IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 21, 2021, accepted March 16, 2021, date of publication March 23, 2021, date of current version April 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068162

Show Auto-Adaptive and Tell: Learned From the SEM Image Challenge

JING SU AND JING LI

Shanghai Ultra-Precision Optical Manufacturing Engineering Center, Department of Optical Science and Engineering, Fudan University, Shanghai 200433, China

Corresponding author: Jing Li (lijing@fudan.edu.cn)

ABSTRACT Scanning electron microscopy (SEM) has been widely used in optical material science. However, a considerable quantity of human resources is required to analyze and describe SEM images. In recent years, the application of computer technology in material science and engineering developed endlessly. Computer science, including data processing, simulation technique, and mathematical model, promotes material science progress tremendously. Moreover, deep learning has been achieved success in image classification and image analysis. In this paper, we propose a novel automatic analysis tool using a triplet neural network called show auto-adaptive and tell to analyze optical SEM images automatically. Firstly, we collected SEM images and corresponding captioning from previous papers and built a database. Then, a triplet neural network with proposed loss function to train the show auto-adaptive and tell model on 60% of the dataset for SEM images analysis, test on 30% and validate on 10%. Finally, experiment on the four metrics index as the evaluation criterion shows that the novel method gets better performance than previous work.

INDEX TERMS Scanning electron microscopy, show, adaptive and tell model, adversarial training.

I. INTRODUCTION

The morphology of the surface plays a critical role in the function of the films. It forces us to consider how the physical material SEM images should bring real essence and underlying information that the material to own. Material science has made significant advances over the years that researchers are looking for the intelligent method [1]— the physical and artificial system combined to find the way out of the physical difficulty. Besides, with the current research results, it is announced that the applications of the artificial intelligent method in the physical world [2]–[4] are widely used. In the artificial intelligence field, to explore the potential of computer vision and natural language processing [5], the implementation of the application can describe the content of the image with the accuracy of expression [6]. The computer can express the image in a brief and concise sentence [7], [8]. Further, the physical property contents depend on many experimental results on corresponding images in various conditions. Images involving a large amount of information analyze the essentials of amplification material structure by artificial intelligence method [9], [10]. It is a fascinating new physical analysis method that is promising in the future. Moreover, we would make it more palatable to researchers in the physical domain by raising analysis ability and details extracted to express semantic information of image – not just object recognition or image classifications [11].

Due to the difficulty of feature expression [12], [13] in the image accurately, our research demonstrates the strong demand for the language model combined visual understanding [14]–[17]. Artificial intelligence methods have been taking settle status in the history of physics. Physicists have grown to dabble the way in first principles digging for information across many different data. It is well known that obtaining the physical database is a relatively expensive process and requires a very long test time to access from the experiment. There are many issues and difficulties at initial dataset-building development, such as the lack of sharing database. Therefore establish the physical database, including data updates and retrieval performing strong demand for artificial intelligence systems, especially in deep learning. The alternative approach to obtain data [18] is calculated by a formula defined from the fundamental theorem. Compared to traditional optimization methods [19], [20], deep learning yields the preferable performance and lower computation resources based on matrix computation. Though, of course,

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

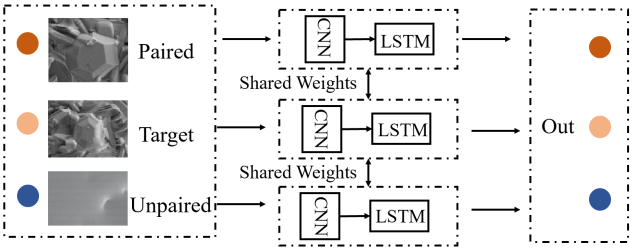

FIGURE 1. The overview of triple network. The images extract information embedding by shared weights CNN and LSTM.

it is impossible to dig deeper or a new view of physical mechanism, primarily to limit the scope of the method. To verify the new branch of physical method correctness and quality, it is a useful mathematical tool, and it turns out the most available data in various optical SEM images [9], [10].

We build a dataset via thesis that would be one of the fundamental quickest, cheapest modes. It is practical and affordable to acquire all researcher experiments images and projects information from the thesis. Draw on the concept of ''paired-associate learning'' the way has an independent contribution of image analysis. The overview of a triple network, a shared weight network to process paired, unpaired, and target images, is shown in figure.1. Information that can be paired, like the shape and kinds of material, works well for machine learning. To break through the apparent limitation of single image simplicity, a complete in-depth understanding of all aspects of the images is realized by a 'cross-domain' image learning style. In such a way, it would be much more likely to learn the common characteristics in the paired images and the distinguishing feature in the unpaired images. Hashing is the most popular and influential technology is used to decode image for retrieving [21], [22].

For training and evaluating, we create the new database of SEM images from around 1000 thesis by other researchers' similar works [16], [23] as the cross-domain and paired images. Our method owns the outstanding consistent performance on test data by learning extensive scale train data (60% for training, 10% for valid, and 30% for testing). Our contributions of the paper are summarized as follows:

1. We built the new dataset for SEM images and Captioning from approximately 1000 thesis.

2. An approach is proposed to the automatic paired image instead of manually in preprocess.

3. A novel method with a triplet neural network and a new loss function exploits the relation between the different domain attributes.

4. The experiment shows that our method gets better performance than the exciting method [24]–[28] on the four metrics index as the evaluation criterion.

II. RELATED WORK

A. ARTIFICIAL INTELLIGENCE WITH PHYSICAL DEEP LEARNING

Artificial neural network (ANN) is a fundamental approach predicting reinforcement on the physical determined by inherent characteristics. As the ANN inherent characteristic that can take a suitable format as output and input, respectively. Another application – the empirical model developed by physically utilizing ANN expresses the dynamics process interpreted as the transition between two formations [18]. In recent years, there have been many attempts in ANN to explore the novel branch of physics. The ANN learning, similar to physics learning, is studying physics conclusion experiments data to conclude. Therefore the approach would do a smooth unexpected beneficial effect.

B. DEEP LEARNING

The deep learning technology through the complex algorithm turns to be a realistic mapping to achieve better results. How to use deep learning to promote physical understanding depends on a significant study. A large amount of data needed at once quite costly limits producing the physical data and makes it not practical in the system. Whatever the structure of the network we devise, enough data is key to the deep learning available. A critical fact lies in the data is obtained by empirical to reduce complexity and resource consumption. On a deeper level, while we cannot learn well on the way diving into a poorly-understood and unknown physical region. The network can be significantly utilized to dig the essence when we have enough raw experimental data.

Meanwhile, it is an opportunity for deep learning in physics. The image involves a significant number and complex information with the advantages of easy procurement. We would grab the database, including the image and its corresponding text from other researchers' work that is the particular category of images obtained by the same device to have enough data to train and test.

C. NATURAL LANGUAGE PROCESSING

Recent works mainly focused on generating natural language from visual data, aiming to learn and develop the whole sentences [15]–[17]. Show auto-adaptive and tell method can be divided into three categories: deep hashing as image retrieval technology for pairing, convolutional neural networks as feature extraction for preprocessing, extended short term memory network as the syntactic sentence for decoding. The approach for building the fast running and pairing images model is proposed based on binary format. Besides, the method behind the algorithm is partition using a deep hash method to analyze the images.

In the research of image preprocessing, feature extraction among various methods, the CNNs is a relatively mature and practical approach that the significant contributions have proved the tremendous success for image manipulation. LSTM has been developed to decode the image features into sentences, performing semantic analysis and natural language processing. With the conditional random field (CRF) appearance, it is a giant leap to apply that technology for propulsion sentence-making.

In this paper, we improve the organization of captions by using the CRF module to make the most of every caption

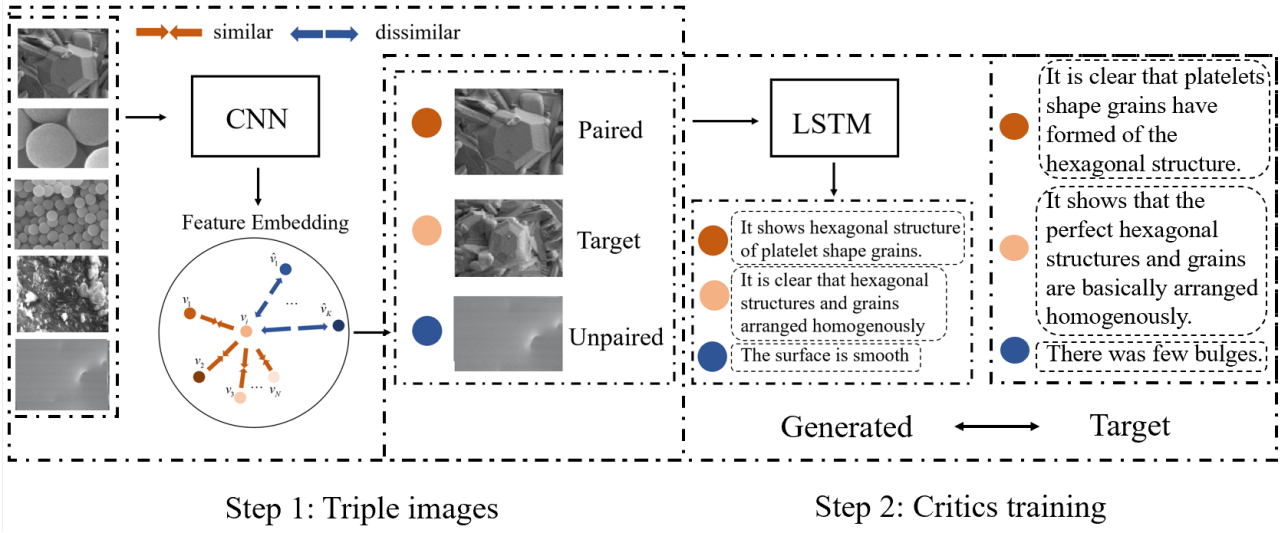

FIGURE 2. The overview of the method. Step 1: The image extract feature embedding by the shared weight CNN and classified into triplets by Euclidean distance. (dissimilar is unpaired, similar is paired). Step 2: The triple images generate three captions through a shared weight LSTM network, and then the loss between the generated captions and the standard captions.

during training. Based on the deep hashing method that pair images would replace creating data manually by automatically. We would retrieve and retain the unpaired and paired images for a particular image. As the system brings together these seem irrelevant algorithms connecting cheek by jowl, the results could lift the project to a new level. Our model can also train with the cascade network to boost the performance further.

III. METHODOLOGY

We can supply critical words for the trained model to form sentences and make it the most out of every phrase. Traditionally, training image subsets are selected manually in the adaptive problem that is very time-consuming. We extracted image features and auto classify paired and unpaired based on the distance between different images. In this paper, the overview of the framework is two parts, the first part is to form triple images, and the second part critic training showed in figure.2. In critics training process, sentences generated from paired images should be similar, sentences of unpaired images should be dissimilar.

The module aims to generate captions composing of the CNN image encoders CNN(x_n), the LSTM language decoders LSTM(d) in Eq(1). Over the view of the method, the adversarial part explains the log-likelihood of the different domain images loss L and the generated amount to explain captioning loss $J(\theta)$ in Eq(2).

$$d = CNN(x^n), \quad y^n = LSTM(d) \tag{1}$$

$$\begin{aligned} J(\theta) &= -log(P_\theta(y^n|x^n,\theta)) \\ &= -\frac{1}{M}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T_m} log(P(y_t^n|x^n, y_1^n \ldots y_{t-1}^n)) \\ L &= l_{\theta,m}(f_m^{generated,n}, f_m^{d,n}, f_m^{paired,n}) \\ &\quad + l_{\theta,d}(f_d^{generated,n}, f_d^{paired,n}, f_d^{unpaired,n}) \end{aligned} \tag{2}$$

where N is the number of images, T is the length of the sentence, x^n is the n^{th} target image, y^n is the n^{th} target caption, d is the image feature captured by CNN, which is the input of LSTM module, θ is the network parameters, $P_\theta(\cdot)$ is the outputs the possibility distribution of each time step t word by the previous time step $t-1$ word. Loss function $l_{\theta,m}$ and $l_{\theta,d}$ is described in the III-B section, captioning loss $J(\theta)$ is described in the III-A section. The critics training framework is shown in figure.3, which is second step of the overall framework in figure.2. The framework comprises a generated model and an adversarial training model for encoding the image and decoding of the sentence. The first model is the standard CNN-LSTM based image features. In this paper, we deal with the captions using policy gradient given a reward of each step. For the second model, to achieve the unpaired or paired images for dissimilarity and similarity, the deep hash technique is an effective method. Given a target image x^1, corresponding paired image x^2, unpaired image x^3 as well as generated target image caption $y_1 = \{y_1^1, y_2^1, \ldots.., y_n^1\}$, y_i^1 is the i_{th} word in target caption, paired caption y^2, unpaired caption y^3, respectively. The difference is the loss expression, are the paired and unpaired (y^1,y^2,y^3) and paired and target (y^1,y^2,y^d), where y^d is the ground-truth, respectively. The novel method proposes through iteratively updating in algorithm 1.

A. GENERATED MODEL

Sentence learning is tasked with learning the function from ground-truth caption words and maximizing the like-hood of the time series P_θ. Generally, this series of captions should be generated in the development with previous prediction words by LSTM. Thus, given the crossing Entropy used as an indicator to train the algorithms, the performance of the

Algorithm 1 Training Process

Require: target, paired and unpaired sentences y^1, y^2, y^3, image x^1, x^2, x^3, initial parameters θ, learning rate η, the number of paired images-sentences in paired as well as unpaired N_m, N_d, image-sentences paired x

Ensure: sentence y

Initial parameters with random and setting learning rate
while $l_{\theta,m} > 10^{-5}$ and $l_{d,m} > 10^{-5}$ **do**
  for $i = 0; i < N_c; ++i$ **do**
    image $x \to y$, where y^n is generated sentence using Eq(1)
    Computing the fusing sentences domain f_m using Eq(7)
    Adam update θ using $l_{\theta,m}$, as in Eq(10)
    image $x \to y$, where y^n is generated sentence using Eq(1)
    Computing the fusing sentences domain f_m using Eq(9)
    Adam update θ using $l_{d,m}$, as in Eq(11)
    Update the learning rate
  end for
  for $i = 0; i < N_d; ++i$ **do**
    image $x \to y$, where y^n is generated sentence using Eq(1)
    for $j = 0; j < T; ++j$ **do**
      $\{x, y\} \to \{Q((x, y^k_{t-1}), t^k_t)\}$ with Monte Carlo using Eq(6)
    end for
    Computing $J(\theta)$ using Eq(5)
    Adam update θ using $\Delta J(\theta)$
  end for
end while

scheme is followed.

$$J_{n,m}(\theta) = -\sum_{t=1}^{T_m} P_\theta(y^n_t(x^n, y^n_{t-1})Q((x^n, y^n_{t-1}), y^n_t)) \quad (3)$$

The primary reward section should replace Loss by the comparable approach throughout the image X and captions Y measuring the state-action $Q((x, y_{t-1}), y_t)$. y_t shows each time step t words, and y_{t-1} shows the previous time step $t-1$ in the sentence. Generated N sentence executes the empirical mean, and T_m represents the length of the generated m^{th} sentence as follows.

$$J(\theta) = \frac{1}{M}\sum_{n=1}^{N}\sum_{m=1}^{M} J_{n,m}(\theta) \quad (4)$$

$$\Delta J(\theta) = \nabla_\theta J(\theta) \quad (5)$$

N represents the batch number, M represents the number of the total sentences, $\Delta J(\theta)$ in Eq(5) represents the entire gradient. Proper policy optimization could determine the sample y^n_t and achieve the system remark performance with computed gradient $\Delta J(\theta)$. Since the vast workload that comes with the new method, we have to round out the Monte Carlo showed as followed.

$$\begin{aligned} Q((x^k, y^k_{t-1}), y^k_t) &\approx \frac{1}{K}\sum_{k=1}^{K} R(\vec{y^k}|\cdot) \\ &\approx P_{\theta,m} + P_{\theta,d} \end{aligned} \quad (6)$$

$R(\vec{y}|\cdot)$ is the reward in Monte Carlo, K represents the sampled complete sentences, $P_{\theta,m}$ and $P_{\theta,d}$ represents loss function of sentences relevant to paired and unpaired images in following. Evaluation index is according to the reward of state-action $Q((x^k, y^k_{t-1}), y^k_t)$.

B. ADVERSARIAL MODEL

Sentences generated can be viewed from two aspects: (1) complete all the analogous tasks together in paired images (2) complete the generated sentences to distinguish them from irrelevant in unpaired images. To get better performance, we should follow the principle of the two rules. The clustering is addressed in our model, as it classifies the 'paired images' or 'unpaired images'. We dump out the raw data by decoding the image by CNN structure spliced together the corresponding image data. Finally, with the fully connected layer and the softmax layer, we can get the probability generated.

The task of $l_{\theta,d}$ and $l_{\theta,m}$ is to distinguish between the target, paired, unpaired and target, generate, paired, respectively. For verifying the relevant classified domain, the imaging procedure can be installed to fall into three categories: 'paired', 'unpaired', or 'generated' data. We can achieve the probability by connecting the fusion layer and the softmax layer. The formulated parameters as the classification training objective can be adjusted and updated by stochastic gradient descent algorithm, thereby applied in a more significant effect training process.

In the network training process, connecting the weight and bias variable should be alternating optimized. The system blends generated and adversarial part from multiple paired image and training events to create an integrated view of each image x and sentences y. Each test confirms that a method produces the expected output when we alternative update. In turn, we trace out of another expectation, owning the small values of $P_{\theta,d}$ and $P_{\theta,m}$. Thus the two training procedures ratio and error can be controlled.

$$f_m = sign(tanh(W_x \cdot x + b_x) \cdot tanh(W_{y1} \cdot y + b_{y1})) \quad (7)$$

$$f_d = sign(tanh(W_{y2} \cdot y + b_{y2})) \quad (8)$$

$$\begin{aligned} &dist(f^1_m, f^2_m) \\ &\quad = (Len - (f^1_m)^T \cdot f^2_m) \end{aligned} \quad (9)$$

$$\begin{aligned} l_{\theta,m} = -\sum_{n=1}^{N_m}(&\alpha - f^{d,n}_m \cdot f^{paired,n}_m - f^{d,n}_m \cdot f^{generated,n}_m \\ &- f^{generated,n}_m \cdot f^{paired,n}_m - log(1 + exp(\alpha - f^{d,n}_m \\ &\cdot f^{paired,n}_m - f^{d,n}_m \cdot f^{generated,n}_m - f^{generated,n}_m \cdot f^{paired,n}_m))) \end{aligned} \quad (10)$$

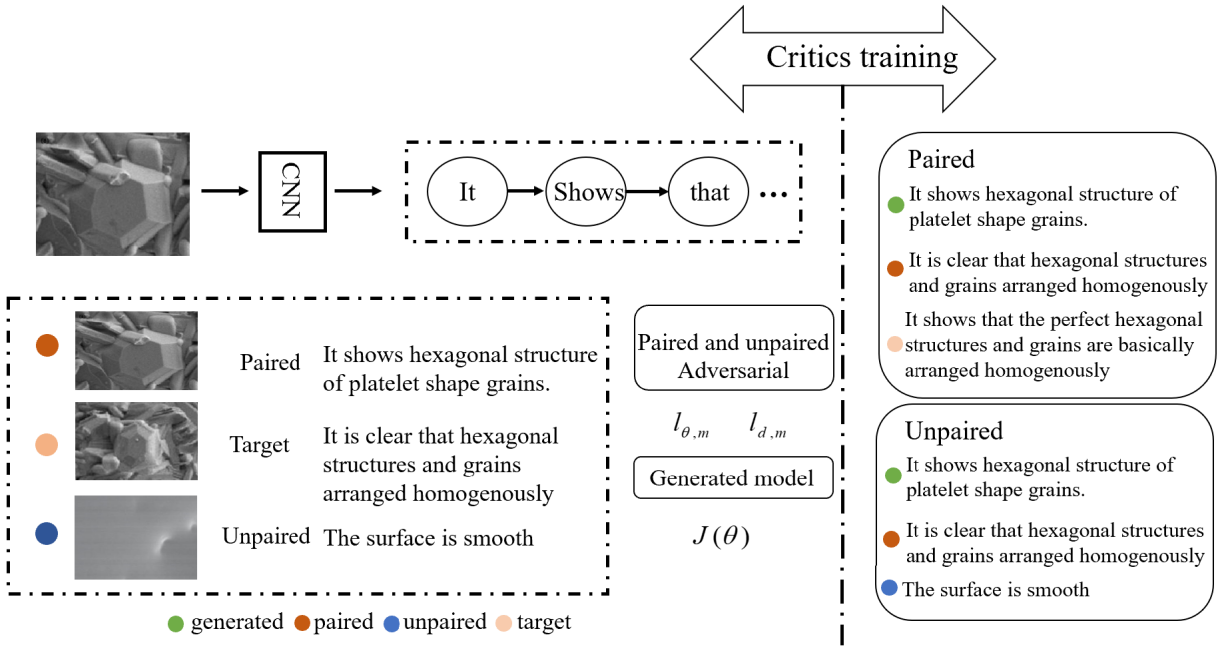

FIGURE 3. The overview of the critics training process. Left panel: the image encode, and the sentence generated section. Right panel: the critics observe the sentences from the paired and unpaired images and discriminate from the target sentence. During the training, both left and right are iteratively updated to achieve the goals in algorithm 1.

$$l_{\theta,d} = -\sum_{n=1}^{N_d}(\alpha - f_d^{paired,n} \cdot f_d^{unpaired,n} - f_d^{paired,n} \cdot f_d^{generated,n} - f_d^{generated,n} \cdot f_d^{unpaired,n} - log(1 + exp(\alpha - f_d^{paired,n} \cdot f_d^{unpaired,n} - f_d^{paired,n} \cdot f_d^{generated,n} - f_d^{generated,n} \cdot f_d^{unpaired,n}))) \quad (11)$$

where $W_x, b_x, W_{y1}, b_{y1}, W_{y2}, b_{y2}$ are parameters to be learned, $\cdot$ denotes element-wise multiplication, f_m, f_d is the hash code over three classes: target, paired, and generated data. α is the constant, Len is the length of the hash codes. N_m, N_d is the number of paired image-sentences in paired as well as unpaired, respectively. In Eq(7), the image x and sentence y representations are fused via element-wise multiplication; otherwise, in Eq(8), pass the representation to generate a different domain. In Eq(9) is the Hamming distance between the two binary codes. Eq(10) and Eq(11) can be formulated as the negative log-likelihood as loss functions.

IV. EXPERIMENTS

To prove the practical model, we give the datasets of SEM. We show that our model can describe the majority of physical images correctly. We can get better performance by combining paired and unpaired images. The deep learning network has already contributed enormously to the construction of the physical region.

A. DATA PREPROCESSING

To gain access to images and sentences that are obtained from thesis getting from other researchers. Therefore, the method can save experimental resources and improve work efficiency. We use the same vocabulary in the process, including special begin-of-sentences and end-of-sentences tokens. To set the target vocabulary datasets, we can remove non-dominant words in all sentences.

We first pre-train the caption on the dataset. Next, we use the deep hash to pair the images automatically. Then, we update the parameters from the training set in the domain by combined training. Finally, we evaluate the method on the different targets to represent the results.

B. BASELINE

We re-implement show and tell and show adaptive and tell as our baseline methods. Show and tell consists of the image features extractor and a sentences generator. The format is a CNN model to extract, and the latter is an LSTM model to generate. Image comprehension is then a process while retrieving the meaning of features based on image recognition and forms whole sentences by syntactic and semantic parsing. Finally, we enforce the model by adding an adaptive part. The precise nature of critic is crucial and shows adaption and has been put forward to account for the difference between the unpaired and paired images. As a new practical approach, critics-based planning is used extensively and effectively in the visual region.

We further categorize the domain shift automatically to the target image. The advantage of the present scheme lies in a significant reduction in human resources.

C. ANALYSIS OF SENTENCES

The model devotes the microscopic image, with sharpness and detail in images that are greatly exceeded the traditional

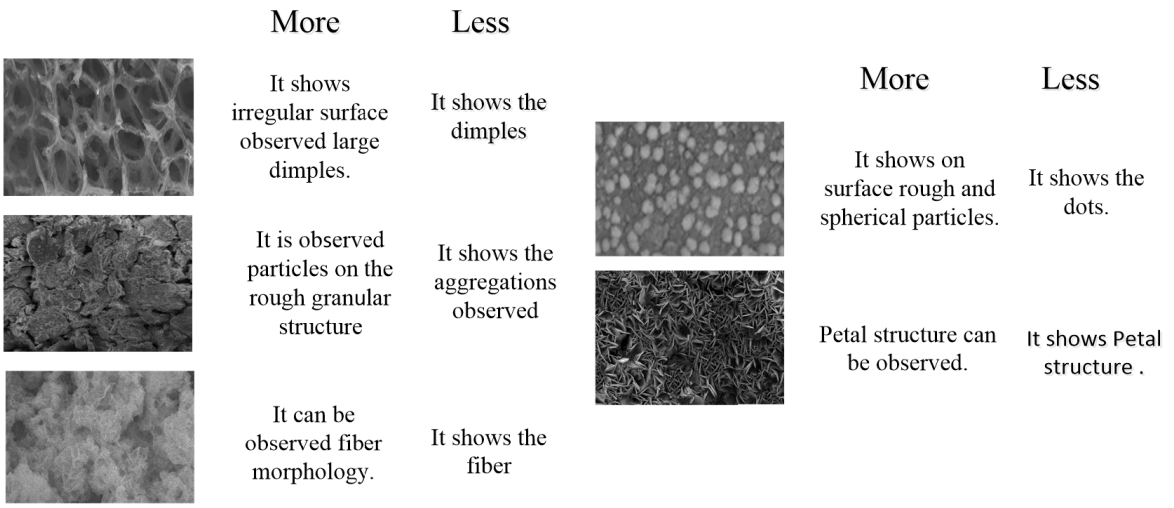

FIGURE 4. Examples of different domain images datasets for two words limitations.

TABLE 1. Results of SEM images datasets based on four baseline methods.

The captions number	target	Bleu	ROUGE	Meteor	CIDEr
/	$CUB-200$[5-6]	91.4%	58.6%	27.6%	24.8%
/	$Oxford-102$[16]	85.6%	72.1%	36.4%	29.3%
/	$TGIF$[22]	47.5%	37.0%	14.5%	22.2%
/	$Flickr30k$[9]	62.1%	42.1%	16.7%	32.6%
$More$	SEM	66.0%	67.7%	30.3%	46.67%
$Less$	SEM	58.3%	62.6%	36.1%	45.92%

TABLE 2. Results of SEM images datasets based on different methods.

method	The captions number	target	Bleu	ROUGE	Meteor	CIDEr
Show and tell	/	SEM	32.3%	26.6%	21.1%	22.3%
Show attention and tell	/	SEM	29.6%	33.2%	35.1%	30.1%
Show adaptive and tell	/	SEM	35.8%	51.2%	28.6%	33.2%
Ours	$More$	SEM	66.0%	67.7%	30.3%	46.67%
Ours	$Less$	SEM	58.3%	62.6%	36.1%	45.92%

classifier in the visual region. A depiction of the image is represented in figure.4. The features and relationship of each detail are tucked away in no real consistency images. To analyze the effectiveness of the method, we make an ablation comparison with the different word constraints. The results are precise that the restriction of the word in objective results is most likely to find critical features in the image and deliver the intended outcome. It would take further experiments to describe the key features of the well-functioning system, adopting the word constraint in sentences. We can find ''the large dimple'', ''irregular'' and ''particles'' by making accurate observations and ''particles'' by the rough statement. The word constraint helps the model differentiate and captures the subtleties and the main features in the surfaces. A depiction of the image is represented in figure.5 for different methods. We find that triplets method can absorb and capture the details ''flowerlike'' between similar images, and distinguish the dissimilar between different images.

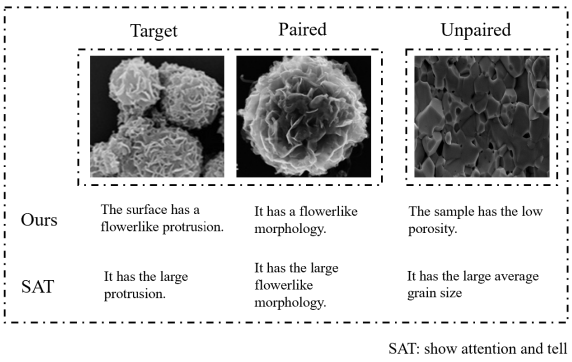

FIGURE 5. Examples of different domain images datasets for different methods.

D. ADAPTION ACROSS FOUR TARGET METHODS

According to the SEM image dataset, the metrics aim at testing the target domain by comparing the different metrics to prove its effectiveness and practicality. As its generalization, the table results as a guide for (Belu 65.92%, ROOGE 67.65%, Meteor 30.31%, CIDER 46.67%) the other datasets to verify the approach could be applicable in a particular scene. In sum, our method shows the slight floating for the various attributes ($\pm 10\%$) across different datasets, despite large domain shift to demonstrate the dataset and the method with feasibility and effectiveness descriptions in details vs. compress difference in adjective usage

auto-paired study. The different words constraint between analysis detailed and compressed together reveals that it is the most challenging scenario with the same high precision and recall rate in the model. The approach gives more detailed expressions of attributes with higher precision and lower recall rate than a less complicated expression with just the opposite.

As the image is shown in the figure.4, descriptions usually are the main character with fewer words and the more characters with more words describing surface and particles with fine-grained object attribute. There are more adjective expressions in complex images, while more adjectives are applied to the visual computer with a high rate of repeated codes and wrong cedes. On top of that, we would enhance the constraints with the number of words to lower tolerance fault. Examples in figure.5 show that our model can accurately describe the details on the surface and general shape of the object. Table 1 shows that our method also significantly improves over the main index parameters. For comparison, the table results as a guide for the other methods to verify the approach could be more competitive. Table 2 shows that our method also significantly improves over the main methods.

We have proposed the auto-paired procedure with the deep hashing method (unsupervised method) without labeling them manually. To analyze the effectiveness of the method, we make an ablation comparison with either one. We argue that automatic pre-processes are vital for saving human resources physical resources and simplifying the production, with few minutes instead of few days setting manually.

V. CONCLUSION

We propose the novel training procedure for captioning and new datasets with physical images for the visual computer to describe the images. The novel automatic triple training method is applied to the new datasets introduced to further improve the caption generation process consistently on the four challenging metrics for judging the results. Furthermore, it is pretty tricky to handle the details of output with word constraints. In the future, the method is inching towards perfection to the other dataset by deeply analyzed means.

REFERENCES

[1] B. Karanov, M. Chagnon, F. Thouin, T. A. Eriksson, H. Bülow, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end deep learning of optical fiber communications," *J. Lightw. Technol.*, vol. 36, no. 20, pp. 4843–4855, Oct. 15, 2018.

[2] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 127–135.

[3] Y. Jo, S. Park, J. Jung, J. Yoon, H. Joo, M.-H. Kim, S.-J. Kang, M. C. Choi, S. Y. Lee, and Y. Park, "Holographic deep learning for rapid optical screening of anthrax spores," *Sci. Adv.*, vol. 3, no. 8, Aug. 2017, Art. no. e1700606.

[4] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha, "Unsupervised deep learning for optical flow estimation," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

[5] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3137–3146.

[6] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting CNN models for apparent personality trait regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 55–63.

[7] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," 2017, *arXiv:1712.02051*. [Online]. Available: http://arxiv.org/abs/1712.02051

[8] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[9] R. Li, T. Zeng, H. Peng, and S. Ji, "Deep learning segmentation of optical microscopy images improves 3-D neuron reconstruction," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1533–1541, Jul. 2017.

[10] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Comput. Methods Biomechanics Biomed. Eng., Imag. Visualizat.*, vol. 6, no. 3, pp. 283–292, May 2018.

[11] O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, Jun. 2016.

[12] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[13] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.

[14] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2407–2415.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[16] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 521–530.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[18] A. Prša, E. Guinan, E. Devinney, M. DeGeorge, D. Bradstreet, J. Giammarco, C. Alcock, and S. Engle, "Artificial intelligence approach to the determination of physical properties of eclipsing binaries. I. The EBAI project," *Astrophys. J.*, vol. 687, no. 1, p. 542, 2008.

[19] Y. D. Valle, G. K. Venayagamoorthy, S. Mohagheghi, J.-C. Hernandez, and R. G. Harley, "Particle swarm optimization: Basic concepts, variants and applications in power systems," *IEEE Trans. Evol. Comput.*, vol. 12, no. 2, pp. 171–195, Apr. 2008.

[20] B. Du, Q. Wei, and R. Liu, "An improved quantum-behaved particle swarm optimization for endmember extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6003–6017, Aug. 2019.

[21] X. Wang, Y. Shi, and K. M. Kitani, "Deep supervised hashing with triplet labels," in *Proc. Asian Conf. Comput. Vis.* Springer, 2016, pp. 70–84.

[22] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," 2015, *arXiv:1511.03855*. [Online]. Available: http://arxiv.org/abs/1511.03855

[23] J. Peters and J. A. Bagnell, "Policy gradient methods," *Scholarpedia*, vol. 5, no. 11, p. 3698, 2010.

[24] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," Tech. Rep., 2011.

[26] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

[27] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[28] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo, "TGIF: A new dataset and benchmark on animated GIF description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4641–4650.