Knowledge-Based Systems 239 (2022) 108006

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

CJE-TIG: Zero-shot cross-lingual text-to-image generation by Corpora-based Joint Encoding

Han Zhang [a], Suyi Yang [b], Hongqing Zhu [a,*]

[a] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*
[b] *Department of Mathematics, Natural, Mathematical & Engineering Sciences, King's College London, Strand, London WC2R 2LS, United Kingdom*

ARTICLE INFO

Article history:
Received 28 August 2021
Received in revised form 14 December 2021
Accepted 17 December 2021
Available online 24 December 2021

Keywords:
Cross-lingual pre-training
Text-to-image synthesis
Universal contextual word vector space
Semantic alignment
Joint adversarial training

ABSTRACT

Recently, text-to-Image (T2I) generation has been well developed by improving synthesis authenticity, text-consistency and generation diversity. However, large amount of pairwise image–text data required restricts generalization of synthesis models only to its pre-trained language. In this paper, a cross-lingual pre-training method is proposed to adapt target low-resource language to pre-trained generative models. As far as we known, this is the first time that arbitrary input languages could access T2I generation. This joint encoding scheme fulfills both universal and visual semantic alignment. With any prepared GAN-based T2I framework, pre-trained source encoder model could be easily fine-tuned to construct target encoder model and hence entirely enable transfer of T2I synthesis ability between languages. After that, a semantic-level alignment independent of source T2I structure is established to guarantee optimal text consistency and detail generation. Different from monolingual T2I methods that apply discriminator to enhance generation quality, we use an adversarial training scheme that optimizes the sentence-level alignment along with the word-level alignment with a self-attention mechanism. Considering of training for low-resource languages lack of parallel texts in practice, target input embedding is designed available for zero-shot learning. Experimental results prove robustness of the proposed cross-lingual T2I pre-training on multiple downstream generative models and target languages applied.

1. Introduction

Generative adversarial network (GAN) based Text-to-Image (T2I) synthesis is a powerful generation system that directly converts natural language into vivid photo. However, GAN-based T2I models are highly restricted by training language where generative model would loss synthesis capacity whenever input embedding space differs from pre-trained text embedding space. So far, almost all existing GAN-based T2I architectures are trained with English text encoder [1–4], which incapacitates them for globalized propagation. This is reasonable since existing T2I databases [5–7] present that merely several well-resource languages could access lingual-specific pre-training. In addition, building dataset for any possible input language is costly and impractical. Subsequent GAN training may also suffer from more uncertainty such as unstable training behavior [8].

Applying additional machine translation is one alternative for cross-lingual tasks. However, this causes more reliance and cost on external models which is unsustainable and is unfair in terms of resource level. In addition, translation artifact arises in sentence translation commonly even by professional translation machine of which degree varies between languages [9]. In fact, cross-lingual pre-training in comparison with alternatives on downstream tasks and multiple languages including even very low-resource ones are still frequently researched. Schuster et al. [10] testify the performance of three cross-lingual transfer methods on task-oriented dialog and discover that translating training data provides the worst results. Optimal results are achieved by cross-lingual contextual word representation than alternative cross-lingual pre-trained embedding. In addition, Schuster et al. [11] apply cross-lingual alignment of contextual word representation to zero-shot dependency parsing. Conneau et al. [12] also verify the effectiveness of cross-lingual language model on cross-lingual classification and find its advantage especially for low-resource languages (e.g. Nepali, Thai, etc.). It is also presented that GAN-based language-universal feature representation outperforms machine translating on question answering with fewer linguistic resource [13]. As our cross-lingual pre-training strategy seeks for the same goal of obtaining aligned universal vector space ahead of any other task-specific procedure, the conclusion is somehow transferable to our vision task.

The necessity of researching end-to-end cross-lingual framework is for its advantage as a sustainable one-time solution

[*] Corresponding author.
E-mail address: hqzhu@ecust.edu.cn (H. Zhu).

https://doi.org/10.1016/j.knosys.2021.108006
0950-7051/© 2021 Elsevier B.V. All rights reserved.

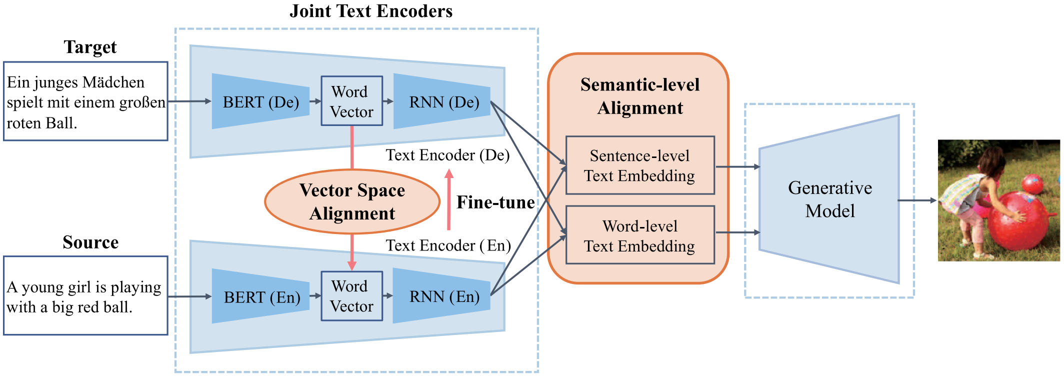


Fig. 1. Illustration of the proposed dual-path joint encoding scheme with multi-level alignment for cross-lingual T2I synthesis.

especially for T2I synthesis which the task itself arouses interest for its practical usability. In comparison with alternatives such as applying additional sentence-level translation for each single input, our framework eliminates all additional expense and delay once training is done. In practice, selecting the pre-trained language model and T2I model we prefer, the proposed framework fine-tunes T2I encoders in joint encoder mode with only several batches, and the testing data need just be delivered to the target encoder branch which is equivalently easy as starting a monolingual T2I with its original encoder. To maintain good generalization on GAN-based T2I at the same time, replacement concentrates on T2I encoder and adversarial training scheme such that the rest of the structure with diversified creation taken place in mainstream T2I methods could still be entirely compatible with. Aiming to provide equal access to qualified T2I synthesis, this paper establishes a widely applicable joint embedding scheme for general T2I GAN architectures and input languages unrestricted by training data availability.

In this paper, a Corpora-based Joint Encoding scheme for Text-to-Image Generation (CJE-TIG) is proposed as the first T2I targeted cross-lingual pre-training model. Based on methodology of unifying semantic layout between target and source languages during embedding phase with generative model fixed, the proposed dual-path encoder structure is designed easily substituted for classical T2I text encoders. As shown in Fig. 1, the proposed scheme pre-trains language model that aligns bi-lingual vector spaces and fine-tunes GAN-based encoders to attribute visually discriminative features. Additionally, T2I specific fine-tuning strategy by several semantic-level alignment approaches is also proposed for better adaption to downstream image generation. Target RNN is initialized with parameters of English decoder, and then fine-tuned using the proposed alignment strategy. In specific, a new adversarial training strategy is designed to jointly optimizes bi-lingual encoders for sentence-level alignment, and a self-attention mechanism is proposed for word-level alignment. The proposed framework weakly depends on the selection of languages and is widely applicable for mainstream T2I architectures. Experiments verify the robustness of the proposed joint embedding scheme on several GAN-based T2I methods and target languages.

The main contributions of this paper are as follows:

- A bi-lingual joint embedding scheme is designed to be a baseline of cross-lingual GAN-based T2I synthesis. To our best knowledge, restriction of input language in T2I generation is removed entirely for the first time.
- A dual-path encoder architecture assisting target language embedding is established substituting T2I text encoders.
- For better text consistency and detail generation, discriminator is used to jointly optimize sentence-level alignment and self-attention word-level alignment in bi-lingual encoder.
- Zero-shot learning is conducted for target lingual encoding by which the purpose of cross-lingual synthesis for low-resource language lacking parallel training data is verified feasible.
- The proposed scheme is testified applicable to mainstream T2I architectures and the style of cross-lingual generation consists with source language generation.

2. Related works

Text-to-image synthesis The branch of T2I synthesis algorithm using GAN [14] is firstly explored by Reed et al. [15] based on the framework of cGAN [16]. Subsequent researches [17–19] raise synthesis resolution to 256 × 256 by progressive generation. Then, AttnGAN [1] efficiently improves synthesizing text-consistent images in fine-grained level and following researches, such as MirrorGAN [4], DM-GAN [2] and ControlGAN [3], build their synthesis models upon AttnGAN's framework. However, even monolingual T2I mostly relies on a fully labeled dataset and triggers semi-supervised training over unlabeled data [20]. In addition to T2I, problems on data resources frequently exist and initially inspire cross-media methods on vision tasks such as training image classifier with less effort on labeling image data by leveraging their textual information [21]. Later, tasks that evolve interaction between modalities also suffer from domain-variance of specific modalities, and thus corresponding solutions in the field of image–text retrieval [22] and image captioning [23] have been raised. However, language-invariance property still has not been explored for T2I synthesis while current research mainly focuses on improving generation quality.

Word representations model Word representation models are firstly established for monolingual [24] in early years. In specific, static word embedding methods have been conducted without supervision on large corpora for question answering [25] and document classification [26,27]. Then, contextualized word representation is adopted by ELMo [28] which comprehensively represents complex word characteristics and their variances across linguistic context. Sequence transduction model replacing recurrent layers with attention mechanism firstly enables continuous word representation [29]. GPT [30] firstly obtains universal word representations learned by Transformer while each token could attend to context of one direction by constrained self-attention. Recently, BERT [31] is proposed as a deep bi-directional Transformer that adopts self-attention to both sides. Later, language models are applied to multiple downstream tasks by which most general architectures are accepted as word extraction pre-trained models followed by fine-tuning for downstream tasks.

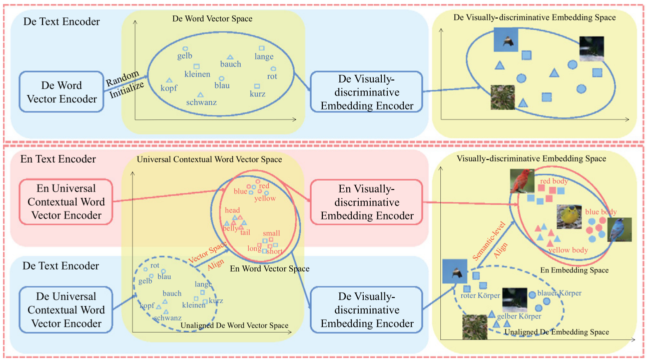

De Text Encoder
De Word Vector Space
De Visually-discriminative Embedding Space
De Word Vector Encoder
Random Initialize
gelb
bauch
lange
kleinen
rot
kopf
blau
kurz
schwanz
De Visually-discriminative Embedding Encoder
En Text Encoder
Universal Contextual Word Vector Space
Visually-discriminative Embedding Space
En Universal Contextual Word Vector Encoder
blue
red
yellow
head
belly
tail
small
long
short
En Visually-discriminative Embedding Encoder
red body
blue body
yellow body
Vector Space
Align
En Word Vector Space
Semantic-level
Align
En Embedding Space
De Text Encoder
rot
gelb
blau
lange
bauch
kopf
kleinen
kurz
schwanz
Unaligned De Word Vector Space
De Universal Contextual Word Vector Encoder
De Visually-discriminative Embedding Encoder
roter Körper
blauer Körper
gelber Körper
Unaligned De Embedding Space

Fig. 2. Illustration of bi-lingual embedding space. Up: Encoding phase of unsuccessful synthesis process by target language on pre-trained GAN-based T2I model. Down: The proposed cross-lingual encoding framework.

Cross-lingual alignment Cross-lingual alignment for sequential word vectors generated from language model is widely applied to many downstream tasks [32]. Early study demonstrates that potential geometric similarity exists between bi-lingual word vectors, and transfers by linear map [33] or orthogonal linear transform [34]. MUSE [35] realizes unsupervised word translation which builds a bi-lingual dictionary by adversarial learning. For contextual word vector embedding, subsequent works explore alignment of word pairs in the same context with or without parallel sentences [36,37]. CLBT [38] collects training pairs with unsupervised bi-directional word alignment and uses off-line word vector transformation that projects target vectors to the space of source domain. Compositional sentence model [39] is firstly applied to multi-lingual representation for sentence-level semantic alignment. Recently, a global–local sequence alignment strategy (T2VLAD) for text-video retrieval task is proposed [40] which measures the global and local similarities between text and videos. In particular, both T2VLAD and our method conduct two-scale alignment in different fine-grained levels to align source and target domains, but their local alignment is in supervised manner with topic labeled while ours is in unsupervised self-attention manner.

Cross-lingual transfer learning Cross-lingual transfer learning is extending to various natural language processing (NLP) tasks. Such approaches help to expand the range of languages available especially for insufficient task-specific data in low-resourced languages. Recently, mainstream transfer learning approach fine-tunes pre-trained ELMo [28], BERT [31], ToD BERT [41], LaBSE [42], XLM-RoBERTa [43], etc. models for specific tasks. For example, Marko et al. [44] verify the effectiveness of cross-lingual word embedding on Twitter sentiment classification on 13 languages of different resource levels. Tian et al. [45] propose a rumor detection model that utilizes pre-trained BERT and self-training loop to adapt the model to target language. Cross-lingual transfer learning scheme for Dialog Act (DA) recognition is reported by Martinek et al. [46]. They transfer model trained on a standard English DA corpus to two other languages, German and French. Abad et al. [47] introduce an adaptation strategy of acoustic model by applying cross-language model, in which multiple CNN layers are shared between languages, so that the domain adaptation transformation learned for a resource-rich language can be applied to low-resource languages. Similar cross-lingual transfer on multiple downstream tasks, includes natural language generation [48], machine reading comprehension [49], cross-lingual multi-modal retrieval [22] and cross-lingual image captioning [23], but no T2I synthesis approaches involve cross-lingual scenario thus far.

3. Methodology

The proposed cross-lingual T2I synthesis scheme follows the ideology of attaining highly similar semantic layout available for fixed pre-trained generative models. Therefore, the replaced bi-lingual encoder structure differs from classical T2I language models [1–4]. It could be seen from the upper row in Fig. 2 that randomly initialized word vectors are not identifiable for pre-trained T2I models. By the proposed scheme as shown in lower row, parallel input fulfills domain adaption during encoding phase and is directly applicable to pre-trained generative models. The architecture of the proposed CJE-TIG contains two stages: universal contextual word vector alignment stage and T2I task-specific visual semantic alignment stage as depicted in Fig. 3. Specifically, language model is trained on corpora in Stage I and T2I specific training is implemented in Stage II.

3.1. Universal contextual word vector alignment

In stage I, a competitive pre-trained language model BERT [31] is used for contextual word representation essential for T2I sentence encoding. Firstly, unlabeled sequential words in grouped sentence from Parallel Sentence Set S [50] are input into target and source BERT. Then, universal contextual word vector x^s and x^t are generated by pre-trained BERT models. Vector space distributions of two languages are in potential geometric similarities and vector space transfer is contributing in T2I specific training. Considering even distant language could be well transferred by distribution transformation using dictionary on monolingual data [33], we adopted it on the unlabeled corpora data to adapt inter-lingual transfer to T2I sentence pair robust of input languages. In specific, parallel pairs (x^s, x^t) are collected by dictionary D to learn matrix $W^{t\to s}$ for bi-lingual T2I input training.

$$X^s = W^{t\to s}X^t. \tag{1}$$

where X^s and X^t respectively represent source and target lingual word vector space. A visualization of universal contextual word vector alignment is provided in Fig. 5.

Considering of the direction fine-tuned on target T2I pre-trained encoder, mapping direction is set to $t \to s$ in this article opposite to previous studies [33,35]. $W^{t\to s}$ is optimized using stochastic gradient descent to minimize:

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N}\left\|W^{t\to s}x_i^t - x_i^s\right\|^2, \tag{2}$$

where x_i^s and x_i^t belong to collected word pairs $\{x_i^s, x_i^t\}_{i=1}^N$. Then, $W^{t\to s}$ at the best alignment is freezed for our downstream task of T2I synthesis.

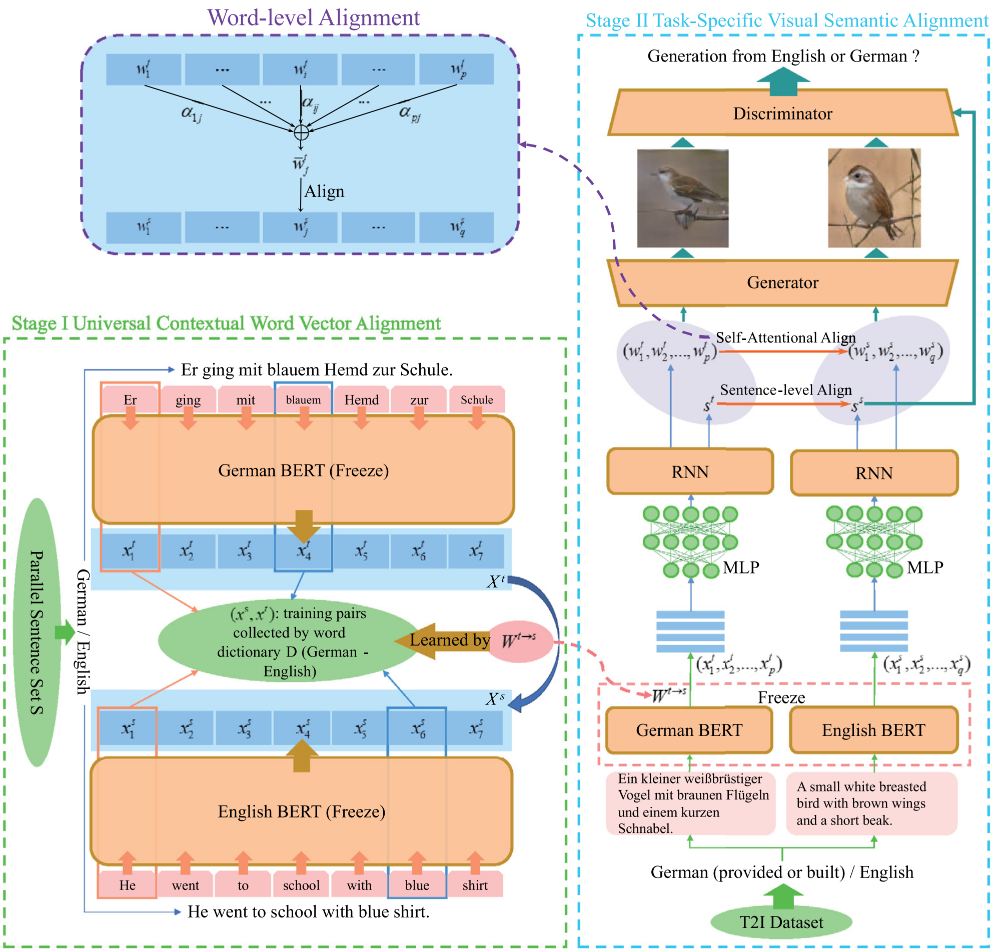

Word-level Alignment
w_1^t ... w_i^t ... w_p^t
α_{1j} α_{ij} α_{pj}
$\bar{w}_j^t$
Align
w_1^s ... w_j^s ... w_q^s
Stage II Task-Specific Visual Semantic Alignment
Generation from English or German ?
Discriminator
Generator
Self-Attentional Align
$(w_1^t, w_2^t, ..., w_p^t)$
$(w_1^s, w_2^s, ..., w_q^s)$
Sentence-level Align
s^t
s^s
RNN
RNN
MLP
MLP
$(x_1^t, x_2^t, ..., x_p^t)$
$(x_1^s, x_2^s, ..., x_q^s)$
$W^{t\to s}$
Freeze
German BERT
English BERT
Ein kleiner weißbrüstiger Vogel mit braunen Flügeln und einem kurzen Schnabel.
A small white breasted bird with brown wings and a short beak.
German (provided or built) / English
T2I Dataset
Stage I Universal Contextual Word Vector Alignment
Er ging mit blauem Hemd zur Schule.
Er ging mit blauem Hemd zur Schule
German BERT (Freeze)
x_1^t x_2^t x_3^t x_4^t x_5^t x_6^t x_7^t
X^t
(x^s, x^t): training pairs collected by word dictionary D (German - English)
Learned by $W^{t\to s}$
X^s
x_1^s x_2^s x_3^s x_4^s x_5^s x_6^s x_7^s
English BERT (Freeze)
He went to school with blue shirt
He went to school with blue shirt.
Parallel Sentence Set S
German / English

Fig. 3. Pre-trained language model and transfer matrix $W^{t\to s}$ learned on corpora data in Stage I are applied to Stage II on T2I Dataset. Then, initial pre-trained T2I encoders are fine-tuned and visually discriminative embedding is aligned both sentence-level and word-level. The figure also presents details of self-attentional mechanism for word-level alignment.

3.2. T2I task-specific visual semantic alignment

Although highly similar target distribution to source language can be obtained from Stage I, visual semantic level features are remained undistinguished which is vital for T2I generation task. Hence, dual-path encoders are set for two languages which differs from the original GAN-based T2I approaches with single text encoder. As only the target language path of the encoder would run to generate embedding in practice, we design a T2I specific joint encoding scheme which the source encoder aids the training of the target encoder. Dual-path encoder has been established in various structures such as image–text matching network [51] that learns visual and textual feature representation. We establish a dual-path encoder that learns shared bilingual textual representations by fine-tuning one from the other. Because mainstream T2I encoder composes of MLP layer and RNN to encode sequential word vectors $(x_1^t, x_2^t, \dots, x_p^t)$ and $(x_1^s, x_2^s, \dots, x_q^s)$, we adopt it to see the validity of our fine-tuning scheme. Starting with two identical pre-trained encoders, the one for target language would be fine-tuned by optimization based on judgment of whether a shared bilingual textual representation could be perfectly learned after the final multi-level alignment. To achieve text consistent and fine-grained synthesis simultaneously, sentence-level embedding s and word-level embedding w_t are both obtained from the final hidden state and each internal state of RNN respectively as:

$$s = h_n = RNN(MLP(x_n), h_{n-1}), \tag{3}$$

$$w_t = h_t = RNN(MLP(x_t), h_{t-1}), \; 1 \le t \le n, \tag{4}$$

where h_t is the hidden state at the tth time step. To transfer synthesis ability on visual semantic features for downstream T2I generative models, two alignment approaches are introduced to fine-tune source lingual embedding encoder. Nonetheless, other GAN-based T2I generation models which only take sentence-level embedding as input could also adapt to the proposed framework by simply abandoning word-level alignment without substantial impact on synthesis behavior.

Sentence-level alignment To improve global semantic consistency in joint embedding phase, an adversarial training strategy is designed to accurately optimize overall sentence alignment. Easily available for multiple downstream architectures, we use a conditional discriminator D_v on generation side providing adversarial visual alignment instead of conventional generator optimization. Architecture of the conditional discriminator is shown in Fig. 4. Sentence-level embedding is concatenated with lower-dimensional image feature maps for source/target prediction. In specific, image generated from source lingual text is regarded as real sample in D_v training, while image generated from target lingual text is regarded as fake sample. Meanwhile, D_v accepts source sentence-level embedding s^s as criterion to distinguish source of text input. D_v is optimized by minimizing the binary

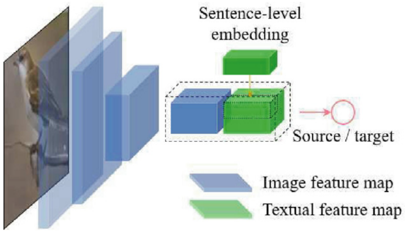


Fig. 4. Architecture of the conditional discriminator.

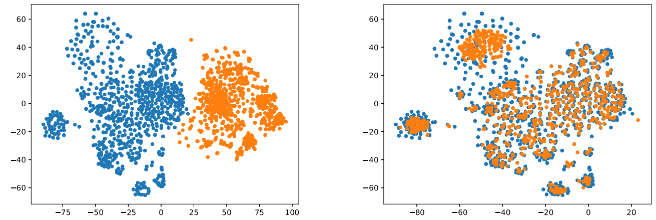


Fig. 5. Universal contextual word vector space before (left) and after alignment (right) where source and target languages are denoted in blue and orange respectively.

cross-entropy loss defined as:

$$\begin{aligned}\mathcal{L}_{D_v} = &-\mathbb{E}_{s^s\sim p_{s^s},w^s\sim p_{w^s}}[\log(D_v(G(s^s,w^s),s^s))]\\ &-\mathbb{E}_{s^t\sim p_{s^t},w^t\sim p_{w^t}}[1-\log(D_v(G(s^t,w^t),s^s))].\end{aligned} \tag{5}$$

Since generation model is fixed while only the proposed encoding model is optimized, adversarial training scheme under this proposed architecture is ultimately devoted to sentence-level embedding alignment. The text encoder competes with D_v by optimizing:

$$\mathcal{L}_s^{adv} = -\mathbb{E}_{s^t\sim p_{s^t},w^t\sim p_{w^t}}[\log(D_v(G(s^t,w^t),s^s))], \tag{6}$$

where D_v back propagates encoder without any self-parameter update on generator. In this case, s^t and w^t should ultimately approximate s^s and w^s, and thus target lingual text encoder gains sophisticated synthesis capability visually on generative models. In order to accelerate global semantic alignment considering of the instable training of GAN-based T2I synthesis essentially, squared Euclidean distance (mean squared error, MSE) between source sentence-level embedding s^s and target sentence-level embedding s^t is used as a loss function:

$$\mathcal{L}_s^{mse} = \frac{1}{M}\sum_{i=1}^{M}\left\|s_i^s - s_i^t\right\|^2. \tag{7}$$

Overall sentence-level alignment loss function is formulated as:

$$\mathcal{L}_s = \mathcal{L}_s^{mse} + \mathcal{L}_s^{adv}. \tag{8}$$

Word-level alignment Although sentence-level embedding realizes text-consistent image generation fundamentally, we also provide word-level alignment strategy for downstream models which input word-level embedding. Therefore, another attention mechanism [52] is used and jointly optimized with sentence-level alignment. For the jth word-level embedding w_j^s in source lingual sentence, we calculate its corresponding target lingual word-level embedding $\bar{w}_j^t$ as a weighted sum of all target lingual word-level embedding w_i^t as follows:

$$\bar{w}_j^t = \sum_{i=1}^{p}\alpha_{ij}w_i^t, \tag{9}$$

in which source-lingual word-level embedding would gain the most information from target-lingual word-level embedding of largest correspondence manipulated by the weight α_{ij} on each target embedding:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{p}\exp(e_{kj})}, \tag{10}$$

where e_{ij} is designed to measure similarity between the ith word of target lingual sentence and the jth word of source lingual sentence using dot-product operator as:

$$e_{ij} = w_i^t \cdot w_j^s. \tag{11}$$

Similar to sentence-level alignment, we minimize the squared Euclidean distance between $\bar{w}_j^t$ and w_j^s to align word-level embedding. The word-level alignment loss is formulated as:

$$\mathcal{L}_w = \frac{1}{M}\sum_{k=1}^{M}\left(\frac{1}{q}\sum_{j=1}^{q}\left\|\bar{w}_j^t - w_j^s\right\|^2\right). \tag{12}$$

Finally, bi-lingual encoder loss consists of sentence-level alignment and word-level alignment loss is formulated as:

$$\mathcal{L} = \mathcal{L}_s + \gamma\mathcal{L}_w, \tag{13}$$

where parameter γ is used to control the weight of the word-level alignment loss.

Normally, query text should have included parallel texts from specific T2I dataset. However, we form parallel target sentence by word-by-word translation for a more propagable perspective. In practice, labeled T2I text descriptions for a majority of languages are hardly obtainable. Based on consideration that unordered word input is of high possibility affects expression in T2I tasks limitedly, feasibility of zero-shot ideology is initially approved. For example, ill-formed sentence "Bird the yellow wing and white belly" is interpreted equivalently as "Bird with yellow wing and white belly" in semantic space. Even still, replacement-built sentence still suffers from mild deviation from true semantic distribution, while deviation enlarges when queried dictionary is insufficient in parallel T2I sentence generation. Overall, experimental results show that this training scheme achieves qualified results although true parallel sentence input performs slightly better as expected.

4. Experimental results

4.1. Datasets and evaluation metrics

In universal contextual word vector alignment stage, MUSE [35] bilingual dictionary and parallel sentences from the Europarl corpora [50] are used in which 10,000 sentence pairs are randomly picked. In stage II, GAN-based T2I synthesis approaches are pre-trained on three T2I datasets.

CUB dataset [5] contains 11,788 images from 200 bird species. Referring to [15], CUB is split into 8855 training images from 150 species and 2933 test images from 50 other species with 10 English text description each. These English descriptions are translated into German, and thus 88,550 English–German train sentence pairs and 29,330 test pairs are engaged.

Multi30K dataset [53] is a multi-lingual multi-modal dataset containing 31,014 images. It provides 31,014 English–German parallel image description sentences, where the German translations are created by professional translators. Multi30K divides these English–German sentence pairs into 30,014 for training and 1000 for test.

COCO dataset [7] contains images with multiple objects and various backgrounds. In this paper, COCO's training set containing 82,783 images with 5 English captions and each is directly

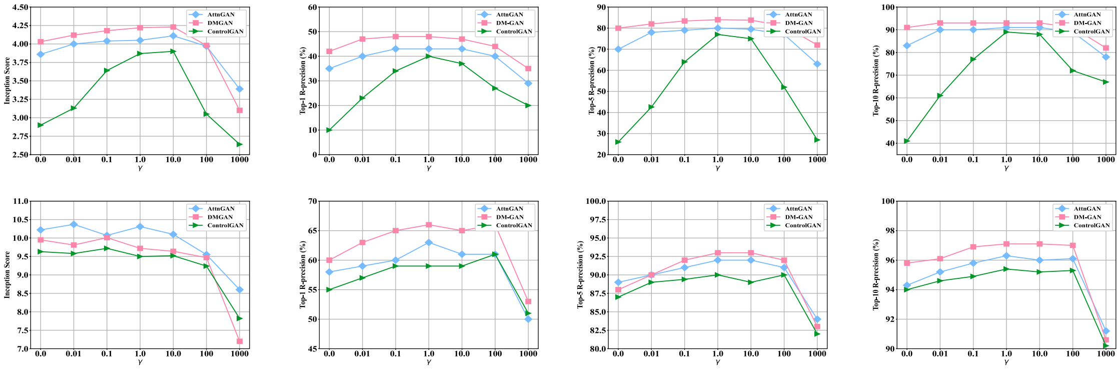


Fig. 6. Inception scores and R-precision rates at different γ by three T2I models on CUB (up) and Multi30K (down) datasets.

used for training. A multilingual version of COCO test subset by Rajendran et al. [54] collects French and German translations for all 5 English captions of the 1000 test images is used for test. Totally, 413,915 English train sentences and 5000 English–German–French test sentences from COCO participate in our experiment.

In this framework, images correspond to each text description from CUB, Multi30K and COCO are not necessarily required. Besides, only English sentences from CUB, Multi30K and COCO sentence pairs are used for zero-shot learning, while only parallel sentences for test are all utilized in verification and comparison experiment.

We use two quantitative metrics commonly used in T2I synthesis tasks [1–4], Inception Score (IS) and R-precision, to evaluate the visual quality, image diversity and text consistency.

4.2. Implementing details

In stage I, uncased English BERT model pre-trained on BookCorpus and English Wikipedia available at [55] learns universal word vectors initialized at 768 dimensions. Pre-trained uncased BERT models on two target languages, German and French, are available at [56] and [57]. Then, we use SGD solver to minimize $W^{t\rightarrow s}$. In stage II, we use bi-directional LSTM as an RNN pre-trained according to [1] to extract 256-dimension text embeddings. For the fine-tuned source lingual text encoders for joint embedding, three downstream generative models optimized based on their original strategies, AttnGAN [1], DM-GAN [2] and ControlGAN [3], are re-trained to testify the robustness of the proposed alignment method. All these retrained generative models are available at [58]. In stage II, encoder and discriminator are optimized using ADAM solver.

4.3. Word/sentence-level alignment evaluation

To visualize word-level attentional weights α_{ij} in (10) for all word pairs in paired input sentence, an attention map is illustrated in Fig. 7. In specific, lexes are listed by the order in input texts, and box in deeper color indicates larger α_{ij} value. It could be observed that only very few pairs are highly matched with each other while other pairs hardly occupy any weight. Some lexes may relate to several bi-lingual words or spread equal attention among all bi-lingual components. Different languages aligned may present varying α_{ij} distribution in the grid, while attention maps are mostly diagonally concentrated in En-De alignment.

Fig. 8 presents t-SNE plots where distribution before and after sentence-level alignment is visualized. It could be observed from initial stage that bi-lingual embeddings are clustered in completely different orientation of the space, while essential vector space transfer is accomplished as shown in Fig. 8 (right). Aligned vector space shows that less spare monolingual points are distributed around the space, and geometric characteristics of clustering could be observed. It is also likely that fine-grained word-level detail from CUB could not be possibly represented well by global sentence-level embedding which hinders precise pairwise alignment to some extent.

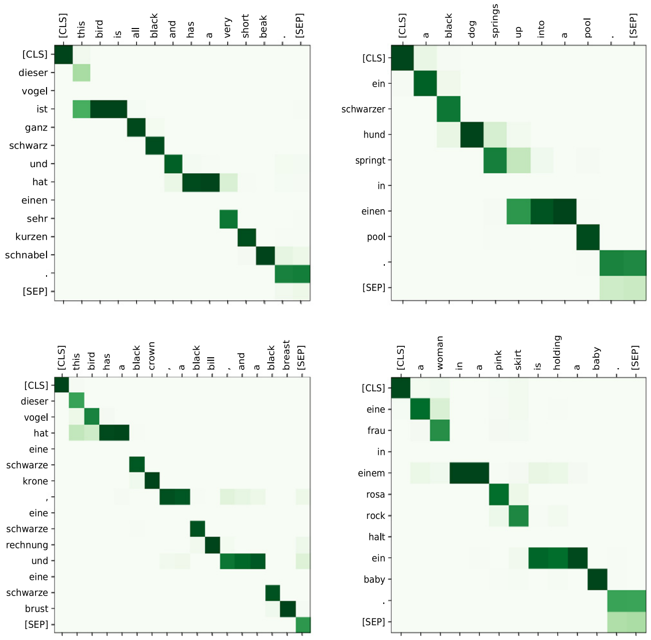


Fig. 7. Visualization of word-level attention weights α_{ij}, where deeper color indicates closer correspondence of the bi-lingual word pair. Horizontal: English sentence, vertical: German sentence. Left: CUB, right: Multi30K.

4.4. Impact of word-level alignment loss function

Considering that word-level alignment is conducted in unsupervised manner, experiment on γ is implemented to verify the most appropriate proportion between sentence-level and word-level alignments. IS and R-precision on CUB and Multi30K datasets with seven γ values in loss function (13) from 0 to 1000 are shown in Fig. 6. It could be seen that excessive γ value (≥ 100) leads to dramatic decrease of IS and R-precision. Taking value $0 \leq \gamma \leq 10$, IS and R-precision all stay at relatively good range. It could also be observed that different T2I models are in different sensitivity to the level of fine-grained feature alignment. In addition, quantitative results are more sensitive to word-level alignment on CUB than on Multi30K in terms of γ. Since CUB dataset collects bird images of numerous categories,

Table 1
Inception scores by T2I models with different lingual texts.

T2I methods	CUB		Multi30K		COCO		
	English	German	English	German	English	German	French
AttnGAN [1]	4.30 ± 0.04	4.05 ± 0.05	10.64 ± 0.89	10.31 ± 0.88	19.82 ± 0.88	19.58 ± 0.99	19.32 ± 1.05
DMGAN [2]	4.44 ± 0.04	4.22 ± 0.05	10.32 ± 0.88	9.72 ± 1.10	19.83 ± 1.66	18.44 ± 1.18	18.52 ± 1.51
ControlGAN [3]	4.15 ± 0.04	3.87 ± 0.04	10.27 ± 0.71	9.50 ± 0.63	18.95 ± 0.90	18.17 ± 1.05	18.20 ± 1.36

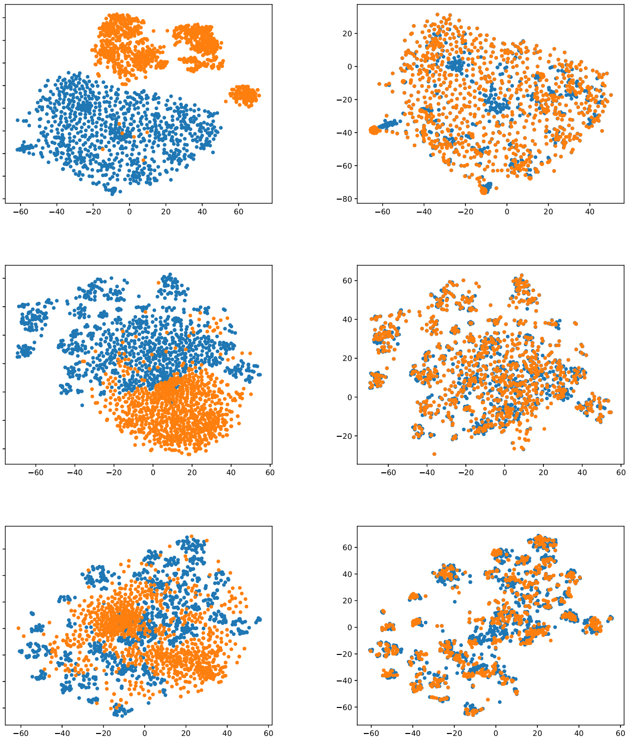


Fig. 8. Visualization of sentence-level embedding distribution before (left) and after (right) alignment on CUB, Multi30K, and COCO datasets (upper to lower) in 2D projection by t-SNE. Orange: English (source) embedding, Blue: German (target) embedding.

fine-grained features are appreciated to synthesize distinguishable bird species. This indicates indispensable impact of word-level alignment on fine-grained targeted synthesis. Despite from structural difference that fluctuates quantitative results, γ value between 1 and 10 achieves the most appropriate tradeoff between sentence-level and word-level alignments in T2I synthesis. For computation simplicity, γ is set to 1 in following experiments.

4.5. Evaluation on sentence-level alignment loss function

In this experiment, contribution of losses in sentence-level loss function would be examined. In specific, $\mathcal{L}_s^{adv}$ is optimized to deceive discriminator for judging target or source language generation. $\mathcal{L}_s^{mse}$ is used to narrow the gap between target and source language sentence-level embedding distribution fast and stably. As shown in Fig. 10, w/o $\mathcal{L}_s^{mse}$ leads to noticeably slower convergence of MSE. Similarly, one could observe from Fig. 11 that visual semantic space of the target domain is distinctively distributed than source domain if not applying $\mathcal{L}_s^{mse}$. As shown in Table 5, ablating either of the losses leads to large decrease in quantitative results especially on R-precision which mainly measures text consistency. As expected, this might be because of the sentence-level alignment that essentially guarantees sentence consistent generation results of target language than of pre-trained source generative model.

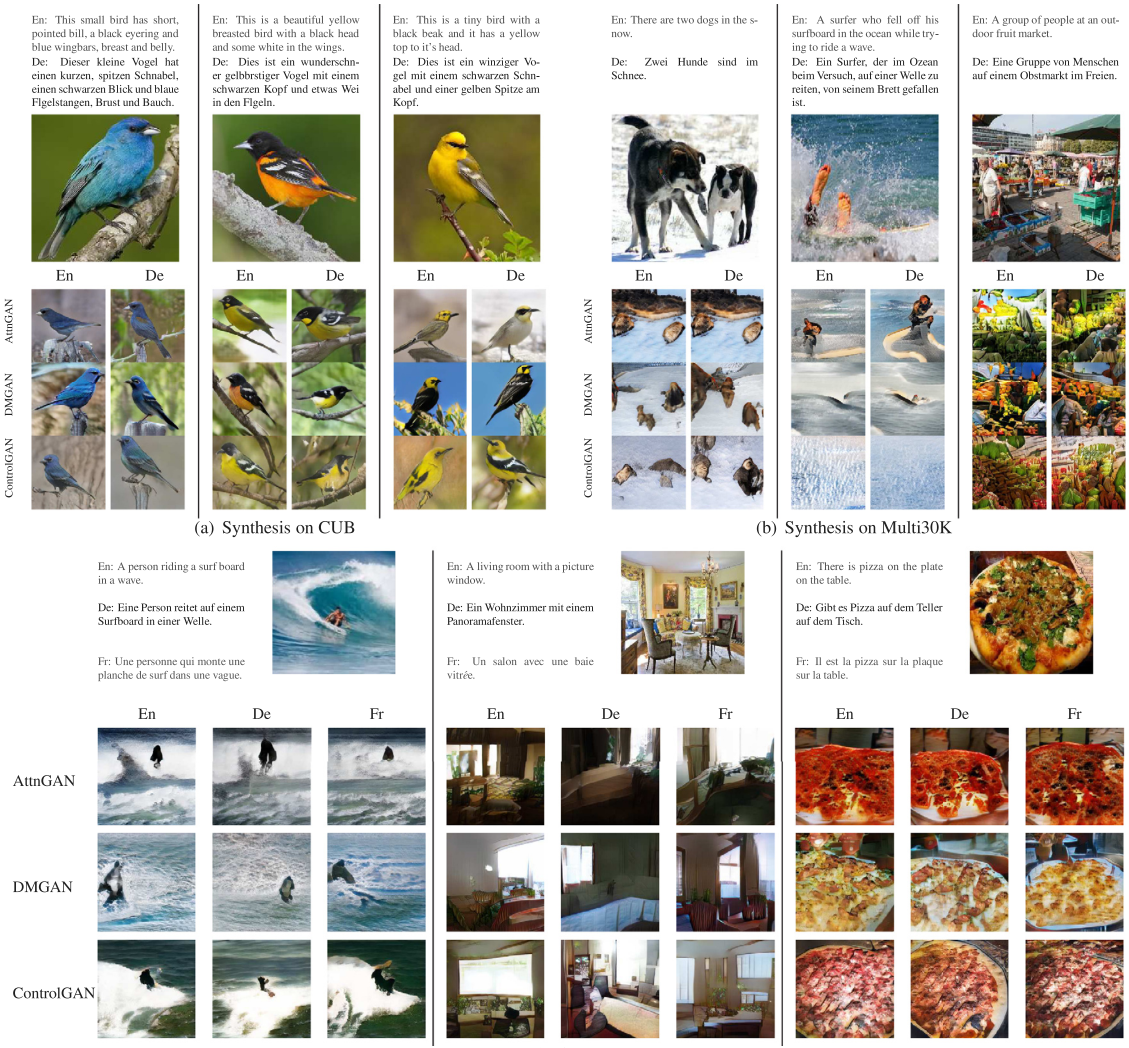
4.6. Selection of general structure

In terms of the general module in this cross-lingual pre-training framework, RNN in joint encoder is selected as general GAN-based T2I methods all adopt RNN in their text encoder. Since the principle of this framework is to only fine-tune source language generative models after applying a language model, replacing for others would turn this task into more complicated procedure. However, alternatives for BERT that establishes contextual word embedding space may include random initialization, mBERT, GPT, etc. Quantitative results for AttnGAN by these options are testified on Multi30K dataset as shown in Table 6. As contextual world vector space is not the final latent space for T2I generative models, alternative models present similar results on generation quality. Compared to parallel BERT, mBERT, and GPT, random initializing word vector could not construct contextual vector space. Hence, randomly distributed space input to downstream RNN would bring larger difficulty to pre-trained generative model fine-tuning. In general, adopting mBERT and BERT achieve relatively competitive synthesis performance.

4.7. Quality and quantity evaluation

In this subsection, the proposed framework is testified on three downstream T2I generation models. Fig. 9 shows generated images by AttnGAN [1], DM-GAN [2] and ControlGAN [3] on texts from CUB, Multi30K and COCO datasets. Left column of each parallel sentence pair lists images synthesized from original generative models trained on source language, where rest column lists images produced by the proposed method. In general, the proposed framework on three models all synthesize images that match English and German text descriptions fundamentally. Besides, individual generative models perform diversely in terms of visual quality and text consistency due to their architectural difference. However, a similar synthesis style seems to transfer from source language to target such as images for "There is pizza on the plate on the table" in Fig. 9(c). This verifies the effect of the proposed joint embedding scheme independent of downstream model capability. However, some unsatisfied cases still remain, such as examples in Fig. 9(b) that "felling surver" is not fully reflected. Hence, cross-lingual synthesis performance of the proposed model still highly relies on basic functionality of pre-trained T2I model.

Quantitative comparison of three downstream models on three datasets is also discussed in Tables 1,2. As expected, generated images on English T2I models achieve better results since pre-trained generators fit data space from source embedding space superior than target. In specific, generation on source language achieves slightly higher IS and R-precision. It could be interpreted that bi-lingual generation might not evidently affect image quality and generation diversity, while text distinctive generation based on individual text would be weaken. Although performance decline by cross-lingual synthesis seems to be inevitable, the proposed model adapts wide range of T2I models and languages by no significant variance in performance degradation. For example, all levels of R-precisions for German and French synthesis on COCO (Table 2) only varies for approximately 1%. Nonetheless, different generative models accommodate to cross-lingual synthesis differently, for example,



En: This small bird has short, pointed bill, a black eyering and blue wingbars, breast and belly.
De: Dieser kleine Vogel hat einen kurzen, spitzen Schnabel, einen schwarzen Blick und blaue Flgelstangen, Brust und Bauch.

En: This is a beautiful yellow breasted bird with a black head and some white in the wings.
De: Dies ist ein wunderschner gelbbrstiger Vogel mit einem schwarzen Kopf und etwas Wei in den Flgeln.

En: This is a tiny bird with a black beak and it has a yellow top to it's head.
De: Dies ist ein winziger Vogel mit einem schwarzen Schnabel und einer gelben Spitze am Kopf.

En: There are two dogs in the snow.
De: Zwei Hunde sind im Schnee.

En: A surfer who fell off his surfboard in the ocean while trying to ride a wave.
De: Ein Surfer, der im Ozean beim Versuch, auf einer Welle zu reiten, von seinem Brett gefallen ist.

En: A group of people at an outdoor fruit market.
De: Eine Gruppe von Menschen auf einem Obstmarkt im Freien.

(a) Synthesis on CUB

(b) Synthesis on Multi30K

En: A person riding a surf board in a wave.
De: Eine Person reitet auf einem Surfboard in einer Welle.
Fr: Une personne qui monte une planche de surf dans une vague.

En: A living room with a picture window.
De: Ein Wohnzimmer mit einem Panoramafenster.
Fr: Un salon avec une baie vitrée.

En: There is pizza on the plate on the table.
De: Gibt es Pizza auf dem Teller auf dem Tisch.
Fr: Il est la pizza sur la plaque sur la table.

(c) Synthesis on COCO

Fig. 9. Examples of images synthesized from different languages and downstream generative networks on (a) CUB, (b) Multi30K and (c) COCO datasets. English generation refers to benchmarks that reflects the performance of pre-trained T2I models, while German (or French) generation presents performance of the proposed CJE-TIG.

Table 2
Top-k R-precision (%) by T2I models with different lingual texts.

		CUB			Multi30K			COCO		
		AttnGAN	DMGAN	ControlGAN	AttnGAN	DMGAN	ControlGAN	AttnGAN	DMGAN	ControlGAN
En	k@1	59.00 ± 2.53	61.58 ± 2.82	53.69 ± 2.81	84.96 ± 2.63	87.90 ± 2.02	82.87 ± 4.83	84.58 ± 1.40	85.48 ± 1.43	81.24 ± 1.54
	k@5	91.97 ± 1.24	93.10 ± 0.94	89.46 ± 1.25	99.10 ± 0.74	99.27 ± 0.70	99.00 ± 0.77	99.52 ± 0.32	99.70 ± 0.16	99.22 ± 0.36
	k@10	97.46 ± 0.67	97.91 ± 0.55	96.27 ± 0.71	99.80 ± 0.40	99.80 ± 0.04	99.60 ± 0.66	99.92 ± 0.13	99.96 ± 0.02	99.94 ± 0.09
Ge	k@1	43.02 ± 3.22	47.25 ± 3.09	39.59 ± 2.87	63.25 ± 4.46	65.80 ± 4.36	58.86 ± 4.93	70.76 ± 1.66	71.81 ± 1.87	64.80 ± 2.06
	k@5	79.98 ± 3.37	83.88 ± 2.06	77.51 ± 2.85	91.64 ± 2.90	93.11 ± 2.71	90.02 ± 1.59	95.98 ± 0.76	96.26 ± 0.96	94.34 ± 0.75
	k@10	90.64 ± 2.22	93.07 ± 1.24	89.16 ± 1.99	96.13 ± 2.31	96.90 ± 1.04	95.03 ± 1.28	98.22 ± 0.66	98.34 ± 0.60	97.66 ± 0.49
Fr	k@1	–	–	–	–	–	–	70.61 ± 1.97	70.28 ± 1.51	66.56 ± 1.75
	k@5	–	–	–	–	–	–	95.68 ± 1.00	95.98 ± 1.00	94.32 ± 1.61
	k@10	–	–	–	–	–	–	97.88 ± 0.62	97.96 ± 0.80	97.52 ± 0.98

cross-lingual synthesis by AttnGAN on COCO dataset (En to De) only drops for 1.7% from source language generation at top-10 R-precision, while ControlGAN drops 2.28%. In comparison with top-1 R-precision, top-5 and top-10 R-precision are affected less noticeably. For instance, cross-lingual synthesis by DMGAN on Multi30K only decreases by 6.16% and 2.9% in terms of top-5 and top-10 R-precision. Overall, the proposed scheme is applicable to various languages and text descriptions with relatively reliable performance.

4.8. Zero-shot vs. parallel performance

Comparison on cross-lingual embedding with zero-shot (zs) and true parallel required (pr) text input is conducted to verify the feasibility of the proposed methods on low-resource language training in practice. Tables 3,4 report the quantitative evaluation results on CUB and Multi30K datasets. Quantitative results on true parallel sentence input are slightly higher than built parallel, which might be explained by the imperfect representation of

Table 3
Inception Score of T2I models in zero-shot (zs) and parallel-required (pr) cases.

T2I	CUB			Multi30K		
	En	De (zs)	De (pr)	En	De (zs)	De (pr)
AttnGAN	4.30	4.05	4.23	10.64	10.31	10.34
DMGAN	4.44	4.22	4.31	10.32	9.72	10.02
ControlGAN	4.15	3.87	3.98	10.27	9.50	9.86

Table 4
R-precision (%) of T2I models in zero-shot (zs) and parallel-required (pr) cases.

Language		CUB			Multi30K		
		AttnGAN	DMGAN	ControlGAN	AttnGAN	DMGAN	ControlGAN
English	k@1	59.00 ± 2.53	61.58 ± 2.82	53.69 ± 2.81	84.96 ± 2.63	87.90 ± 2.02	82.87 ± 4.83
	k@5	91.97 ± 1.24	93.10 ± 0.94	89.46 ± 1.25	99.10 ± 0.74	99.27 ± 0.70	99.00 ± 0.77
	k@10	97.46 ± 0.67	97.91 ± 0.55	96.27 ± 0.71	99.80 ± 0.40	99.80 ± 0.04	99.60 ± 0.66
German (zs)	k@1	43.02 ± 3.22	47.25 ± 3.09	39.59 ± 2.87	63.25 ± 4.46	65.80 ± 4.36	58.86 ± 4.93
	k@5	79.98 ± 3.37	83.88 ± 2.06	77.51 ± 2.85	91.64 ± 2.90	93.11 ± 2.71	90.02 ± 1.59
	k@10	90.64 ± 2.22	93.07 ± 1.24	89.16 ± 1.99	96.13 ± 2.31	96.90 ± 1.04	95.03 ± 1.28
German (pr)	k@1	52.19 ± 2.29	56.24 ± 2.41	50.18 ± 2.97	69.84 ± 3.73	71.02 ± 3.37	67.31 ± 4.34
	k@5	87.73 ± 1.98	88.62 ± 1.74	85.16 ± 1.64	94.81 ± 2.44	95.62 ± 2.56	94.16 ± 3.04
	k@10	95.42 ± 1.07	96.35 ± 1.52	94.12 ± 1.63	98.30 ± 1.26	98.90 ± 1.42	97.12 ± 1.21

Table 5
Quantity evaluation with different sentence-level loss functions on CUB dataset.

Items	IS	R-precision (%)		
		k@1	k@5	k@10
w/o $\mathcal{L}_s^{mse}$	3.44 ± 0.03	17.73 ± 2.00	50.15 ± 4.68	67.27 ± 4.14
w/o $\mathcal{L}_s^{adv}$	3.84 ± 0.04	34.66 ± 2.57	73.55 ± 3.72	86.62 ± 2.66
CJE-TIG	4.05 ± 0.05	43.02 ± 3.22	79.98 ± 3.37	90.64 ± 2.22

Table 6
Quantitative results on German subset of Multi30K by different options for word embedding.

Items	IS	R-precision (%)		
		k@1	k@5	k@10
Random initialization	9.27 ± 0.94	54.23 ± 4.72	83.26 ± 3.21	89.14 ± 1.92
BERT	10.31 ± 0.88	63.25 ± 4.46	91.64 ± 2.90	96.13 ± 2.31
mBERT	10.25 ± 0.63	63.31 ± 3.97	90.96 ± 2.62	96.02 ± 1.94
GPT	10.11 ± 0.91	62.13 ± 4.12	91.06 ± 2.78	95.89 ± 2.46

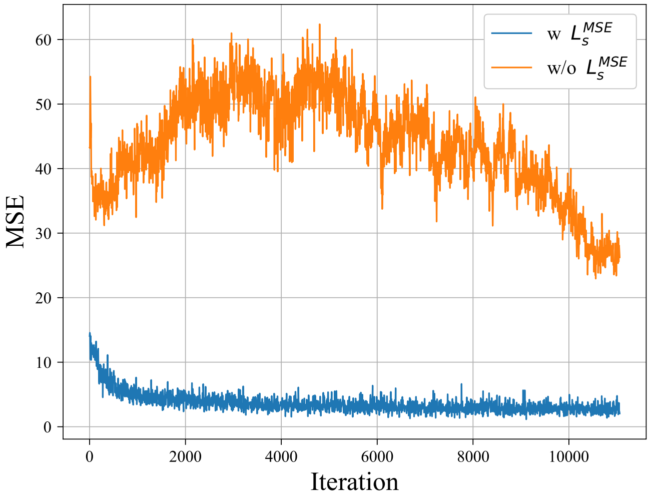


Fig. 10. Impact of ablating $\mathcal{L}_s^{mse}$ on MSE.

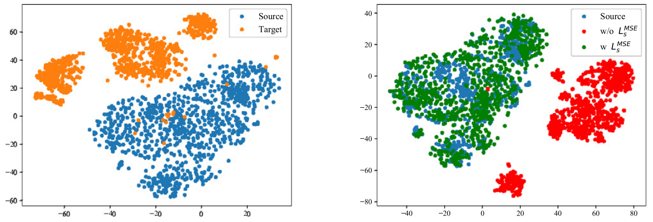


Fig. 11. Alignment of sentence-level embedding when optimizing w/ $\mathcal{L}_s^{mse}$ and w/o $\mathcal{L}_s^{mse}$. Target (orange) and source (blue) language embedding before alignment are shown in the left. Sentence-level embedding space after alignment are shown in the right where red dots refer to w/o $\mathcal{L}_s^{mse}$, green dots refer to w/ $\mathcal{L}_s^{mse}$.

sentence embedding obtained by word-by-word translation. For instance, the IS of De (zs) on CUB and Multi30K by AttnGAN dropped to 4.05 and 10.31 in comparison with 4.23 and 10.34 of De (pr). R-precision for De (zs) by AttnGAN still achieves 63.25%, 91.64%, 96.13%, which is 6.59%, 3.17%, 2.17% lower than embedded by true parallel sentence on Multi30K. Top *k*th R-precision of DM-GAN decreases by 8.99%, 4.74%, 3.28% on CUB, and 5.22%, 2.51%, 2.00% on Multi30K. In general, experimental results show that the proposed scheme is sufficient for cross-lingual generation even by the most basic zero-shot learning. In addition, cross-lingual synthesis by parallel text training would be selected in priority if provided, which is also an alternative for re-training original T2I models. However, there is no restriction for applicable language by this proposed cross-lingual T2I generation methodology.

5. Conclusions

In this paper, a cross-lingual T2I pre-training approach CJE-TIG is proposed that firstly eliminates restriction of accessing GAN-based T2I synthesis models for arbitrary input languages. This framework changes classical language-specific training pattern of previous T2I generation methods. Based on general T2I GAN architectures, we replace text encoder with a bi-lingual joint encoder, apply discriminator for encoder optimization, and directly use original generative models for generation. Experimental results show that the cross-lingual pre-training method could well adapt multiple generative models even under zero-shot scenario. However, cross-lingual performance of this method partially depends on reliability of the starting dictionary, and word translation performance of different language pairs may differ to some extent. Apart from the architecture, we adopt mostly generalized modules among synthesis approaches and nature language processing to verify the ideology in general purpose and provides chance for further modification. Future works might continue exploring specific performance of different target languages, more approximating performance to pre-trained models, different zero-shot strategies, and other alternatives to joint encoding scheme in cross-lingual synthesis.

CRediT authorship contribution statement

Han Zhang: Conceptualization, Methodology, Experiment, Software, Investigation, Writing – original draft. **Suyi Yang:** Conceptualization, Validation, Analysis, Writing – review & editing. **Hongqing Zhu:** Supervision, Revision and finalizing of the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61872143.

References

[1] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1316–1324.

[2] M. Zhu, P. Pan, W. Chen, Y. Yang, DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5802–5810.

[3] B. Li, X. Qi, T. Lukasiewicz, P. Torr, Controllable text-to-image generation, in: Advances in Neural Information Processing Systems, 2019, pp. 2063–2073.

[4] T. Qiao, J. Zhang, D. Xu, D. Tao, MirrorGAN: Learning text-to-image generation by redescription, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200–2011 Dataset, Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[6] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008, pp. 722–729.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.

[8] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 214–223.

[9] M. Artetxe, G. Labaka, E. Agirre, Translation artifacts in cross-lingual transfer learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 7674–7684.

[10] S. Schuster, S. Gupta, R. Shah, M. Lewis, Cross-lingual transfer learning for multilingual task oriented dialog, 2018, arXiv preprint arXiv:1810.13327.

[11] T. Schuster, O. Ram, R. Barzilay, A. Globerson, Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing, 2019, arXiv preprint arXiv:1902.09492.

[12] A. Conneau, G. Lample, Cross-lingual language model pretraining, Adv. Neural Inf. Process. Syst. 32 (2019) 7059–7069.

[13] C. Lee, H. Lee, Cross-lingual transfer learning for question answering, 2019, arXiv preprint arXiv:1907.06042.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[15] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text-to-image synthesis, in: Proceedings of the 33rd International Conference on Machine Learning, 2016.

[16] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.

[17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5907–5915.

[18] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, StackGAN++: Realistic image synthesis with stacked generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1947–1962, http://dx.doi.org/10.1109/TPAMI.2018.2856256.

[19] D. Zhou, K. Sun, M. Hu, Y. He, Image generation from text with entity information fusion, Knowl.-Based Syst. 227 (2021) 107200, http://dx.doi.org/10.1016/j.knosys.2021.107200.

[20] Z. Ji, W. Wang, B. Chen, X. Han, Text-to-image generation via semi-supervised training, in: 2020 IEEE International Conference on Visual Communications and Image Processing, IEEE, 2020, pp. 265–268.

[21] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, D. Xu, Image classification by cross-media active learning with privileged information, IEEE Trans. Multimed. 18 (12) (2016) 2494–2502, http://dx.doi.org/10.1109/TMM.2016.2602938.

[22] J. Wehrmann, D.M. Souza, M.A. Lopes, R.C. Barros, Language-agnostic visual-semantic embeddings, in: Proceedings of the International Conference on Computer Vision, 2019, pp. 5804–5813.

[23] W. Lan, X. Li, J. Dong, Fluency-guided cross-lingual image captioning, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1549–1557.

[24] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013.

[25] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100, 000+ questions for machine comprehension of text, in: The 2016 Conference on Empirical Methods on Natural Language Processing, 2016, pp. 2383–2392.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[27] X.L. ana Qingcai Chen, Y. Liu, J. Siebert, B. Hu, X. Wu, B. Tang, Decomposing word embedding with the capsule network, Knowl.-Based Syst. 212 (2021) 106611, http://dx.doi.org/10.1016/j.knosys.2020.106611.

[28] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017.

[30] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[31] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[32] S. Li, R. Pan, H. Luo, X. Liu, G. Zhao, Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling, Knowl.-Based Syst. 218 (2021) 106827, http://dx.doi.org/10.1016/j.knosys.2021.106827.

[33] T. Mikolov, Q.V. Le, I. Sutskever, Exploiting similarities among languages for machine translation, 2013, arXiv preprint arXiv:1309.4168.

[34] C. Xing, D. Wang, C. Liu, Y. Lin, Normalized word embedding and orthogonal transform for bilingual word translation, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1006–1011.

[35] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data, in: International Conference on Learning Representations, 2018.

[36] H. Aldarmaki, M. Diab, Context-aware cross-lingual mapping, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. pp. 3906–3911.

[37] T. Schuster, O. Ram, R. Barzilay, A. Globerson, Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 1599–1613.

[38] Y. Wang, W. Che, J. Guo, Y. Liu, T. Liu, Cross-Lingual BERT transformation for zero-shot dependency parsing, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 5725–5731.

[39] K.M. Hermann, P. Blunsom, Multilingual distributed representations without word alignment, 2013, ArXiv Preprint 2013: arXiv:1312.6173.

[40] X. Wang, L. Zhu, Y. Yang, T2VLAD: Global-Local sequence alignment for text-video retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 5079–5088.

[41] C.-S. Wu, S. Hoi, R. Socher, C. Xiong, TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020, pp. 917–929.

[42] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, 2020, ArXiv Preprint 2020: arXiv:2007.01852.

[43] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, 2020, pp. 8440–8451.

[44] M. Robnik-Sikonja, K. Reba, I. Mozetic, Cross-lingual transfer of twitter sentiment models using a common vector space, 2020, ArXiv E-Prints, 2020: arXiv:2005.07456.

[45] L. Tian, X. Zhang, J.H. Lau, Rumour detection via zero-shot cross-lingual transfer learning, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2021, pp. 603–618.

[46] J. Martínek, C. Cerisara, P. Král, L. Lenc, Cross-lingual transfer learning for dialogue act recognition, 2020, ArXiv E-Prints, 2020: arXiv:2005.09260.

[47] A. Abad, P. Bell, A. Carmantini, S. Renais, Cross lingual transfer learning for zero-resource domain adaptation, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 6909–6913.

[48] Z. Chi, L. Dong, F. Wei, W. Wang, X.-L. Mao, H. Huang, Cross-lingual natural language generation via pre-training, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, pp. 7570–7577.

[49] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, G. Hu, Cross-Lingual machine reading comprehension, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, in: 1586–1595.

[50] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: MT Summit, Vol. 5, 2005, pp. 79–86.

[51] Z. Zhedong, Z. Liang, G. Michael, Y. Yi, X. Mingliang, S. Y.-Dong, Dual-path convolutional image-text embeddings with instance loss, ACM Trans. Multimed. Comput. Commun. Appl. 16 (02) (2020) 1–23, http://dx.doi.org/10.1145/3383184.

[52] D. Bahdanau, K.H. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015.

[53] Multi30K: Multilingual english-german image descriptions, in: Proceedings of the 5th Workshop on Vision and Language, 2016, pp. 70–74.

[54] J. Rajendran, M.M. Khapra, S. Chandar, B. Ravindran, Bridge correlational neural networks for multilingual multimodal representation learning, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 171–181.

[55] English Uncase BERT Model, https://huggingface.co/bert-base-uncased.

[56] German BERT Model, https://github.com/dbmdz/berts.

[57] Multilingual BERT Model, https://github.com/google-research/bert/blob/master/multilingual.md.

[58] Pre-trained Downstream Generators, https://drive.google.com/file/d/1_UTX4wnThj99ysYCAiT-T7SXtVz4vC4u/view?usp=sharing.