train a GAN model [21] using their own city images and the conversion results are distinct: in New York, most of the streets in New York are straight and the road network is reconstructed accurately. But in the southeast Asian city, the road network is reconstructed defectively with messy patterns and wrongly connected...
training but with additional semantic regulation. The major way, the overall system follows a standard adversarial satellite image. These two generated images are naturally outputs a semantic estimation from the Semantic-Estimator GPS-Render of satellite images and geographic data. The GAN (GAN) is shown in Fig. 2. In this model, the input consists of our proposed the regions with sparse geographic data. The architecture generated during the translation, which occur frequently for dilution is also introduced to further reduce the noisy patterns in a conditional GAN structure. A high-level semantic regulation is also introduced to reduce the visual artefacts generated during the translation, which occur frequently for the regions with sparse geographic data. The architecture of our proposed Semantic-regulated Geographic GAN (SG-GAN) is shown in Fig. 2. In this model, the input consists of satellite images and geographic data. The GPS-Render creates a GPS image from the crowdsourced geo-data. The Semantic-Estimator outputs a semantic estimation from the satellite image. These two generated images are naturally embedded into a GAN modeling as additional layers. In such a way, the overall system follows a standard adversarial training but with additional semantic regulation. The major contributions of this work are summarized as below:

1) This is the first work to embed the geographic information into the GAN structure, for satellite-to-map image conversion. A novel semantic-level regulation is also introduced to reduce the visual artefacts generated during the traditional pixel-wise conversion.

2) The proposed method is versatile to be easily embedded into various backbone GAN structures for the image-to-image translation.

3) Extensive experiments have been carried out and the results indicate promising improvements both quantitatively and qualitatively. On average, the Inception Score, Frechet Inception Distance Score and SSIM improve from the state-of-the-art Pix2Pix [21] over 20.76% and 27.91%, and 26.61%, respectively.

In the rest of this paper, we briefly summarize the related work in Section II. The proposed SG-GAN is introduced in Section III. We conduct experimental evaluation in Section IV, followed by a conclusion in Section V.

II. RELATED WORK

A. MAP IMAGE GENERATION

The conventional methods generate map images based on pre-collected geospatial data for each spatial entity within a geographic region. The geospatial data describes e.g., the position of spatial entities, shape, and mutual relationship, etc. For example, patent [28] has a map painter module to draw an image for an area with the desired appearance based on the stored geographic data. In one particular instance, map drawing generates an image of a target area (e.g., Midwestern United States or the city of San Francisco, or some other geographical area) from vector data defining points, lines, and areas of geographical features such as roads and railways, cities and parks. Patent [31] introduces a map image rendering technique at a client device in a server-client setting. To achieve this, a map server selects map data from a geospatial database for a certain geographic area, generates multiple map image layers using the selected map data, and transmit them separately to the client device.

The geo-spatial data is typically collected by professionals following certain data collection plan and protocol (e.g., which region, building, street to be collected.), using specially designed devices and tools (e.g., a survey car with GPS and cameras) [24]. A key concern is that field data collection is an expensive, time consuming, and cumbersome task. So alternative low-cost resources such as remote sensing images becomes a new resource for map generation. Usually, the professionals manually identify the objects (e.g., outline the object contours, assign the categories, etc) from high-resolution remote sensing imagery and the identified data is further verified and processed for study. However, these works still need to extract the geospatial data for each individual object and treat the map rendering as a separate work where large manual efforts are included.

To overcome the above challenges, Generative Adversarial Network (GAN) is a new method to directly convert the satellite images to map images [21], [40]. These methods do not require extract the geospatial data for each object and hence greatly reduce the manual efforts that spent in traditional methods. Additionally, satellite images are provided by various businesses around the world (e.g., Sentinel [5], MODIS [4], etc) and the resource identification efforts are reduced as well. Last but not least, satellite images are updated frequently (e.g., daily update from MODIS. This helps to reduce the efforts in latest information identification.

Most of studies propose a general image to image translation framework and satellite-to-map conversion is an application. Thus the results are not consistently good in various settings as seen in the introduction. Different from the above,
this work embeds some commonly used geographic information into the GAN models and therefore generates more promising results. Another related work modifies the classic GAN for this task [15]. But many implementation details are missing in the manuscript and we could not reproduce the architecture at this moment.

B. SATELLITE IMAGE APPLICATIONS

Roads are backbones of a map and they are the essential modes of transportation [36]. So in this part, we review the existing literature which extracts the roads from satellite images. The early approaches focus on feature-engineering to represent a road or road structure [36]. Various deep models have been proposed for this task such as the restricted Boltzmann machines [25], FCN with UNet [39] and D-LinkNet [32]. A few works conduct road segmentation using adversarial training. Zhang et al. [38] propose an improved GAN to extract the road via an end-to-end framework only requiring a few samples for training. Costea et al. [13] present a two-stage method where the first stage uses a dual-hop GAN to segment roads and their intersections. In the second stage, a road-graph is reconstructed by applying a smoothing-based graph optimization. These studies well extract the road information from the image data. However, the roads are not represented in a human readily-readable form.

C. GENERATIVE ADVERSARIAL NETWORK (GAN)

Our work is based on GAN models and we show some preliminary background in this part. The basic GAN model contains two major components, a generator \( G \) and a discriminator \( D \). The generator generates new samples in the targeting domain while the discriminator tries to distinguish whether this generated sample is from the real or the generated distribution. The objective of the generator is to confuse the discriminator by generating samples that are difficult to differentiate from the real ones. Denoting a random noise vector as \( z \) and the source input as \( x \), then the adversarial loss of a GAN is shown in Eqn. 1

\[
L_{\text{GAN}}^{\text{adversarial}} = E_x[\log D(x)] + E_z[\log(1 - D(G(z)))]
\]

where \( G \) tries to minimize this objective against the adversarial \( D \) that tries to maximize it. Under the basic GAN modeling, there is not much control of the relation between the input and output. Thus, various factors, could be further conditioned to it, which is termed as conditional GAN (cGAN) and the objective is converted to minimize the adversarial loss in Eqn. 2

\[
L_{\text{GAN}}^{\text{adversarial}} = E_{x,y} [\log D(x, y)] + E_{x,z} [\log(1 - D(x, G(x, z)))]
\]

where \( y \) is an expected label for the input \( x \). In an image translation problem, \( x \) is an image from the original domain and \( y \) is an image from the target domain.

III. METHODOLOGY

Our task is to translate a satellite image to a map-style image using a conditional GAN. In this work, we take Pix2Pix [21] as the backbone conditional GAN model, which is one of the well known image translation works and its objective is to minimize a combination of adversarial loss and contents loss as in Eqn. 3

\[
\mathcal{L} = L_{\text{GAN}}^{\text{adversarial}} + \lambda_c \cdot L_{\text{contents}}
\]

A common choice for the \( L_{\text{contents}} \) is L1 loss calculated by the following equation:

\[
L_{\text{contents}} = E_{x,y,z} [||y - G(x, z)||_1]
\]
A. GEOGRAPHIC-EMBEDDED ADVERSARIAL LOSS

As shown in the introduction, Pix2Pix faces the challenges when it is deployed to the regions with complex road networks and hence a more tailored solution for this specific geographic task is needed.

To address the above challenge, we integrate the external crowd-sourced vehicle GPS data into the conditional GAN. GPS data has contributed greatly to many geo-applications such as map inference [12] and map matching [11]. With the popularity of ride-sharing services such as Uber, GPS sequence is collected easily during individual driving. And a collective of such GPS data becomes a natural and powerful indicator for the underlying road networks. The crowd-sourced GPS is collected continuously over time and hence it has a good potential to reflect the latest geospatial information.

Let us denote the crowd-sourced GPS data as $g$, the adversarial loss of our conditional GAN is modified to a new version as in Eqn. 5

$$L^G_{adversarial} = E_{x,y}[\log D(x,y)] + E_{x,z}[\log(1 - D(x, G(x, g, z)))]$$ (5)

B. GPS IMAGE LAYER RENDERING

So a key question goes to how to embed the GPS data into the GAN architecture and use it together with the satellite images. GPS data is recorded in a textual form and each discrete GPS sample contains a pair of latitude and longitude numbers, i.e., $l = (lat, lng)$. Jointly leveraging both textual and visual information has been commonly used in geo-applications. Since the raw GPS coordinates are very fine-grained level place indicators, they can not be effectively used directly [33]. Among many existing geo-applications, it is the very first step to convert the raw GPS into other textual features [23], [33], [34], [37].

Such processing has two problems: 1) the hand-crafted GPS feature extraction requires additional computation and might not be optimized for our GAN structure. 2) The GPS feature and the satellite/map images come from different modalities, although multi-modality deep fusion is feasible but in general not an easy task [16].

In this work, we convert the GPS data to an image and use it together with the satellite images in the GAN for end-to-end training. First, given a satellite image $S$, we determine its covered geographic region ($R$) and divide it into equally sized grids. Next, for each grid $R_{ij}$, we retrieve its GPS data and denote them as $L_{ij} = \{l[i] \in R_{ij}\}$. Lastly, we create a matrix $P$ with the same size as $R$, and each of its element $P_{ij}$ is a binary value indicating if there exists any GPS in this grid by Eqn. 6. This binary matrix $P$ can be rendered as an image to represent the geographic information in a visual form and we name it as a GPS image.

$$P_{ij} = \begin{cases} 1, & \text{if } |L_{ij}| > 0. \\ 0, & \text{otherwise.} \end{cases}$$ (6)

C. SEMANTIC-REGULATED GAN REFINEMENT

Following the above formulation, GPS works as natural guidance to recover the road structure and hence improves the output’s quality. A key characteristics of crowd-sourced GPS data is their density levels vary from region to region. Based on our experiments, the conversion results usually contain visual artefacts for the regions with sparse or no GPS. A plausible reason might come from the pixel-wise translation of GAN, where the high-level structure is not sufficiently addressed [22]. This may benefit some applications where rich and diverse details are required. In our task, a map is a symbolic depiction emphasizing relationships between elements of some space, such as objects, regions, or themes [2]. The generated noises will weaken the visual division among different spaces, and hence heavily reduce the map quality.

To alleviate the problem, we bring in a semantic estimator before the generator to reduce the noises as in Fig. 2. This module approximates a semantic estimation for a satellite image, which is a region-alike division in geo-space. The satellite image’s semantics is referred to as its containing land-types and we introduce a semantic loss to monitor this estimation process. The land-type indicates which areas of a city are used for which purpose. It identifies parts of a city and the major activities (land use) that happen there. The land-type categories are usually determined by the target map themes. In our application, the categories include the some basic and widely used instances such as ground and street and we show the details in Section IV-C. Mathematically, for each satellite image’s pixel $s \in S$, its ground truth land-type is denoted as $t$. The semantic estimator assigns each pixel a land-type $\hat{t}$ and its objective is to minimize the prediction errors over all pixels. In this work, we use the multi-class cross-entropy to measure the semantic error for image $S$ by

$$I = S \oplus P$$ (7)

where the symbol $\oplus$ indicates a channel-wise image concatenation. Only the generator requires GPS data while the discriminator does not.
Eqn. 8:

$$T(S) = -\sum_{s_i \in S} \sum_{c=1}^{C} t_i^c \cdot \log(t_i^c)$$ (8)

where $C$ is the total number of land-types and $N$ is the total number of pixels inside that image. The semantic loss from all satellite images is defined as below:

$$\mathcal{L}_{\text{Semantic}} = \mathbb{E}_x[T(x)]$$ (9)

The overall loss for the proposed SG-GAN is a combination of adversarial loss, contents loss and semantic loss.

$$\mathcal{L}_{\text{SGGAN}} = \mathcal{L}_{\text{Adversarial}} + \lambda_\text{s} \cdot \mathcal{L}_{\text{Contents}} + \lambda_\text{s} \cdot \mathcal{L}_{\text{Semantic}}$$ (10)

where the parameters $\lambda$ is a loss weight. Under this refined loss, the low level pixel-wise translation will also be regulated at a region-wise high level and the visual noises are reduced. We can monitor its effect in the semantic estimator’s output mask and the generator’s output image as detailed in the experiments section. This predicted semantic mask ($M$) is layered together with the satellite image and the GPS image as the generator’s input:

$$I = S \oplus P \oplus M$$ (11)

D. MODEL STRUCTURE

Here we provide the model structure details for the generator, discriminator and the semantic estimator. For the generator and discriminator, the underlying model structures are kept the same as that in Pix2Pix. Such configurations help us to verify that the improvement, if exists, is contributed from the geographic data rather than any model structure changes.

Specifically, 1) For the generator, it uses a U-Net model which has 7 encoding layers and 7 decoding layers with skip connections. For each convolutional (or transpose) layer, we use $4 \times 4$ sized filters. The number of filter for each convolutional layers is 64, 128, 256, 512, 512, 512, 512 during the encoding phase and in reverse order during the decoding phase. Each convolutional (or transpose) layer is followed by batch normalization and ReLU activation. 2) For the discriminator, it has 6 convolutional layers. The filter size is $4 \times 4$ and the number of filters are 64, 128, 256, 512, 512 and 1 in each convolutional layer. The output of the last layer is activated by Sigmoid function. 3) The semantic estimator has the same structure as the generator but it replaces the ReLU to Softmax activation to generate a probability for each semantic class. Fig. 4 shows input and output difference between the semantic estimator and the generator.

E. DISCUSSION

Image segmentation is another potential way for satellite-to-map image conversion but with a few differences from a GAN-based method in the following aspects: 1) The objective of segmentation is to partition an image into sub-regions so that pixels with similar characteristics are grouped together. It outputs a mask with pixel-wise labels but will not address the color or texture mapping between two image domains. Additional manual efforts are required to convert the segmentation mask to a map-style image. 2) A map image usually contains dozens of categories [3] while some of them might not frequently appear. If we want to generate a good map from a predicted segmentation mask, we require most of these categories are encoded in the segmentation training. In SG-GAN, the semantic estimator works as a coarse level of segmentation. It roughly outlines the major objects in geospace so that the generator converts the image towards a more region-alike way. The semantic-richness in the final results is not restricted by the encoded land-types but determined by the backbone generator itself. Thus, we do not require all the categories to be included in the estimation. We will show in the experiments that a two-class encoded SG-GAN generates more realistic map images than a five-class encoded segmentation method.

IV. EXPERIMENTS

Satellite and Map Images In the evaluation, we collect the satellite images and the map images using HERE API [11] for a city covering a 24.14 $\times$ 10.97 km$^2$ area. The zoom level (resolution) is set at level 16 at which the streets of the area are able to be represented [7]. The zoom level indicates the satellite image’s spatial resolution and is used to align with the GPS data. All the images are saved at same resolution 256 $\times$ 256 pixels, which are the default size for most GANs [21].

GPS Dataset We randomly selected GPS records from a large and recent open dataset Grab-Posisi [20] that is collected during the ride-sharing service. In total, it has 2 million GPS samples from 31K drivers. Fig. 5 shows some GPS records from Grab-Posisi. The modern GPS not only includes latitude and longitude coordinates but also an accuracy level as shown in the last column. The accuracy unit is the meter and a smaller number indicates the GPS coordinate is more accurate. The GPS is collected with either Android or iOS devices. With Android devices, the accuracy level refers to the radius within which the location confidence is 68%. In iOS devices, the reported latitude and longitude identify the center of a circle with a radius of the reported accuracy level. The true location is assumed to be randomly distributed inside the circle. We plot the GPS accuracy histogram and CDF in Fig. 6. Most GPS is very accurate with an error less than 10 meters, thus we can safely use them. If such accuracy
levels are missing, an alternative way to ensure their correctness is firstly infer the road-network via the wisdom of the crowd. This task is known as map inference. For example, we can use kernel-density estimation to infer the road structure [14]. But this is another research task beyond the scope of this work.

Automatic Semantic Mask Preparation To obtain the semantic ground truth mask for the satellite images, a traditional way is to manually label all object’s contour which is very time-consuming. For our task, as the training image comes in satellite-map pair and there is a known color legend list for the map, the ground truth mask for satellite images can be inferred automatically from their corresponding map images. Specifically, we identify the color for the major objects in maps and then mask all map pixels from that color. In this work, the default number of land-type is two, where road pixels are set as one and rest as zero. Note that, the road might have multiple types (with different colors in the map), but we mark all these fine-grained road types as the same. Thus, our semantic mask is a coarse-level map labeling, rather than a mapping from the full map-legend list.

Training vs Testing The satellite and map images without GPS samples will be discarded. In total, we have collected 722 satellite-map-GPS images with ground truth masks and Fig. 7 shows an example. In the southeastern Asian cities, the geographical distribution of geospatial objects differs greatly from region to region and hence the image correlation is very low (examples in Figs. 7, 10 to 13. So it is safe to select the training and test data randomly. We randomly choose 80% data for training and the rest 20% is used for testing and Fig. 8 shows the data distribution in the geographic region.

Data Statistics We show some statistics of our experimental data in Fig. 9. GPS density counts the pixel-percentage in a satellite image where GPS data exists. Road density counts the percentage of road-pixels in a satellite image. Overlap density counts the pixel-percentage in a satellite image where GPS falls on a road. From the plots, GPS only exists for a small percentage of an image, mostly below 5% while road density follows a normal-alike distribution with the center around 15%. The gap between a higher road density and a lower GPS density indicates that the GPS itself cannot fully reflect the whole road network structure and thus we can not rely on GPS only to generate a full-version map. We also find that the overlap between the GPS and road is mostly over 80%, meaning the GPS is mostly accurate. Note that the GPS might not be a subset of the road-pixels. For example, a newly-appeared road can be identified from GPS but not from an out-of-date map.

Training and Parameters The semantic estimator and generator are jointly trained at the same time via the discriminator. At each epoch end, we calculate the final loss and update the parameters accordingly. The batch size is 32 and the epoch is 100. Adam optimizer is set with a learning rate of 0.0002 and a decay of 0.5. Parameter $\lambda_c = 100$ as...
FIGURE 9. Left: GPS density counts the pixel-percentage in a satellite image where GPS data exists. Middle: road density counts the pixel-percentage in a satellite image where the land-type is a road. Road density is higher than GPS density, indicating the GPS itself can not reflect a complete road network structure. Right: overlap density counts the pixel-percentage in a satellite image where GPS falls on a road. The overlap is mostly over 85%, indicating most of the GPS data are reliable.

FIGURE 10. Example (1). Plots (a)–(c) is satellite image, ground truth map image, and rendered GPS image, respectively. Plots (d)-(g) are converted maps using 6 different GAN-based approaches. Plots (j)-(l) are converted maps using 3 different segmentation-based approaches. This example is a very typical case in an Southeast Asian city which contains a mix of straight/curved, long/short, main/sub streets. The GPS integration obviously improves the map quality as in (e) and (g) but noisy patterns (highlighted in red box) appear in (e) if with sparse GPS.

suggested by Pix2Pix [21] and $\lambda_s = 10$. For every 10 epochs, we save the training model to generate the maps for all the unseen test data.

**Baselines** To evaluate the performance of our proposed SG-GAN, we firstly conduct the comparison with the following GAN based solution:

- **Pix2Pix [21]**. As introduced, it is one of the well-cited works in the image translation domain.
- **G-GAN**: In this model, we remove the semantic estimator from SG-GAN while the rest keeps the same. This is to evaluate the contribution of semantic regulation.
- **S-GAN**: In this model, we remove the GPS rendering from SG-GAN. The generator’s input is the satellite image and the intermediate semantic estimation image. This is to evaluate the contribution of GPS data.
- **CycleGAN [40]** This is a more powerful GAN for image translation. CycleGAN has a two-step transformation of source domain image - first by trying to map it to target domain and then map back to the source domain and aims to reduce the gap between the reconstructed image and the original image.
- **S-CycleGAN [35]** It is a very recent solution that targets for remote sensing image conversion. It leverage the cycle loss similar to CycleGAN and MSE loss similar to Pix2Pix to achieve an enhanced image to image translation. We follow the same parameters used in the original paper.

We also implement three well-known image segmentation based methods. The default segmentation classes include ground, road, building, greenery, and water, which are the top five land-types in the dataset. Based on the pixel-wise predicted categories, we manually do a color mapping from the segmentation mask to the targeting map.

- **U-Net [29]**: It is one of the most classic works to conduct image segmentation. Its architecture adopts an encoder-decoder framework with skip connections.
- **Vgg U-Net**: It replaced U-Net’s underlying structure to Vgg16 model. We choose it as its experimental results look better than U-Net.
- **SegNet [9]**: A state-of-the-art segmentation model and one of its key applications is to segment city-scape images which is similar to our task.
A. QUALITATIVE COMPARISON

To intuitively compare the results, we show some examples from Fig. 10 to Fig. 13. They are very common cases in a southeast Asian city. For each satellite image, the underlying road structures are quite irregular and the buildings are with various shapes or colors. Most objects are small in size and have a tight arrangement in space. Moreover, the similarity among different images is relatively low, which may impose further difficulty in effective training as the shareable information is limited. Examples (2)-(4) are partially covered with clouds, which is a very common condition among satellite images as 67% of Earth’s surface is typically covered by clouds [26]. Example (2) also contains a small shadow region at the left-bottom corner. This might affect the visual identification of the covered objects.

Let’s first look at the results from six GAN-based approaches in the sub-figures (d)-(g) in each example where a few observations are made:

- Overall, the four GAN-based approaches well convert the satellite image to map image in terms of color and structure whereas the details and quality are varied.
- Pix2Pix is prone to generate small lines that roughly outline most of the roads and buildings. A key problem is that these small lines are not well connected so that the road structure is not well reflected in the maps.
- G-GAN improves the road-connectivity obviously from Pix2Pix. Comparing the GPS image in sub-figure (c) and the G-GAN results in sub-figure (e), these improvements appear the most when the region contains sufficient GPS data. This benefits the scenarios when the objects in satellite images are blocked by environmental factors.
such as cloud and shadow. But for the regions with sparse GPS, G-GAN outputs noisy visual patterns frequently, as highlighted by the red boxes.

- SG-GAN well reduces the noises from G-GAN as the semantic estimator regulates the conversion working towards a region-alike manner.

- S-GAN looks the worst among the GAN methods. Compared to Pix2Pix, it has a larger model due to the involvement of the semantic estimation module; compared to SG-GAN, it has fewer input resource. These two disadvantages make it less effective when trained with the same amount of data.

- CycleGAN looks very different from the rest GANs by identifying a more number of blocky items rather than lines. So the road network structure is not clearly reflected.

- S-CycleGAN is a kind of combination from Pix2Pix and CycleGAN. The original S-CycleGAN paper uses more than 10k image pairs during generation and it might not work well on our small training dataset.

For the three segmentation-based approaches as shown in subplot (h)-(j) in each example, they detect most of the objects but partially. A good thing using these segmentation is that the objects are identified in a more region-alike way rather than a pixel-alike way. But most of these results look not very like a map-style. Thus, we believe GAN-based solutions still show more advantages in this translation task.

### B. QUANTITATIVE COMPARISON

Next, we present a quantitative evaluation using three widely-adopted metrics.

**Inception Score (IS)** is a popular used metric to evaluate the quality of image generative models. It correlates well with human scoring of the realism of generated images from the CIFAR-10 dataset. The score seeks to capture both image quality and image diversity of a collection of generated images. As shown in Tbl. 1, SG-GAN has a much higher IS value than the rest baselines among most scenarios. Compared to other natural image problems [10], our IS value is relatively low because the map images do not belong to the default categories of CIFAR-10 dataset. But the real map can be taken as the best targeting image and by comparing their IS value, we find that SG-GAN has the smallest gap from the real map IS score among most scenarios.

| Epoch | Segmentation-based Methods | GAN-based Methods |
|-------|---------------------------|-------------------|
|       | U-Net | Vgg U-Net | SegNet | Pix2Pix | G-GAN | S-GAN | SG-GAN | CycleGAN | S-CycleGAN |
| 40    | 1.5214 | 1.8773 | 1.3937 | 1.7783 | 2.2066 | 1.7504 | 2.0776 | 1.7312 | 1.5995 |
| 60    | 1.3393 | 1.7287 | 1.3345 | 1.6326 | 2.2743 | 1.7901 | 2.3864 | 2.1100 | 1.6716 |
| 70    | 1.6256 | 1.8023 | 1.5172 | 1.5479 | 1.9111 | 1.7514 | 2.3562 | 1.9194 | 1.5972 |
| 80    | 1.4202 | 1.7304 | 1.4183 | 1.9461 | 2.1523 | 1.9815 | 2.4595 | 2.2828 | 1.5777 |
| 90    | 1.4969 | 1.7148 | 1.4018 | 2.0235 | 2.1695 | 1.8401 | 2.3337 | 2.2622 | 1.7546 |
| 100   | 1.3483 | 1.5741 | 1.4854 | 2.1996 | 2.4944 | 1.9409 | 2.4308 | 2.2500 | 1.7350 |

Score of real map images = 2.8231
TABLE 2. FID comparison at various epochs on test data. A lower FID value indicates a better quality.

| Epoch | Segmentation-based Methods | GAN-based Methods |
|-------|---------------------------|-------------------|
|       | U-Net | Vgg U-Net | SegNet | Pix2Pix | G-GAN | S-GAN | SG-GAN | CycleGAN | S-CycleGAN |
| 40    | 112523 | 95246 | 133573 | 21954 | 14195 | 36147 | 14882 | 74872 | 24489 |
| 60    | 112980 | 98201 | 137297 | 18458 | 13290 | 19351 | 12691 | 36777 | 20382 |
| 70    | 102335 | 95235 | 89117 | 18734 | 15840 | 18585 | 12783 | 17030 | 23311 |
| 80    | 86972  | 97603 | 99136 | 19912 | 19829 | 15578 | 13824 | 19635 | 28252 |
| 90    | 102685 | 91942 | 96672 | 18726 | 12786 | 15251 | 14966 | 21625 | 31528 |
| 100   | 106365 | 100958 | 90002 | 16893 | 14523 | 16303 | 13498 | 22652 | 34180 |

TABLE 3. SSIM comparison at various epochs on test data. A higher SSIM value indicates a better quality.

| Epoch | Segmentation-based Methods | GAN-based Methods |
|-------|---------------------------|-------------------|
|       | U-Net | Vgg U-Net | SegNet | Pix2Pix | G-GAN | S-GAN | SG-GAN | CycleGAN | S-CycleGAN |
| 40    | 0.4188 | 0.4030 | 0.4473 | 0.0202 | 0.0318 | 0.0373 | 0.0357 | 0.0194 | 0.0172 |
| 60    | 0.4245 | 0.4056 | 0.4532 | 0.0239 | 0.0317 | 0.0320 | 0.0460 | 0.0224 | 0.0176 |
| 70    | 0.4029 | 0.4060 | 0.3693 | 0.0231 | 0.0376 | 0.0278 | 0.0413 | 0.0193 | 0.0170 |
| 80    | 0.3749 | 0.3907 | 0.4084 | 0.0272 | 0.0317 | 0.0300 | 0.0459 | 0.0212 | 0.0172 |
| 90    | 0.4049 | 0.4010 | 0.3887 | 0.0253 | 0.0311 | 0.0334 | 0.0470 | 0.0200 | 0.0150 |
| 100   | 0.4148 | 0.4143 | 0.3960 | 0.0316 | 0.0326 | 0.0310 | 0.0430 | 0.0207 | 0.0171 |

Frechet Inception Distance (FID) reflects the distance between the real and generated samples [18]. A lower FID is better, indicating a more similar result to the real sample. From the Table 2, in majority cases, SG-GAN reduces the distance. And the three segmentation-based methods work much worse than GAN-based solution.

Structural Similarity (SSIM) Map image is highly structured contents and we further evaluate the structural similarity between the generated and real map-image. SSIM comes from a perception-based model that considers image degradation as the perceived change in structural information [6]. From Table 3, it is interesting to see that segmentation-based approaches now bypass the GAN-solutions. This is not surprised as structural information is reflected from the pixels with strong inter-dependencies especially when they are spatially close. While the objective of segmentation is to group the spatially closed pixels if they share similar characteristics, it directly highlights the structural information to be well presented. Among the four GAN-based works, SG-GAN clearly improved maps’ structure quality.

Fig. 15 visualizes the quantitative performance difference among four GAN-based solutions by normalizing the score of the best approach as value one where we have two key observations. First, SG-GAN and G-GAN work better than the others, indicating the effectiveness of GPS data. Second, SG-GAN further improves G-GAN, indicating the effectiveness of semantic regulation. This is reflected in its 24.23% increase in SSIM, which is 3-4 times better than the increase in FID (9%) and IS (6%).
The influence of parameter $\lambda$ to the generated semantic mask and map image for the example in Fig. 10. Plots (a)-(e) corresponds to five different $\lambda$ values and plot (f) is the ground truth mask and map image.

![Fig. 16](image1)

FIGURE 16. The influence of parameter $\lambda$ to the generated semantic mask and map image for the example in Fig. 10. Plots (a)-(e) corresponds to five different $\lambda$ values and plot (f) is the ground truth mask and map image.

The influence of parameter $C$ to the generated semantic mask and map image for the example in Fig. 10. $C = 2$ includes road and non-road; $C = 3$ includes road, building and ground; $C = 5$ includes road, building, ground, water and greenery.

![Fig. 17](image2)

FIGURE 17. The influence of parameter $C$ to the generated semantic mask and map image for the example in Fig. 10. $C = 2$ includes road and non-road; $C = 3$ includes road, building and ground; $C = 5$ includes road, building, ground, water and greenery.

| $C$   | Average Time | Standard Deviation |
|-------|--------------|--------------------|
| 2     | 0.0197       | 0.0003             |
| 3     | 0.0204       | 0.0005             |
| 5     | 0.0240       | 0.0003             |

TABLE 4. Map image generation time comparison (unit: second per image).

Time Cost: Lastly we compare the map generation time among GAN methods in Table 4. All experiments are done on a Linux platform with a Quadro RTX 6000 Graphics Card. Pix2Pix, G-GAN, S-CycleGAN have very similar time and they are the fastest methods among all methods due to their compact model size. SG-GAN is not the most efficient one but still quite fast on an average of around 0.02 seconds per image. Since this application does not have a high realtime requirement and considering the improvements in image quality, we believe the proposed solution still has some merits. But we will improve the network design to compact the model size with better efficiency in future work.

C. MODEL PARAMETERS

Parameter $\lambda$ controls the semantic contribution in the adversarial training. Fig. 16 (a)-(e) shows the results when $\lambda$ is set at five different values ($1, 5, 10, 20, 40$) for the example in Fig. 10. Each result contains a binary mask and a generated map image. The binary mask labels a pixel in white if its semantic type is predicted as a road and black otherwise. From the masks, when $\lambda$ value is small, the semantic-distinguish among pixels cannot be sufficiently addressed. This leads the generated map to depend more on the GPS layer. When the value goes larger, semantic regulation works better and the route structure is identified much clearer. Hence the generated map integrates the semantics and the GPS layer with a better balance. When the value is set very large as in sub-figure (e) $\lambda = 40$, the model emphasizes the semantic loss strongly. Along with an advantage that noises in a mask are reduced, there is a disadvantage that some weak connections between road segments are also removed. This makes the map quality decreases again. As seen from the semantic mask, some long roads are divided into short segments and the objects are more clustered. Such an overemphasize on semantics rather than GPS also results in an over-fitting and the test performance drops quickly.

Parameter $C$ is the number of semantic categories to be encoded in the training. Fig. 17 shows the results when this parameter is set with three different values ($2, 3, 5$). A larger value indicates more categories will be included. By default, we use $C = 2$ including road and non-roads. When the parameter value increases to $C = 3$, we further refine the
non-roads to ground and building so that we have three categories including road, ground and building. The parameter $C = 5$ includes road, ground, building, water and greenery. With different $C$ values, the difference lies not only in their grayscale masks but also the generated images. For example, when $C$ increases from 2 to 3, the newly added category (building) is clearly labeled with light-gray color in Fig. 17(b)'s mask. However, it is interesting to see that, although the building category is not included when $C = 2$, the map in Fig. 17(a) still includes these objects to some extent. So the semantic richness in a generated map will not be restricted to the encoded semantic category whereas the clarity level will be affected. Another finding is that the map quality may not increase with a larger $C$ value for two reasons. When the parameter goes larger as in $C = 5$, the semantic estimator outputs a five-layer mask to be appended to the generator’s input and this reduces the relative contribution from the singly-layered GPS image. Secondly, a larger value increases the training difficulty as more categories should be predicted. Given the same amount of training data, segmentation over-fitting occurs easily and it explains why the generated map quality drops.

V. CONCLUSION

In this work, we proposed a GAN model for satellite-to-map image conversion. Considering the challenge that some objects might not be easily identified visually from the satellite image, we novelty integrate the external geographic data into a GAN structure to build the conversion. The leverage of GPS data is proved to benefit the scenarios when a region has a complex road network, small and irregular routes, underpass, or with occlusion from cloud and shadow. We also bring in a semantic regulation to estimate the satellite images’ high-level information. This estimated semantics help the translation working towards region-alike and hence reduce many pixel-wise noises. The proposed GPS-integration and semantic-estimation can be easily embedded into various backbone GAN structures. Extensive experiments have been carried out where both qualitative and quantitative improvements are observed.

ACKNOWLEDGMENT

(Ying Zhang and Yifang Yin contributed equally to this work.)

REFERENCES

[1] HERE Maps API. Accessed: 2020. [Online]. Available: https://developer.here.com
[2] Map. Accessed: 2020. [Online]. Available: https://en.wikipedia.org/wiki/Map
[3] Map Key—OpenStreetMap. Accessed: 2020. [Online]. Available: https://www.openstreetmap.org/key
[4] MODIS. Accessed: 2020. [Online]. Available: https://modis.gsfc.nasa.gov/
[5] Sentinel. Accessed: 2020. [Online]. Available: https://sentinel.esa.int/web/sentinel/home
[6] Structural Similarity. Accessed: 2020. [Online]. Available: https://en.wikipedia.org/wiki/Structural_similarity
[7] Zoom Levels. Accessed: 2019. [Online]. Available: https://wiki.openstreetmap.org/wiki/Zoom_levels
[8] (2019). Google Maps Update. [Online]. Available: https://www.techjunkie.com/how-often-google-maps/
[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[10] B. Barratt and R. Sharma, “A note on the inception score,” 2018, arXiv:1801.01973. [Online]. Available: http://arxiv.org/abs/1801.01973
[11] M. Bierlaire, J. Chen, and J. Newman, “A probabilistic map matching method for smartphone GPS data,” Transp. Res. C, Emerg. Technol., vol. 26, pp. 78–98, Jan. 2013.
[12] C. Chen, C. Lu, Q. Huang, Q. Yang, D. Gunopulos, and L. Guibas, “City-scale map creation and updating using GPS collections,” in Proc. KDD, 2016, pp. 1465–1474.
[13] D. Costea, A. Marcu, M. Leordeanu, and E. Slusanschi, “Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 2100–2109.
[14] J. J. Davies, A. R. Beresford, and A. Hopper, “Scalable, distributed, real-time map generation,” IEEE Pervas. Comput., vol. 5, no. 4, pp. 47–54, Oct. 2006.
[15] S. Ganguli, P. Garzon, and N. Glaser, “GeoGAN: A conditional GAN with reconstruction and style loss to generate standard layer of maps from satellite images,” 2019, arXiv:1902.05611. [Online]. Available: http://arxiv.org/abs/1902.05611
[16] S. Ghosh, “Challenges in deep learning for multimodal applications,” in Proc. ACM Int. Conf. Multimodal Interaction (ICMI), 2015, pp. 611–615.
[17] J. Gu, Y. Zhang, W. Zhang, H. Yu, S. Wang, W. Yang, and L. Wang, “Aerial image and map synthesis using generative adversarial networks,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., Jul. 2019, pp. 9903–9806.
[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6626–6637.
[19] J. Huang, M. Deng, J. Tang, S. Hu, H. Liu, S. Wariyo, and J. He, “Automatic generation of road maps from low quality GPS trajectory data via structure learning,” IEEE Access, vol. 6, pp. 71965–71975, 2018.
[20] X. Huang, Y. Yin, S. Lim, G. Wang, B. Hu, J. Varadarajan, S. Zheng, A. Bulusu, and R. Zimmerman, “Grab-Posisi: An extensive real-life GPS trajectory dataset in southeast Asia,” in Proc. PredictGIS, 2019, pp. 1–10.
[21] P. Ioia, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.
[22] X. Liang, H. Zhang, L. Lin, and E. Xing, “Generative semantic manipulation with mask-contrasting GAN,” in Proc. ECCV, 2018, pp. 558–573.
[23] S. Liao, X. Li, H. T. Shen, Y. Yang, and X. Du, “Tag features for geo-aware image classification,” IEEE Trans. Multimedia, vol. 17, no. 7, pp. 1058–1067, Jul. 2015.
[24] K. Lwin, M. Hashimoto, and Y. Murayama, “Real-time geospatial data collection and visualization with smartphone,” J. Geograph. Inf. Syst., vol. 6, no. 2, pp. 99–108, 2014.
[25] V. Mnih and G. E. Hinton, “Learning to detect roads in high-resolution aerial images,” in Proc. ECCV, 2010, pp. 210–223.
[26] NASA. Cloudy Earth. Accessed: 2015. [Online]. Available: https://earthobservatory.nasa.gov/images/85843/cloudy-earth
[27] O. Pink and C. Stiller, “Automated map generation from aerial images for precise vehicle localization,” in Proc. IEEE ITS Conf., Sep. 2010, pp. 1517–1522.
[28] J. E. Rasmussen, L. E. Rasmussen, and S. Ma, “Generating and serving tiles in a digital mapping system,” U.S. Patent 7 962 281, Jun. 14, 2011.
[29] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, “Leveraging crowdsourced GPS data for road extraction from aerial imagery,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7509–7518.
[30] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev, “Improving image classification with location context,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1008–1016.
[34] G. Wang, D. Hoiem, and D. Forsyth, “Building text features for object image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 1367–1374.

[35] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu. “SAR-to-optical image translation using supervised cycle-consistent adversarial networks,” IEEE Access, vol. 7, pp. 129136–129149, 2019.

[36] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, and P. Eklund. “A review of road extraction from remote sensing images,” J. Traffic Transp. Eng., vol. 3, no. 3, pp. 271–282, 2016.

[37] Y. Yin, Z. Liu, and R. Zimmermann, “Geographic information use in weakly-supervised deep learning for landmark recognition,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2017, pp. 1015–1020.

[38] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual U-Net,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 749–753, May 2018.

[39] Y. Yin, Z. Liu, and R. Zimmermann, “Aerial image road extraction based on an improved generative adversarial network,” Remote Senses., vol. 11, no. 8, p. 930, 2019.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2223–2232.

**YING ZHANG** (Member, IEEE) received the B.E. degree in computer science from Northwestern Polytechnical University, Xi’an, China, in 2009, and the Ph.D. degree in computer science from the National University of Singapore (NUS), in 2014. She is currently a Senior Research Fellow of the NUS. Before that, she was the Head of the Data Security Unit and a Research Scientist with the Department of Cyber Security and Intelligence, Institute for Infocomm Research (I2R), A*STAR, Singapore. Her research interests include machine learning, spatiotemporal data mining, and multimedia.

**YIFANG YIN** received the B.E. degree from the Department of Computer Science and Technology, Northeastern University, Shenyang, China, in 2011, and the Ph.D. degree from the National University of Singapore (NUS), Singapore, in 2016. She is currently a Research Fellow of the Grab-NUS AI Laboratory, NUS. She worked as a Research Intern at the Incubation Center, Research, and Technology Group, Fuji Xerox Company Ltd., Japan, from October 2014 to March 2015. Her research interests include machine learning, spatiotemporal data mining, and multimodal analysis in multimedia.

**ROGER ZIMMERMANN** (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the University of Southern California (USC), in 1994 and 1998, respectively. He is currently a Professor with the Department of Computer Science, National University of Singapore (NUS), Singapore. He is also the Deputy Director of the Smart Systems Institute (SSI) and was the Co-Director of the Centre of Social Media Innovations for Communities, NUS. He has coauthored a book, more than 250 conference publications, journal articles, and book chapters. He holds seven patents. His research interests include streaming media architectures, distributed systems, mobile and geo-referenced video management, the applications of machine/deep learning, and spatial data management. He is an Associate Editor of IEEE MULTIMEDIA, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Multimedia Tools and Applications (MTAP) (Springer), and the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY (OJ-COMS). He is a Distinguished Member of the ACM.

**GUANFENG WANG** received the B.E. degree in software engineering from Zhejiang University, China, in 2010, and the Ph.D. degree from the National University of Singapore, in 2014. He is currently a Software Engineer Lead with GrabTaxi Holdings. His research interests include mobile media analysis, mobile sensing, geostatistics, and multimedia systems.

**JAGANNADAN VARADARAJAN** received the M.Sc. degree in mathematics and the M.Tech. degree in computer science from Sri Sathya Sai University, India, in 2003 and 2005, respectively, and the Ph.D. degree from the Idiap Research Institute and the Ecole Polytechnique Federal de Lausanne, Switzerland, in 2012. He worked as a Scientist at Hewlett-Packard Labs and GE Global Research, India, from 2005 to 2008. His research interests include computer vision and machine learning. He is currently the Head of Data Science for Machine Learning and the Maps Team, GrabTaxi Holdings.

**SEE-KIONG NG** received the B.S. degree in computer science from Carnegie Mellon University (CMU), the M.S. degree from the University of Pennsylvania, and the Ph.D. degree in computer science from the CMU.

He is currently a Professor of practice with the Department of Computer Science, School of Computing, National University of Singapore, where he is also the Director of translational research with the Institute of Data Science. He has authored over 100 articles in diverse and cross-disciplinary research topics from bioinformatics to smart cities based on data science.