NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024

北京交通大學
BEIJING JIAOTONG UNIVERSITY

DenoiseRep: Denoising Model for Representation Learning

Zhengrui Xu[1†]
zrxu23@bjtu.edu.cn

Guan'an Wang[† ‡]
guan.wang0706@gmail.com

Xiaowen Huang[1,2,3 *]
xwhuang@bjtu.edu.cn

Jitao Sang[1,2,3]
jtsang@bjtu.edu.cn

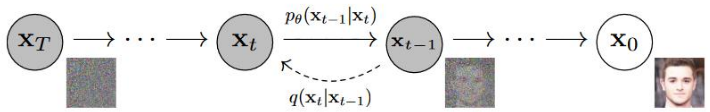
[1]School of Computer Science and Technology, Beijing Jiaotong University
[2]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University
[3]Key Laboratory of Big Data & Artificial Intelligence in Transportation(Beijing Jiaotong University), Ministry of Education

[†]Equal Contribution.
[‡]Project Lead.
[*]Corresponding Author.

Contents

Introduction

- **Diffusion Model**
  - **A powerful generative model**
  - **Generate images through denoising**



- **Representational Learning**
  - **Representation learning is important in discriminative tasks(classification, detection, retrieval, segmentation...)**

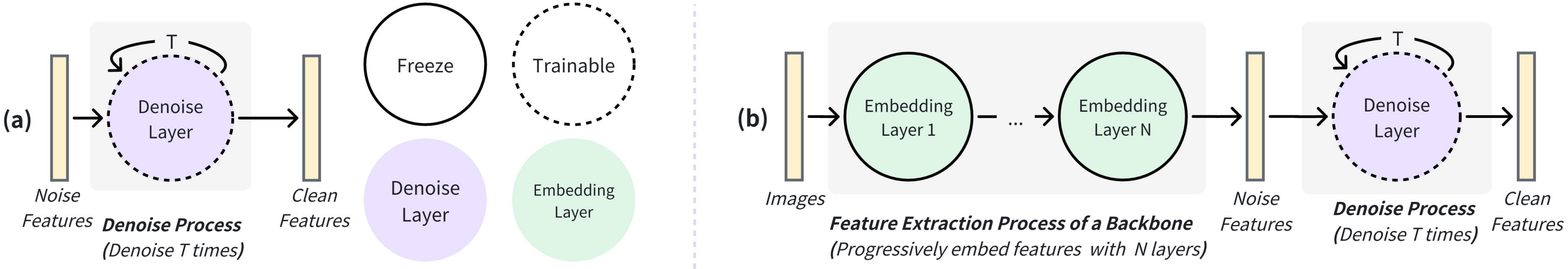
- **Contribution**
  - **We propose a novel Denoising Model for Representation Learning (*DenoiseRep*).**
  - **We propose a denoising method without additional inference time.**
  - **We validate the effectiveness and generalization of our algorithm on multiple datasets and various discrimination tasks.**

Contents

北京交通大學
BEIJING JIAOTONG UNIVERSITY

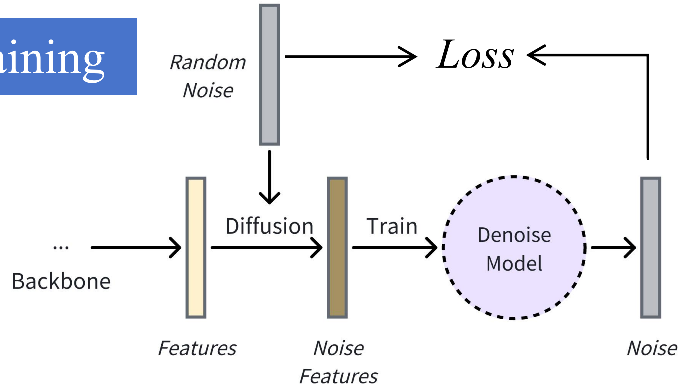
Methodology

■ Joint Feature Extraction and Feature Denoising (*DenoiseRep*$^{-}$)

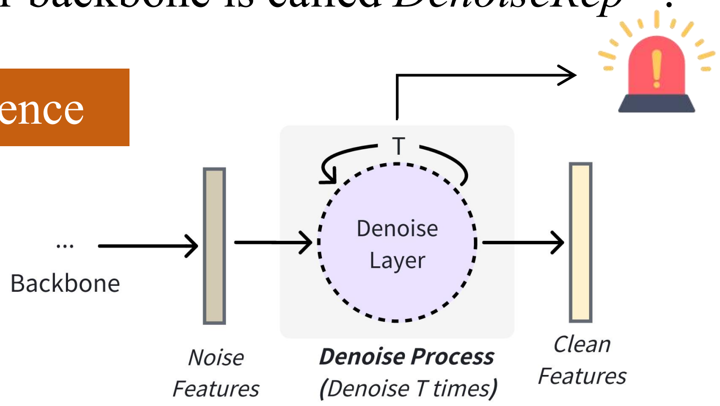


We refer to the diffusion modeling approach to denoise the noisy features through T-steps to obtain clean features. The method of adding denoising process after backbone is called *DenoiseRep*$^{-}$.

Training



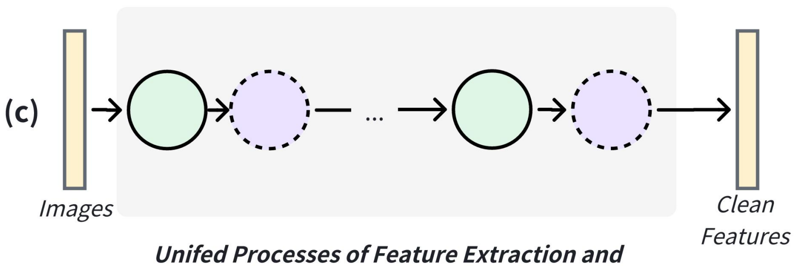
Inference



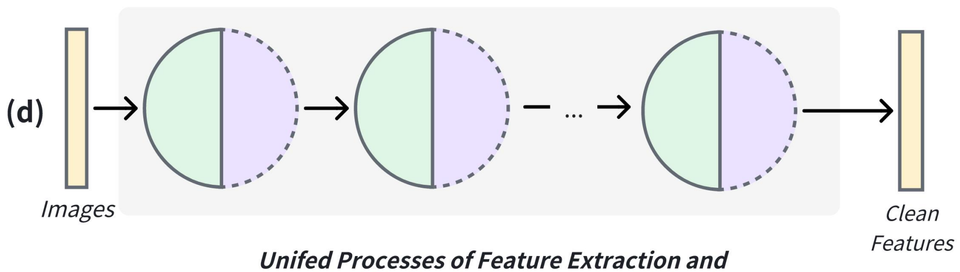
Too much inference cost

Methodology

- Fuse Feature Extraction and Feature Denoising (*DenoiseRep*)



Unifed Processes of Feature Extraction and Denoising with extra N denoise layers
(Progressively embed N layers and denoise N times, require more computation)



Unifed Processes of Feature Extraction and Denoising with denoise layers fused
(Progressively embed N layers and denoise N times, no extra computation)

We propose a novel Denoising Model for Representation Learning (*DenoiseRep*) that adds denoising layers to each embedding layer in the backbone network and fuses the parameters of the denoising layers into the parameters of the corresponding embedding layers and theoretically demonstrates their equivalence.

? How to fuse feature extraction and feature denoising?

DDPM: $X_{t-1} = \frac{1}{\sqrt{a_t}}(X_t - \frac{1-a_t}{\sqrt{1-\bar{a}_t}} D_\theta(X_t, t)) + \sigma_t z$

$$\frac{1}{\sqrt{a_t}} X_t - X_{t-1} = \frac{1-a_t}{\sqrt{a_t}\sqrt{1-\bar{a}_t}} D_\theta X_t - \sigma_t z$$

Feature extraction: $Y = WX + b$

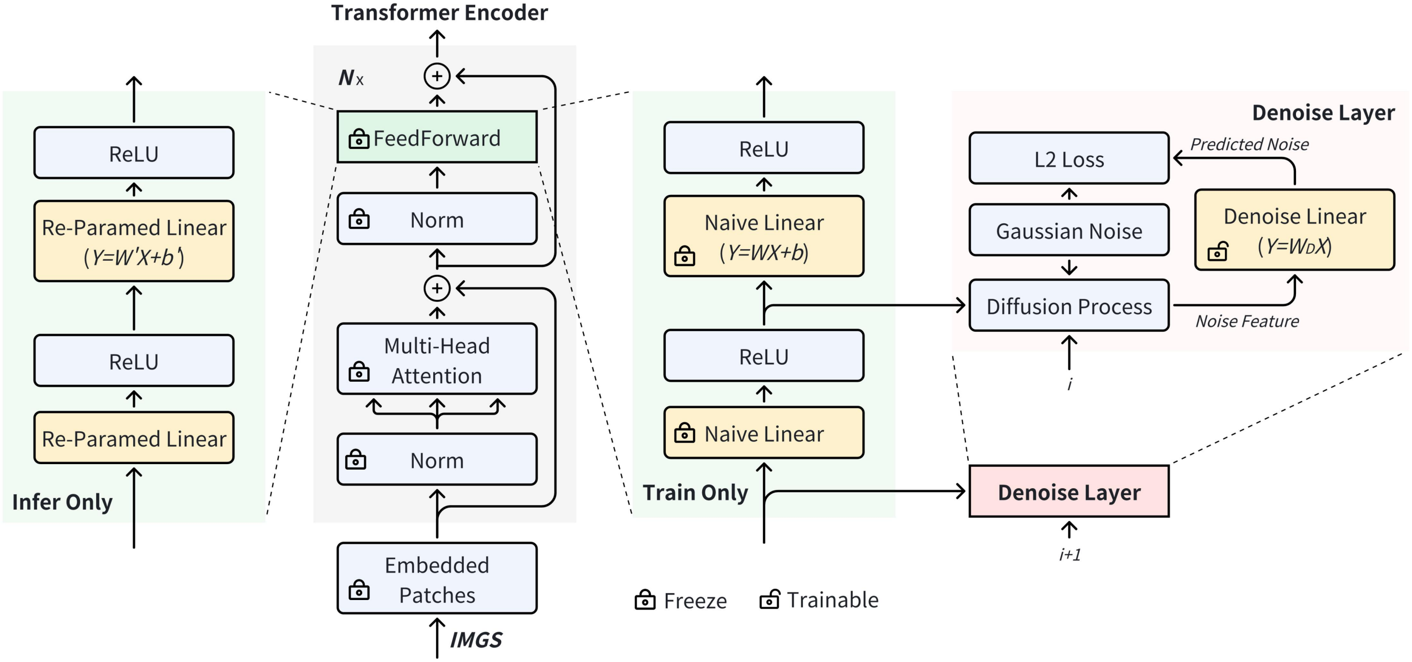
$$\frac{1}{\sqrt{a_t}} Y_t - Y_{t-1} = \frac{1-a_t}{\sqrt{a_t}\sqrt{1-\bar{a}_t}} D_\theta Y_t - \sigma_t z$$

$$Y_{t-1} = [W - C_1(t) W W_D] X_t + W C_2(t) C_3 + b$$

$$C_1(t) = \frac{1-a_t}{\sqrt{a_t}\sqrt{1-\bar{a}_t}} \quad C_2(t) = \frac{1-\bar{a}_{t-1}}{1-\bar{a}_t}\beta_t \quad C_3 = Z \sim N(0, I)$$

Methodology

■ Pipeline of our proposed *DenoiseRep*



Training loss:

$$Loss_p = \sum_{i=1}^{N} |\epsilon_i - D_{\theta_i}(X_{t_i}, t_i)|$$

$$Loss = (1-\lambda)Loss_l + \lambda Loss_p$$

- In training: We freeze the original network parameters, train only the denoising layer parameters, and input the diffusion features into DenoiseRep for prediction.

- In inference: We merge the parameters of the feature layer and the denoising layer, merging the two branches into one without additional inference time.

Contents

Experiments

■ Analysis of Generalization Ability

Task	Model	Backbone	Dataset	Metric	*Baseline*	***+DenoiseRep***
Classification	SwinT [39]	SwinV2-T	ImageNet-1k	acc@1	81.82%	82.13%
Person-ReID	TransReID-SSL [41]	ViT-S	MSMT17	mAP	66.30%	67.33%
Detection	Mask-RCNN [19]	SwinV2-T	COCO	AP	42.80%	44.30%
Segmentation	FCN [40]	ResNet-50	ADE20K	BIoU	28.70%	29.90%

Our proposed *DenoiseRep* is a versatile method that can be incrementally applied to various discriminative tasks. The table demonstrates that *DenoiseRep* yields stable and substantial improvements across image classification, object detection, image segmentation, and person re-identification.

北京交通大學
BEIJING JIAOTONG UNIVERSITY

Experiments

■ Analysis of Generalization Ability

Method	Datasets	Param	acc@1		acc@5	
			Baseline	***+DenoiseRep***	*Baseline*	***+DenoiseRep***
SwinV2-T [39]	ImageNet-1k	28M	81.82%	82.13%	95.88%	96.06%
SwinV2-S [39]	ImageNet-1k	50M	83.73%	83.97%	96.62%	96.86%
SwinV2-B [39]	ImageNet-1k	88M	84.20%	84.31%	96.93%	97.06%
Vmanba-T [38]	[illegible]et-1k	30M	82.38%	82.51%	95.80%	95.89%
Vmanba-S [38]	ImageNet-1k	50M	83.12%	83.27%	96.04%	96.22%
Vmanba-B [38]	ImageNet-1k	89M	83.83%	83.91%	96.55%	96.70%
ViT-S [14]	ImageNet-1k	22M	83.87%	84.02%	96.73%	96.86%
ViT-B [14]	ImageNet-1k	86M	84.53%	84.64%	97.15%	97.23%
ResNet50 [18]	ImageNet-1k	26M	76.13%	76.28%	92.86%	92.95%
ViT-S [14]	Cifar-10	22M	96.13%	96.20%	–	–
ViT-B [14]	Cifar-10	87M	98.02%	98.31%	–	–

Classification

Method	Backbone	MSMT17		Market1501		DukeMTMC		CUHK03-L	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
MGN [59]	ResNet-50	–	–	86.90	95.70	78.40	88.70	67.40	68.00
OSNet [74]	OSNet	52.90	78.70	84.90	94.80	73.50	88.60	–	–
BAT-net [15]	*GoogLeNet*	*56.80*	*79.50*	*87.40*	*95.10*	*77.30*	*87.70*	*76.10*	*78.60*
ABD-Net [8]	ResNet-50	60.80	82.30	88.30	95.60	78.60	89.00	–	–
RGA-SC [68]	ResNet-50	57.50	80.30	88.40	96.10	–	–	77.40	81.10
ISP [76]	HRNet-W32	–	–	88.60	95.30	80.00	89.60	74.10	76.50
CDNet [29]	CDNet	54.70	78.90	86.00	95.10	76.80	88.60	–	–
Nformer [60]	ResNet-50	59.80	77.30	91.10	94.70	83.50	89.40	78.00	77.20
TransReID [20]	ViT-base-ics	67.70	85.30	89.00	95.10	82.20	90.70	84.10	86.40
TransReID	ViT-base	61.80	81.80	87.10	94.60	79.60	89.00	82.30	84.60
TransReID-SSL [41]	ViT-small	66.30	84.80	91.20	95.80	80.40	87.80	83.50	85.90
TransReID-SSL	ViT-base	75.00	89.50	93.10	96.52	84.10	92.60	87.80	89.20
CLIP-REID [32]	ViT-base	75.80	89.70	90.50	95.40	83.10	90.80	–	–
TransReID + ***DenoiseRep***	ViT-base-ics	68.10	85.72	89.56	95.50	82.35	90.87	84.15	86.39
TransReID + ***DenoiseRep***	ViT-base	62.23	82.02	87.25	94.63	80.12	89.33	82.44	84.61
TransReID-SSL + ***DenoiseRep***	ViT-small	67.33	85.50	92.05	96.68	81.22	88.72	84.11	86.47
TransReID-SSL + ***DenoiseRep***	ViT-base	75.35	89.62	**93.26**	**96.55**	**84.31**	**92.90**	**88.08**	**89.29**
CLIP-REID + ***DenoiseRep***	ViT-base	**76.30**	**90.60**	91.10	95.80	83.70	91.60	–	–

Person-ReID

Methods	Backbones	AP		AP_{50}		AP_{75}	
		Baseline	***+DenoiseRep***	*Baseline*	***+DenoiseRep***	*Baseline*	***+DenoiseRep***
Mask-RCNN	SwinV2-T	42.8%	44.3%	65.1%	67.1%	47.0%	48.6%
	SwinV2-S	48.2%	49.0%	69.9%	70.9%	52.8%	53.8%
	ResNet-50	42.6%	43.2%	63.7%	65.0%	46.4%	46.8%
Faster-RCNN	ResNet-50	37.4%	38.3%	58.1%	58.8%	40.4%	41.0%
ATSS	ResNet-50	39.4%	39.9%	57.6%	58.2%	42.8%	43.2%
YOLO	DarkNet-53	27.9%	28.4%	49.2%	50.3%	28.3%	27.8%
DETR	ResNet-50	39.9%	40.8%	60.4%	59.9%	41.7%	42.9%
CenterNet	ResNet-50	40.2%	40.6%	58.3%	59.1%	43.9%	44.0%

Detection

Methods	Backbones	aAcc		B-IoU		mIoU	
		Baseline	***+DenoiseRep***	*Baseline*	***+DenoiseRep***	*Baseline*	***+DenoiseRep***
FCN [40]	ResNet-50	0.774	0.779	0.287	0.299	0.359	0.365
FCN	ResNet-101	0.793	0.796	0.306	0.316	0.396	0.404
SegFormer [63]	mit_b0	0.782	0.788	0.292	0.297	0.374	0.381
SegFormer	mit_b1	0.812	0.816	0.341	0.348	0.422	0.425

Segmentation

Experiments

BEIJING JIAOTONG UNIVERSITY

■ Analysis of Label Informations

Method	DukeMTMC(%)	MSMT17(%)	Market1501(%)	CUHK-03(%)
TransReID-SSL	80.40	66.30	91.20	83.50
+*DenoiseRep* (label-free)	80.92 (↑ 0.52)	66.87 (↑ 0.57)	91.82 (↑ 0.62)	83.72 (↑ 0.22)
+*DenoiseRep* (label-aug)	81.22 (↑ 0.82)	67.33 (↑ 1.03)	92.05 (↑ 0.85)	84.11 (↑ 0.61)
+*DenoiseRep* (merged ds)	80.98 (↑ 0.58)	66.99 (↑ 0.69)	91.80 (↑ 0.60)	83.86 (↑ 0.36)

- As shown in the table line2, compared with baseline method (line1), the baseline method performs better after adding our label-free method.
- Introducing supervised training label information can further improve performance.
- Since our method can perform unsupervised feature denoising, adding more data to train the model can further improve its performance.

Experiments

■ Analysis of Parameter Fusion

Method	DukeMTMC	MSMT17	Market1501	CUHK-03	Inference Time
TransReID-SSL	80.40%	66.30%	91.20%	83.50%	**0.34s**
+***DenoiseRep***$^{-}$	80.76%	66.81%	91.07%	83.59%	0.39s (+15%)
+***DenoiseRep***	81.22%	67.33%	92.05%	84.11%	**0.34s** (+0%)

- The proposed *DenoiseRep* is computation-free. We proved by theoretical derivation that inserting our denoising layer into each feature layer and fusion it does not introduce additional computation.
- Adding *DenoiseRep* $^{-}$ is able to improve the the performance but brings extra inference latency (about 15%).
- Adopting *DenoiseRep* achieves a greater increase, it denoise the features on each layer, which can better remove noise at each stage. And *DenoiseRep* does not take extra inference latency cost.

Contents

北京交通大學
BEIJING JIAOTONG UNIVERSITY

Conclusion

- In this work, we demonstrate that the diffusion model paradigm is effective for feature level denoising in discriminative model, and propose a computation-free and label-free method: *DenoiseRep*.

- It utilizes the denoising ability of diffusion models to denoise the features in the feature extraction layer, and fuses the parameters of the denoising layer and the feature extraction layer, further improving retrieval accuracy without incurring additional computational costs.

- We validate the effectiveness of the *DenoiseRep* method on multiple common image discrimination task datasets.

北京交通大學
BEIJING JIAOTONG UNIVERSITY

Thank you

Reporter: Zhengrui Xu
Email: zrxu23@bjtu.edu.cn