Equivariant Representation Learning for Augmentation-based Self-Supervised Learning via Image Reconstruction

Qin Wang[1] **Kai Krajsek**[2] **Hanno Scharr**[1]
[1]IAS-8: Data Analytics and Machine Learning, Forschungszentrum Jülich, Germany
[2]Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Germany
{qi.wang, k.krajsek, h.scharr}@fz-juelich.de

Abstract

Augmentation-based self-supervised learning methods have shown remarkable success in self-supervised visual representation learning, excelling in learning invariant features but often neglecting equivariant ones. This limitation reduces the generalizability of foundation models, particularly for downstream tasks requiring equivariance. We propose integrating an image reconstruction task as an auxiliary component in augmentation-based self-supervised learning algorithms to facilitate equivariant feature learning without additional parameters. Our method implements a cross-attention mechanism to blend features learned from two augmented views, subsequently reconstructing one of them. This approach is adaptable to various datasets and augmented-pair based learning methods. We evaluate its effectiveness on learning equivariant features through multiple linear regression tasks and downstream applications on both artificial (3DIEBench) and natural (ImageNet) datasets. Results consistently demonstrate significant improvements over standard augmentation-based self-supervised learning methods and state-of-the-art approaches, particularly excelling in scenarios involving combined augmentations. Our method enhances the learning of both invariant and equivariant features, leading to more robust and generalizable visual representations for computer vision tasks.

1 Introduction

Popular augmentation-based self-supervised learning methods [1]–[5] have shown remarkable success in their domain, primarily focusing on learning invariant features across different views of the same image. While these approaches have proven to be effective for many tasks, their performance is limited for downstream applications that require equivariant behavior [6].

Equivariance in feature learning ensures that a model's learned representations remain consistent under various transformations, including 2D or 3D translations, rotations, scaling, and changes in color or illumination [7]. Mathematically, this property implies that the model's transformation commutes with the transformation acting on both the input and feature spaces. In practical terms, an equivariant model's response to an object in an image remains stable regardless of its position, orientation, or other imaging conditions, potentially leading to better generalization on unseen data.

Recent work, such as SIE (Split Invariant and Equivariant) [8], has attempted to address the limitations of invariance-focused learning by introducing a split between invariant and equivariant features. During the pretraining process, SIE [8] uses known transformations (e.g. rotation and colour jittering) including their parameters to learn a linear mapping between equivariant features from two views. However, this approach faces several challenges:

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

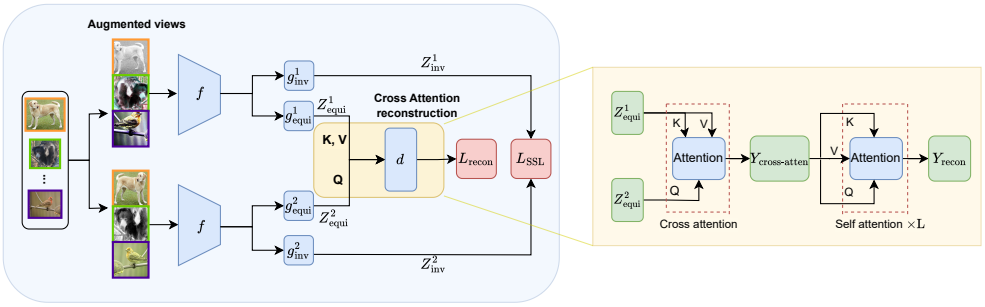

Figure 1: **Illustration of proposed equivariant reconstruction mechanism.** Cross-attention reconstruction decoder is composed with one cross attention layer, where Key and Value from first view and Query from the second view are mixed for computing the attention matrix. Subsequently, L× self-attention layer is added for reconstructing the image.

- It has been tested only on small, artificial datasets (3DIEBench [8]), limiting its proven applicability to real-world scenarios.
- It requires prior knowledge of transformations to learn equivariant features, which may not always be available or easily determinable.
- It struggles when dealing with images that have undergone unknown transformations.

To overcome these limitations, based on recent success in self-supervised learning using reconstruction to learn image features [9], [10], we propose a novel equivariance learning method based on image reconstruction, which leverages a cross-attention mechanism to facilitate the neural network in learning equivariant features. By utilizing image reconstruction, our approach enables the network to better capture the relationships between transformed images, leading to improved learning of equivariant representations. This method can be applied to all natural images without requiring prior knowledge of transformations, such as object motion tracking tasks, addressing a key limitation of previous work.

Our contributions are as follows:

- We introduce reconstruction as an auxilary task to learn equivariance, addressing the limitations of augmentation-based self-supervised learning.
- We demonstrate the effectiveness of our method on both artificial (3DIEBench [8]) and natural (ImageNet [11]) datasets, showing comparable (3DIEBench) and improved performance (ImageNet) compared to existing baselines.
- We provide extensive evaluations on various image transformations, including rotation, color jittering, translation, and scaling, demonstrating the robustness of our learned representations.

2 Method

Our proposed method integrates learning equivariant features with invariant augmentation-based self-supervised learning, as illustrated in Figure 1. The method builds upon the SIE framework (Split Invariant and Equivariant) [8]. The SIE framework divides the representations extracted from the encoder into two parts: one invariant and the other equivariant. The invariant part uses augmentation-based SSL loss as VICReg [4] to encourage the network to learn invariant features. Meanwhile, the equivariant part first encodes the transformations, enabling the construction of a linear predictor that maps equivariant representations from the first view to the second view.

In our approach, for the equivariant part, instead of building a linear predictor, the equivariant features are used to compute the reconstruction loss L_{recon} as the equivariance loss (see details in 2). This auxiliary reconstruction task encourages the network to learn robust equivariant features without requiring additional transformation encoding, as is needed in SIE. The final loss is a linear combination of the augmentation-based SSL loss and reconstruction losses, given by $L = \lambda_{\text{SSL}} L_{\text{SSL}} + \lambda_{\text{recon}} L_{\text{recon}}$.

Split Invariant and Equivariant Representations The original RGB images, each of size $X \times Y$, in a minibatch of size N, denoted as $\mathrm{I} \in \mathbb{R}^{N \times 3 \times X \times Y}$, are augmented to generate two different views, v_1 and v_2. These views are fed into an encoder, f, which shares weights between both inputs. The encoder outputs are then split into two parts: Y_{inv}, which contains invariant information, and Y_{equi}, which contains equivariant information. In our experiments, the output dimension for each view is 512. We split this 512-dimensional vector into two 256-dimensional vectors. These two representations are processed by separate heads, g_{inv} and g_{equi}, which produce embeddings $Z_{\text{inv}} \in \mathbb{R}^{N \times 192}$ and $Z_{\text{equi}} \in \mathbb{R}^{N \times 192}$, representing the invariant and equivariant features, respectively.

Cross-Attention Reconstruction To facilitate the learning of *equivariant features* from the images, we introduce an auxiliary reconstruction task. The reconstruction is performed using a decoder, d, which consists of a *cross-attention layer* followed by L self-attention layers. In the cross-attention layer, the *Key* (K) and *Value* (V) are derived from Z^1_{equi} of the first view, v_1, while the *Query* (Q) is derived from Z^2_{equi} of the second view, v_2. The cross-attention mechanism is defined as:

$$Y_{\text{cross-atten}} = \textbf{softmax}((W_Q Z^2_{\text{equi}})(W_K Z^1_{\text{equi}})^{\mathrm{T}})(W_V Z^1_{\text{equi}}) \tag{1}$$

The output of the cross-attention layer, $Y_{\text{cross-atten}}$, maintains the same dimensionality as the input feature Z_{equi}. This output is then passed through L self-attention layers, yielding the reconstructed images Y_{recon}. These reconstructed images are then used to compute the pixel-wise mean squared error (MSE) as the reconstruction loss L_{recon}, with v_2 serving as the target.

3 Experimental results

Experiment settings. We employ the ViT-Small architecture as the base encoder, which is trained with a batch size of $N = 2048$ for 800 epochs. The optimizer used is Adam, with a linearly scaled learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. The decoder consists of 6 blocks, each with an embedding dimension of 192. Notably, only the first block of the decoder incorporates a cross-attention layer. All experiments adhere to these settings. Under these conditions, pretraining on natural images (ImageNet) takes approximately 23 hours using 16 A100 GPUs.

Representation evaluation metrics. We follow the same evaluation metrics as SIE [8], applying a linear classifier on top of the pretrained frozen encoder to predict the transformations. The representations from the two augmented views are fed into a 3-layer MLP, which is trained to regress the transformations between the two views.

For all transformation predictions, performance is evaluated using the coefficient of determination, R^2, which quantifies how well predicted values approximate true values. Specifically, R^2 indicates the proportion of variance in the true values that is explained by the predictions, where $R^2 = 1$ represents perfect prediction, and $R^2 = 0$ indicates that the model explains none of the variance.

3DIEBench	Classification	Rotation Prediction	Color Prediction
SIE(rot)	**0.820**	**0.724**	0.054
SIE(rot+color)	<u>0.809</u>	0.502	**<u>0.980</u>**
Ours	0.782	<u>0.554</u>	0.954

Table 1: Comparison of different methods on the 3DIEBench [8] dataset. Bold values indicate overall best results, underlined values indicate the better results within direct comparison of Ours and SIE [8] with combined augmentations (rotation and color jittering).

Evaluation on 3DIEBench [8] dataset. We evaluate our method using the 3DIEBench dataset, which provides transformation parameters. In contrast to the SIE method, which relies on the knowledge of the augmentation transformation parameters, our approach does not require any information about the transformations involved.

From Table 1, the SIE model with rotation knowledge excels in Classification and Rotation Prediction but performs poorly in Color Prediction without color prior knowledge. Incorporating color augmentation in SIE(rot+color) greatly improves Color Prediction but reduces performance in Rotation Prediction. The Ours model strikes a balance, performing well across all tasks, making it a versatile choice without any knowledge of transformation involved.

Evaluation on natural images. We augment the image to create two views. SIE needs the augmentation parameters as prior knowledge in the pretraining process. We create augmentation from the first view to the second view, and provide augmentation parameters.

ImageNet	Rotation	Color	Blur radius	Translation	Crop prediction	Flip
SIE(rot)	**0.990**	0.867	0.042	0.540	0.266	0.532
SIE(color)	0.078	**0.890**	0.097	0.355	0.178	0.333
SIE(blur)	0.153	0.883	**0.941**	0.189	0.412	0.415
SIE(trans)	0.213	0.885	0.023	**0.978**	0.368	0.511
SIE(crop)	0.273	0.819	0.018	0.450	**0.922**	0.485
SIE(flip)	0.155	0.798	0.056	0.312	0.266	**0.993**
VICReg[4]	0.318 ± 0.005	0.804 ± 0.016	0.101 ± 0.023	0.333 ± 0.008	0.423 ± 0.140	0.872 ± 0.070
SIE(all)	0.331 ± 0.007	0.899 ± 0.003	0.211 ± 0.005	$\underline{0.925 \pm 0.002}$	0.835 ± 0.008	0.945 ± 0.004
SIE(all, single each time)	0.435 ± 0.011	0.907 ± 0.009	0.377 ± 0.004	0.922 ± 0.010	0.829 ± 0.005	0.939 ± 0.007
Ours	$\underline{0.862 \pm 0.004}$	$\underline{\mathbf{0.921 \pm 0.006}}$	$\underline{0.823 \pm 0.003}$	0.853 ± 0.005	$\underline{0.912 \pm 0.002}$	$\underline{0.952 \pm 0.008}$

Table 2: Performance comparison on ImageNet for different prediction tasks. Bold values indicate overall best results, underlined values indicate the better results within direct comparison of Ours and SIE [8]. The results of VICReg are obtained by using a pretrained model with a 3-layer MLP for finetuning, specifically for evaluating equivariance.

In Table 2 we see, that SIE models excel when pretrained with specific single transformations, such as rotation, achieving the best results for rotation prediction. However, their performance drops significantly for other transformations. Even with all transformation information (as in SIE(all)), their performance remains lower than ours. Pretraining with randomly selected transformations (as in SIE(all, single each time)) improves results compared to SIE(all) but still falls short of our method.

We further evaluate transfer learning on smaller classification datasets and segmentation tasks, i.e. ADE20K[12]. The results are attached in A.2.

Cifar10	Rotation	Color	Blur Radius	Translation
Supervised	0.214	0.229	0.437	0.386
SIE(all)	0.402	0.395	0.511	0.479
Ours	**0.815**	**0.879**	**0.944**	**0.878**

Table 3: Comparison on partial CIFAR10 data. The bold values indicate the overall best.

Utilisation of unknown transformations for learning equivariant representations We evaluate the effectiveness of our method for learning equivariant representations only with knowledge of parts of the augmentation transformations and compare its performance with that of SIE. With the CIFAR10 dataset [13] we denote 80% of the training data as data subject to unknown transformations and for 20% the transformations including their parameters are known. Since SIE need to know the transformations, it can only use the 20% of the data. Therefore, SIE as well as *supervised* are trained exclusively on the remaining 20% data with known transformations, whereas our method leverages the entire dataset. Both models are evaluated on the validation set. As shown in Table 3, our method significantly outperforms SIE, and thus emphasises its ability to efficiently use data that has been subjected to unknown transformation to generate robust equivariant representations.

4 Conclusions

Our proposed method integrates equivariant representation learning into augmentation-based self-supervised learning through a reconstruction task, demonstrating potential for enhancing the generalization capabilities of invariant augmentation-based self-supervised learning. In this paper, we evaluate our approach on multiple datasets and downstream tasks to measure its impact on the equivariant properties of pretrained networks. Our method matches the performance of SIE [8] on the 3DIEBech dataset and surpasses it on natural image datasets.

Our experiments are currently limited to using smaller backbones and datasets for testing. In the future, we plan to explore larger network architectures and datasets to further evaluate the effectiveness of our method across a broader range of augmentation-based self-supervised learning techniques. Additionally, we will investigate alternative image reconstruction methods for learning equivariant representations.

Acknowledgements The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[2] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735. DOI: 10.1109/CVPR42600.2020.00975.

[3] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[4] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2022.

[5] J. Zhou *et al.*, "iBOT: Image BERT pre-training with online tokenizer," *International Conference on Learning Representations (ICLR)*, 2022.

[6] H. Lee, K. Lee, K. Lee, H. Lee, and J. Shin, "Improving transferability of representations via augmentation-aware self-supervision," in *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[7] M. Weiler and G. Cesa, "General $E(2)$-equivariant steerable CNNs," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[8] Q. Garrido, L. Najman, and Y. LeCun, "Self-supervised learning of split invariant equivariant representations," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[9] A. Gupta, J. Wu, J. Deng, and L. Fei-Fei, "Siamese masked autoencoders," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988. DOI: 10.1109/CVPR52688.2022.01553.

[11] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[12] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5122–5130. DOI: 10.1109/CVPR.2017.544.

[13] A. Krizhevsky, *Learning multiple layers of features from tiny images*, https://api.semanticscholar.org/CorpusID:18268744(visited on 2024-10-26), 2009.

[14] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 446–461. DOI: https://doi.org/10.1007/978-3-319-10599-4_29.

[15] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010. DOI: 10.1109/CVPR.2010.5539970.

[16] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3606–3613. DOI: 10.1109/CVPR.2014.461.

[17] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3498–3505. DOI: 10.1109/CVPR.2012.6248092.

[18] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, *Fine-grained visual classification of aircraft*, 2013. arXiv: 1306.5151 [cs-cv].

[19] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 845–853. DOI: 10.1109/CVPR.2016.98.

A Appendix / supplemental material

A.1 Experiments on CIFAR10

Similar to the ImageNet dataset discussed in Section 3, we also perform transformation prediction on smaller datasets, such as CIFAR10. The conclusions drawn from these smaller datasets are consistent with those observed for ImageNet.

Cifar10	Rot Prediction	Color Prediction	Blur Radius	Trans Prediction
SIE(rot)	**0.989**	0.887	0.836	0.911
SIE(color)	0.813	**0.921**	0.825	0.822
SIE(blur)	0.814	0.833	**0.990**	0.807
SIE(trans)	0.876	0.812	0.810	**0.987**
SIE(all)	0.845	0.864	0.889	0.886
Ours	0.826	0.906	0.972	0.890

Table 4: Comparison of different prediction methods on the CIFAR10 dataset. Bold values indicate overall best results, underlined values indicate the better results within direct comparison of Ours and SIE

A.2 Transfer learning on downstream tasks

Transfer learning on classification tasks. We follow the standard self-supervised learning evaluation pipeline, where the pretrained network is frozen and only the linear head is fine-tuned on downstream tasks. From Table 5, we observe that our method performs well on most classification datasets, with the exception of the Pets dataset, when compared to SIE. In the case of the Aircraft dataset, SIE outperforms other methods due to its rotation prior, which better accommodates the rotation-invariant nature of the images.

Methods	Cifar10 [13]	Cifar100 [13]	Food101 [14]	SUN397 [15]	DTD [16]	Pets [17]	Aircraft [18]
SIE(rot)	71.56	46.88	55.48	43.11	64.22	81.51	**50.21**
SIE(color)	67.99	48.78	57.19	42.32	60.87	80.27	41.15
SIE(crop)	80.84	49.35	59.24	52.38	61.82	84.63	47.35
Supervised	80.99	50.66	59.32	52.98	62.03	83.59	47.83
SIE(all)	79.91±0.18	53.12±0.05	58.42±0.20	56.11±0.08	63.56±0.11	**85.34±0.19**	46.88±0.23
Ours	**81.12±0.11**	**54.22±0.10**	**59.21±0.14**	**59.53±0.13**	**67.66±0.12**	84.32±0.09	49.75±0.22

Table 5: Transfer learning on classification tasks. Bold values indicate best results within direct comparison of Ours vs. SSL methods, underlined values overall best.

Transfer learning on segmentation task. We use an encoder with HyperNet [19] on top of the encoder for the segmentation task. The table below presents the results of transfer learning experiments on the ADE20K dataset, evaluating three different methods: Supervised, SIE(all), and Ours. The metrics used are mean Intersection over Union (mIOU), mean Accuracy (mAcc), and overall Accuracy (aAcc). In this comparison, our method outperforms both the Supervised and SIE(all) approaches across all metrics, achieving the highest mean Intersection over Union (mIOU), mean Accuracy (mAcc), and overall Accuracy (aAcc).

ADE20K	mIOU	mAcc	aAcc
Supervised	0.268	0.328	0.751
SIE(all)	0.292	0.356	0.774
Ours	**0.312**	**0.379**	**0.802**

Table 6: Transfer learning on segmentation tasks. Bold values indicate best results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in abstract accurately reflect the paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In conclusion section, we discuss our current limitation and future work for it.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our research is experimental research, on theoretical results are provided.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included all information needed to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will publish the code in the paper once the paper got accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: For the training details, we include all the necessary settings in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Some important results we report the error bars, but some experiments, limited by time and computing resources, the statistical information is neglected.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide our information of computer resources along with our settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: We will review the NeurIPS Code of Ethics and refine our code to fully comply with its requirements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed in our paper.

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not relate to the safe risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have used the existed library and code, we also cite them in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We will documnet our model and code in the paper later when the decision is made.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: Our paper does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.