https://doi.org/10.31871/IJNTR.9.1.7
International Journal of New Technology and Research (IJNTR)
ISSN: 2454-4116, Volume-9, Issue-1, January 2023 Pages 06-10

A Review on Self-Supervised Learning

Athul Raj, Srinjoy Dutta

***Abstract*—In recent years, the field of AI has made great strides in developing AI systems that learn from massive amounts of carefully labeled data. This supervised learning model has proven to be successful in training specialized models to perform exceptionally well on the task for which they were trained. Unfortunately, there is a limit to what the field of AI can go with supervised learning alone.**

Supervised learning is a bottleneck for building smarter general-purpose models that can multitask and learn new skills without the need for large amounts of labeled data. Practically speaking, it is impossible to label everything in the world. There are also some tasks that don't have enough labeled data, such as training a translation system for resource-limited languages, or data that requires experts to label, such as medical data. If AI systems can gather deeper and more nuanced insights into reality beyond what is specified in the training dataset, they will be more useful and ultimately bring AI closer to human-level intelligence.

Self-supervised learning (SSL), also known as self-supervision, is an emerging solution to the challenge posed by data labeling. By building models autonomously, self-supervised learning reduces the cost and time to build machine learning models.

In our paper, we first look at SSL and how it can solve the challenge of data labeling. Then we look at some approaches to SSL that have been developed through the past years. And we finally conclude with what the future holds for SSL in the domain of Artificial Intelligence.

***Index terms*- Self-supervised learning, Artificial Intelligence**

I. INTRODUCTION

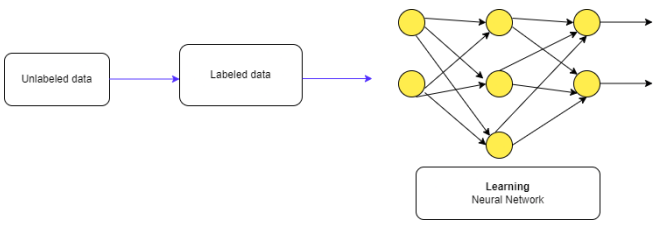

Fig. 1. Supervised Learning - Learning from labelled data

Supervised learning is the category of machine learning algorithms that require annotated training data. Deep Learning is the idea of building a system by assembling parameterized modules into a computation graph. Deep learning can be applied to different learning paradigms including supervised learning, reinforcement learning, as well as unsupervised or self-supervised learning. The majority of deep learning algorithms that have found their way into practical applications are based on supervised learning models. Image classifiers, facial recognition systems, speech recognition systems, and many of the other AI applications we use every day have been trained on millions of labeled examples.

Despite the huge contributions of deep learning to the field of artificial intelligence, there's something very wrong with it: It requires huge amounts of data. In fact, deep learning didn't emerge as the leading AI technique until a few years ago because of the limited availability of useful data and the shortage of computing power to process that data. It is tedious and costly to label such a huge amount of data. Also, for data in fields such as medicine, we need experts to label the data. Reducing the data-dependency of deep learning is currently among the top priorities of AI researchers. Self-supervised learning is one of several plans to create data-efficient artificial intelligence systems.

II. SELF-SUPERVISED LEARNING

Self-Supervised learning is a form of supervised learning where the data provides supervision. Generally, we withhold some part of the data and the network is tasked to predict it. The task defines a proxy loss and the network is forced to learn the part that we really care about, e.g., semantic representation, in order to solve it.

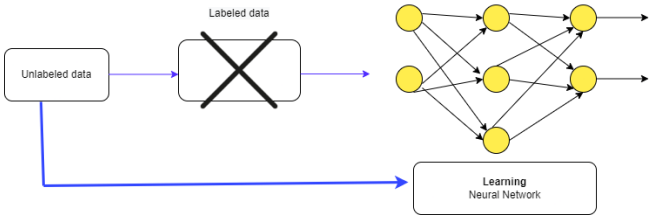

Fig. 2. Self-Supervised Learning - Learning from unlabeled data

III. WHY SELF-SUPERVISED LEARNING?

Some of the reasons we prefer self-supervised learning are: 1. It is expensive to create new datasets for each new task, 2. Some areas are supervision-starved, such as the medical field, where it is difficult to annotate every unlabeled data 3. Availability of a large number of unlabeled data such as audio, images, and videos from social media sites such as Facebook and YouTube.

Athul Raj, Electronics and Communication, National Institute of Technology, Silchar, India
Srinjoy Dutta, Electronics and Communication, National Institute of Technology, Silchar, India

IV. APPROACHES TO SELF-SUPERVISED LEARNING

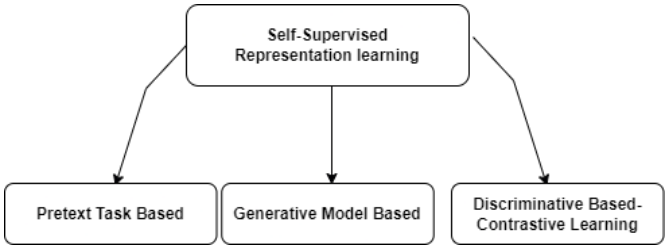

Fig. 3. Approaches to Self-Supervised Learning

From our reading and understanding of the different papers on self-supervised learning, the approach to self-supervised learning can be broadly classified into three different categories. 1. Pretext Task-Based 2. Generative Model-Based 3. Discriminative-based (or) Contrastive Learning

A. **PRETEXT TASKED-BASED:** Pretext task means the proxy task that we design for the model to learn from the unlabeled data. We will not give any explicit labels but some proxy tasks that the model can learn from the data. Once the model learns from it, we can transfer it to the main task. One such proxy task is predicting the rotation.

Let us look at some papers on pretext task-based approach to self-supervised learning.

1. **THE RELATIVE POSITIONING:** [1] The Relative Positioning approach was proposed by Carl Doersch, Abhinav Gupta1, and Alexei A. Efros in their paper Unsupervised Visual Representation Learning by Context Prediction. In this paper, the network is trained to predict relative position of two regions in the same Image. Here sampling of random pairs of patches in one of eight spatial configurations is done, and present each pair to a machine learner, providing no information about the patches original position within the image. The algorithm must then guess the position of one patch relative to the other. The underlying speculation is that doing properly on this project requires understanding scenes and objects, i.e., a suitable visible illustration for this assignment will need to extract objects and their components in order to purpose about their relative spatial region

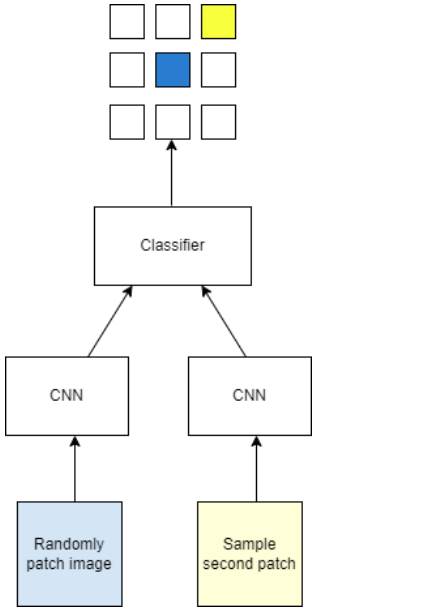

Fig. 4. Pretext task for predicting relative positioning of patches

2. **ROTATION:** Rotation approach was proposed by Spyros Gidaris, Praveer Singh, and Nikos Komodakis in their paper Unsupervised Representation Learning by Predicting Image Rotations. In their work, a novel formulation for self-supervised feature learning that trains a ConvNet model to be able to recognize the image rotation that has been applied to its input images. Despite the simplicity of this self-supervised task, it successfully forces the ConvNet model trained on it to learn semantic features that are useful for a variety of visual perception tasks, such as object recognition, object detection, and object segmentation. They exhaustively evaluated their method in various unsupervised and semi-supervised benchmarks and achieved in all of them state-of-the-art performance. This specifically approach manages to drastically improve the state-of-the-art results on unsupervised feature learning for ImageNet classification, PASCAL classification, PASCAL detection, PASCAL segmentation, and CIFAR-10 classification, surpassing prior approaches by a significant margin and thus drastically reducing the gap between unsupervised and supervised feature learning.

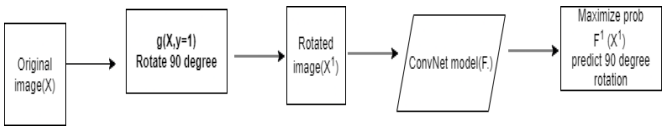

Fig. 5. Illustration of the self-supervised task that we propose for semantic feature learning. Given four possible geometric transformations, the 90-degree rotation, we train a ConvNet model F(.) to recognize the rotation that is applied to the image that it gets as input. F y (Xy) is the probability of rotation transformation y predicted by model F(.) when it gets as input an image that has been transformed by the rotation transformation y

3. **MULTIPLE PRETEXT:** [2] This paper presented a novel pretext-task for SSL called image enhanced rotation prediction (IE-Rot), which combines Rotation and IEs to learn useful representations focusing on not only information of object shapes but also information of textures. They

confirmed that IE-Rot with Rotation and Solarization improves the target performance across various datasets, tasks, and network architectures. Although this work focuses on improving Rotation to preserve the simplicity of the pretext-task, the idea of capturing both object shapes and textures can be extended to other pretext-tasks. As an important future direction, the idea to contrastive learning such as MoCo has been applied [3] IEs modify information of textures and are often used for data augmentation along with geometric transformations like rotation.

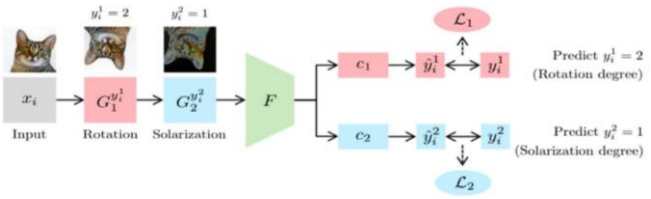

Fig. 6. Illustration of IE-Rot in the combination of Rotation and Solarization

This implies that IEs induce informative differences that are useful for training CNNs. Furthermore, in contrast to rotation, IEs hardly change geometric information of objects in images; this means IEs and Rotation have little or no interference with each other. Thus, IEs are suitable for combining with Rotation. Through the simultaneous prediction of rotation capture the information of not only object shapes but also textures.

4. **JIGSAW PUZZLE:** Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, by Mehdi Noroozi and Paolo Favaro [4] introduced a Context-Free Network (CFN), a CNN that can easily transfer features between detection/classification and puzzle reassembly tasks. They build a convolutional neural network (CNN) that can be trained to solve Jigsaw puzzles as a pretext task, which requires no manual labeling, and then later re-purposed to solve object classification and detection. They built a learning plan that generated an average of 69 puzzles for 1.3 million images and converged in just 2.5 days. To maintain the compatibility across tasks we introduce the context-free network (CFN), a siamese-ennead CNN. The CFN takes image tiles as input and explicitly limits the receptive field (or context) of its early processing units to one tile at a time. The CFN includes fewer parameters than AlexNet while preserving the same semantic learning capabilities. By training the CFN to solve Jigsaw puzzles, we learn both a feature mapping of object parts as well as their correct spatial arrangement. The learned features are evaluated by both classification and detection, and the experimental results show that they are superior to the previous best model. More importantly, the performance of these features closes the gap with those studied while observing.

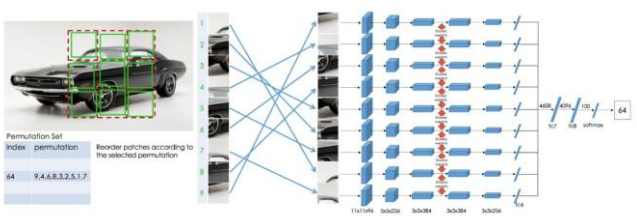

Fig. 7. Context Free Network. The figure illustrates how a puzzle is generated and solved.

B. **GENERATIVE MODEL BASED:** A generative model is a probabilistic rather than deterministic. The model is merely a fixed calculation, such as taking the average value of each pixel in the dataset. It is not generative because the model produces the same output every time. The model must include a stochastic (random) element that influences the individual samples generated by the model. A generative model describes how a dataset is generated, in terms of a probabilistic model. By sampling from this model, we are able to generate new data.

1. **AUTOENCODERS:** Autoencoding is a data compression algorithm where the compression and decompression functions are 1. Data-specific - they will only be able to compress data similar to what they have been trained on 2. Lossy - the decompressed outputs will be degraded compared to the original inputs 3. Learned automatically from data examples - means that it is easy to train specialized instances of the algorithm that will perform well on a specific type of input. It doesn't require any new engineering, just appropriate training data. To build an autoencoder, we need three things: an encoding function, a decoding function, and a distance function between the amount of information loss between the compressed representation of your data and the decompressed representation (i.e., a" loss" function).

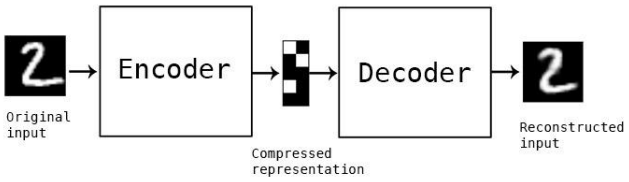

Fig. 8. Autoencoders

The encoder and decoder will be chosen to be parametric functions (typically neural networks), and to be differentiable with respect to the distance function, so the parameters of the encoding/decoding functions can be optimized to minimize the reconstruction loss, using Stochastic Gradient Descent. Two interesting practical applications of autoencoders are data denoising [5] and dimensionality reduction for data visualization [6]. Autoencoders have attracted so much research and attention are because they have long been thought to be a potential avenue for solving the problem of unsupervised learning, i.e., the learning of useful representations without the need for labels. They are a self-supervised technique.

2. **CONTEXT ENCODERS**: Context Encoders is a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. It was proposed by Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros of the University of California, Berkeley in their paper Context Encoders: Feature Learning by Inpainting [7]. In this paper, they cut off a patch of the image, and task the network to generate the area which was missing.

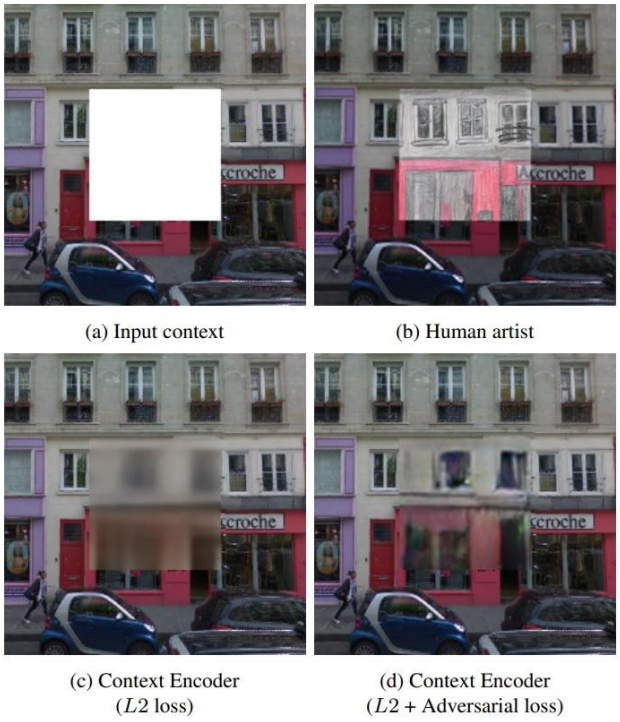

Fig. 9. The first image (a) is the input image. (b) is the image in which the patch is drawn by a human artist. (c) is the image they got from using Context Encoder with L2 loss and (d) is the image they got from using Context encoder with L2 and Adversarial loss. Their paper shows that their model is competitive with other models trained with auxiliary supervision.

The idea of context encoder was further extended by Suriya Singh, Anil Batra, Guan Pang, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and C. V. Jawahar in their paper Self-Supervised Feature Learning for Semantic Segmentation of Overhead Imagery [8]. Their paper was done on the satellite images, where instead of cropping out one patch, they cropped out multiple patches and tasked the network to fill out those patches.

3. **BiGANS:** Bidirectional Generative Adversarial Networks (BiGANs) were proposed by Jeff Donahue, Trevor Darrell, and Philipp Krähenbüh in their paper ADVERSARIAL FEATURE LEARNING [9]. They proposed BiGANS as a means of learning inverse mapping, and demonstrate that the resulting learned feature representation is useful for auxiliary supervised discrimination tasks, competitive with contemporary approaches to unsupervised and self-supervised feature learning.

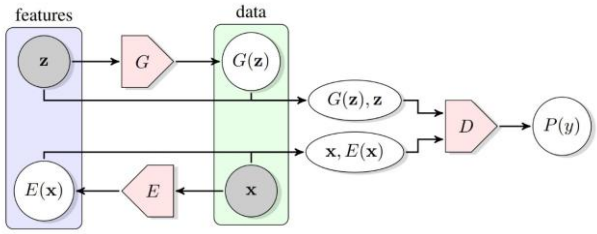

Fig. 10. Structure of BiGANS

The above figure depicts the structure of BiGANS. In addition to the generator G from the standard GAN framework (Goodfellow et al., 2014) [10], BiGAN includes an encoder E which maps data x to latent representations z. The BiGAN discriminator D discriminates not only in data space (x versus G(z)) but jointly in data and latent space (tuples (x, E(x)) versus (G(z), z)), where the latent component is either an encoder output E(x) or a generator input z.

C. **CONTRASTIVE LEARNING OR DISCRIMINATIVE LEARNING:** If we have an image and we take two different views of the same image and if the representations are taken from the same image, then both these representations should be close to each other. On the other hand, if we have a view of another image say chair and a view of another image, say a dog, then the representation learned from these two different images should be far apart. This is known as contrastive learning. Contrastive Learning was first proposed in the paper SimCLR.

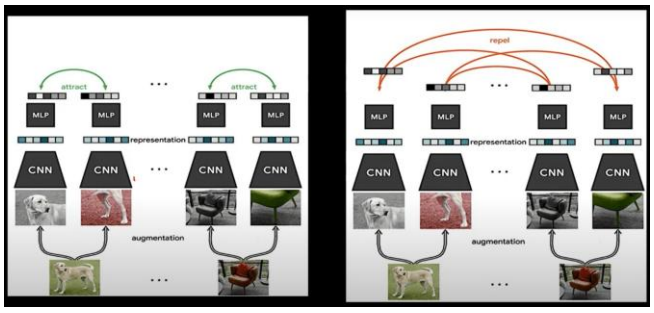

Fig. 11. Contrastive learning or Discriminative Based approach to Self-Supervised Learning

1. **SimCLR:** This paper, proposed by Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton presents SimCLR: A simple framework for contrastive learning of visual representations [11].
In this work, they present a simple framework and its instantiation for contrastive visual representation learning. They carefully study its components, and show the effects of different design choices. By combining their findings, they improve considerably over previous methods for self-supervised, semi-supervised, and transfer learning. Their approach differs from standard supervised learning on ImageNet only in the choice of data augmentation, the use of a nonlinear head at the end of the network, and the loss function.

2. **SimCLR V2**: SimCLR Version 2 was proposed by Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey Hinton in their paper Big Self-Supervised Models are Strong Semi-Supervised Learners [12]. This is now the state of art algorithm in terms of ImageNet top one percent accuracy. Here, the network is bigger, the ResNet size was both wider and has more depth and the MLP projection head has a greater number of layers. They follow the three step of learning 1. Do unsupervised pre-training in a task independent way 2. Then they do a supervised fine tuning on the small amount of labelled data set 3. Once fine-tuned, they do a self-training on the unlabeled data in a task specific way.

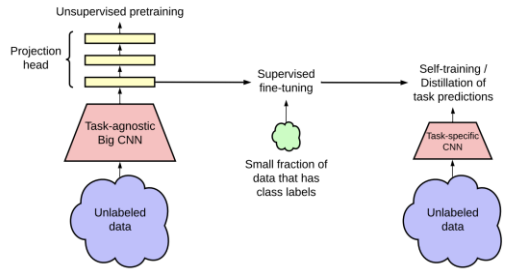

Fig. 12. Three steps of SimCLR V2

3. SwAV: [13] Unsupervised image representations have substantially decreased the space with supervised pre-training, particularly with the latest achievements of contrastive learning methods. These contrastive strategies commonly work on-line and depend on a massive variety of specific pairwise function comparisons, which is computationally difficult.

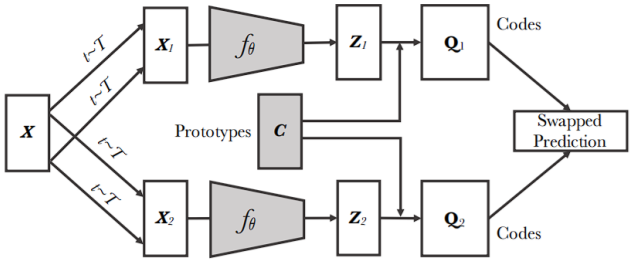

Fig. 13. Swapping Assignments between Views

In SwAV, first codes are obtained by assigning features to prototype vectors. The next step is solving of swapped prediction problem, wherein the codes obtained from one data augmented view are predicted using the other view. Thus, SwAV does not directly compare image features. Prototype vectors are learned along with the ConvNet parameters by backpropagation. Features are being learned by Swapping Assignments between multiple Views of the same image (SwAV). The features and the codes are learned online, allowing our method to scale to potentially unlimited amounts of data. In addition, SwAV works with small and large batch sizes and does not need a large memory bank or a momentum encoder.

V. CURRENT AND FUTURE APPLICATIONS OF SELF-SUPERVISED LEARNING

One of the best-known examples of SSL for natural language processing is Grammarly, an automated writing helper that provides better syntax and paraphrase methods. It has to analyze thousands of sentences to understand the context. Open AI's GPT-3 is another great example of self-supervised learning. This model analyzed half the internet and no team can manually label this amount of data. Instead, GPT-3 learns to create new content by understanding data structures such as text or code. We can see more examples of the use of self-supervised learning in manufacturing, especially computer vision. Today, self-supervised learning is used for face recognition, cancer diagnosis, text interpretation and writing. In the future, this technology will be used in more products such as medical and industrial robots, virtual assistants, and software systems of all kinds. SSL has the potential to revolutionize the autonomous vehicle market. After spending thousands of hours on the track, you can accumulate knowledge and navigate in unfamiliar situations.

VI. CONCLUSION

In our paper, we have reviewed Self-Supervised Learning, how it solves the challenge of data labeling and the various approaches to SSL. We have also discussed some practical applications where SSL is being used. As most deep learning systems today depend on supervised learning, they are all affected by the challenge of data labeling. But there is a large availability of unlabeled data. If we can successfully use them in our models, we can create models far superior to our present models. It will be closer to Artificial General Intelligence (AGI), which is the goal of all AI researchers. We firmly believe that self-supervised learning is the right step towards this goal and are very excited to continue our research in this domain.

REFERENCES

[1] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," 2016.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
[2] S. Yamaguchi, S. Kanai, T. Shioda, and S. Takeda, "Image enhanced rotation prediction for self-supervised learning," in 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 489–493.
[3] 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.
[4] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in European conference on computer vision. Springer, 2016, pp. 69–84.
[5] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," in Advances in Neural Information Processing Systems, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.
[6] X. Jiang, J. Gao, X. Hong, and Z. Cai, "Gaussian processes autoencoder for dimensionality reduction," in Advances in Knowledge Discovery and Data Mining, V. S. Tseng, T. B. Ho, Z.-H. Zhou, A. L. P. Chen, and H.-Y. Kao, Eds. Cham: Springer International Publishing, 2014, pp. 62–73.
[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544
[8] S. Singh, A. Batra, G. Pang, L. Torresani, S. Basu, M. Paluri, and C. V. Jawahar, "Self-supervised feature learning for semantic segmentation of overhead imagery," in BMVC, 2018.
[9] J. Donahue, P. Krahenb ¨ uhl, and T. Darrell, "Adversarial feature learn- ¨ ing," arXiv preprint arXiv:1605.09782, 2016.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.
[11] 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.
[12] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," Advances in neural information processing systems, vol. 33, pp. 22 243–22 255, 2020.
[13] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2021.