Response to the Review for GRSL-00665-2022: A Semi-Supervised Image-to-Image Translation Framework for SAR-Optical Image Matching

Dear Editors and Reviewers,

We now submit a revised version of the manuscript titled "A Semi-Supervised Image-to-Image Translation Framework for SAR-Optical Image Matching" to your reputed IEEE Geoscience and Remote Sensing Letters. Our manuscript's ID is GRSL-00665-2022.

We thank the editors and the reviewers for the valuable comments that really improve the quality of our manuscript. We have made substantial modifications according to these valuable comments. In particular, we highlight the major changes in this response as follows:

1. We describe more details about our framework to show our advantages compared to the two well-known image-to-image translation methods, i.e., Pix2pix and CycleGAN.
2. We highlight that we only focus on the SAR-optical image matching methods based on image-to-image translation.
3. We remove the name of proposed method, i.e., "SPC-GAN", throughout the revised manuscript.

In the following, we provide a point-by-point response and a manuscript highlighted the revisions. We hope the new revision can address all concerns of editors and reviewers. Once again, we would like to express our sincere appreciations to both editors and reviewers for their valuable comments.

Best Regards,

Wen-Liang Du (E-mail: wldu@cumt.edu.cn)

Yong Zhou

Hancheng Zhu

Jiaqi Zhao

Zhiwen Shao

Xiaolin Tian

Response to Reviewer 1

The authors would like to express sincere appreciations to the reviewer for the thorough and constructive comments. We present a detailed response to the individual points raised by the reviewer. In this reply, we indicate the revised parts with blue color when directly quote them from the revised manuscript.

Concern # 0: Image to image translation in the context of SAR-optical translation is useful in remote sensing for many applications including registration, fusion, and change detection. This manuscript revisits this problem. The two most popular image to image translation frameworks are pix2pix (that needs aligned SAR and optical images) and CycleGAN (that does not need aligned SAR and optical images). The manuscript claims to combine them to create a new semi-supervised setting. However, the claims of the manuscript are possibly erroneous. Here are more detailed comments:

Response: We thank the reviewer for the valuable comments that really improve the quality of our manuscript. We next present a point-by-point response as follows.

Concern # 1: Potentially erroeneous claim: Pix2pix is generally conditioned on the input image. As detailed in Page 3 of pix2pix original paper (https://arxiv.org/pdf/1611.07004.pdf), the idea is to learn a mapping from observed image x and random noise vector z to an output image y. Objective of pix2pix (Equation 1 in pix2pix paper) is conditioned on x. Instead, if we do not condition on x, it simply reduces to Equation 2 in pix2pix paper, which the authors in this manuscript claim to use in Section II.A. However, this merely reduces to CycleGAN (please see Equation 1 in the CycleGAN paper). In other words, in this manuscript the authors are simply using CycleGAN throughout, with the additional proposition that some of the samples are aprior known to be paired.

Response: We thank the reviewer for the constructive comment. We are sorry that we are failed to represent our framework. We will give replies in two aspects.

1. In Tab. 1, we show the results of simply using CycleGAN throughout trained with 4,000 paired training data and in the "–data_mode"[1] of "aligned". In this setting, CycleGAN obtains **no qualified matching**, while CycleGAN trained with 4,000 unpaired data and in the "–data_mode" of "unaligned" obtained at least 20 qualified matchings from 900 SAR-optical image matchings. In contrast, our framework obtains much more qualified matchings than "simply using CycleGAN throughout" (see Tab. 1).

We then tshow three pair of SAR-to-optical and optical-to-SAR translation results of CycleGAN and our framework in Fig. 1. Obviously, CycleGAN still assigns wrong gray values in the supervised settings (see the water in the bottom of urban examples). Thanks to the Pix2pix structure, our framework assigns correct gray values in both supervised and semi-supervised settings.

[1] "–data_mode" is a training option in the CycleGAN's source codes. "–data_mode aligned" means random samplings of SAR data and paired with its corresponding aligned optical data; "–data_mode unaligned" means random samplings of SAR and optical data.

We further show a certain training iteration of CycleGAN and our framework in Fig. 2. Obviously, CycleGAN can only train one pair of data in one iteration, while our framework can train two pairs of data in one iteration. Hence, for our semi-supervised version, we can feed one pair of aligned data to the supervised module and one pair of unaligned data to the unsupervised module. The experiments results shown in our manuscript (Fig. 3-8) indicates that our framework indeed combines the benefits of Pix2pix (correct gray value assignment) and CycleGAN (good edge-preserving), and avoids the disadvantage of the two methods (i.e., Pix2pix obtains blur results; CycleGAN assigns wrong gray value to some features)

Therefore, we are **NOT** "simply using CycleGAN throughout, with the additional proposition that some of the samples are aprior known to be paired".

Table 1: NOQMs obtained by different settings of CycleGAN and our framework.

Methods	Rural (300)	Semi-urban (300)	Urban (300)
CycleGAN-sup+4,000 aligned data	0	0	0
CycleGAN-unsup+4,000 unaligned data	4	3	13
Ours-sup+4,000 aligned data	90	89	94
Ours-semi+1,000 aligned data and 3,000 unaligned data	76	67	79

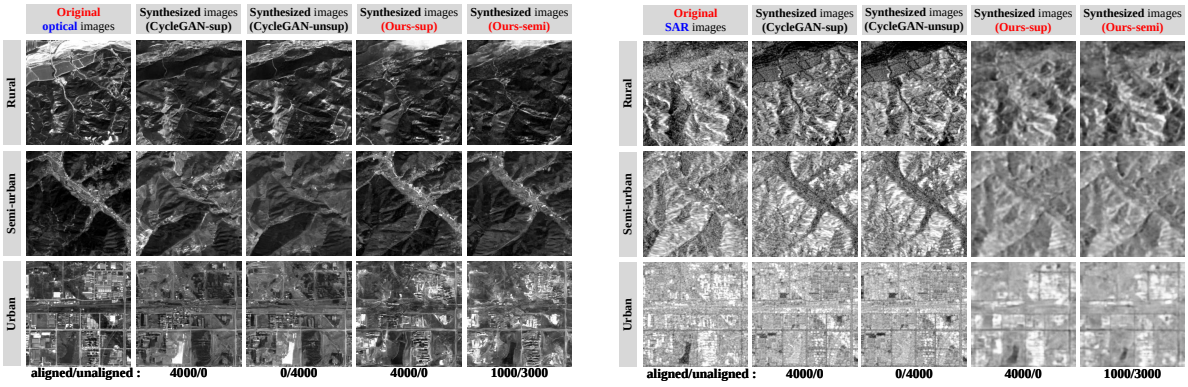

Figure 1: SAR-to-optical and optical-to-SAR translation results of CycleGAN and our framework.

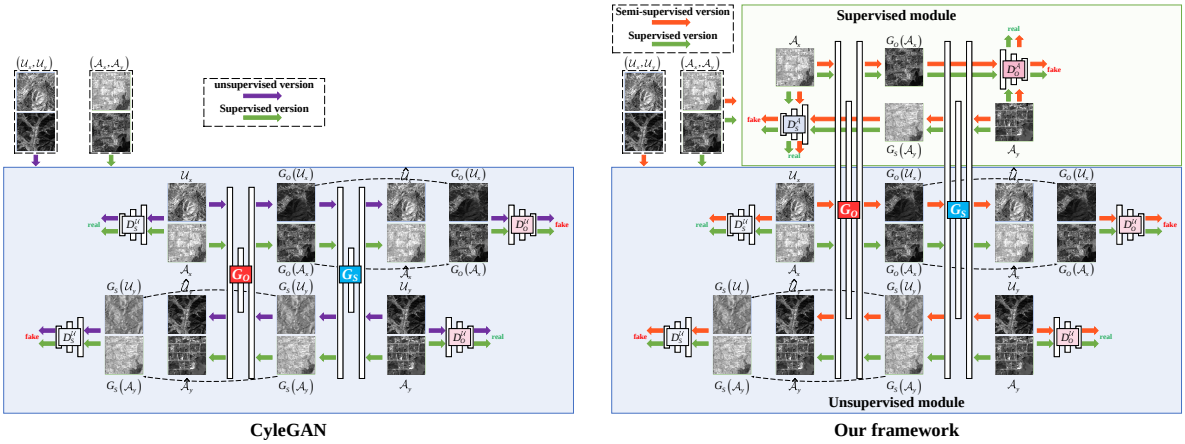

Figure 2: Illustrations of a training iteration of CycleGAN and our framework.

2. There is only a little difference between Pix2pix in conditional and unconditional manners for the SAR-optical image matching task (see Tab. 2). Therefore we perform our supervised module **in unconditional manner**. In this way, our framework is much simpler, and the equations of adversarial losses in our supervised and unsupervised modules are unified.

Table 2: NOQMs of Pix2pix in conditional and unconditional manners.

Methods	Rural (300)	Semi-urban (300)	Urban (300)
conditional Pix2pix (L1=50)	86	70	59
Unconditional Pix2pix (L1=50)	78	68	56

We describe more details about our framework in the revised manuscript. We will also upload Tab. 1 and 2 and Fig. 1 to Github (*https://github.com/WenliangDu/Semi-I2I*) along with our source codes in the future. We hope this could address your concern. We except the major revisions as follows.

In the Section I. Introduction,

> "...In this work, we propose a simple yet effective framework for semi-supervised image-to-image translation, which merges two well-known supervised and unsupervised image-to-image translation methods, i.e., Pix2pix [1] and CycleGAN [2]. Our framework combines the benefits of Pix2pix (correct gray value assignment) and CycleGAN (good edge-preserving), and avoids the disadvantage of the two methods (i.e., Pix2pix obtains blur results; CycleGAN assigns wrong gray value to some features)..."

In the Section II. Method,

> "...For *the semi-supervised version*, our goal is to learn two mapping functions between SAR ($\mathcal{X}$) and optical ($\mathcal{Y}$) images given K aligned SAR-optical image pairs ($\mathcal{A}_{\mathcal{X}} = \{\mathcal{A}_x^k\}_{k=1}^K \in \mathcal{X}$, $\mathcal{A}_{\mathcal{Y}} = \{\mathcal{A}_y^k\}_{k=1}^K \in \mathcal{Y}$) while given unaligned M SAR ($\mathcal{U}_{\mathcal{X}} = \{\mathcal{U}_x^i\}_{i=1}^M \in \mathcal{X}$) and N optical image pairs ($\mathcal{U}_{\mathcal{Y}} = \{\mathcal{U}_y^j\}_{j=1}^N \in \mathcal{Y}$). Hence, in each training iteration of the *semi-supervised version*, a pair of aligned and unaligned data are fed into the unsupervised and supervised modules, respectively..."

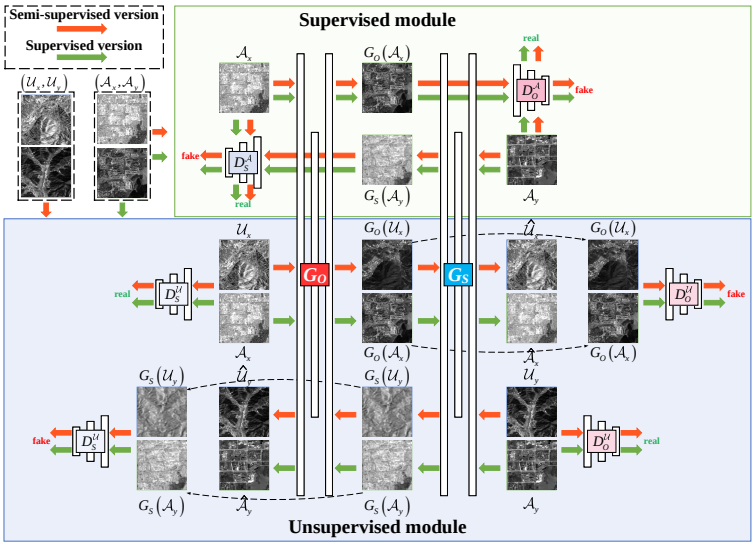

Figure 2: An illustration of a training iteration of our framework.

Concern # 2: Potentially erroneous claim: The authors claim to be outperforming Pix2pix in supervised setting, however there is nothing in the method to have outperformed Pixpix, except that they are actually using CycleGAN throughout, which means a network twice complex and it may have outperformed pix2pix, to no surprise.

Response: We thank the reviewer for the valuable comments. We are sorry that we are failed to represent our framework. As we replied in the former response, we are not "using CycleGAN throughout, and the CycleGAN performs worse when trained with aligned data" (see Fig. 2).

Specifically, in our supervised version, a pair of aligned data is fed into both supervised module (Pix2pix module in unconditional manner) and unsupervised module (CycleGAN module). While Pix2pix only trains one pair of aligned data in one iteration. Hence, we indeed consider more cycle constraint and adversarial constraint than Pix2pix.

In addition, The public implementation of Pix2pix is based on U-net, but the SAR-optical matching results of Pix2pix performed on U-net (initial settings as Pix2pix) are much worse than Pix2pix performed on ResNet (the same network architecture as CycleGAN, see Tab.3). Hence, the experimental results of Pix2pix shown in our manuscript are performed by the same network architecture as our framework (i.e., ResNet blocks). That is, we share the same generator's network complexity with Pix2pix, **not the twice complex** to Pix2pix. However, our framework uses four discriminators, while CycleGAN and Pix2pix only use two.

Table 3: NOQMs of Pix2pix based on U-net and ResNet.

Methods	Rural (300)	Semi-urban (300)	Urban (300)
conditional Pix2pix + U-net (L1=50)	1	0	0
conditional Pix2pix + ResNet (L1=50)	86	70	59

Besides, for a certain experiment, to transform SAR-optical image matching into SAR-SAR and optical-optical image matchings by Pix2pix, we train Pix2pix twice to get SAR-to-optical and optical-to-SAR mapping functions. In other words, the Pix2pix is also trained in a Cycle-like scheme in our manuscript. But *the supervised version* of our framework performs better in the SAR-optical image matching task than the **twice trained Pix2pix**.

Therefore, we believe that our framework indeed outperforms Pix2pix in the supervised setting.

We describe more details about *the supervised version* of our framework in the revised manuscript and about the training scheme of Pix2pix. We will also upload Tab. 3 to Github (`https://github.com/WenliangDu/Semi-I2I`) along with our source codes in the future. We hope this could address your concern. We except the major revisions as follows.

In the Section II. Method,

> "...While for *the supervised version*, only K aligned SAR-optical image pairs $(\mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}})$ are given to learn the two mapping functions. Hence, in each training iteration of *the supervised version*, the unsupervised and supervised modules are fed by the same pair of aligned data..."

In the Section III.-B. *Supervised, Unsupervised, and Semi-Supervised Baselines*,

> "...For a fair comparison, we implement Pix2pix and CycleGAN using the same network architecture and details as our framework. We use the public implementation of TCR due to fundamental differences in architecture..."

In the Section IV.-D. *Comparison of Image Matching*,

> "...Note that because SAR-optical image matching results in this work are based on both SAR-SAR and optical-optical matchings, Pix2pix and TCR are trained twice to obtain SAR-to-optical and optical-to-SAR translations..."

Concern # 3: Redundant content: CycleGAN has been already used many times by now in remote sensing literature. The entire method is repetition of it and and equations 1,2..4,5..6,7 are just repetitions. I would have expected only 2 out of those 6 equations. At most 4. However, not 6.

Response: We thank the reviewer for the constructive comments. We use the same repetitions as the CycleGAN paper to reduce the original equations 1, 2, 4, 5, 6, and 7 to equation 1 in the revised manuscript. Note that $\mathcal{L}_{\text{L1}}\left(G_O, G_S\right)$, $\mathcal{L}_{\text{cyc}}\left(G_O, G_S\right)$, and $\mathcal{L}_{\text{cyc}_{\text{S}}}\left(G_O, G_S\right)$ is simplified to $\mathcal{L}_{\text{L1}}$, $\mathcal{L}_{\text{cyc}}$, and $\mathcal{L}_{\text{cyc}_{\text{S}}}$ due to the limitation of the letter.

We except the major revisions as follows.

In the Section II. Method,

> "...The adversarial loss shows the key idea of GAN—training a generator and a discriminator in a minimax two-player game. The adversarial loss of $\mathcal{A}_{\mathcal{X}} \rightarrow \mathcal{A}_{\mathcal{Y}}$ is expressed as:
>
> $$\begin{aligned}\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{A}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) =& \mathbb{E}_{\mathcal{A}_{\mathcal{Y}}}\left[\log D_O^{\mathcal{A}}\left(\mathcal{A}_{\mathcal{Y}}\right)\right] + \\ & \mathbb{E}_{\mathcal{A}_{\mathcal{X}}}\left[\log\left(1 - D_O^{\mathcal{A}}\left(G_O\left(\mathcal{A}_{\mathcal{X}}\right)\right)\right)\right],\end{aligned} \tag{1}$$
>
> where $D_O^{\mathcal{A}}$ learns to differentiate real and synthesized optical images from aligned data; $\mathbb{E}_{(\cdot)} \triangleq \mathbb{E}_{(\cdot)\sim p_{\text{data}}(\cdot)}$ and $\mathbb{E}_{(\cdot)\sim p_{\text{data}}(\cdot)}\left[f\left(\cdot\right)\right]$ returns the expectation of $f(\cdot)$ with respect to the data-generating distribution $p_{\text{data}}(\cdot)$. Similarly, the adversarial loss of $\mathcal{A}_{\mathcal{Y}} \rightarrow \mathcal{A}_{\mathcal{X}}$ is introduced as $\mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{A}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right)$, where $D_S^{\mathcal{A}}$ learns to differentiate real and synthesized SAR images from aligned data..."

> "...For *the semi-supervised version*, the adversarial losses of $\mathcal{U}_{\mathcal{X}} \rightarrow \mathcal{U}_{\mathcal{Y}}$ and $\mathcal{U}_{\mathcal{Y}} \rightarrow \mathcal{U}_{\mathcal{X}}$ are introduced as: $\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{U}_{\mathcal{X}}, \mathcal{U}_{\mathcal{Y}}\right)$ and $\mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{U}_{\mathcal{Y}}, \mathcal{U}_{\mathcal{X}}\right)$, where $D_O^{\mathcal{U}}$ and $D_S^{\mathcal{U}}$ discriminate real and synthesized images, and G_O and G_S are the same as the generators in the supervised module.
>
> The adversarial losses in *the supervised version* of unsupervised module are introduced as: $\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right)$ and $\mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right)$. The similarities between these two adversarial losses and the adversarial losses in supervised module are that they are all performed on the same aligned image pairs ($\mathcal{A}_{\mathcal{X}}$, $\mathcal{A}_{\mathcal{Y}}$), and they share the same two generators. The different is that they use different discriminators..."

> "...The full objectives of the *semi-supervised* and *supervised* versions of our framework are respectively expressed as:
>
> $$\begin{aligned}&\mathcal{L}_{\text{semi}}\left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right) = \lambda_1\mathcal{L}_{\text{L1}} + \lambda_2\mathcal{L}_{\text{cyc}} + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{A}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{A}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right) + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{U}_{\mathcal{X}}, \mathcal{U}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{U}_{\mathcal{Y}}, \mathcal{U}_{\mathcal{X}}\right)\end{aligned} \tag{2}$$
>
> $$\begin{aligned}&\mathcal{L}_{\text{sup}}\left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right) = \lambda_1\mathcal{L}_{\text{L1}} + \lambda_2\mathcal{L}_{\text{cyc}_{\text{sup}}} + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{A}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{A}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right) + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right)\end{aligned} \tag{3}$$

where λ_1 and λ_2 are hyper-parameters that control the relative importance of the L1 loss and cycle consistence, respectively. The goals of the two versions are to solve:

$$\min_{(G_O,G_S)} \max_{(D_O^{\mathcal{A}},D_S^{\mathcal{A}},D_O^{\mathcal{U}},D_S^{\mathcal{U}})} \mathcal{L}_{\text{semi}} \left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right), \tag{4}$$

$$\min_{(G_O,G_S)} \max_{(D_O^{\mathcal{A}},D_S^{\mathcal{A}},D_O^{\mathcal{U}},D_S^{\mathcal{U}})} \mathcal{L}_{\text{sup}} \left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right). \tag{5}$$

..."

Concern # 4: Lack of evidence: Experimental result is not simulating enough to be convinced that this setting has indeed brought any advantage. It would make sense to show result on more downstream application (e.g., change detection or time-series analysis where some of the optical observations are missing and are filled by synthesized SAR observations and it improves the result).

Response: We thank the reviewer for the valuable comments. We are sorry that we are failed to represent our experimental results. In fact, SAR-optical image matching is already a well downstream application to measure SAR-to-optical and optical-to-SAR image translation methods. Specifically, as we can see in the Fig. 5 in our manuscript, it's hard to measure the image quality of synthesized images by a single metric. However, if an image-to-image translation method can synthesize clear images and assign correct gray values, the SAR-optical image matching based on this method should achieve better results and vice versa.

We highlight that SAR-optical image matching is a well downstream application to measure SAR-to-optical and optical-to-SAR image translation methods in the revised manuscript. We except the major revisions as follows.

In the Section IV.-D. *Comparison of Image Matching*,

"...Hence, SAR-optical image matching can also be used as a downstream application to measure SAR-to-optical and optical-to-SAR image translation methods..."

References

[1] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, pages 5967–5976, 2017.

[2] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, pages 2242–2251, 2017.

Response to Reviewer 2

The authors would like to express sincere appreciations to the reviewer for the thorough and constructive comments. We present a detailed response to the individual points raised by the reviewer. In this reply, we indicate the revised parts with blue color when directly quote them from the revised manuscript.

> **Concern # 0:** The authors propose an algorithm for SAR-Optical image matching relying on a combination of Pix2pix and CycleGAN. The first one being a supervised method and the second one being an unsupervised algorithm, the combination of the two results in a semi-supervised training strategy. Given the wide amount of unaligned SAR-optical image data, being able to exploit them in order to complement aligned data has a huge potential and is a promising research direction.

Response: We thank the reviewer for the valuable comments that really improve the quality of our manuscript.

We next present a point-by-point response as follows.

> **Concern # 1:** However, the contribution of this article is not clear. The authors state that this is the first semi supervised algorithm for SAR-optical image matching, reducing the need for aligned data. Nevertheless, this thesis is not supported by experiments. Moreover, in the literature, a semi-supervised strategy for SAR-optical image matching already exists: see Hughes, L. H., and M. Schmitt. "A semi-supervised approach to SAR-optical image matching." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4 (2019): 71-78.
>
> Therefore, I recommend major revisions. Detailed comment on the submitted manuscript are given below.

Response: We thank the reviewer for the valuable comments. We are sorry that we failed to represent our contributions in the previous manuscript. In fact, we didn't mention that we are the first semi-supervised algorithm for SAR-optical image matching. To the best of our knowledge, we are just the first to develop a semi-supervised **image-to-image translation framework** on the SAR-optical image matching task.

In addition, our method is totally different from the method proposed by Hughes et al. [5]. Specifically, they develop a network to determine whether a pair of SAR and optical images are matched or not. Hence, their method belongs to a deep learning based classification task. However, our method belongs to a deep learning based image-to-image translation task.

In the revised manuscript, we highlight that we only focus on the SAR-optical image matching methods based on image-to-image translation. We hope this could address your concern. We excerpt the major revisions as follows.

In the second paragraph of Section I. Introduction,

> "...Deep learning emerges as a powerful tool for SAR-optical image matching. Recent researches

can be roughly classified into two categories: 1. construct learning-based metrics; 2. eliminate the difference between SAR and optical images. The former tries to develop learning-based metrics to replace the handcraft feature descriptors or similarity metrics [1, 6, 5]. The latter tries to unify the textures between SAR and optical images by image-to-image translation [9, 4, 2, 3] or style-transfer methods [12], such that the SAR-optical image matching is transformed into SAR-SAR or optical-optical matchings (Stage 2 in Fig. 1). Then the well-designed single-mode image matching methods, e.g., the Scale-Invariant Feature Transform (SIFT) [8] method, could be applied to the transformed single-mode matching (Stage 3 in Fig. 1). In this work, we focus on the SAR-optical image matching methods based on image-to-image translation..."

Concern # 2: The introduction paragraph is not actually providing an introduction to the subject of image matching: a list of references is given, without pointing at the differences between existing approaches. This looks more like a list to me. For instance, when mentioning that "Then, the optimization, opportunities, and limits of using cGAN-based CycleGAN [3] for SAR-optical image translation on remote sensing tasks have been comprehensively analyzed in [4]", you are not pointing at these limitations. When citing unsupervised learning, the authors could give a brief explanation to the reader on how this is applied to the task of SAR-optical image matching.

Response: We thank the reviewer for the constructive comments. We are sorry that we failed to represent the references in the introduction. We rewrite most of the introduction in the revised manuscript. We excerpt the major revisions as follows.

In the Section I. Introduction,

"...The potential of conditional generative adversarial networks (cGANs) for SAR-optical image matching has been explored in [10]. Their method outperforms the well-known single-mode matching methods in SAR-optical image matching. Then, Ref. [4] proposed a cGAN-based CycleGAN trained with aligned data. They concluded that the results derived by CycleGAN is better for rural/semi-urban areas but worse for urban areas. Meanwhile, the assigned gray values by CycleGAN might lead to false conclusions. To obtain better results in urban areas, Ref. [2] enhanced cGAN by segmentation constraint. Nevertheless, training supervised learning methods need large collections of aligned SAR-optical images, and creating these large aligned datasets requires a great deal of engineering effort [11]

Unsupervised learning can be trained with only unaligned images. Ref. [3] applied feature matching loss in CycleGAN to enforce feature matching consistency. They obtained 2.6 times more qualified SAR-optical matchings than CycleGAN. However, similar to CycleGAN, they still assigned wrong gray values to some features..."

Concern # 3: How to you set the hyper-parameters lambda 1 and 2 in equation 11? This should be stated.

Response: We thank the reviewer for the valuable comments. We set the cycle consistent loss's weight (λ_1) to 10, the same as the public implementation of CycleGAN [13]. We rough-tune the L1 loss's weight to 50, the half of the public implementation of Pix2pix [7]. The comparisons between settings of L1 loss of Pix2pix performed in an unconditional manner are shown in Tab. 1.

We show the results of Tab. 1 in our first submission. However, due to the space limitation of the letter, we remove the analysis of settings of hyper-parameters. We will upload the analysis to Github (*https://github.com/WenliangDu/Semi-I2I*) along with our source codes in the future. We add simple notations to the settings of hyper-parameters in the revised manuscript. We excerpt the major revisions as follows.

Table 1: NOQMs obtained by Pix2pix in different L1 losses.

Methods	Rural (300)	Semi-urban (300)	Urban (300)
Pix2pix (L1=50)	**86**	**70**	**59**
Pix2pix (L1=100)	74	59	45

In the Section III.-A. *Datasets and Set-Up*,

> "...We set the L1 loss's weight (λ_1) to 50 (half in Pix2pix [7]) and set the cycle consistent loss's weight (λ_2) to 10 (same to CycleGAN [13])..."

Concern # 4: Figure 3 and 4 would benefit by adding a reference image to check how accurate the translation is. Without reference image in the transformed domain, it is difficult to evaluate the quality of the generator.

Response: We thank the reviewer for the constructive comments. We are sorry that we failed to represent the reference images in Figures 3 and 4. We highlight the reference images in the revised manuscript. We excerpt the major revisions as follows.

In the Section III.-C. *Comparison of Image-to-Image Translation*,

> "...The three original aligned SAR-optical pairs (in the left-most column of Figures 3 and 4) are from rural, semi-urban, and urban areas, respectively..."

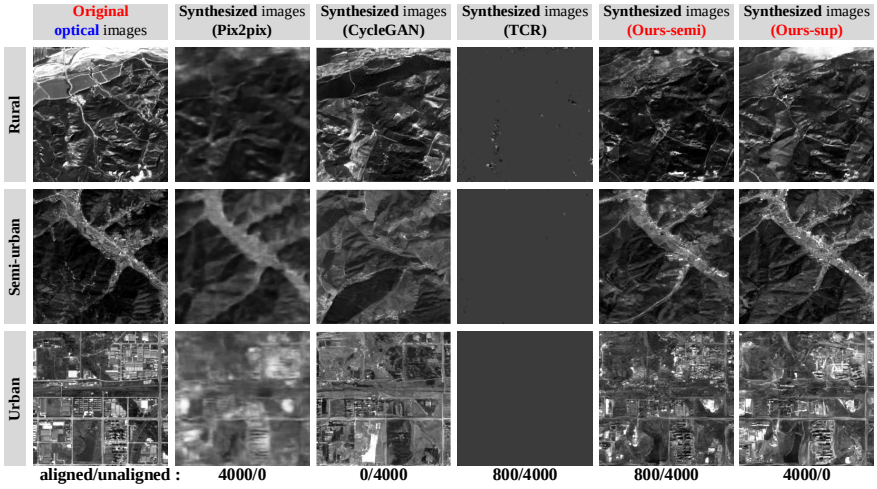

Figure 3: SAR-to-optical translation results of baselines and our framework.

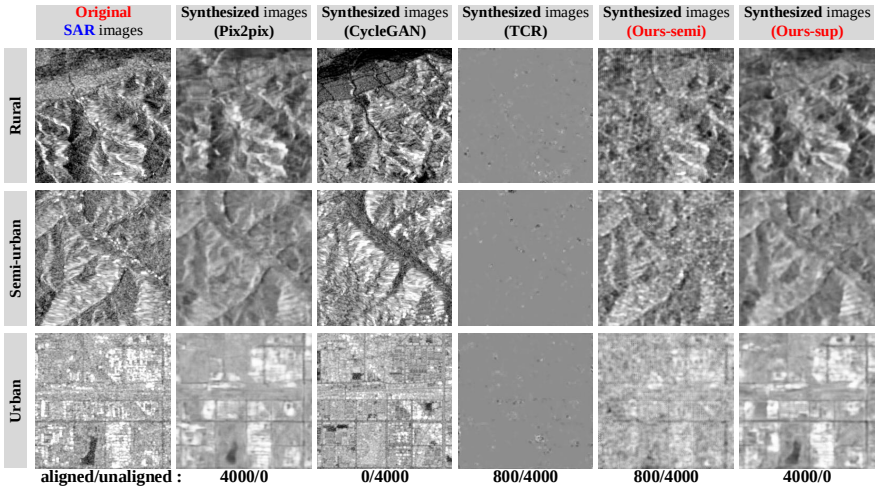

Figure 4: Optical-to-SAR translation results of baselines and our framework.

Concern # 5: The conclusions one can draw from Figure 5 are not clear. From the PSNR confidence interval graph, for optical image matching it seems that Pix2pix is performing better than SPC-GANs. Relying on PIQE, CycleGAN is performing better than both SPC-GAN training settings. There is a discrepancy between PSNR and PIQE that makes it difficult to objectively tell which method is best performing. Clear conclusions cannot be drawn, neither relying on the images in Figure 3 and 4, nor on the provided quality scores. You could discuss why PSNR and PIQE behaves differently. Which one is more reliable?

Response: We thank the reviewer for the valuable comments. Because it's hard to measure the image quality by a single metric, we introduce PSNR and PIQE to measure the quality of synthesized images. The higher PSNR means the synthesized images are more similar to the real (reference) images, while the lower PIQE indicates better perceptual quality. Hence, a image with high quality should achieve both high PSNR and low PIQE. We excerpt the major revisions as follows.

In the Section III.-C. *Comparison of Image-to-Image Translation*,

> "...We then provide the quantitative results of different methods in Figure 5. Because it's hard to measure the image quality by a single metric, we introduce a full-reference quality metric, i.e., Peak Signal-to-Noise Ratio (PSNR), and an no-reference image quality score, i.e., Perception-based Image Quality Evaluator (PIQE) to measure the quality of synthesized images. The higher PSNR means the synthesized images are more similar to the real (reference) images, while the lower PIQE indicates better perceptual quality. Hence, a image with high quality should achieve both high PSNR and low PIQE.
>
> Again only our framework, both the supervised and semi-supervised versions, can achieve high PSNR and low PIQE values simultaneously, indicating that our framework indeed takes advantages of both Pix2pix (high PSNR but high PIQE) and CycleGAN (low PIQE but low PSNR)..."

Concern # 6: The discussion around Figure 8 is not clear. It seems like only semi-urban ares benefit from a semi-supervised training with unaligned images. This would make the contribution too weak. On the other hand, the computation of NOQMs on all areas seems to indicate the the semi-supervised methods works better globally. The authors should clarify this point and draw conclusions confirmed

by the experiments.

Response: We thank the reviewer for the constructive comments. Thanks to the significant growth in semi-urban areas and the slight growth in urban areas, the results of NOQMs in all areas are significantly better than Pix2pix. In the revised manuscript, we describe our effects of in urban areas and our limitations in rural areas. We also limit our contribution in semi-urban and urban areas. We excerpt the major revisions as follows.

In the Section III.-D. *Comparison of Image Matching*,

> "...In figure 8, we evaluate our semi-supervised version on various amounts of unaligned training image pairs. As shown in the semi-urban area, the NOQMs of our framework trained with 4,000 unaligned data are 1.3 times more than the NOQMs trained with 400 unaligned data. However, the NOQMs show slight growth in the urban areas and are nearly unchanged in the rural areas, indicating the limitation of our framework in rural areas. These results also suggest that, compared to the fully supervised learning, adding a small number of unaligned image pairs in our framework (e.g., 10%) substantially improves the SAR-optical image matching in urban areas..."

In the last paragraph of Section I. Introduction,

> "...We extensively study the properties of the proposed framework on SAR-optical image matching task and achieve significant performance improvements in semi-urban and urban areas..."

Concern # 7: To complement the comparison, the authors could include methods specific to SAR-optical image matching such as: "Hughes, L. H., Marcos, D., Lobry, S., Tuia, D., and Schmitt, M. (2020). A deep learning framework for matching of SAR and optical imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 169, 166-179."

Response: We thank the reviewer for the valuable comments. We didn't compare with Ref. [6] in the revised manuscript, because the schemes between our method and Ref. [6] are different. Specifically, Ref. [6] develop an end-to-end architecture, constructed by three networks, to locate and match to-be-matched patches. While our method transforms the SAR-to-optical image matching into SAR-SAR and optical-optical matchings by image-to-image translation and then matches the SAR and optical images by SIFT. As a result, the type of training data is also different between Ref. [6] and our method.

We highlight the scope of our research in the revised manuscript. We hope this could address your concern. We excerpt the major revisions as follows.

In the second paragraph of Section I. Introduction,

> "...Deep learning emerges as a powerful tool for SAR-optical image matching. Recent researches can be roughly classified into two categories: 1. construct learning-based metrics; 2. eliminate the difference between SAR and optical images. The former tries to develop learning-based metrics to replace the handcraft feature descriptors or similarity metrics [1, 6, 5]. The latter tries to unify the textures between SAR and optical images by image-to-image translation [9, 4, 2, 3] or style-transfer methods [12], such that the SAR-optical image matching is transformed into SAR-SAR or optical-optical matchings (Stage 2 in Fig. 1). Then the well-designed single-mode image matching methods, e.g., the Scale-Invariant Feature Transform (SIFT) [8] method, could be applied to the transformed single-mode matching (Stage 3 in Fig. 1). In this work, we focus on the SAR-optical image matching methods based on image-to-image translation..."

References

[1] Song Cui, Ailong Ma, Liangpei Zhang, Miaozhong Xu, and Yanfei Zhong. Map-net: Sar and optical image matching via image-based convolutional network with attention mechanism and spatial pyramid aggregated pooling. *IEEE Trans. on Geosci. and Remote Sens.*, 60:1–13, 2022.

[2] W.-L. Du, Y. Zhou, J. Zhao, and X. Tian. K-means clustering guided generative adversarial networks for sar-optical image matching. *IEEE Access*, 8:217554–217572, 2020.

[3] Wen-Liang Du, Yong Zhou, Jiaqi Zhao, Xiaolin Tian, Zhi Yang, and Fuqiang Bian. Exploring the potential of unsupervised image synthesis for sar-optical image matching. *IEEE Access*, 9:71022–71033, 2021.

[4] Mario Fuentes Reyes, Stefan Auer, Nina Merkle, Corentin Henry, and Michael Schmitt. Sar-to-optical image translation based on conditional generative adversarial networksoptimization, opportunities and limits. *Remote Sens.*, 11(17), 2019.

[5] Lloyd H. Hughes and Michael Schmitt. A semi-supervised approach to sar-optical image matching. *in ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2019.

[6] Lloyd Haydn Hughes, Diego Marcos, Sylvain Lobry, Devis Tuia, and Michael Schmitt. A deep learning framework for matching of sar and optical imagery. *ISPRS J. of Photogramm. and Remote Sens.*, 169:166–179, 2020.

[7] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, pages 5967–5976, 2017.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Comput. Vis.*, 60(2):91–110, 2004.

[9] N. Merkle, S. Auer, R. Mller, and P. Reinartz. Exploring the potential of conditional adversarial networks for optical and sar image matching. *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, 11(6):1811–1820, June 2018.

[10] N. Merkle, S. Auer, R. Mller, and P. Reinartz. Exploring the potential of conditional adversarial networks for optical and sar image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1811–1820, June 2018.

[11] M. Schmitt, L. H. Hughes, and X. X. Zhu. The sen1-2 dataset for deep learning in sar-optical data fusion. *in ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-1:141–146, 2018.

[12] J. Zhang, W. Ma, Y. Wu, and L. Jiao. Multimodal remote sensing image registration based on image transfer and local features. *IEEE Geosci. Remote Sens. Lett.*, 16(8):1210–1214, Aug 2019.

[13] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, pages 2242–2251, 2017.

Response to Reviewer 3

The authors would like to express sincere appreciations to the reviewer for the thorough and constructive comments. We present a detailed response to the individual points raised by the reviewer. In this reply, we indicate the revised parts with blue color when directly quote them from the revised manuscript.

> **Concern # 0:** Compared to the previous submission, the authors have made substantial revisions based on the reviewers' comments.

Response: Many thanks for your positive and valuable comments that really improve the quality of our manuscript. We next present a point-by-point response as follows.

> **Concern # 1:** In the first-round review, the reviewer points out the novelty of this work is weak. Unfortunately, in the reply, the author does not defense the innovation of the work, but mentions the results of the method again, as well as the newly added comparison method.

Response: We thank the reviewer for the constructive comments. The novelty of this work is that, to the best of our knowledge, we are the first to develop semi-supervised image-to-image translation framework on SAR-optical image matching task. Besides, our framework indeed combines the benefits of Pix2pix (correct gray value assignment) and CycleGAN (good edge-preserving), and avoids the disadvantage of the two methods (i.e., Pix2pix obtains blur results; CycleGAN assigns wrong gray value to some features). Meanwhile, our framework is simple yet effective.

In the revised manuscript, we introduce more contributions compared to Pix2pix and CycleGAN and highlight that our framework is simple yet effective. We excerpt the major revisions as follows.

In Section-I. Introduction,

> "...In this work, we propose a simple yet effective framework for semi-supervised image-to-image translation, which merges two well-known supervised and unsupervised image-to-image translation methods, i.e., Pix2pix [1] and CycleGAN [2]. Our framework combines the benefits of Pix2pix (correct gray value assignment) and CycleGAN (good edge-preserving), and avoids the disadvantage of the two methods (i.e., Pix2pix obtains blur results; CycleGAN assigns wrong gray value to some features)..."

In Abstract,

> "...To this end, we combine the benefits of both supervised and unsupervised well-known image-to-image translation methods, i.e., Pix2pix and CycleGAN, and propose a simple yet effective semi-supervised image-to-image translation framework..."

In Section-I. Introduction,

> "...In this work, we propose a simple yet effective framework for semi-supervised image-to-image translation, which merges two well-known supervised and unsupervised image-to-image translation methods, i.e., Pix2pix [1] and CycleGAN [2]..."

In Section-IV. Conclusion,

> "...In this work, we proposed a simple yet effective semi-supervised image-to-image translation framework integrated by Pix2pix and CycleGAN...."

Concern # 2: It may be unfair for TCR which is chose as a comparison method. After all, TCR directly performs cross-domain translation, while the proposed method converts the cross-domain translation into the co-domain translation. The methods are completely different in nature.

Response: We thank the reviewer for the valuable comments. We agree with the reviewer that our framework has co-domain losses (cycle consistency loss), which TCR hasn't. Hence, we think this is also an advantage of our framework compared TCR.

In addition, SAR-optical image matching results in this work are based on both SAR-SAR and optical-optical matchings, i.e., TCR is trained twice for a certain experiment to obtain SAR-to-optical and optical-to-SAR translation models. In contrast, our framework is trained once a time to get the two translation models. Hence, we think the comparison between our framework and TCR is fair in our manuscript.

In the revised manuscript, we highlight the fairness of the comparison. We excerpt the major revisions as follows.

In Section-III-D. Comparison of Image Matching,

> "...Note that because SAR-optical image matching results in this work are based on both SAR-SAR and optical-optical matchings, Pix2pix and TCR are trained twice to obtain SAR-to-optical and optical-to-SAR translations...."

Concern # 3: Finally, this is not a huge problem, but the reviewer still wants to bring the authors to their attention. About the abbreviation (SPC-GAN) to the name of the proposed method, i.e., Semi-supervised-Pix2pix-Cycle-GAN. Although it is difficult to give a more appropriate name, I still think the current name is inappropriate. From the point of view of word formation, "Semi-supervised" denotes the learning strategy, that is to say, it is neither fully supervised nor fully unsupervised, but both which is the combination of a supervised method Pix2pix and an unsupervised method "CycleGAN". The hyphen between SPC and GAN tends to confuse readers into thinking that this is a new GAN method, namely SPC GAN, which is obviously not what the authors expect.

Response: We thank the reviewer for the constructive comments. We remove all the "SPC-GAN" from the revised manuscript. Please refer to our revised manuscript (the highlighted version) for more details.

Concern # 4: "number of qualied matchings" in the text of section III.D, it is written as NOQMs, while in Fig.7 and Fig.8, it is NQMs.

The use of parentheses is confusing between line10 to line 13 in page 2. Please recheck and correct them.

Response: We thank the reviewer for the constructive comments. We correct all the "NQMs" to "NOQMs" in the revised manuscript, and we correct the parentheses in page 2. We excerpt the major revisions as follows.

In Section II. Method,

> "...($\mathcal{A}_{\mathcal{X}} = \{\mathcal{A}_x^k\}_{k=1}^K \in \mathcal{X}$, $\mathcal{A}_{\mathcal{Y}} = \{\mathcal{A}_y^k\}_{k=1}^K \in \mathcal{Y}$) while given unaligned M SAR ($\mathcal{U}_{\mathcal{X}} = \{\mathcal{U}_x^i\}_{i=1}^M \in \mathcal{X}$) and N optical image pairs ($\mathcal{U}_{\mathcal{Y}} = \{\mathcal{U}_y^j\}_{j=1}^N \in \mathcal{Y}$). Hence, in each training iteration of the *semi-supervised version*, a pair of aligned and unaligned data are fed into the unsupervised and supervised modules, respectively..."

In Figures 7 and 8,

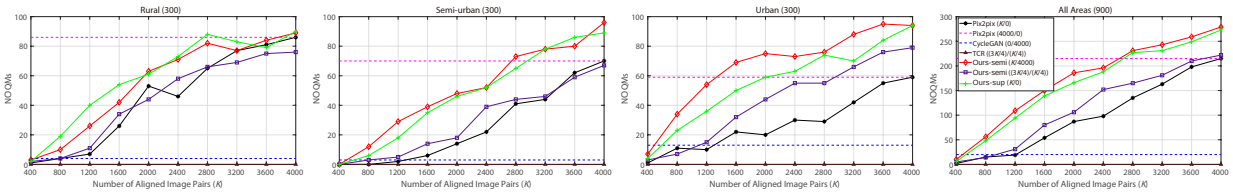

Figure 7: NOQMs for each method on 900 test image pairs as the amount of aligned training data varies.

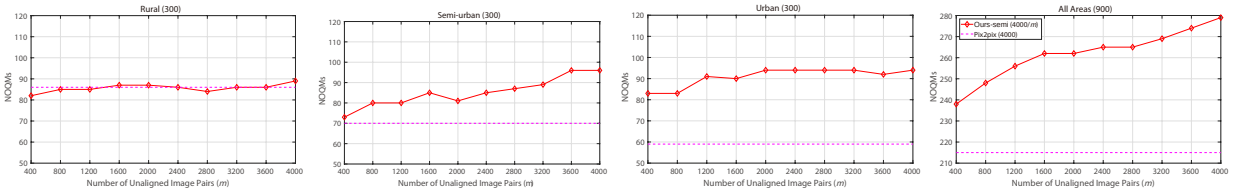

Figure 8: NOQMs for our framework on 900 test image pairs as the amounts of unaligned training data varies.

References

[1] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, pages 5967–5976, 2017.

[2] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV*, pages 2242–2251, 2017.

A Semi-Supervised Image-to-Image Translation Framework for SAR-Optical Image Matching

Wen-Liang Du, *Member, IEEE,* Yong Zhou, Hancheng Zhu, Jiaqi Zhao, *Member, IEEE,* Zhiwen Shao, and Xiaolin Tian

***Abstract*—Synthetic Aperture Radar (SAR) and optical image matching aims to acquire correspondences from a certain pair of SAR and optical images. Recent advances in the image-to-image translation provided a way to simplify the SAR-optical image matching into the SAR-SAR or optical-optical image matchings. Existing image-to-image translations mainly focus on supervised or unsupervised learning. However, gathering sufficient amounts of aligned training data for supervised learning is challenging, while unsupervised learning cannot guarantee enough correct correspondences. In this work, we investigate the applicability of semi-supervised image-to-image translation for SAR-optical image matching such that both aligned and unaligned SAR-optical images could be used. To this end, we combine the benefits of both supervised and unsupervised well-known image-to-image translation methods, i.e., Pix2pix and CycleGAN, and propose a simple yet effective semi-supervised image-to-image translation framework. Through extensive experimental comparisons to baseline methods, we verify the effectiveness of the proposed framework in both semi-supervised and fully-supervised settings. Our codes are available at *https://github.com/WenliangDu/Semi-I2I*.**

***Index Terms*—Image matching, semi-supervised-image-synthesis, synthetic aperture radar (SAR), generative adversarial networks (GANs).**

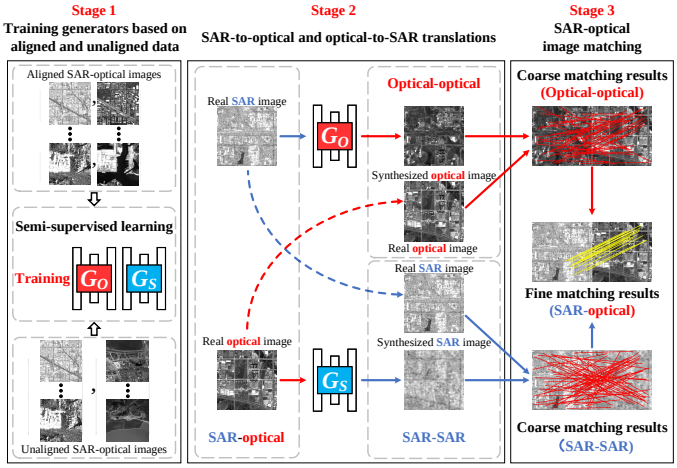

Fig. 1. SAR-optical image matching based on semi-supervised image-to-image translation. G_O and G_S represent generators of SAR-to-optical and optical-to-SAR translations, respectively. Red and yellow lines in the Stage 3 represent correspondences derived from coarse and fine matching.

I. Introduction

SYNTHETIC Aperture Radar (SAR) and optical (SAR-optical) image matching is the task of establishing correspondences between SAR and optical images of the same scene. The quality of SAR-optical image matching will affect the accuracy of subsequent remote sensing applications such as image registration, image fusion, and etc. Nevertheless, the SAR-optical image matching remains challenging because of the significant global geometric distortions and non-linear intensity differences between SAR and optical images.

Wen-Liang Du, Yong Zhou, Hancheng Zhu, Jiaqi Zhao, and Zhiwen Shao are with the School of Computer Science and Technology, China University of Mining and Technology, 221116, Xuzhou, China, and they are also with Engineering Research Center of Mine Digitization, Ministry of Education of the Peoples Republic of China, Xuzhou 221116, China (email: wldu@cumt.edu.cn, yzhou@cumt.edu.cn, zhuhancheng@cumt.edu.cn, jiaqizhao@cumt.edu.cn, and zhiwen_shao@cumt.edu.cn).

Xiaolin Tian is with State Key Laboratory of Lunar and Planetary Sciences, Macau University of Science and Technology, Macau, China (email: xltian@must.edu.mo).

Manuscript received XXXX XX, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62002360, 62101555, 61806206, and 62106268; in part by the Science and Technology Development Fund of Macau under Grant 0038/2020/A1; in part by the opening fund of State Key Laboratory of Lunar and Planetary Sciences (Macau University of Science and Technology) (Macau FDCT grant No. 119/2017/A3), and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20201346 and BK20210488 (*Corresponding author: Yong Zhou*).

Deep learning emerges as a powerful tool for SAR-optical image matching. Recent researches can be roughly classified into two categories: 1. construct learning-based metrics; 2. eliminate the difference between SAR and optical images. The former tries to develop learning-based metrics to replace the handcraft feature descriptors or similarity metrics [1]–[3]. The latter tries to unify the textures between SAR and optical images by image-to-image translation [4]–[7] or style-transfer methods [8], such that the SAR-optical image matching is transformed into SAR-SAR or optical-optical matchings (Stage 2 in Fig. 1). Then the well-designed single-mode image matching methods, e.g., the Scale-Invariant Feature Transform (SIFT) [9] method, could be applied to the transformed single-mode matching (Stage 3 in Fig. 1). In this work, we focus on the SAR-optical image matching methods based on image-to-image translation.

Lately, extensive research has leveraged supervised and unsupervised image-to-image translations for SAR-optical image matching. The potential of conditional generative adversarial networks (cGANs) for SAR-optical image matching has been explored in [10]. Their method outperforms the well-known single-mode matching methods in SAR-optical image matching. Then, Ref. [5] proposed a cGAN-based CycleGAN trained with aligned data. They concluded that the results derived by CycleGAN is better for rural/semi-urban areas but worse for urban areas. Meanwhile, the assigned gray values by Cycle-

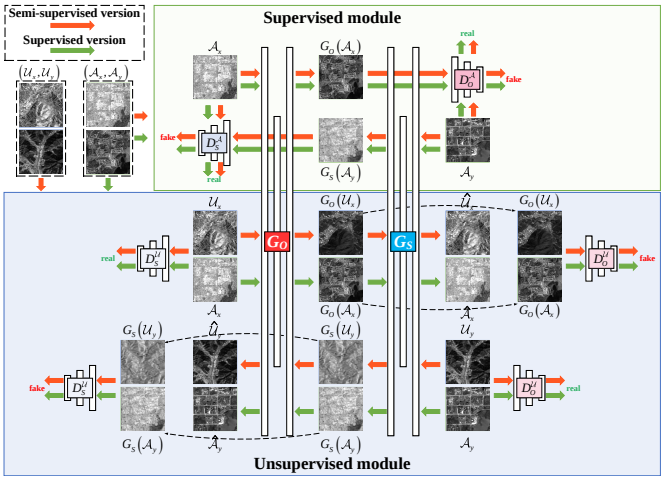

Fig. 2. An illustration of a training iteration of our framework.

GAN might lead to false conclusions. To obtain better results in urban areas, Ref. [6] enhanced cGAN by segmentation constraint. Nevertheless, training supervised learning methods need large collections of aligned SAR-optical images, and creating these large aligned datasets requires a great deal of engineering effort [11].

Unsupervised learning can be trained with only unaligned images. Ref. [7] applied feature matching loss in CycleGAN to enforce feature matching consistency. They obtained 2.6 times more qualified SAR-optical matchings than CycleGAN. However, similar to CycleGAN, they still assigned wrong gray values to some features.

Semi-supervised learning trades off the supervised and unsupervised results by training on both aligned and unaligned images. Ref. [12] introduced transformation consistency regularization in a semi-supervised image-to-image translation model to enforce the model's predictions for unlabeled data to be invariant to a diverse set of geometric transformations. However, image-to-image translation remains unexplored by semi-supervised learning methods [12].

In this work, we propose a simple yet effective framework for semi-supervised image-to-image translation, which merges two well-known supervised and unsupervised image-to-image translation methods, i.e., Pix2pix [13] and CycleGAN [14]. Our framework combines the benefits of Pix2pix (correct gray value assignment) and CycleGAN (good edge-preserving), and avoids the disadvantage of the two methods (i.e., Pix2pix obtains blur results; CycleGAN assigns wrong gray value to some features). Our main contributions are as follows:

1) To the best of our knowledge, we are the first to develop a semi-supervised image-to-image translation framework for the SAR-optical image matching task, reducing the amount of required aligned training data.
2) We extensively study the properties of the proposed framework on SAR-optical image matching task and achieve significant performance improvements in semi-urban and urban areas.

II. Method

Our framework consists of two modules: a supervised and an unsupervised module (Figure 2). The supervised module can be only trained with aligned image pairs ($\mathcal{A}$), while the unsupervised module can be trained with either unaligned ($\mathcal{U}$) or aligned image pairs ($\mathcal{A}$). As a result, our framework can be trained in two versions: *the semi-supervised version*, training with both aligned and unaligned image pairs, and *the supervised version*, training with only aligned image pairs.

For *the semi-supervised version*, our goal is to learn two mapping functions between SAR ($\mathcal{X}$) and optical ($\mathcal{Y}$) images given K aligned SAR-optical image pairs ($\mathcal{A}_{\mathcal{X}} = \left\{\mathcal{A}_x^k\right\}_{k=1}^{K} \in \mathcal{X}$, $\mathcal{A}_{\mathcal{Y}} = \left\{\mathcal{A}_y^k\right\}_{k=1}^{K} \in \mathcal{Y}$) while given unaligned M SAR ($\mathcal{U}_{\mathcal{X}} = \left\{\mathcal{U}_x^i\right\}_{i=1}^{M} \in \mathcal{X}$) and N optical image pairs ($\mathcal{U}_{\mathcal{Y}} = \left\{\mathcal{U}_y^j\right\}_{j=1}^{N} \in \mathcal{Y}$). Hence, in each training iteration of the *semi-supervised version*, a pair of aligned and unaligned data are fed into the unsupervised and supervised modules, respectively.

While for *the supervised version*, only K aligned SAR-optical image pairs ($\mathcal{A}_{\mathcal{X}}$, $\mathcal{A}_{\mathcal{Y}}$) are given to learn the two mapping functions. Hence, in each training iteration of *the supervised version*, the unsupervised and supervised modules are fed by the same pair of aligned data.

We use G_O and G_S to represent SAR-to-optical ($\mathcal{X} \rightarrow \mathcal{Y}$) and optical-to-SAR ($\mathcal{Y} \rightarrow \mathcal{X}$) mappings, respectively. For both of the two versions of our framework, the supervised and unsupervised modules share the same G_O and G_S, and the two generators are trained simultaneously.

A. Supervised Module

We create the supervised module based on the Pix2pix framework [13] but in an unconditional manner. Hence the supervised objective contains two terms: *adversarial losses* and *L1 loss*. Recall that the supervised module is only trained with aligned image pairs ($\mathcal{A}_{\mathcal{X}}$, $\mathcal{A}_{\mathcal{Y}}$) in both supervised and semi-supervised version of our framework.

1) Adversarial Loss in Supervised Module: The adversarial loss shows the key idea of GAN—training a generator and a discriminator in a minimax two-player game. The adversarial loss of $\mathcal{A}_{\mathcal{X}} \rightarrow \mathcal{A}_{\mathcal{Y}}$ is expressed as:

$$\begin{aligned}\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{A}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) =& \mathbb{E}_{\mathcal{A}_{\mathcal{Y}}}\left[\log D_O^{\mathcal{A}}\left(\mathcal{A}_{\mathcal{Y}}\right)\right] + \\ & \mathbb{E}_{\mathcal{A}_{\mathcal{X}}}\left[\log\left(1 - D_O^{\mathcal{A}}\left(G_O\left(\mathcal{A}_{\mathcal{X}}\right)\right)\right)\right],\end{aligned} \tag{1}$$

where $D_O^{\mathcal{A}}$ learns to differentiate real and synthesized optical images from aligned data; $\mathbb{E}_{(\cdot)} \triangleq \mathbb{E}_{(\cdot)\sim p_{\text{data}}(\cdot)}$ and $\mathbb{E}_{(\cdot)\sim p_{\text{data}}(\cdot)}\left[f\left(\cdot\right)\right]$ returns the expectation of $f(\cdot)$ with respect to the data-generating distribution $p_{\text{data}}(\cdot)$. Similarly, the adversarial loss of $\mathcal{A}_{\mathcal{Y}} \rightarrow \mathcal{A}_{\mathcal{X}}$ is introduced as $\mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{A}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right)$, where $D_S^{\mathcal{A}}$ learns to differentiate real and synthesized SAR images from aligned data.

2) L1 Loss: We introduce L1 loss to guarantee that the synthesized images generate similar content to the corresponding real images. The L1 loss is expressed as:

$$\begin{aligned}\mathcal{L}_{\text{L1}} =& \mathbb{E}_{(\mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}})}\left[\left\|\mathcal{A}_{\mathcal{Y}} - G_O(\mathcal{A}_{\mathcal{X}})\right\|_1\right] + \\ & \mathbb{E}_{(\mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}})}\left[\left\|\mathcal{A}_{\mathcal{X}} - G_S(\mathcal{A}_{\mathcal{Y}})\right\|_1\right].\end{aligned} \tag{2}$$

B. Unsupervised Module

Our unsupervised module is built based on the CycleGAN framework [14]. Therefore, the unsupervised objective contains two terms: *adversarial losses* and *cycle consistency loss*. In addition, each loss has two types according to the semi-supervised and supervised versions of our framework. Recall that the unsupervised module is trained with unaligned image pairs ($\mathcal{U}_{\mathcal{X}}$, $\mathcal{U}_{\mathcal{Y}}$) in *the semi-supervised version* or trained with aligned image pairs ($\mathcal{A}_{\mathcal{X}}$, $\mathcal{A}_{\mathcal{Y}}$) in *the supervised version*.

1) Adversarial Loss in Unsupervised Module: For *the semi-supervised version*, the adversarial losses of $\mathcal{U}_{\mathcal{X}} \rightarrow \mathcal{U}_{\mathcal{Y}}$ and $\mathcal{U}_{\mathcal{Y}} \rightarrow \mathcal{U}_{\mathcal{X}}$ are introduced as: $\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{U}_{\mathcal{X}}, \mathcal{U}_{\mathcal{Y}}\right)$ and $\mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{U}_{\mathcal{Y}}, \mathcal{U}_{\mathcal{X}}\right)$, where $D_O^{\mathcal{U}}$ and $D_S^{\mathcal{U}}$ discriminate real and synthesized images, and G_O and G_S are the same as the generators in the supervised module.

The adversarial losses in *the supervised version* of unsupervised module are introduced as: $\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right)$ and $\mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right)$. The similarities between these two adversarial losses and the adversarial losses in supervised module are that they are all performed on the same aligned image pairs ($\mathcal{A}_{\mathcal{X}}$, $\mathcal{A}_{\mathcal{Y}}$), and they share the same two generators. The different is that they use different discriminators.

2) Cycle Consistency Loss: We adopt the cycle consistent loss to favor $\hat{\mathcal{U}}_{\mathcal{X}} \approx \mathcal{U}_{\mathcal{X}}$, $\hat{\mathcal{U}}_{\mathcal{Y}} \approx \mathcal{U}_{\mathcal{Y}}$, $\hat{\mathcal{A}}_{\mathcal{X}} \approx \mathcal{A}_{\mathcal{X}}$, and $\hat{\mathcal{A}}_{\mathcal{Y}} \approx \mathcal{A}_{\mathcal{Y}}$, where $(\cdot)\hat{}_{\mathcal{X}}$ and $(\cdot)\hat{}_{\mathcal{Y}}$ represent the reconstructed SAR and optical images respectively. More specifically, $(\cdot)\hat{}_{\mathcal{X}} = G_S(G_O((\cdot)_{\mathcal{X}}))$ and $(\cdot)\hat{}_{\mathcal{Y}} = G_O(G_S((\cdot)_{\mathcal{Y}}))$.

The cycle consistent losses for the *semi-supervised* and *supervised* versions are respectively expressed as:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\left\|\mathcal{U}_{\mathcal{X}} - \hat{\mathcal{U}}_{\mathcal{X}}\right\|_1\right] + \mathbb{E}_{\mathcal{U}_{\mathcal{Y}}}\left[\left\|\mathcal{U}_{\mathcal{Y}} - \hat{\mathcal{U}}_{\mathcal{Y}}\right\|_1\right], \quad (3)$$

$$\mathcal{L}_{\text{cyc}_{\text{sup}}} = \mathbb{E}_{\mathcal{A}_{\mathcal{X}}}\left[\left\|\mathcal{A}_{\mathcal{X}} - \hat{\mathcal{A}}_{\mathcal{X}}\right\|_1\right] + \mathbb{E}_{\mathcal{A}_{\mathcal{Y}}}\left[\left\|\mathcal{A}_{\mathcal{Y}} - \hat{\mathcal{A}}_{\mathcal{Y}}\right\|_1\right]. \quad (4)$$

C. Full Objective

The full objectives of the *semi-supervised* and *supervised* versions of our framework are respectively expressed as:

$$\begin{aligned}&\mathcal{L}_{\text{semi}}\left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right) = \lambda_1\mathcal{L}_{\text{L1}} + \lambda_2\mathcal{L}_{\text{cyc}} + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{A}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{A}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right) + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{U}_{\mathcal{X}}, \mathcal{U}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{U}_{\mathcal{Y}}, \mathcal{U}_{\mathcal{X}}\right) \quad (5)\end{aligned}$$

$$\begin{aligned}&\mathcal{L}_{\text{sup}}\left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right) = \lambda_1\mathcal{L}_{\text{L1}} + \lambda_2\mathcal{L}_{\text{cyc}_{\text{sup}}} + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{A}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{A}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right) + \\ &\mathcal{L}_{\text{adv}}\left(G_O, D_O^{\mathcal{U}}, \mathcal{A}_{\mathcal{X}}, \mathcal{A}_{\mathcal{Y}}\right) + \mathcal{L}_{\text{adv}}\left(G_S, D_S^{\mathcal{U}}, \mathcal{A}_{\mathcal{Y}}, \mathcal{A}_{\mathcal{X}}\right) \quad (6)\end{aligned}$$

where λ_1 and λ_2 are hyper-parameters that control the relative importance of the L1 loss and cycle consistence, respectively. The goals of the two versions are to solve:

$$\min_{(G_O, G_S)} \max_{(D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}})} \mathcal{L}_{\text{semi}}\left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right), \quad (7)$$

$$\min_{(G_O, G_S)} \max_{(D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}})} \mathcal{L}_{\text{sup}}\left(G_O, G_S, D_O^{\mathcal{A}}, D_S^{\mathcal{A}}, D_O^{\mathcal{U}}, D_S^{\mathcal{U}}\right). \quad (8)$$

D. Network Architecture and Implementation Details

We use the same generator and discriminator architectures as those in CycleGAN. Specifically, our discriminator networks are used as 70×70 PatchGANs; we use LSGAN for stable training; we use the Adam solver to train our framework and baselines from scratch.

E. SAR-Optical Image Matching

Based on the G_O and G_S trained by our framework, the original SAR-optical image matching is transferred into SAR-SAR and optical-optical image matchings (see Figure 1). The coarse matching results of the two single-mode image matchings are shared and then the fine matching (eliminate false matches) is performed on the shared coarse matching results [7].

III. Experimental Results and Analysis

A. Datasets and Set-Up

We train all the models on the SEN1-2 dataset [11] (aligned SAR-optical image pairs in 256×256 pixels), choosing 4,000 and 900 aligned SAR-optical image pairs for conducting aligned training and test datasets, respectively. It worthy noting that the test dataset cover rural, semi-urban, and urban areas, where each area occupies 300 test image pairs. We also construct an unaligned training dataset by 4,000 unaligned SAR-optical image pairs. In particular, SAR images in the unaligned training dataset are the same as the SAR images in the aligned training dataset; optical images in the unaligned training dataset are chosen as the optical images that are not aligned to the corresponding SAR images.

We set the L1 loss's weight (λ_1) to 50 (half in Pix2pix [13]) and set the cycle consistent loss's weight (λ_2) to 10 (same to CycleGAN [14]). All the experiments are performed on an NVIDIA GeForce RTX 3080Ti GPU, and the GPU takes 34 hours to train 200 epochs of our framework on 4,000 aligned and 4,000 unaligned SAR-optical image pairs.

B. Supervised, Unsupervised, and Semi-Supervised Baselines

We compare our our framework with Pix2pix and CycleGAN, and with a state-of-the-art semi-supervised method: Transformation Consistency Regularization (TCR) [12].

For a fair comparison, we implement Pix2pix and CycleGAN using the same network architecture and details as our framework. We use the public implementation of TCR due to fundamental differences in architecture.

C. Comparison of Image-to-Image Translation

We firstly show three qualitative examples comparing our results with baselines for SAR-to-optical and optical-to-SAR translations in Figures 3 and 4. The three original aligned SAR-optical pairs (in the left-most column of Figures 3 and 4) are from rural, semi-urban, and urban areas, respectively. As for the three baselines, Pix2pix produced blurry results; CycleGAN assigned wrong gray values (see the water in the bottom of urban examples); TCR is failed in both SAR-to-optical and optical-to-SAR translations, indicating that TCR

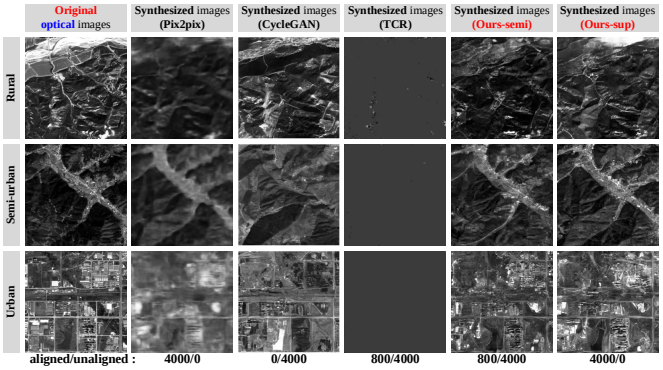

Fig. 3. SAR-to-optical translation results of baselines and our framework.

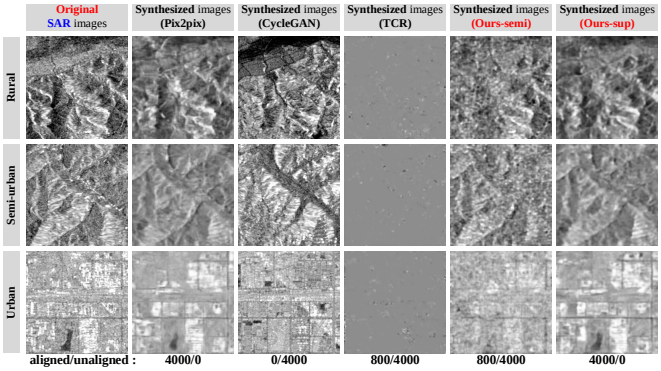

Fig. 4. Optical-to-SAR translation results of baselines and our framework.

does not apply to the translation between domains that have excessive differences. In contrast, our framework synthesized clear images and with the correct gray values by only 20% aligned training image pairs. Although the supervised version of our framework produced a few blurry results, the results are still much clearer than Pix2pix produced.

We then provide the quantitative results of different methods in Figure 5. Because it's hard to measure the image quality by a single metric, we introduce a full-reference quality metric, i.e., Peak Signal-to-Noise Ratio (PSNR), and an no-reference image quality score, i.e., Perception-based Image Quality Evaluator (PIQE) to measure the quality of synthesized images. The higher PSNR means the synthesized images are more similar to the real (reference) images, while the lower PIQE indicates better perceptual quality. Hence, a image with high quality should achieve both high PSNR and low PIQE.

Again only our framework, both the supervised and semi-supervised versions, can achieve high PSNR and low PIQE values simultaneously, indicating that our framework indeed takes advantages of both Pix2pix (high PSNR but high PIQE) and CycleGAN (low PIQE but low PSNR).

D. Comparison of Image Matching

We firstly illustrate SAR-optical coarse image matching results of baselines and our framework on the former three examples in Figure 6. In particular, the feature points are extracted by SIFT method, and the coarse matching is done by ratio method [9]. Because SAR-optical image pairs in

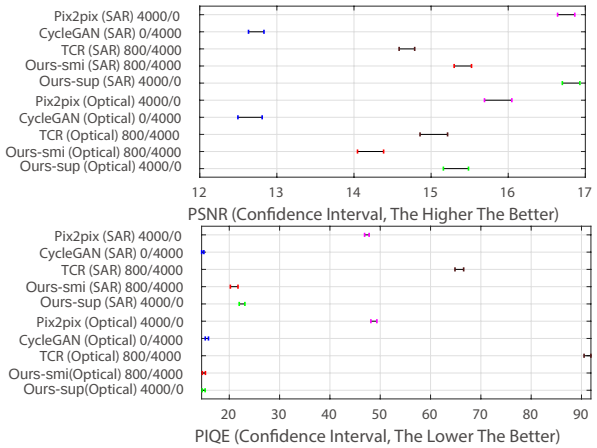

Fig. 5. Confidence intervals (confidence level is 95%) of PSNR and PIQE achieved by the baselines and our framework from the 900 test image pairs.

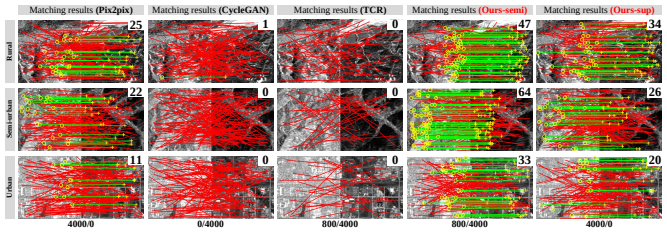

Fig. 6. SAR-optical coarse image matching results of baselines and our framework. Green and red lines represent correct and incorrect correspondences, respectively; yellow circles and crosses are the feature-points of correct correspondences; the digits on the top-right corner of each matching are the numbers of corresponding correct correspondences; all the SAR and optical images in presenting are the real SAR and optical images.

our test dataset are spatially aligned, we define the correct correspondence as the correspondence whose corresponding feature points' Euclidean distance is less than three pixels. Hence, SAR-optical image matching can also be used as a downstream application to measure SAR-to-optical and optical-to-SAR image translation methods.

Obviously, for the three examples, the semi-supervised version of our framework (trained by only 20% aligned image pairs) achieved the most correct correspondences, indicating that better fine matching results could be obtained.

We then provide quantitative comparisons against baselines using the 900 aligned test data. The quantitative evaluations are performed by the number of qualified matchings (NOQMs), where a qualified matching is a SAR-optical coarse image matching that contains eight or more correct correspondences. Note that because SAR-optical image matching results in this work are based on both SAR-SAR and optical-optical matchings, Pix2pix and TCR are trained twice to obtain SAR-to-optical and optical-to-SAR translations. Figures 7 shows NOQMs of different methods when varying the amount of aligned training image pairs. As for the three baselines, the NOQMs got by Pix2pix shows a linear growth trend when increasing the amount of aligned training image pairs; CycleGAN obtained very few NOQMs in the three areas; TCR obtained no NOQMs in all the experiments, because it failed in SAR-to-optical and optical-to-SAR translations.

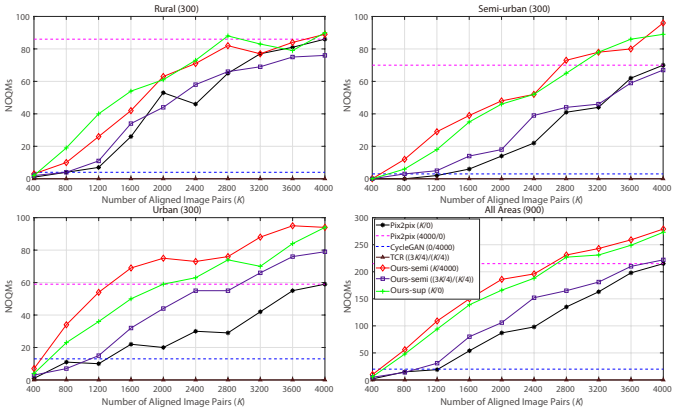

Fig. 7. NOQMs for each method on 900 test image pairs as the amount of aligned training data varies.

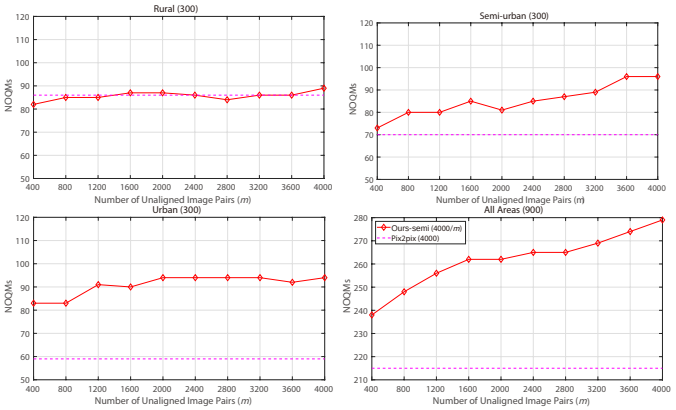

Fig. 8. NOQMs for our framework on 900 test image pairs as the amounts of unaligned training data varies.

As for our framework, we firstly evaluate the semi-supervised version when trained with 4,000 unaligned data and various aligned data (red lines with diamond markers). For the whole 900 test data, our framework shows a substantial performance boost than Pix2pix when the amount of aligned image pairs is over 800 (20% of 4,000). Especially in the urban area, our framework trained with only 30% aligned training data (1,200) obtains NOQMs comparable to Pix2pix trained with 100% aligned training data (4,000), indicating that our framework effectively enhances SAR-optical image matching with less aligned training data.

We further evaluate the semi-supervised version supplied with the same amount of training data as Pix2pix (purple lines with square marks). Specifically, these experiments were trained with $3k/4$ aligned and $k/4$ unaligned image pairs, where k represent the amount of aligned training data used by Pix2pix. As can be seen that, with the same amount of training data, our framework performs significantly better than Pix2pix in the urban area, and is comparable to Pix2pix in the rural and semi-urban areas.

We also report the performance of our supervised version in Figure 7 (green lines with cross marks). Obviously, our supervised version obtains significantly more NOQMs than Pix2pix in semi-urban and urban areas.

In figure 8, we evaluate our semi-supervised version on various amounts of unaligned training image pairs. As shown in the semi-urban area, the NOQMs of our framework trained with 4,000 unaligned data are 1.3 times more than the NOQMs trained with 400 unaligned data. However, the NOQMs show slight growth in the urban areas and are nearly unchanged in the rural areas, indicating the limitation of our framework in rural areas. These results also suggest that, compared to the fully supervised learning, adding a small number of unaligned image pairs in our framework (e.g., 10%) substantially improves the SAR-optical image matching in urban areas.

IV. Conclusion

In this work, we proposed a simple yet effective semi-supervised image-to-image translation framework integrated by Pix2pix and CycleGAN. Our framework combines the benefits and avoids the disadvantages of the two methods. We have applied our framework to SAR-optical image matching where we achieve comparable performance with less aligned training data compared to baseline methods. In addition, the supervised version of our framework significantly outperforms the Pix2pix method in semi-urban and urban areas.

References

[1] S. Cui, A. Ma, L. Zhang, M. Xu, and Y. Zhong, "Map-net: Sar and optical image matching via image-based convolutional network with attention mechanism and spatial pyramid aggregated pooling," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 60, pp. 1–13, 2022.

[2] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of sar and optical imagery," *ISPRS J. of Photogramm. and Remote Sens.*, vol. 169, pp. 166–179, 2020.

[3] L. H. Hughes and M. Schmitt, "A semi-supervised approach to sar-optical image matching," *in ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2019.

[4] N. Merkle, S. Auer, R. Mller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and sar image matching," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 11, no. 6, pp. 1811–1820, June 2018.

[5] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "Sar-to-optical image translation based on conditional generative adversarial networksoptimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, 2019.

[6] W.-L. Du, Y. Zhou, J. Zhao, and X. Tian, "K-means clustering guided generative adversarial networks for sar-optical image matching," *IEEE Access*, vol. 8, pp. 217 554–217 572, 2020.

[7] W.-L. Du, Y. Zhou, J. Zhao, X. Tian, Z. Yang, and F. Bian, "Exploring the potential of unsupervised image synthesis for sar-optical image matching," *IEEE Access*, vol. 9, pp. 71 022–71 033, 2021.

[8] J. Zhang, W. Ma, Y. Wu, and L. Jiao, "Multimodal remote sensing image registration based on image transfer and local features," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1210–1214, Aug 2019.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[10] N. Merkle, S. Auer, R. Mller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and sar image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1811–1820, June 2018.

[11] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The sen1-2 dataset for deep learning in sar-optical data fusion," *in ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. IV-1, pp. 141–146, 2018.

[12] M. Aamir and M. Rafa, "Transformation consistency regularization – a semi-supervised paradigm for image-to-image translation," in *Proc. ECCV*, 2020, pp. 599–615.

[13] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.

[14] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2242–2251.