Research Proposal

Xin Cai

Title: Neural Earth Observer

1. Introduction

An increasing number of satellites have been monitoring dynamic spatial-temporal processes on the Earth's surface and continuously generating massive amounts of data. For instance, the Sentinel 2 multispectral satellite constellation acquires data at up to 10m resolution in 13 spectral bands every two to five days. Despite the rapid development of remote sensing systems, the abundance of generated data has not yet been fully exploited. A large body of 7.76 TiB Sentinel 2 data was published on a daily basis. Still, only 7.6% of all published images in 2018 have actually been downloaded. Accordingly, 12 out of 13 published images remain unused. Similar phenomena can be observed for the Sentinel 1, 3, and 5 missions. Traditional methods used to deal with remotely sensed data are heavily reliant on hand-crafted features, such as Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Brightness Index (BI), Inverted Red-Edge Chlorophyll Index (IRECI), and Enhanced Vegetation Index (EVI), followed by machine learning algorithms such as Random Forest (RF) and Support Vector Machine (SVM) for making predictions. Despite the achieved satisfactory performance, the preprocessing pipeline requires extensive domain-specific expertise and causes excessive computational burden, impeding the exploitation of the full potential of continuously growing satellite data.

Recent years have witnessed the impressive achievements made by deep learning models in many research fields, such as Computer Vision (CV), Natural Language Processing (NLP) and Automatic Speech Recognition (ASR). The dominant feature of deep learning models is to integrate feature extraction and task-specific prediction into a unified architecture, allowing joint optimization and obviating the need for laborious manual feature engineering. In particular, this characteristic also implies the suitability of employing deep learning techniques to tackle challenges in the field of remote sensing, given the fact that models that do not strictly require extensive data preprocessing would facilitate the efficient utilization of publicly available satellite data. In fact, there has been a growing interest in applying deep learning techniques to address remote sensing problems in recent years, such as land use/land cover classification[1, 2], crop type classification[3, 4, 5, 6, 7], change detection[8, 9], saliency detection[10], super-resolution[11, 12], and multi-modality information fusion[11, 12, 13]. Despite the attained promising results, there are still many challenges needed to be addressed for the successful application of deep learning in the field of remote sensing. Firstly, unlike the universal applicability of ResNet[14] in computer vision tasks, neural architectures appropriate for processing satellite data have not yet been established. For instance, researchers have proposed to use Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), the combination of these two architectures, or the attention mechanism to tackle the classification problem of Satellite Image Time Series (SITS) data. Recently, it has been demonstrated in [3] that temporal information is much more important than spatial features in the classification problem of SITS data, because the resolution of sentinel 2 images restricts the richness of texture information, thereby damaging the logic of CNNs. Therefore, model complexity should be largely allocated to the temporal component when designing hybrid models. Due to the characteristics of satellite data, such as multi-spectral and multi-temporal, exploring neural architectures which are more suitable for processing them is still of significant importance. Besides, the remarkable success of current deep learning-based remote sensing systems has been achieved in a fully supervised fashion, requiring large amounts of annotated data. Semi-supervised or unsupervised representation learning has received increasing attention in other research areas, particularly benefiting from advances in deep generative models. However, it has been rarely explored in the field of remote sensing. Several primitive attempts [15, 16] relied mainly on using autoencoders to reconstruct the input data, which serve as auxiliary supervision to compensate the scarcity of annotated data. Recent research [17] has presented a framework which combines Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs) to distil disentangled feature representations from SITS data in an unsupervised manner, showing the great potential of unsupervised pre-training and setting a precedent for further exploring this research direction. Last but not least, multi-modality information fusion plays a vital role in remote sensing, as it allows for the use of complementary multi-modal information captured by different types of sensors, in the visible spectrum or not, from satellites or planes, with various spatial precision. Multi-modal image registration techniques are at the core of realizing multi-modality information fusion. While it has long been a predominant topic in the field of remote sensing, deep learning-based multi-modal image registration methods specifically designed for remotely sensed data has rarely been explored. Moreover, how to combine the registration network with its subsequent networks for downstream tasks in an end-to-end fashion is still an open problem. Recently, several pioneering works have explored to integrate the registration component explicitly [11] or implicitly [12] into a unified architecture for Multi-Frame Super-Resolution (MFSR), demonstrating the great potential of multi-modality information fusion. Due to the broad spectrum of applications in the field of remote sensing, there are inevitably many challenges specific to a particular application scenario, which may be further considered in my future research.

To summarize, the primary research objectives are as follows:

- exploring the neural network design space to identify more effective neural architectures which accommodate the characteristics of remote sensing data;
- developing unsupervised representation learning methods to fully exploit large amounts of unlabelled data, especially by leveraging advances in generative deep learning;
- developing deep learning-based multi-modal image registration models for fusion

multi-source remotely sensed data;

- developing unified neural architectures which realize image registration through their built-in components and can be trained in an end-to-end fashion for the exploitation of multi-modality remote sensing data.

2. Related Work

2.1 Various Neural Architecture Designs

Recently, a wide range of deep neural network architectures has been proposed to deal with remote sensing data. Generally, CNNs, RNNs, self-attention networks, and their combined variants have been successfully applied to address many problems in the field of remote sensing, such as classification of SITS data[4, 2, 5, 6] and change detection[8, 9]. Most of the work simply adapted well-established deep learning models in CV, NLP or ASR to process remote sensing data, failing to accommodate their characteristics. Several recent works have shed light on architecture designs based on comprehensive empirical evaluations. In [3], the authors have discovered that the temporal structure of Sentinel 2 data is richer than the spatial structure for crop type classification, and consequently proposed to allocate most of the parameters (up to 90%) to model the temporal structure when designing hybrid models. In [18], the authors have observed that preprocessing can consistently improve the overall classification accuracy for all models, which challenges the superiority of deep learning models in processing remote sensing data, that is removing the need of heavily relying on domain-specific expertise (preprocessing techniques). However, they further discovered that self-attention networks and RNNs performed competitively well on raw data compared to preprocessed data. Based on these observations, researchers in[7] have proposed a novel architecture to process SITS data, leveraging advances in 3D point cloud processing. Specifically, they regarded medium-resolution (10m per pixel) Sentinel 2 images as sets of unordered elements and consequently introduced the pixel-set encoder inspired by the work [19] as an efficient alternative to CNN encoders, followed by the bespoke transformer architecture[20] for modelling temporal relations.

2.2 Unsupervised Representation Learning for Satellite Data

The impressive performance that has been attained by existing deep learning-based remote sensing models can be largely attributed to supervised learning, which requires enormous amounts of annotated data. Due to the laborious annotation process of multispectral satellite data and the overfitting problem caused by supervised training, it is beneficial to explore unsupervised representation learning methods to enhance the efficient utilization of data and the generalization ability of extracted feature representations. Recently, deep learning-based unsupervised representation learning has attracted increasing attention, which can be generally subsumed under three categories: (i) deep generative learning-based methods, (ii) self-training-based methods, and (iii) self-supervised learning-based methods.

There has been rapid progress made in deep generative models, such as GANs[21], VAEs[22], deep autoregressive models[23], normalizing flow-based models[24] and their hybrid variants[25, 26, 27]. The core motivation of using generative models to perform unsupervised representation learning is that the capability of capturing rich and complex distributions of modelled data arises from identifying intrinsic and meaningful structures of modelled data, therefore being able to benefit downstream tasks. For example, it has been demonstrated that adversarial autoencoder (AAE) [26] can be employed to perform semi-supervised learning, disentangling style and content of images, and unsupervised clustering. Closely related to unsupervised representation learning, there has been an emerging trend to learn disentangled feature representations using deep generative models, especially using various adversarial losses to regularize the feature extraction process. Identifying factors of variation in the modelled dataset is beneficial for extracting task-specific features and isolating undesirable noise factors. For example, InfoGAN [28] and β-VAE [29] are introduced to learn interpretable factorized features in an unsupervised manner. A two-step disentanglement method[30] is used to extract label relevant information for image classification. Moreover, disentangled representation learning has enabled great success in Image-to-Image (I2I) translation[31, 32, 33, 34] and Unsupervised Domain Adaptation (UDA)[35, 36]. UDA has been gaining increasing attention in image classification, object detection, and semantic segmentation[37], aiming at improving the generalization ability of deep learning models on unseen scenarios. In the pioneering work [17], researchers have presented a framework combining VAE and GAN methods to learn disentangled representations for SITS data in an unsupervised manner. The disentangled representations can isolate common information of the entire time series data from exclusive information specific to each image and have proven to be useful for several downstream tasks, such as image classification, image retrieval, image segmentation and change detection.

Apart from deep generative learning-based methods, there have been other types of research attempts to tackle unsupervised representation learning. Self-training-based methods[38, 39] chiefly follow the paradigm of alternately performing clustering and supervised learning, in which clustering algorithms are used to generate pseudo labels to serve as supervisory signals, leading to general-purpose feature representations. Additionally, self-supervised learning as an emerging research field has been actively studied in recent years[40, 41, 42], where discriminative approaches have been adopted rather than generative methods based on the assumption that pixel-level generation is computationally costly and may not be necessary for representation learning. One of the most common strategies for self-supervised learning is to predict future, missing or contextual information. For example, recent work in unsupervised learning has successfully used these ideas to learn word representations by predicting neighbouring words[43]. For images, predicting colour from grey-scale[44] or the relative position of image patches[45] has also been shown useful for extracting high-level features.

2.3 Multi-Modal Image Registration and Information Fusion

Multi-modality information fusion is of prime importance for a variety of applications in remote sensing, as multi-modal sensors allow gathering a wide range of physical properties, which are complementary and yield richer scene representations. The realization of multi-modal information fusion relies on reliable image registration techniques. Classic (mono- or multi-modal) image registration techniques attempt to warp a source image to match a target one via a non-linear optimization process, seeking to maximize a predefined similarity metric [46]. The slow optimization process and the difficulty of manually designing local descriptors and similarity metrics have led to the recent development of deep regression models. Concretely, mono-modal image registration can be decomposed as the rigid transformation which can be captured by a global affine transformation matrix and local non-rigid deformations which are generally characterized by dense displacement vector fields. Early research methods have attempted to learn the parameters related to linear and non-linear transformations through supervised learning [47, 48, 13]. These methods require ground truth deformation fields which are extremely hard to collect and thus synthetically generated. The downside of these supervised methods is that synthetic deformation ground truths are based on human prior knowledge or conventional image registration algorithms, meaning that these methods cannot fully capture the diversity of real-world correspondence or the performance is limited by those conventional methods. Therefore, many research efforts have been devoted to unsupervised approaches [49, 50, 51], i.e., learning spatial transformations guided by similarity metrics and with smoothness regularization. Multi-modal image registration presents more challenges than its mono-modal counterpart due to the difficulty of measuring similarity across different modalities, which may cause significant appearance variations. Several recent research efforts [52, 53] have shown promising results by leveraging advances in deep generative models, especially those related to I2I translation [31, 32, 33, 34]. The principle of these methods is to employ I2I translation techniques to convert the moving image (to be registered) to the modality identical to the reference/fixed image, thereby circumventing the difficulty of manually devising multi-modal similarity metrics. Such registration techniques are focused primarily on medical image registration tasks, which have rarely been explored in remote sensing setting.

An important application in the field of remote sensing is MFSR, which aims to reconstruct hidden high-resolution details from multiple low-resolution views of the same scene. While Single Image Super-Resolution (SISR) has attracted much attention in the computer vision and deep learning communities [54, 55] in the last decade, not much work has explored the end-to-end deep learning system for the more general setting of MFSR, which needs to address the additional challenges of co-registration and fusion of multiple low-resolution images. Recently, there are several pioneering works in MFSR with remote sensing data. In [12], the authors have presented the first deep learning architecture–HighRes-net, which learns its sub-tasks in an end-to-end fashion: (i) co-registration, (ii) fusion, (iii) up-sampling, and (iv) registration-at-the-loss. DeepSUM [11] is also a recently proposed approach that exploits both spatial and temporal correlations to perform MFSR, which consists of three components: (i) the SISR network, (ii) the registration network, and (iii) the fusion network.

3. Methodology

The literature review section 2 has covered representative research work in remote sensing using deep learning techniques, providing a solid foundation on which my future research will be conducted.

3.1 Neural Architecture Design

As stated in 2.1, recent research[3] has shown that the relatively low spatial resolution of multi-temporal satellite images may challenge the justification of adopting CNNs as spatial encoders as texture and shape information in these satellite images is limited. In [7], researchers have proposed an innovative alternative the pixel-set encoder by regarding satellite images as sets of unordered elements. Indeed, significant advances have been made in 3D point cloud processing[19, 56, 57], giving rise to various powerful set encoders and decoders. Besides, a closely related and emerging research field is Graph Convolutional Networks (GCNs)[58, 59, 60], aiming at generalizing convolutions to non-Euclidean data. As a result, exploring the potential of adapting such set encoders/decoders and GCNs to process satellite data, especially SITS data, is a promising research direction. Furthermore, these novel architectures may allow for designing operators that can simultaneously process spatial-temporal data rather than using separate components adopted by most existing methods.

3.2 Joint Discriminative and Generative Learning

The second primary objective for my future research is to develop unsupervised representation learning methods for remote sensing data. The section 2.2 has given an overview of current representative solutions, including (i) deep generative learning-based methods, (ii) self-training-based methods, and (iii) self-supervised learning-based methods. Firstly, integrating deep generative components into current discriminative frameworks to achieve the efficient utilization of satellite data, especially unlabelled data, is a promising research direction. Besides, it has been reported that deep learning models trained on the source domain are likely to encounter dramatic performance degradation when deployed on a novel target domain in a variety of applications, such as semantic segmentation[61, 62] and person re-identification[63, 64] in CV, which has spurred the development of UDA techniques[35, 36]. While this problem has rarely been studied in the remote sensing community so far, it is reasonable to assume the existence of such a phenomenon. For example, satellite images captured in different regions and seasons may exhibit significant appearance differences caused by various factors, such as meteorological conditions, illuminations, and terrain conditions. The domain shift problem coupled with heterogeneous sensors (i.e., multi-modal information fusion, which will be discussed in the next section) used to capture these data would complicate the situation even further. Taking it a step forward, it is of significant importance to study the generalization ability of current deep learning-based remote sensing systems. One of the most promising strategies to enhance the robustness and generalization ability of feature representations is to distil disentangled feature representations. For example, the pioneering work[17] has demonstrated the usefulness of decomposing feature representations of SITS data into the common component encoding information shared by the entire time series and the exclusive component encoding information specific to each image in the time series. Designing various adversarial losses to regularize the feature embedding space is a feasible scheme to achieve feature disentanglement, which also suggests the importance of injecting generative components into discriminative frameworks. Lastly, self-training and self-supervised learning techniques are beneficial for extracting robust feature representations without annotated data. To the best of my knowledge, leveraging advances in these research fields to devise deep learning-based remote sensing systems has rarely been explored so far.

3.3 Multi-Modal Information Fusion

In remote sensing, images of the Earth can be acquired by different types of sensors, in the visible spectrum or not, from satellites or planes, with various spatial precision. The analysis of these images captured by multi-modal sensors allows the monitoring of ecosystems and their evolution (drought monitoring, natural disasters and associated help planning), urban growth, as well as the automatic creation of maps or more generally digitizing the Earth. As stated in the section 2.3, multi-modal image registration is indispensable for multi-modality information fusion. Currently, learning-based mono-modal image registration techniques[49, 50, 51] have shown superiority over traditional registration methods because of replacing costly optimization with expeditious inference (i.e., the forward-pass of neural networks). Multi-modal image registration has also gained increasing attention since several early research attempts [52, 53] discovered that I2I translation techniques can be used to realize cross-modality conversion, which allows training the registration network using simple and reliable mono-modality similarity metrics. Given the fact that the majority of advances has been made in the setting of medical image registration, it is worthwhile to adapt these approaches for remote sensing applications by adhering to the same principle. Furthermore, as demonstrated by recent work in MFSR [11, 12], there are many challenges needed to be addressed to devise an end-to-end deep learning model which incorporates image registration as its subcomponent. Therefore, the third objective for my future research is to develop unified neural architectures which can realize image registration through their built-in components and be trained in an end-to-end fashion for the exploitation of multi-modality remote sensing data.

4. Timeline

Generally, I plan to divide the period of three years of working towards a PhD degree into two parts: 1) the first two years will focus on publishing three papers on international conferences or journals with potential research topics surrounding around those stated in section 3.1, 3.2, 3.3, respectively; 2) the final year will be used to complete my PhD dissertation by summarizing the previous work.

The detailed timetable of the first two years are outlined as follows:

- $1 \sim 4$ months: conducting a comprehensive literature review and selectively reproducing results of some representative research work, revolving around the central topic: developing novel neural network architectures by leveraging advances in 3D point cloud processing and GCNs;
- $5 \sim 8$ months: proposing my own methods and publishing on international conferences or journals, striving to establish a simple but strong baseline model for processing SITS data by conducting rigorous ablation studies;
- $9 \sim 12$ months: conducting a comprehensive literature review and selectively reproducing results of some representative research work, revolving around the central topic: developing unsupervised representation learning methods to fully exploit large amounts of unlabelled satellite data, by leveraging advances in deep generative models, self-training and self-supervised learning;
- $13 \sim 16$ months: proposing my own improvements and publishing on international conferences or journals with a particular focus on extracting robust and generalizable feature representations without using annotated data;
- $17 \sim 20$ months: conducting a comprehensive literature review and selectively reproducing results of some representative research work, revolving around the central topic: developing unified neural architectures which can realize image registration through their built-in components and be trained in an end-to-end fashion for the exploitation of multi-modality remote sensing data;
- $21 \sim 24$ months: proposing my own methods and publishing on international conferences or journals;

References

[1] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy, “M3fusion: A deep learning architecture for multi-{Scale/Modal/Temporal} satellite data fusion,” *ArXiv*, vol. abs/1803.01945, 2018.

[2] R. Interdonato, D. Ienco, R. Gaetano, and K. Ose, “Duplo: A dual view point deep learning architecture for time series classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 91–104, 2019.

[3] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, “Time-space tradeoff in deep learning models for crop classification on satellite multi-spectral image time series,” in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 6247–6250.

[4] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, “3d convolutional neural networks for crop classification with multi-temporal remote sensing images,” *Remote Sensing*, vol. 10, no. 1, p. 75, 2018.

[5] M. Rußwurm and M. Körner, “Convolutional lstms for cloud-robust segmentation of remote sensing imagery,” *arXiv preprint arXiv:1811.02471*, 2018.

[6] M. Rußwurm and M. Korner, “Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[7] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, “Satellite image time series classification with pixel-set encoders and temporal self-attention,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 325–12 334.

[8] H. Chen and Z. Shi, “A spatial-temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[9] L. Mou, L. Bruzzone, and X. X. Zhu, “Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 924–935, 2018.

[10] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, “Dense attention fluid network for salient object detection in optical remote sensing images,” *IEEE Transactions on Image Processing*, 2020.

[11] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, “Deepsum: Deep neural network for super-resolution of unregistered multitemporal images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2019.

[12] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, “Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery,” *arXiv preprint arXiv:2002.06460*, 2020.

[13] A. Zampieri, G. Charpiat, N. Girard, and Y. Tarabalka, “Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 657–673.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] R. Kemker, R. Luu, and C. Kanan, “Low-shot learning for the semantic segmentation of remote sensing imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 6214–6223, 2018.

[16] L. Mou, P. Ghamisi, and X. X. Zhu, “Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 391–406, 2017.

[17] E. H. Sanchez, M. Serrurier, and M. Ortner, “Learning disentangled representations of satellite image time series,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 306–321.

[18] M. Rußwurm and M. Körner, “Self-attention for raw optical satellite time series classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 421–435, 2020.

[19] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *arXiv preprint arXiv:1406.2661*, 2014.

[22] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[23] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” *arXiv preprint arXiv:1601.06759*, 2016.

[24] L. Dinh, D. Krueger, and Y. Bengio, “Nice: Non-linear independent components estimation,” *arXiv preprint arXiv:1410.8516*, 2014.

[25] D. J. Rezende and S. Mohamed, “Variational inference with normalizing flows,” *arXiv preprint arXiv:1505.05770*, 2015.

[26] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, “Adversarial autoencoders,” *arXiv preprint arXiv:1511.05644*, 2015.

[27] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, “Variational approaches for auto-encoding generative adversarial networks,” *arXiv preprint arXiv:1706.04987*, 2017.

[28] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, pp. 2172–2180, 2016.

[29] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[30] N. Hadad, L. Wolf, and M. Shahar, "A two-step disentanglement method," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 772–780.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[33] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *arXiv preprint arXiv:1711.11586*, 2017.

[34] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[35] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[36] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[37] L. Zhang, "Transfer adaptation learning: A decade survey," *arXiv preprint arXiv:1903.04687*, 2019.

[38] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," *Advances in Neural Information Processing Systems*, vol. 29, pp. 2110–2118, 2016.

[39] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[40] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[41] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[44] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[45] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[46] B. Zitova and J. Flusser, “Image registration methods: a survey,” *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[47] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, “Nonrigid image registration using multi-scale 3d convolutional neural networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 232–239.

[48] N. Schneider, F. Piewak, C. Stiller, and U. Franke, “Regnet: Multimodal sensor registration using deep neural networks,” in *2017 IEEE intelligent vehicles symposium (IV)*. IEEE, 2017, pp. 1803–1810.

[49] H. Li and Y. Fan, “Non-rigid image registration using fully convolutional networks with deep self-supervision,” *arXiv preprint arXiv:1709.00799*, 2017.

[50] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “An unsupervised learning model for deformable medical image registration,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9252–9260.

[51] C. Stergios, S. Mihir, V. Maria, C. Guillaume, R. Marie-Pierre, M. Stavroula, and P. Nikos, “Linear and deformable image registration with 3d convolutional neural networks,” in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 13–22.

[52] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, “Unsupervised deformable registration for multi-modal images via disentangled representations,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 249–261.

[53] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, “Unsupervised multi-modal image registration via geometry preserving image-to-image translation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 410–13 419.

[54] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[55] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[56] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *arXiv preprint arXiv:1706.02413*, 2017.

[57] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, “Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5589–5598.

[58] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[59] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” *arXiv preprint arXiv:1706.02216*, 2017.

[60] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” *arXiv preprint arXiv:1710.10903*, 2017.

[61] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, “Fully convolutional adaptation networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6810–6818.

[62] Z. Wang, M. Yu, Y. Wei, R. Feris, J. Xiong, W.-m. Hwu, T. S. Huang, and H. Shi, “Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 635–12 644.

[63] Z. Zhong, L. Zheng, S. Li, and Y. Yang, “Generalizing a person retrieval model hetero-and homogeneously,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–188.

[64] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 2138–2147.