Manifold Constraint Regularization for Remote Sensing Image Generation

Xingzhe Su*, Changwen Zheng*, Wenwen Qiang, Fengge Wu, Junsuo Zhao, Fuchun Sun, *Fellow, IEEE*, Hui Xiong, *Fellow, IEEE*

***Abstract*—Generative Adversarial Networks (GANs) have shown notable accomplishments in remote sensing domain. However, this paper reveals that their performance on remote sensing images falls short when compared to their impressive results with natural images. This study identifies a previously overlooked issue: GANs exhibit a heightened susceptibility to overfitting on remote sensing images. To address this challenge, this paper analyzes the characteristics of remote sensing images and proposes manifold constraint regularization, a novel approach that tackles overfitting of GANs on remote sensing images for the first time. Our method includes a new measure for evaluating the structure of the data manifold. Leveraging this measure, we propose the manifold constraint regularization term, which not only alleviates the overfitting problem, but also promotes alignment between the generated and real data manifolds, leading to enhanced quality in the generated images. The effectiveness and versatility of this method have been corroborated through extensive validation on various RS datasets and GAN models. The proposed method not only enhances the quality of the generated images, reflected in a 3.13% improvement in Frechet Inception Distance (FID) score, but also boosts the performance of the GANs on downstream tasks, evidenced by a 3.76% increase in classification accuracy. The source code is available at https://github.com/rootSue/Manifold-RSGAN.**

***Index Terms*—Image Generation, Generative Adversarial Networks, Remote Sensing, Data Manifold.**

I. Introduction

In recent years, the field of artificial intelligence has witnessed the emergence of Generative Adversarial Networks (GANs) [1], a paradigm-shifting technology that has significantly advanced the capabilities of image generation. Among the myriad domains benefiting from this technology, remote sensing (RS) imagery stands out as a particularly promising area. Here, GANs have demonstrated their potential in data generation or augmentation [2, 3, 4, 5], haze or cloud removal [6, 7, 8, 9], and image super-resolution [10, 11, 12, 13]. In this paper, our primary focus is on the application of GANs to RGB images within the RS domain.

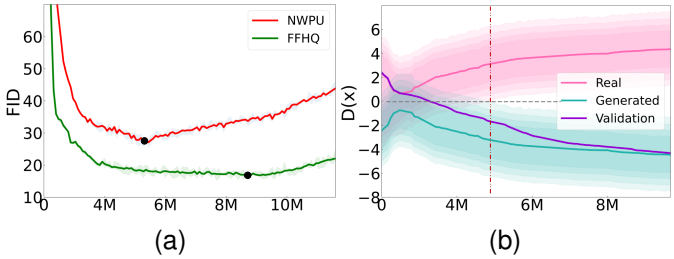

Fig. 1. The horizontal axis indicates the training process (the number of real images shown to the discriminator). (a) Training curves of StyleGAN2 on NWPU and FFHQ datasets. We randomly sample 30,000 training images from these two datasets, respectively. The GAN model diverges earlier when trained on NWPU dataset. (b) The outputs of the discriminator during training on NWPU dataset. As training progresses, the validation set of real images is misclassified as generated images, highlighting the discriminator's overfitting issue on NWPU dataset.

Current GAN methods achieve remarkable success with natural images. However, our research reveals a performance gap when applied to RS images, as shown in Fig.1(a). The vertical axis is the Fréchet Inception Distance (FID) score [14], a commonly-used image quality evaluation metric, with lower FID scores indicating higher image quality. We train the same GAN model on datasets of identical size: a natural image dataset (FFHQ [15]) and an RS image dataset (NWPU [16]). Interestingly, the model diverges earlier when trained on RS data compared to natural data (Fig. 1(a)), leading to lower quality generated RS images. To understand the reasons behind, we analyze the discriminator's outputs for real and generated images during training on NWPU dataset. Initially, the distributions of outputs for real and fake images overlap (Fig.1(b)). However, as the discriminator becomes more confident, these distributions gradually drift apart. Additionally, the accuracy of the discriminator on a separate validation set decreases as training progresses. Based on these observations, we conclude that GANs are more susceptible to overfitting on RS image dataset compared to natural image dataset. This overfitting leads to earlier divergence during training, ultimately resulting in lower quality generated RS images. More experiments are avaliable in Section III-B. As far as we know, this is the first study that identify the overfitting issue of GANs on RS data.

To understand why GANs are more prone to overfitting on RS data, we analyze the key differences between RS and natural image datasets. Compared to natural images, RS

* These authors contributed equally.

Xingzhe Su, Changwen Zheng, Wenwen Qiang, Fengge Wu and Junsuo Zhao are with the National Key Laboratory of Space Integrated Information System, Institute of Software Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China. E-mail: {xingzhe2018, changwen, qiangwenwen, fengge, junsuo}@iscas.ac.cn.

Fuchun Sun is with the National Key Laboratory of Space Integrated Information System, Department of Computer Science and Technology, Tsinghua University, Beijing, China. E-mail: fcsun@tsinghua.edu.cn.

Hui Xiong is with the Hong Kong University of Science and Technology, China. E-mail: xionghui@ust.hk.

Corresponding author: Wenwen Qiang (qiangwenwen@iscas.ac.cn).

images typically cover larger area, encompassing a wider variety of scenes and richer content. Consequently, we hypothesize that the intrinsic dimension of the RS dataset is larger than that of natural dataset. Highlighting a concept from [17], learning a manifold requires a number of samples that grows exponentially with the manifold's intrinsic dimension. Therefore, the discriminator might need more training images from the RS dataset, compared to natural images, to effectively capture the underlying data manifold. Our findings in Section III-B support this hypothesis, confirming that the intrinsic dimension of RS data is indeed higher than that of natural images. Consequently, with datasets of the same size, GANs are more likely to overfit on RS data compared to natural data.

The overfitting problem can significantly hinder the performance of GANs. When the discriminator becomes overfit to the training samples, its feedback to the generator becomes less meaningful, leading to training divergence, excessive memorization, and limited generalization [18, 19]. Consequently, the performance of GANs on tasks like data augmentation could also be compromised. While current research on GANs for RS data focuses on modifying network architectures and loss functions to improve performance on downstream tasks [20, 21], the issue of overfitting in the discriminator specifically for RS data is overlooked. This gap in research presents a unique opportunity to improve the overall effectiveness of GANs for RS applications.

Motivated by our findings, we propose to leverage the real data manifold to mitigate the overfitting problem of GANs on RS data. Specifically, we propose the manifold constraint regularization method (MCR) and integrate the regularization term into GANs' loss functions. MCR presents the discriminator with a more challenging task: capturing the underlying manifold of the real data, rather than simply memorizing the limited variations within the training dataset. This approach helps to eliminate complex model that performs well solely on training data and promotes model that performs well across the entire data manifold, thereby mitigating overfitting. Additionally, by minimizing MCR term, the generator aims to align its output manifold with that of the real images, capturing the authentic characteristics of the data and enhancing generative performance. Our methods is computationally efficient and integrates seamlessly within existing GAN architectures, requiring no major network modifications.

In summary, the contributions of this paper are:

- We identify a previously overlooked issue that the discriminator in GANs is prone to overfitting on RS image datasets, compared to natural images.
- We empirically demonstrate that RS datasets have higher intrinsic dimension than natural datasets, which leads to the discriminator easily overfitting to the RS datasets.
- To address overfitting, we propose MCR method, which consists of two components: feature distribution compactness measure and data manifold evaluation function.
- We theoretically prove that MCR can evaluate the distance between different manifolds. Extensive experiments across multiple RS datasets and GAN models verifies the effectiveness of our method.

In the remainder of this paper, we commence with a review of related works in Section II, and analyze plausible reasons for the subpar results of GAN models on RS images in Section III. Then we provide details of our proposed method in Section IV, followed by theoretical analysis of our approach in Section V. We show the substantially-improved performances of our method over standard GAN models and related techniques for solving overfitting in Section VI. Finally, we wrap up with a discussion in Section VII.

II. Related Work

A. Generative adversarial networks

GANs are notorious for training instability and mode collapse. Various adversarial losses have been proposed to stabilize the training or improve the convergence of the GAN models [1, 22, 23]. Additionally, numerous efforts have been made to address this issue using regularization methods [24, 25, 26, 27], or modifying network architectures [15, 28, 29, 30, 31, 32, 33]. Other than these problems, the overfitting of the discriminator is also a common challenge.

The overfitting problem occurs when the discriminator becomes overly complex with a large number of parameters, resulting in memorization of the training data rather than learning the underlying data distribution. To mitigate this issue, several strategies have been proposed, which can be divided into the following categories. The first category is data augmentation methods [18, 34, 35, 36], which utilizes traditional data augmentation methods, such as rotation and color transformation, to increase the amount of training data. The second type is regularization, which adds regularization term to the loss function of the discriminator [19, 37], allowing it to learn more discriminative representations under limited training data. InsGen [37] proposes a contrastive learning objective to enhance the adversarial loss in the few-shot generation setting. AdaptiveMix [38] narrows down the distance between hard samples and easy samples, where hard samples are regarded as the samples that are difficult for discriminator to classify. The third category is model architecture improvement. FastGAN [39] introduces a self-supervised discriminator and a Skip-Layer channel-wise Excitation module for efficient few-shot image synthesis. MoCA [40] proposes prototype-based memory modulation module to improve the generator network of a GAN. The proposed method in this paper falls into the second category by introducing manifold constraint regularization, a novel approach that addresses the overfitting problem of GANs on RS data from the manifold perspective for the first time.

Previous works [41, 42] introduce geometry constraints to GANs loss functions. These methods utilize statistical mean and radius to approximate the geometry of the real data, but they lack accurate constraints on the data manifold and may lose important geometrical information. Other approaches [43, 44, 45] offer more precise control, but they require modifications to the network architecture. In contrast, our approach avoids the limitations of prior work by being applicable to any GAN model without requiring network architecture changes, offering both efficiency and effectiveness.

B. GAN in the RS field

In this paper, our primary focus is on RGB images in the RS domain. Existing GAN models in the RS field can be categorized into two main types based on their applications.

The first type revolves around data generation or augmentation [2, 3, 4, 5, 46]. For instance, Lin *et al.* introduced MARTAGAN [2], marking the pioneering application of GANs to remote sensing images. MARTAGAN, an extension of DCGAN [47], introduce enhancements like a multi-feature layer in the discriminator and feature loss in the generator to improve image representations. This work demonstrated the potential of GANs to augment datasets and enhance unsupervised classification accuracy. Similarly, Yu *et al.* propose Attention GANs [3], designed for unsupervised classification tasks, by integrating attention mechanisms into GANs to bolster the discriminator's representation capabilities. Furthermore, Wu *et al.* [46] propose a comprehensive change detection framework utilizing GANs to tackle various RS change detection tasks. Their approach incorporates an image-to-image generator to capture spectral and spatial variations between multi-temporal images.

The second type pertains to GANs applied in image enhancement tasks, such as image super-resolution [10, 11, 12, 13] and cloud or haze removal [6, 7, 8, 9]. Examples include the work by Wu *et al.* [10], which introduce a GAN-based edge-enhancement network designed to reconstruct high-resolution RS images, employing adversarial learning to restore edge details obscured by noise. Guo *et al.* [12] have refined the Super Resolution Generative Adversarial Network [48] by altering both the internal and external connections of the residual block and modifying the loss function to incorporate the Charbonnier penalty for improved performance. In the realm of haze removal, Hu *et al.* [6] propose edge-sharpening cycle-consistent adversarial network (ES-CCGAN), ubstituting the standard residual network with a dense convolutional network to enhance edge clarity in images. The edge-sharpening loss function of ES-CCGAN is designed to further recover clear ground-object edges. For cloud removal, Li *et al.* [8] develope a semi-supervised technique, CR-GAN-PM, merging GANs with a physical model to address cloud distortions in unpaired images from various regions.

Our contribution aligns with the first category, focusing on addressing the overfitting challenge in GANs on RS image generation. Our goal is to enhance the quality of generated images and improve the discriminator's effectiveness, thereby elevating GAN performance in data generation and augmentation tasks. This paper marks a pioneering effort to specifically tackle issues related to the overfitting in the context of RS image generation, underscoring our novel approach to improving GAN applications within this domain.

III. Preliminary Study and Analysis

In this section, we first introduce the basic framework of GANs. Then we explore the overfitting problem of GAN models on RS image generation tasks. Finally, we analyze plausible reasons about the poor results of the GAN models on RS images and derive the motivation for this paper.

A. Preliminary GANs

The GAN model aims to learn the distribution of training samples. Based on the idea of the zero-sum game, a GAN model consists of a generator G and a discriminator D. The generator aims to generate realistic samples to fool the discriminator, while the discriminator tries to distinguish between real and fake samples. When the model reaches the final equilibrium point, the generator will model the target distribution and produce counterfeit samples, which the discriminator will fail to discern. Let $\mathcal{L}_{\mathrm{D}}$ and $\mathcal{L}_{\mathrm{G}}$ denote the loss functions of the discriminator D and the generator G, respectively. The training of the GAN frameworks can be generally illustrated as follows:

$$\min_{\mathrm{D}} \mathcal{L}_{\mathrm{D}} = -\mathop{\mathbb{E}}_{x\sim P_{\mathrm{data}}}[D(x)] + \mathop{\mathbb{E}}_{z\sim p_{\mathrm{z}}}[D(G(z))] \quad (1)$$

$$\min_{\mathrm{G}} \mathcal{L}_{\mathrm{G}} = -\mathop{\mathbb{E}}_{z\sim p_{\mathrm{z}}}[D(G(z))] \quad (2)$$

where P_{data} denotes the real data distribution, and P_{z} is usually the *normal distribution*.

B. Problem Analysis

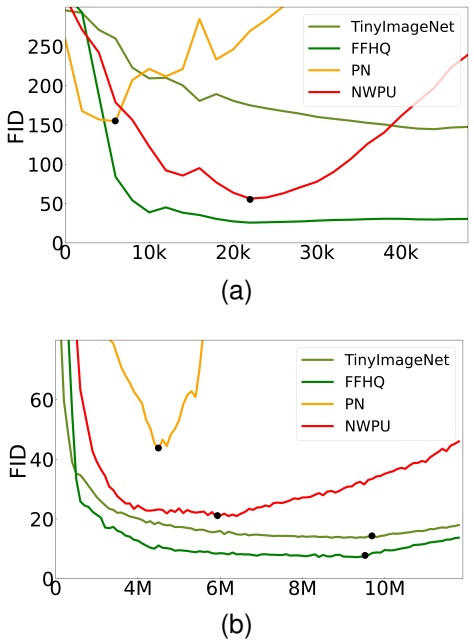

Fig. 2. (a) Training curves of BigGAN on TinyImageNet, FFHQ, NWPU and PN datasets. The horizontal axis is the number of training steps. (b) Training curves of StyleGAN2 on TinyImageNet, FFHQ, NWPU and PN datasets. The horizontal axis indicates the training process (the number of real images shown to the discriminator). These training curves highlight an earlier divergence on RS datasets compared to natural ones, emphasizing our claim about the extent of overfitting in RS data.

As previously stated, GANs are more susceptible to overfitting on RS image dataset compared to natural image dataset. Notably, the GAN model exhibits earlier divergence when trained on the RS images (NWPU) compared to the natural images (FFHQ), as depicted in Fig. 1. To extend our investigation on the prevalence of overfitting across different datasets and GAN architectures, we conduct further experiments using the most popular GAN models: StyleGAN2 [29] and BigGAN

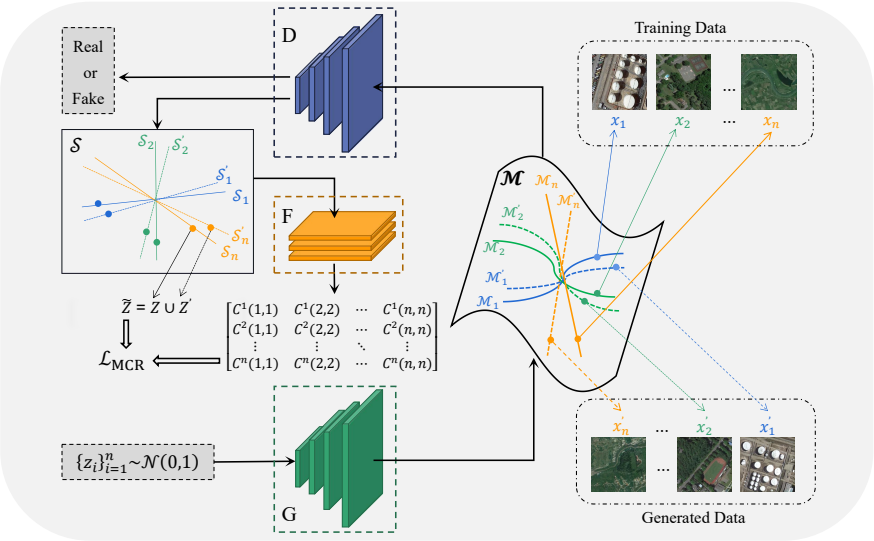

Fig. 3. The overall architecture of our method. G and D are generator and discriminator. F is the MLP network for relationship matrix C. $\mathcal{M}$ represents manifold, and $\mathcal{S}$ represents subspace.

[28]. We employ NWPU and PatternNet (PN) [49] for RS image datasets, and FFHQ256 and TinyImageNet [50] for natural image datasets. Each dataset comprises a random sample of 30,000 images. The experiment results are shown in Fig.2. The training curves of these GAN models underscore an earlier divergence on RS datasets compared to their natural counterparts, reinforcing our inference regarding the severity of overfitting in RS data.

To identify the potential causes underlying this phenomenon, we analyze the inherent differences between RS RGB image and natural image. Compared to their natural counterparts, RS images have a higher spatial resolution and a larger coverage area, encompass a wider variety of scenery, and exhibit richer content. Deep learning has an underlying manifold assumption on the training data, i.e., high-dimensional data can be embedded into low-dimensional manifolds. This inherent assumption allows deep learning models to effectively handle high-dimensional data and achieve remarkable performance, as explicitly confirmed by Pope et al. [51]. It is well-established that learning a manifold requires a number of samples that grows exponentially with the manifold's intrinsic dimension [17]. Hence, we conjecture that GANs are more easily overfitting on RS image generation tasks because of the higher intrinsic dimension of RS images. Under limited training samples, the discriminator struggles to learn the intricate manifold of the RS images, resulting in a tendency to memorize the training data and overfit to the dataset's limited variations.

To validate our hypothesis, we employ the Maximum Likelihood Estimation (MLE) method [52], the same approach used by Pope et al. [51], to estimate the intrinsic dimension of the RS images.

$$m_k = \left[\frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)} \right]^{-1}, \tag{3}$$

TABLE I
THE INTRINSIC DIMENSION ESTIMATED BY MLE.

Dataset	FFHQ	TinyImageNet	NWPU	PN
MLE (k=3)	35	38	52	42
MLE (k=5)	34	36	53	44
MLE (k=10)	33	33	48	44
MLE (k=20)	31	30	44	42

where $T_j(x)$ is the Euclidean (ℓ_2) distance from x to its j^{th} nearest neighbor, and n is the number of samples. We conduct this experiment with different values for $k = 3, 5, 10, 20$ and a fixed sample size $n = 30,000$. The results for the NWPU, PN, FFHQ, and TinyImageNet datasets are presented in Table I. As expected, the intrinsic dimensions of the NWPU and PN datasets are indeed higher compared to those of the FFHQ and TinyImageNet datasets.

Based on our experiments, we can infer that discriminator is more susceptible to overfitting on RS data due to the higher intrinsic dimension of the RS dataset compared to the natural dataset. To address this issue, we propose introducing the data manifold to constrain the discriminator. By encouraging the learned features to align with the structure of the real data manifold, we can guide the discriminator to learn the underlying structure of the data distribution and avoid overfitting to the local features of the training data. This approach favors model that performs well across the entire data manifold, promoting the discriminator to generalize well beyond the training data, thereby mitigating the overfitting problem.

IV. METHODS

In this section, we initially present an assumption regarding the real data manifold, accompanied by three critical attributes that ideal representations of images ought to exhibit. Subsequently, we introduce a novel measure to evaluate the

data manifold of feature distribution. Building upon this, we propose novel regularization term, which we call manifold constraint regularization (MCR). The framework of the proposed method is shown in Fig.3.

A. Preliminary Assumptions on Data Manifold

In practice, it is intractable to learn the data manifold in the high-dimensional ambient space $\mathbb{R}^D$ [51, 53]. Therefore, following the approach in [53, 54], we assume that the real data manifold comprises of a union of low-dimensional nonlinear submanifolds $\cup_{j=1}^{k}\mathcal{M}_j \subset \mathbb{R}^D$, where each submanifold $\mathcal{M}_j$ is of dimension $d_j << D$. Each submanifold $\mathcal{M}_j \subset \mathbb{R}^D$ can be transformed to a linear subspace $\mathcal{S}_j \subset \mathbb{R}^d$, which we refer to the feature space. From this assumption, we can infer that images residing on the same submanifold share similarities, whereas images on different submanifolds exhibit distinct characteristics. Consequently, the ideal features of these images should exhibit the following attributes:

- Between-submanifold Discrepancy: Features of images from different submanifold should be highly uncorrelated.
- Within-submanifold Similarity: Features of images from the same submanifold should be relatively correlated.
- Maximally Variance: Features should have as large dimension as possible to cover all the submanifolds and be variant for the same submanifold.

Therefore, in order to accurately capture the real data manifold, it's desirable for the features from different submanifolds to be as uncorrelated as possible. On the other hand, features from the same submanifold should display a high correlation and coherence. Simultaneously, the features should exhibit maximum variance to cover all possible submanifolds. By learning features that adhere to these properties, the discriminator can be considered to have successfully captured the underlying structure of the real data manifold.

B. Data Manifold Evaluation

Based on the data manifold assumption, features from different submanifolds should have a sparse feature distribution. Conversely, features within the same submanifold should have a compact feature distribution. Our first step is to establish a metric for the compactness of the feature distribution by examining the relationships between features. Secondly, we design data manifold evaluation function based on this measure.

Step I. We propose leveraging singular values to quantify compactness. Recall that singular values (represented by λ_i here) capture the amount of variance explained by each principal component in a data matrix. Singular vectors corresponding to larger singular values represent the principal stretching directions of the data. In simpler terms, a larger number of significant singular values indicates a more uniform distribution, while fewer significant values suggest a more compact distribution. Based on this principle, we propose the following measure:

$$\mathcal{V}(Z) = \sum_{i=1} \lambda_i^2 = \mathrm{Tr}(ZZ^{\mathrm{T}}). \tag{4}$$

A larger value of $\mathcal{V}(Z)$ indicates a broader span of the singular vectors and thus greater uniformity of the data. Considering computational efficiency, we opt for the trace of ZZ^T in our method. In the Section V, we theoretically prove $\mathcal{V}(Z)$ can measure the compactness of a distribution.

Step II. Building upon Eq. 4, we introduce a novel data manifold evaluation function $\mathcal{L}_{\mathrm{Tr}}(Z)$. This function leverages the concept of feature compactness to assess how well the learned features capture the underlying data manifold.

According to the aforementioned assumption, features of different submanifolds should be maximally uncorrelated with each other. Therefore, they together should span a space of the largest possible volume or dimension, and $\mathcal{V}(Z)$ should be as large as possible. Conversely, learned features of the same submanifold should be highly correlated and coherent. Therefore, they should only span a space of a very small volume or dimension. To capture these contrasting properties, our evaluation function, $\mathcal{L}_{\mathrm{Tr}}(Z)$, incorporates both the overall and within-submanifold compactness:

$$\mathcal{L}_{\mathrm{Tr}}(Z) \quad = \quad \frac{1}{2n}(\mathrm{Tr}(ZZ^{\mathrm{T}}) - \sum_{j=1}^{k} \mathrm{Tr}(ZC^jZ^{\mathrm{T}})) \tag{5}$$

where $Z = [z_1, \ldots, z_n] \subset \mathbb{R}^{d\times n}$ denotes the representations of images, $C = \left\{C^j \in \mathbb{R}^{n\times n}\right\}_{j=1}^{k}$ is a set of positive diagonal matrices whose diagonal entries denote the membership of the n samples in the k submanifolds. If the sample x_i belongs to the submanifold j, then $C^j(i,i) = 1$. Otherwise, $C^j(i,i) = 0$, where $i \in \{1, ..., n\}, j \in \{1, ..., k\}$.

The first term, $\mathrm{Tr}(ZZ^{\mathrm{T}})$, measures the overall compactness of the entire feature set, as defined by Eq. 4. The second term, $\sum_{j=1}^{k} \mathrm{Tr}(ZC^jZ^{\mathrm{T}})$, represents the compactness within each individual class. The difference between these two terms reflects the "dispersibility" between features from different submanifolds, and a larger difference signifies better separation. Therefore, a higher value of $\mathcal{L}_{\mathrm{Tr}}(Z)$ indicates that the learned features effectively capture the intrinsic structure of the data manifold, with features from different submanifolds being well-separated and features within a submanifold being well-clustered.

C. Manifold Constraint Regularization

Building on the concept of data manifold evaluation introduced earlier, we propose a novel regularization term, $\mathcal{L}_{\mathrm{MCR}}(Z, Z')$, to assess how well the generated image manifold aligns with the real image manifold.

$$\begin{aligned}\mathcal{L}_{\mathrm{MCR}}(Z, Z') =& \frac{1}{2n}(\mathrm{Tr}(\widetilde{Z}\widetilde{Z}^{\mathrm{T}}) - \frac{1}{2}\sum_{j=1}^{k} \mathrm{Tr}(ZC^jZ^{\mathrm{T}}) \\ &- \frac{1}{2}\sum_{j=1}^{k} \mathrm{Tr}(Z'C'^jZ'^{\mathrm{T}})),\end{aligned} \tag{6}$$

where $\widetilde{Z} = Z \cup Z'$, Z and Z' represent the feature representations of real and generated images, respectively. C^j and C'^j are diagonal matrices encoding submanifold membership information for real and generated data, respectively. k is the number of submanifolds within the data.

The first term of $\mathcal{L}_{\mathrm{MCR}}$ measures the compactness of the joint distribution of real and generated images, i.e. the volume of the space spanned by the real and generated features jointly. The second and third terms measure the compactness of the real and generated distributions individually. If the generated image manifold aligns well with the real image manifold, then the combined distribution should be compact, which translates to a lower value for $\mathcal{L}_{\mathrm{MCR}}$. Conversely, a larger value indicates a significant deviation between the manifolds of real and generated data. We theoretically prove that the $\mathcal{L}_{\mathrm{MCR}}$ could measure the volume of the space between Z and Z'. The proof will be elaborated in the Section V.

D. Loss Functions

We incorporate the data manifold constraint as regularization terms in the loss functions of GAN models, as depicted in Eq.7 and Eq.8, where λ and γ are hyperparameters. Specifically, we optimize the discriminator by maximizing $\mathcal{L}_{\mathrm{MCR}}$, which encourages it to focus on the underlying data manifold and push the manifold of the generated samples to be misaligned with that of the real images. Conversely, minimizing $\mathcal{L}_{\mathrm{MCR}}$ guides the generator towards producing samples that align closely with the real image manifold. This dual strategy sharpens the discriminator's ability to differentiate and guides the generator towards producing higher-quality, more diverse images.

$$\begin{aligned}\min_{\mathrm{D}} \mathcal{L}_{\mathrm{D}} = & \mathop{\mathbb{E}}_{z\sim p_{\mathrm{z}}}[D(G(z))] - \mathop{\mathbb{E}}_{x\sim P_{\mathrm{data}}}[D(x)] \\ & - \lambda \mathcal{L}_{\mathrm{MCR}}(Z, Z')\end{aligned} \tag{7}$$

$$\min_{\mathrm{G}} \mathcal{L}_{\mathrm{G}} = - \mathop{\mathbb{E}}_{z\sim p_{\mathrm{z}}}[D(G(z))] + \gamma \mathcal{L}_{\mathrm{MCR}}(Z, Z') \tag{8}$$

The representations Z and Z' are learned by the discriminator. It has been shown that different network layers are responsible for different levels of detail in the images. Empirically, the latter blocks of the network have more effect on the style (e.g. texture and color) of the image whereas the earlier blocks impact the coarse structure or content of the image [55]. Thus, we choose features from a shallow network layer of the discriminator as representations Z and Z' in our experiments.

While the effectiveness of $\mathcal{L}_{\mathrm{MCR}}$ relies on knowing the submanifold membership matrices C, obtaining this knowledge can be expensive or impractical in unsupervised settings. To address this challenge, we propose a novel unsupervised approach to learn this information directly from the data.

Specifically, we employ a three-layer Multi-Layer Perceptron (MLP) network, denoted as $\boldsymbol{F}$, to learn the relationship between data points in the feature space. This network takes the features Z from the discriminator as input and outputs a matrix $M \in \mathbb{R}^{n\times n}$. The elements of the j-th row of the M matrix are the diagonal elements of the C^j matrix. Hence, $C^j(i,i) = M(j,i)$, where $i \in \{1, ..., n\}$. The key idea is that elements along each row of M correspond to the submanifold membership probabilities for a particular data point. In other words, a higher value $M(j,i)$ indicates a higher likelihood that data points x_i and x_j belong to the same submanifold.

In order to train the network $\boldsymbol{F}$, we employ a pre-trained encoder network, denoted as f_{pre}, to extract features for each data point. These features are represented by the matrix $\bar{Z} = \{\bar{z}_1, ..., \bar{z}_n\}$, where $\bar{z}_i = f_{\mathrm{pre}}(x_i)$, $i \in \{1, ..., n\}$. We define a similarity measure based on the Euclidean distance between feature vectors. Regarding the j-th element $\bar{z}_j$ in $\bar{Z}$ as the anchor, we define that:

$$M^{pro}(j,i) = \exp(-\|\bar{z}_i - \bar{z}_j\|_2^2 \Big/ \tau) \tag{9}$$

where τ is the temperature hyperparameter, $\bar{z}_i$ is the i-th sample in $\bar{Z}$. Then, we can obtain:

$$M_j^{pro} = [M^{pro}(j,1), ..., M^{pro}(j,n)] \tag{10}$$

We in turn treat the samples in $\bar{Z}$ as anchors and obtain $M^{pro} = [M_1^{pro}, ..., M_n^{pro}]^{\mathrm{T}}$. To this end, we give the prior constraint M^{pro} of M based on the similarity of different pairs of samples. The loss function of network $\boldsymbol{F}$ is as follows:

$$\mathcal{L}_{con} = \|M^{pro} - \boldsymbol{F}(Z)\|_2^2 \tag{11}$$

The network $\boldsymbol{F}$ is first pretrained on the training dataset, with Z being extracted by f_{pre}. Then, it is trained together with the discriminator. In the optimal case, the network $\boldsymbol{F}$ will learn the relationship between the samples, and $M(i,j)$ will be close to 1 for data points belonging to the same submanifold, and close to 0 for points from different submanifolds. In our experiments, we find that the choice of the pre-trained encoder has little impact on the final results.

V. Theoretical Analysis

In this section, we present theoretical analysis of the proposed method, MCR. Firstly, we establish a theoretical connection between the compactness measure $\mathcal{V}(Z)$ and information theory. Secondly, we prove that $\mathcal{L}_{\mathrm{Tr}}$ can evaluate the data manifold of features, and $\mathcal{L}_{\mathrm{MCR}}$ can measure the disparity of the manifolds between generated images and real images.

We begin by demonstrating the connection between our proposed measure, denoted by $\mathcal{V}(Z)$, and information theory.

Proposition 1: $\mathcal{V}(Z) = \frac{1}{2n}\operatorname{Tr}\left(ZZ^{\mathrm{T}}\right)$ where $Z = \left[z^1, \ldots, z^n\right] \subset \mathbb{R}^{d\times n}$ can measure the compactness of a distribution from its finite samples Z.

Proof 1:

Based on the first-order Taylor series approximation, $\log\det(\mathbf{C}+\mathbf{D}) \approx \log\det(\mathbf{C}) + \operatorname{Tr}\left(\mathbf{D}^{\mathrm{T}}\mathbf{C}^{-1}\right)$, we can get following equations.

$$\begin{aligned}\frac{1}{2n}\operatorname{Tr}(ZZ^{\mathrm{T}}) &= \frac{1}{2}(\frac{1}{n}\operatorname{Tr}(ZZ^{\mathrm{T}}) + \log\det(\boldsymbol{I})) \\ &\approx \frac{1}{2}\log\det(\boldsymbol{I} + \frac{1}{n}ZZ^{\mathrm{T}})\end{aligned}$$

In information theory, rate distortion can be used to measure the "compactness" of a random distribution [56]. Given a random variable z and a prescribed precision $\epsilon > 0$, the rate distortion $R(z,\epsilon)$ is the minimal number of binary bits needed to encode z such that the expected decoding error is less than ϵ. Given finite samples $Z = \left[z^1, \ldots, z^m\right] \subset \mathbb{R}^{d\times m}$, the average number of bits needed is given by the following expression: $\mathcal{L}(Z,\epsilon) \doteq \left(\frac{m+d}{2m}\right)\log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}ZZ^{\mathrm{T}}\right)$. As the sample size m is large, our approach can be seen as an approximation of the rate distortion, which completes our proof.

Based on the above derivation, the proposed measure can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\mathrm{Tr}} \approx & \frac{1}{2}\log\det(\boldsymbol{I}+\alpha ZZ^{\mathrm{T}}) \\ & -\sum_{j=1}^{k}\frac{\gamma_j}{2}\log\det(\boldsymbol{I}+\alpha_j ZC^j Z^{\mathrm{T}})\end{aligned} \tag{12}$$

where $\alpha=\frac{d}{n\epsilon^2}, \alpha_j=\frac{d}{\mathrm{Tr}(C^j)\epsilon^2}, \gamma_j=\frac{\mathrm{Tr}(C^j)}{n}$ for $j=1,\ldots,k$.

Based on this proposition, we can infer that the optimal solution of Eq.5 have following properties:

Theorem 1: Suppose $\boldsymbol{Z}^*$ is the optimal solution that maximizes the function Eq.5. We have:

- Between-submanifold Discrepancy: If the ambient space is adequately large, the subspaces are all orthogonal to each other, i.e. $\left(\boldsymbol{Z}_i^*\right)^\top \boldsymbol{Z}_j^* = \mathbf{0}$ for $i \neq j$.

- Maximally Variance: If the coding precision is adequately high, i.e., $\epsilon^4 < \min_j\left\{\frac{n_j}{n}\frac{d^2}{d_j^2}\right\}$, each subspace achieves its maximal dimension, i.e. $\mathrm{rank}\left(\boldsymbol{Z}_j^*\right)=d_j$.

The proof for Theorem 1 is provided in the Appendix. In summary, we initially assume that $\left(\boldsymbol{Z}_i^*\right)^\top \boldsymbol{Z}_j^* \neq \mathbf{0}$. Based on the Singular Value Decomposition (SVD) and the condition $\sum_{j=1}^k d_j \leq d$, we infer that $\boldsymbol{Z}^*$ cannot be the optimal solution of Eq.5. This contradiction validates the *Between-submanifold Discrepancy*. Using a similar approach, we demonstrate that the outcome of Eq.5 is related to the singular values of $\boldsymbol{Z}_j^*$. Under the condition $\epsilon^4 < \min_j\left\{\frac{n_j}{n}\frac{d^2}{d_j^2}\right\}$, $\mathrm{rank}\left(\boldsymbol{Z}_j^*\right)=d_j$, which confirms the *Maximally Variance*.

Hence, $\mathcal{L}_{\mathrm{Tr}}$ effectively evaluates the data manifold of the representations. Since the features of each submanifold, Z_j and Z_j', are similar to subspaces or Gaussians, their "distance" can be measured by the rate distortion.

$$\begin{aligned}\mathcal{L}_{\mathrm{inf}}(Z,Z') = & \frac{1}{2nk}\sum_{j=1}^{k}(\mathrm{Tr}(\widetilde{Z}_j\widetilde{Z}_j^{\mathrm{T}}) - \frac{1}{2}\mathrm{Tr}(Z_jZ_j^{\mathrm{T}}) \\ & -\frac{1}{2}\mathrm{Tr}(Z_j'Z_j'^{\mathrm{T}})),\end{aligned} \tag{13}$$

where Z_j and Z_j' represent the feature representations of real and generated images belonging to submanifold j, respectively. $\widetilde{Z}_j = Z_j \cup Z_j'$, $j \in \{1,...,k\}$, and k is the number of submanifolds within the data.

The metric $\mathcal{L}_{\mathrm{inf}}$ quantifies the distance between the submanifolds of real and generated images. A smaller $\mathcal{L}_{\mathrm{inf}}$ value indicates that the submanifolds of the real and generated data are more closely aligned. Given that submanifolds are inherently smaller than the entire manifold, we can say that $\sum_{j=1}^k(\mathrm{Tr}(\widetilde{Z}_j\widetilde{Z}_j^{\mathrm{T}}) \leq \mathrm{Tr}(\widetilde{Z}\widetilde{Z}^{\mathrm{T}})$. This leads to the conclusion that $\mathcal{L}_{\mathrm{inf}} \leq \mathcal{L}_{\mathrm{MCR}}$. Essentially, minimizing $\mathcal{L}_{\mathrm{MCR}}$ during training prompts the generator to produce images that closely align with the real data manifold across all classes. This inherently minimizes the disparity between submanifolds, as indicated by a lower $\mathcal{L}_{\mathrm{inf}}$ value. Conversely, the discriminator aims to maximize the separation between these manifolds by maximizing $\mathcal{L}_{\mathrm{MCR}}$, while concurrently drawing the features within the same submanifold closer together.

VI. Experiments

In this section, we first introduce the datasets used in our experiments. Next, we provide the details of our experimental setup. We then present the results, including both quantitative metrics and qualitative examples, to demonstrate the effectiveness of our proposed method compared to existing approaches. Additionally, we showcase how our method benefits downstream tasks. Finally, we provide more ablation and analysis of different components of our method.

A. Datasets

We use three RS datasets to evaluate our method: the UC Merced Land Use (UCLand) Dataset [57], NWPU-RESISC45 (NWPU) Dataset [16] and PatternNet (PN) Dataset [49]. Their information is shown in Table II.

TABLE II
Details of the Dataset

Attribute	UCLand	NWPU	PN
Images per class	100	700	800
Scene Classes	21	45	38
Resolution (m)	0.3	0.2-30	0.062-4.693
Image Size	256×256	256×256	256×256
Source	United States Geological Survey	Google Earth	Google Earth

The UC Merced Land Use Dataset is one of the most widely used datasets in the field of remote sensing scene classification. It has 21 scene categories, each with 100 images. Each image has the size 256×256 and a spatial resolution of 0.3m. The images in the dataset come from more than 20 cities in the United States, including Las Vegas, Los Angeles, Miami, Santa Barbara, and Seattle.

The NWPU-RESISC45 Dataset has 31,500 images covering more than 100 countries and regions around the world. It has 45 categories with 700 images in each category. Each image is 256×256 pixels in size. The spatial resolution of this dataset is up to 0.2m and the lowest is 30m. The images are varied in lighting, shooting angle, imaging conditions, and so on.

The PatternNet Dataset is a large-scale high-resolution remote sensing dataset. It has 38 categories with 800 images in each category. Each image is 256×256 pixels in size. The spatial resolution of this dataset varies from 0.06m to 4.7m per pixel. The images in PatternNet are collected from Google Earth imagery or via the Google Map API for some US cities.

B. Experiment Setup

Baseline Methods: We evaluate our approach to RS image generation by comparing it with established GAN models, including **BigGAN** [28] and **StyleGAN2** [29]. We also benchmark our method against techniques specifically developed to address GAN overfitting. These include the augmentation methods **ADA** [18] and **APA** [35], regularization methods **LeCam** [19], **AdaptiveMix** [38] and **InsGen** [37], as well

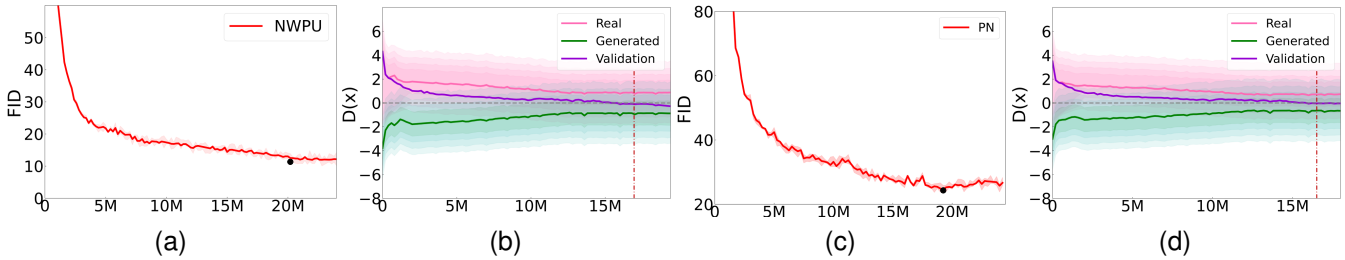

Fig. 4. The horizontal axis indicates the training process (the number of real images shown to the discriminator). (a) Training curves of our method on NWPU dataset. (b) The outputs of the discriminator during training on NWPU dataset. (c) Training curves of our method on PN dataset. (d) The outputs of the discriminator during training on PN dataset. No divergence occurs during training, and the discriminator maintains high accuracy on the validation set. These findings suggest that MCR effectively alleviates the discriminator's overfitting issue.

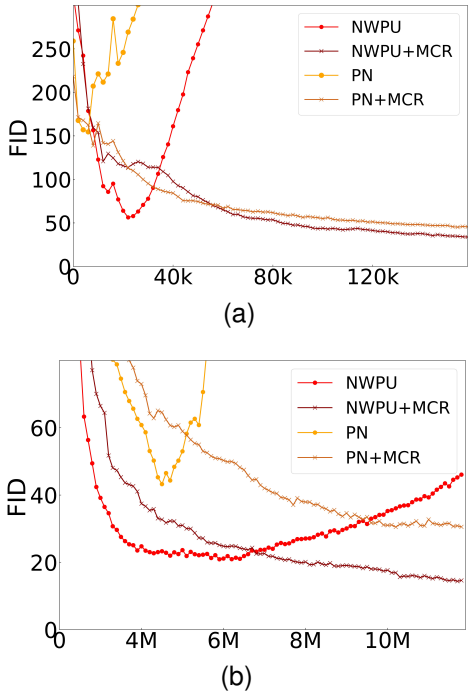

Fig. 5. (a) Training curves of BigGAN. (b) Training curves of StyleGAN2. Our method (MCR) not only reduces discriminator overfitting but also attains superior quality scores.

as few-shot generation methods **FastGAN** [39] and **MoCA** [40].

Implementation Details: We use the official PyTorch implementation of StyleGAN2, ADA, FastGAN, InsGen, MoCA and AdaptiveMix. For BigGAN, APA and LeCam, we use the implementations provided by [58]. Throughout our experiments, we set hyperparameters $\lambda = 1$ and $\gamma = 1$. Further details regarding the ablation study on these hyperparameters can be found in Section VI-E.

Evaluation Metrics: We evaluate our method using Frechet inception distance (FID) [14], as the most commonly-used metric for measuring the quality and diversity of images generated by GAN models. We also employ the Kernel Inception Distance (KID) [59], a metric that remains unbiased by empirical bias [60], to further validate our results.

TABLE III
COMPARISON OF QUALITY SCORES (FID↓) ON THE UCLAND, NWPU AND PN DATASETS

Methods	Backbone	UCLand	NWPU	PN
ADA	StyleGAN2	74.25	11.97	33.53
APA	StyleGAN2	78.54	21.67	33.75
AdaptiveMix	StyleGAN2	70.95	13.53	34.63
LeCam	StyleGAN2	72.66	15.73	31.26
InsGen	StyleGAN2	95.65	10.92	50.76
FastGAN	–	76.63	32.17	51.39
MoCA	StyleGAN2	71.25	10.37	32.79
MCR (Ours)	StyleGAN2	**69.43**	**9.88**	**30.47**

C. RS Image Generation

Training Stability. Fig.4 showcases the training process of our method MCR (based on StyleGAN2) on NWPU and PN datasets. Fig.4 (a) and (c) depict the FID curves during training, indicating no early divergence. Fig.4 (b) and (d) show the discriminator's accuracy on the validation set, consistently remaining high. These observations suggest that the proposed method MCR effectively mitigates the discriminator's overfitting problem.

Quantitative Comparison. Next, we present a quantitative comparison with established baselines. Fig.5 compares the training curves of our method on NWPU and PN datasets against BigGAN (Fig.5(a)) and StyleGAN2 (Fig.5(b)). Our method not only alleviates discriminator overfitting but also achieves superior quality scores.

To further demonstrate the effectiveness of our proposed method MCR, we compare its performance against established methods for addressing discriminator overfitting. These methods typically fall into three categories: data augmentation, regularization, and architectural improvements. Although MCR falls under the regularization category, we compare it with representatives from all three categories. Table III summarizes the experimental results on the UCLand, NWPU, and PN datasets. Our method consistently outperforms all others: achieving a 2.14% improvement in FID on UCLand compared to the well-established method. Similarly, on NWPU, our method achieves an FID of 9.88 compared to MoCA's 10.37, demonstrating

TABLE IV
COMPARISON OF QUALITY SCORES (FID↓, KID($\times 10^{-3}$)↓) ON THE UCLAND, NWPU AND PN DATASETS (THE RED NUMBERS PRESENT OUR IMPROVEMENT)

Methods	UCLand		NWPU		PN	
	FID	KID	FID	KID	FID	KID
BigGAN+ADA	98.09-5.92	52.31-4.24	30.91-3.39	11.13-1.85	63.94-3.59	36.71-1.63
StyleGAN2+ADA	74.25-4.22	37.08-4.08	11.97-1.88	3.27-0.92	33.53-2.96	13.67-2.32
StyleGAN2+APA	78.54-4.39	39.25-3.79	21.67-2.78	8.54-2.03	33.75-4.84	13.34-2.99
BigGAN+ADA+LeCam	92.52-5.27	48.59-3.52	32.65-2.57	12.69-1.87	54.28-4.42	25.89-2.37
StyleGAN2+ADA+LeCam	70.89-3.91	32.21-2.86	14.38-2.34	4.97-1.47	25.87-2.86	9.32-1.73

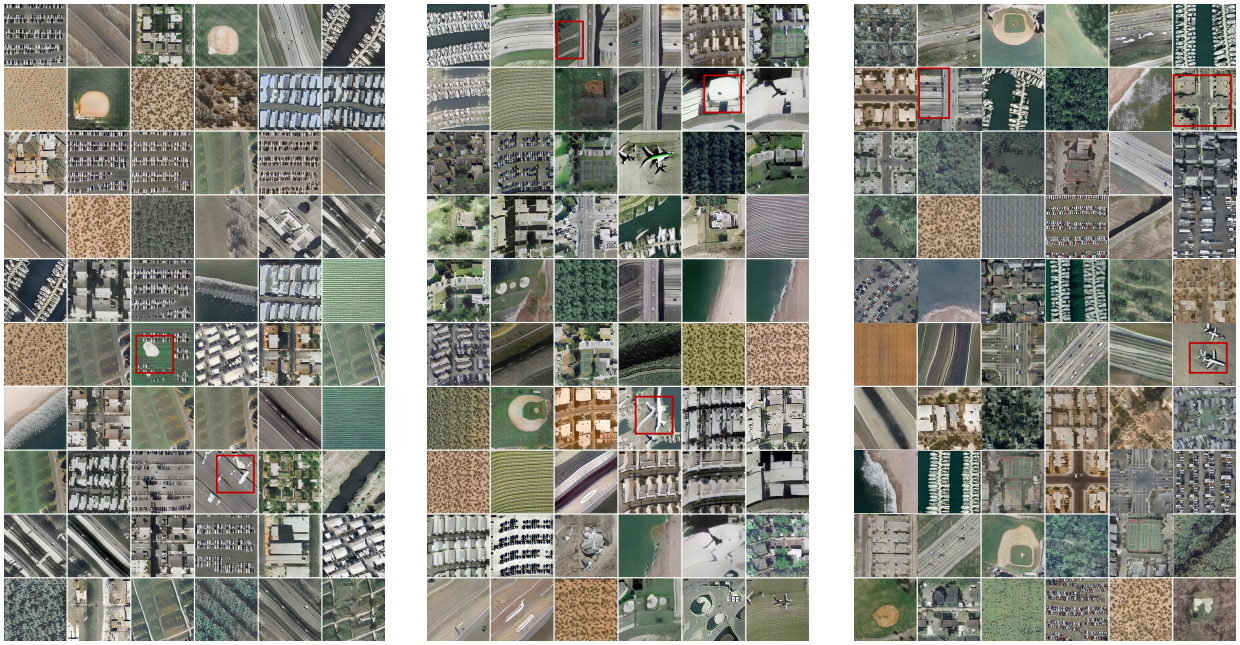

Fig. 6. Experiment results on the UCLand Dataset. The generated images of StyleGAN2 (**left**), StyleGAN2+AdaptiveMix (**middle**) and our method (**right**). The images generated by our method exhibit higher quality and diversity.

MCR's clear advantage. Finally, on the PN dataset, our method surpasses LeCam by 2.53%. It's important to note that these results are based on the same StyleGAN2 architecture. Overall, MCR outperforms the second-best method by approximately 3.13% in terms of FID score. Table III clearly highlights the consistent superiority of our method in addressing the overfitting challenge.

Versatility. To verify the versatility of the method in this paper, we conduct experiments on different combinations of network architectures, enhancement methods and regularization methods. Experimental results are reported in Table IV. The red numbers in Table IV indicate the improvement of the GAN models after using our method. In general, the FID and KID scores for our proposed method indicate a significant and consistent advantage over all the compared methods. In detail, our comparison experiments can be divided into three types. First, under different model structures, BigGAN and StyleGAN2, our method is robust, which proves that our approach can be applied to any model architecture. Second, under different augmentation methods, ADA and APA, our method still performs well. This shows that the proposed approaches can be applied to other GAN models along with existing augmentation approaches. Third, combined with the regularization method LeCam, our method is still effective. The proposed method can be viewed as an effective complement to existing regularization methods.

Qualitative Comparison. To validate the effectiveness of the methodologies presented in this paper, we conduct qualitative experiments on three RS datasets. For each dataset, we select a superior baseline method for benchmarking purposes. Specifically, on the UCLand dataset, our method is compared against StyleGAN2 and its extension, StyleGAN2+AdaptiveMix. On the NWPU dataset, our method is compared against StyleGAN2 and StyleGAN2+MoCA. Lastly, for the PN dataset, we evaluate our method against StyleGAN2 and StyleGAN2+LeCam. The generated images on the UCLand dataset are presented in in Fig. 6.

We randomly select 60 samples from the generated images. The baseline StyleGAN2 model generates similar images, such as multiple parking lot images. In contrast, the imagery

TABLE V
COMPARISON OF UNSUPERVISED CLASSIFICATION ACCURACY(%) ON THE UCLAND DATASET AND NWPU DATASET

Datasets (training ratio)	UCLand(80%)	UCLand(50%)	NWPU(80%)	NWPU(20%)
MartaGAN [2]	94.86±0.80	85.51±0.69	75.43±0.28	75.03±0.28
AttentionGAN [3]	**97.69±0.6**	89.06±0.50	—	77.99±0.19
StyleGAN2	95.29±0.31	87.32±0.22	80.13±0.39	76.53±0.17
StyleGAN2+ADA	95.71±0.42	89.05±0.33	82.40±0.17	78.16±0.23
StyleGAN2+MCR (Ours)	97.33±0.39	**92.72±0.22**	**84.82±0.42**	**80.35±0.13**

TABLE VI
COMPARISON OF SELF-SUPERVISED CLASSIFICATION ACCURACY(%) ON THE UCLAND DATASET AND NWPU DATASET

Method	Backbone	UCLand			NWPU		
		Original	StyleGAN2	MCR(Ours)	Original	StyleGAN2	MCR(Ours)
SimCLR	ResNet18	67.59±0.38	69.61±0.92	**70.48±0.88**	68.86±0.91	69.43±0.22	**71.24±0.19**
	ResNet50	69.39±0.67	70.25±0.28	**71.80±0.35**	69.72±0.52	71.16±0.75	**72.60±0.23**
MAE	ViT-base	93.23±0.41	94.34±0.42	**94.47±0.38**	89.84±0.28	90.21±0.19	**90.29±0.53**
	Swin-T	92.71±0.26	**93.44±0.32**	93.33±0.17	90.21±0.43	89.87±0.59	**91.03±0.22**

produced through our approach demonstrates a more uniform and varied distribution. Upon closer inspection, the images crafted using our method exhibit a higher degree of realism, with the shapes of objects appearing more regular and well-defined, which are marked in the red boxes. Further visual comparisons on the NWPU and PN datasets are presented in the Appendix.

D. Unsupervised Classification

We employ the unsupervised classification task as a downstream experiment, aiming to both validate the effectiveness of MCR and to demonstrate its practical application in real-world tasks. We conduct two primary experiments. The first one operates at the feature level, utilizing the discriminator of the GAN as a feature extractor to perform unsupervised classification based on the extracted features. The second experiment operates at the image level, augmenting the original dataset with images generated by the GAN, and executing a self-supervised classification task based on this augmented dataset. We select the UCLand and NWPU datasets for testing the downstream task due to their inherent challenges for unsupervised classification; one dataset has the least amount of data while the other has the most categories.

In the first experiment, we use the learned representations Z of the discriminator as features and apply a regularized linear L2-SVM classifier, adhering to the methodology used in prior studies [2][3]. The results, presented in Table V, underscore the superior performance of our method compared to StyleGAN2 and StyleGAN2+ADA across all tests. Our approach consistently outperforms in the majority of the experiments, affirming our method's ability to learn enhanced representations that facilitate precise data classification. Specifically, our method surpasses the baseline method by 4.16% on the UCLand dataset. Similarly, on the NWPU dataset, our method outperforms the baseline method by 4.80% and the second-best method by 2.62%. These experiments further validate that by applying the manifold constraint, the discriminator can concentrate on the underlying data manifold and capture its essential characteristics.

In the second experiment, we utilize self-supervised algorithms SimCLR [61] and MAE [62] as classification methods, choosing two distinct network architectures for each algorithm. All these models are initially pre-trained on the ImageNet dataset. We then augment the UCLand dataset with 1,050 generated images and the NWPU dataset with 9,000 generated images. These images are generated using traditional data augmentation, StyleGAN2, and our proposed method, respectively. Following this, we train the self-supervised models on these augmented RS datasets. These trained models are then used as feature extractors, and a regularized linear L2-SVM is employed for classification. As shown in Tables VI, the algorithms trained on the dataset enhanced by our proposed methods consistently outperform those trained on datasets augmented by other approaches in terms of accuracy in most experiments. With an average classification accuracy improvement of 2.32%, these results underscore the effectiveness of our proposed method in downstream tasks.

E. Ablation Study

In this section, we provide further ablation and analysis over different components of our method.

Relationship Matrix C. We employ various methods to construct the relationship matrix C. The first approach leverages SimCLR and K-means. We utilize a pre-trained SimCLR model to extract RS data features, followed by clustering using K-means. The centroids resulting from this clustering process

TABLE VII
ABLATION STUDY ON RELATIONSHIP MATRIX C

Methods	UCLand	NWPU	PN
Baseline	74.25	11.97	33.53
K-means (20)	71.22	11.31	32.78
K-means (40)	72.28	11.02	31.33
learnable (SimCLR)	**70.03**	10.09	**30.57**
learnable (CLIP)	70.98	**9.89**	30.71
learnable (ResNet)	70.81	10.82	30.78

TABLE VIII
ABLATION STUDY ON FEATURES FROM DIFFERENT BLOCKS

Block	4	6	8	10	12
FID	14.92	14.05	**10.09**	16.23	23.19

TABLE IX
ABLATION STUDY ON REGULARIZING GENERATOR VS. DISCRIMINATOR.

Methods	UCLand	NWPU	PN
Baseline	74.25	11.97	33.53
Only G	73.95	11.85	33.06
Only D	71.23	10.44	31.83
Ours	**70.03**	**10.09**	**30.57**

TABLE X
ABLATION STUDY ON THE HYPERPARAMETER λ

λ	UCLand	NWPU	PN
0.1	80.60	16.93	40.09
0.5	74.43	12.91	34.92
0.7	**69.26**	10.68	30.81
1	69.43	**9.88**	**30.47**
3	73.22	12.31	33.78

TABLE XI
ABLATION STUDY ON THE HYPERPARAMETER γ

γ	UCLand	NWPU	PN
0.1	78.29	18.87	41.13
0.5	74.62	13.16	34.92
0.7	71.26	10.73	31.92
1	**69.43**	**9.88**	**30.47**
3	73.24	13.09	35.07

serve as prototypes of the training data. In this experiment, we consider 20 and 40 clustering centroids, respectively. The second approach is a learnable matrix introduced in this paper. We evaluate it using pre-trained encoders such as SimCLR, CLIP [63], and a ResNet50 model (pretrained on ImageNet). We utilize StyleGAN2+ADA as the baseline method. Table VII presents the results of our ablation study on the UCLand, NWPU, and PN datasets. Notably, the choice of the pre-trained encoder has minimal impact on the results. In subsequent experiments, we opt for the learnable approach and the SimCLR model as the pre-trained encoder.

Representations Z. We conduct ablation studies on features from different network layers. As different network layers are related to different levels of details in the generated image, and the earlier blocks of the network impact the coarse structure or content of the image. We conduct experiments on NWPU dataset with StyleGAN2+ADA model, which consists of 14 blocks. We choose 5 blocks for comparison, and the results are shown in Table VIII. We empirically choose the outputs of 8th block as the representations Z.

Regularizing generator vs. discriminator. Our default method add regularization on the loss functions of both generator G and discriminator D. In this experiment, we investigate the effectiveness of separately regularizing G and D. We utilize StyleGAN2+ADA as the baseline method. Table IX presents the results of the ablation study on UCLand, NWPU and PN datasets. The No Regularization version yields poor results as expected. Adding the regularization method on D already brings significant improvement to the model under different datasets. As proposed in our final method, adding the regularization on both G and D achieves the best results.

Hyperparameters. We conduct the ablation study on the hyperparameters λ and γ using StyleGAN2+MCR on UCLand, NWPU and PN datasets. The FID scores are shown in Table X and Table XI. Based on the experiment results, we set $\lambda = 1$ and $\gamma = 1$ in the following experiments.

VII. CONCLUSION

In this study, we aim to address the challenges posed by RS images in the context of GANs. We observe that RS images exhibit higher intrinsic dimensions compared to natural images, resulting in difficulties for the discriminator and an increased risk of overfitting. To mitigate these issues, we introduce a novel measure to capture the real data manifold and propose the MCR method to effectively address the discriminator overfitting while enhancing the generator's performance. We also present innovative learning paradigms for the unsupervised generation of RS images. Our method's effectiveness is confirmed through theoretical analysis and comprehensive experiments on three RS datasets using different GAN models. This demonstrates the adaptability and efficiency of our approach.

As for future works, there are several intriguing avenues for further research. Further exploration into the application of the MCR method to other types of generative models could yield interesting insights. Another valuable direction would be investigating the impact of different RS image

characteristics, such as varying resolutions or spectral ranges, on GAN performance. This could provide crucial information to refine our approach further. By pursuing these lines of inquiry, we aim to continue enhancing the capabilities of GANs in handling the intricacies of RS images.

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst.*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[2] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "Marta gans: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, 2017.

[3] Y. Yu, X. Li, and F. Liu, "Attention gans: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, 2019.

[4] Y. Wei, X. Luo, L. Hu, Y. Peng, and J. Feng, "An improved unsupervised representation learning generative adversarial network for remote sensing image scene classification," *Remote Sens. Lett.*, vol. 11, no. 6, pp. 598–607, 2020.

[5] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.

[6] A. Hu, Z. Xie, Y. Xu, M. Xie, L. Wu, and Q. Qiu, "Unsupervised haze removal for high-resolution optical remote-sensing images based on improved generative adversarial networks," *Remote Sens.*, vol. 12, no. 24, p. 4162, 2020.

[7] L. Zhao, Y. Yin, T. Zhong, and Y. Jia, "Remote sensing image dehazing through an unsupervised generative adversarial network," *Sensors*, vol. 23, no. 17, p. 7484, 2023.

[8] J. Li, Z. Wu, Z. Hu, J. Zhang, M. Li, L. Mo, and M. Molinier, "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 373–389, 2020.

[9] X. Ma, Y. Huang, X. Zhang, M.-O. Pun, and B. Huang, "Cloud-egan: Rethinking cyclegan from a feature enhancement perspective for cloud removal by combining cnn and transformer," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2023.

[10] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, 2019.

[11] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, and W. Xu, "Improved srgan for remote sensing image super-resolution across locations and sensors," *Remote Sens.*, vol. 12, no. 8, p. 1263, 2020.

[12] J. Guo, F. Lv, J. Shen, J. Liu, and M. Wang, "An improved generative adversarial network for remote sensing image super-resolution," *IET Image Proc.*, vol. 17, no. 6, pp. 1852–1863, 2023.

[13] Y. Song, J. Li, Z. Hu, and L. Cheng, "Dbsagan: Dual branch split attention generative adversarial network for super-resolution reconstruction in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, 2023.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 30, 2017.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019, pp. 4401–4410.

[16] C. Gong, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[17] H. Narayanan and S. Mitter, "Sample complexity of testing the manifold hypothesis," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 23, 2010.

[18] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 33, pp. 12 104–12 114, 2020.

[19] H.-Y. Tseng, L. Jiang, C. Liu, M.-H. Yang, and W. Yang, "Regularizing generative adversarial networks under limited data," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 7921–7931.

[20] A. Dash, J. Ye, and G. Wang, "A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines: From medical to remote sensing," *IEEE Access*, 2023.

[21] S. Jozdani, D. Chen, D. Pouliot, and B. A. Johnson, "A review and meta-analysis of generative adversarial networks and their applications in remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 108, p. 102734, 2022.

[22] Y. Mroueh and T. Sercu, "Fisher gan," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 30, 2017.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," jan 2017. [Online]. Available: https://arxiv.org/abs/1701.07875

[24] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *Proc. Int. Conf. Mach. Learn.* PMLR, 2018, pp. 3481–3490.

[25] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," Feb. 2018. [Online]. Available: https://arxiv.org/abs/1802.05957

[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 30, 2017.

[27] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 30, 2017.

[28] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in

Proc. Int. Conf. Learn. Representations, 2018.

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2020, pp. 8110–8119.

[30] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 12 873–12 883.

[31] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 34, pp. 852–863, 2021.

[32] A. Sauer, K. Schwarz, and A. Geiger, "Stylegan-xl: Scaling stylegan to large diverse datasets," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.

[33] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2301.09515/

[34] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency regularization for generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2019.

[35] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Deceive d: Adaptive pseudo augmentation for gan training with limited data," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 34, pp. 21 655–21 667, 2021.

[36] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, "Differentiable augmentation for data-efficient gan training," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 33, pp. 7559–7570, 2020.

[37] C. Yang, Y. Shen, Y. Xu, and B. Zhou, "Data-efficient instance generation from instance discrimination," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 34, pp. 9378–9390, 2021.

[38] H. Liu, W. Zhang, B. Li, H. Wu, N. He, Y. Huang, Y. Li, B. Ghanem, and Y. Zheng, "Adaptivemix: Improving gan training via feature space shrinkage," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2023, pp. 16 219–16 229.

[39] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards faster and stabilized gan training for high-fidelity few-shot image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2020.

[40] T. Li, Z. Li, H. Rockwell, A. Farimani, and T. S. Lee, "Prototype memory and attention mechanisms for few shot image generation," in *Proc. Int. Conf. Learn. Representations*, vol. 18, 2022.

[41] N. Park, A. Anand, J. R. A. Moniz, K. Lee, T. Chakraborty, J. Choo, H. Park, and Y. Kim, "Mmgan: Manifold matching generative adversarial network," 2018. [Online]. Available: https://arxiv.org/abs/1707.08273

[42] Q. Li, B. Kailkhura, R. Anirudh, Y. Zhou, Y. Liang, and P. Varshney, "Mr-gan: Manifold regularized generative adversarial networks," 2018. [Online]. Available: https://arxiv.org/abs/1811.10427

[43] Y. Ni, P. Koniusz, R. Hartley, and R. Nock, "Manifold learning benefits gans," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 11 265–11 274.

[44] D. Bang and H. Shim, "Mggan: Solving mode collapse using manifold-guided training," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 2347–2356.

[45] M. Khayatkhoei, M. K. Singh, and A. Elgammal, "Disconnected manifold learning for generative adversarial networks," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 31, 2018.

[46] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06434

[48] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 4681–4690.

[49] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, 2018.

[50] M. A. mnmoustafa, "Tiny imagenet," 2017. [Online]. Available: https://kaggle.com/competitions/tiny-imagenet

[51] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, "The intrinsic dimension of images and its impact on learning," in *Proc. Int. Conf. Learn. Representations*, 2020.

[52] D. J. MacKay and Z. Ghahramani, "Comments on 'maximum likelihood estimation of intrinsic dimension' by e. levina and p. bickel (2004)," https://www.inference.org.uk/mackay/dimension/, 2005.

[53] B. C. Brown, A. L. Caterini, B. L. Ross, and et al., "Verifying the union of manifolds hypothesis for image data," in *Proc. Int. Conf. Learn. Representations*, 2022.

[54] B. Liu, S. Zhang, G. Song, and et al., "Rectifying the data bias in knowledge distillation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 1477–1486.

[55] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 2085–2094.

[56] T. Cover and J. A. Thomas, "Elements of information theory," 2006.

[57] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[58] M. Kang and J. Park, "Contragan: Contrastive learning for conditional image generation," *Adv. Neural Inf. Process Syst. (NIPS)*, vol. 33, pp. 21 357–21 369, 2020.

[59] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," in *Proc. Int. Conf. Learn.*

Representations, 2018.

[60] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Q. Weinberger, "An empirical study on evaluation metrics of generative adversarial networks," 2018. [Online]. Available: https://arxiv.org/abs/1806.07755

[61] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2020, pp. 1597–1607.

[62] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 16 000–16 009.

[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.