Media Synthesis & Forensics Lab

MAKE THE WORLD MORE CREDIBLE

Progressive Open Space Expansion for Open-Set Model Attribution

Tianyun Yang, Danding Wang*, Fan Tang, Xinying Zhao, Juan Cao, Sheng Tang

Media Synthesis and Forensics Lab, Institute of Computing Technology, CAS

University of Chinese Academy of Sciences

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Tag: WED-PM-334

Problem

- **Model Attribution:** Identify the source model of generated contents.

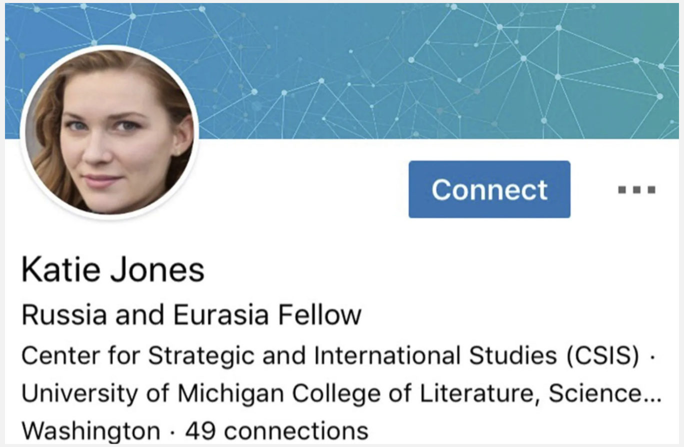
Black-box Model

Which Model?

Applications

Intellectual Protection

Malicous Content Supervision

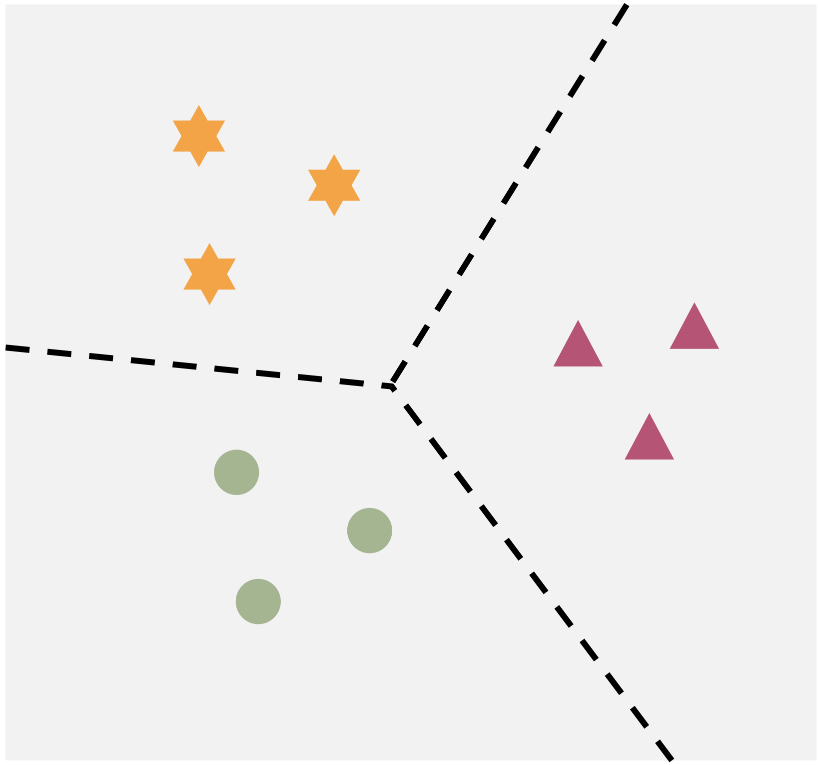


1 Protrit of Edmond Belamy, 2018, created by GAN (Generative Adversarial Network).
2 An AI-Generated Picture Won an Art Prize.
3 Raphael S. (2019, Jul 14). Experts: A spy reportedly used an AI-picture to connect with sources on LinkedIn.

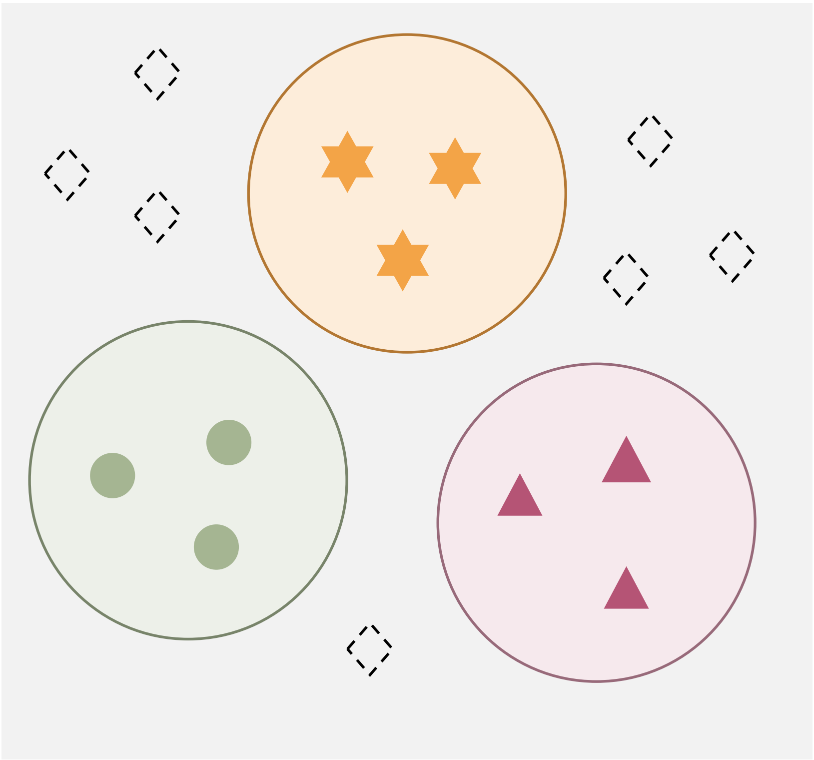
Open-Set Model Attribution

- Attribute images to known models and identify those from unknown ones.

Prior works:
Closed-set model attribution

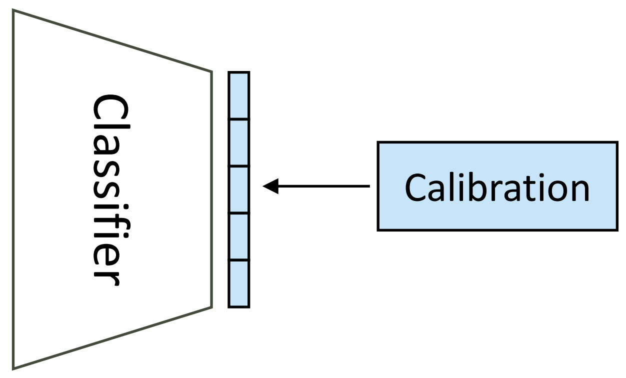


Our work:
Open-set model attribution



Existing Works on OSR

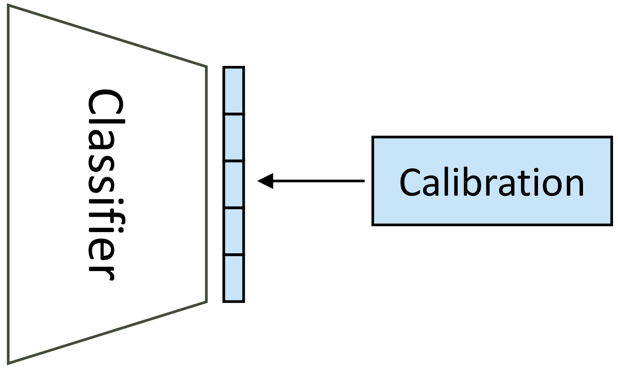
Discriminative-Based



Drawback: The performance depends heavily on the closed-set classifier.

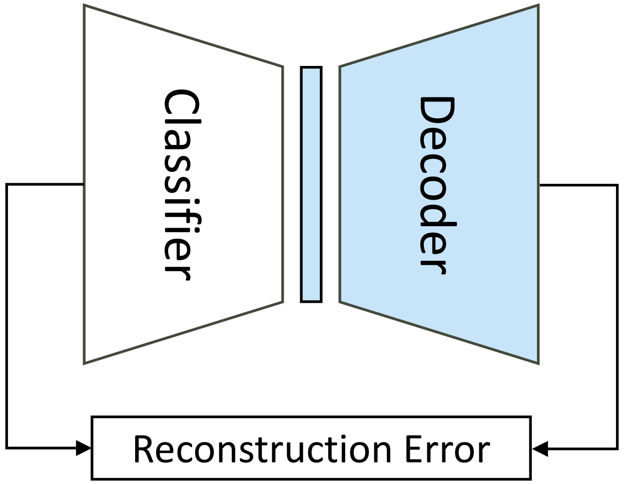
Existing Works on OSR

Discriminative-Based



Drawback: The performance depends heavily on the closed-set classifier

Generative-Based

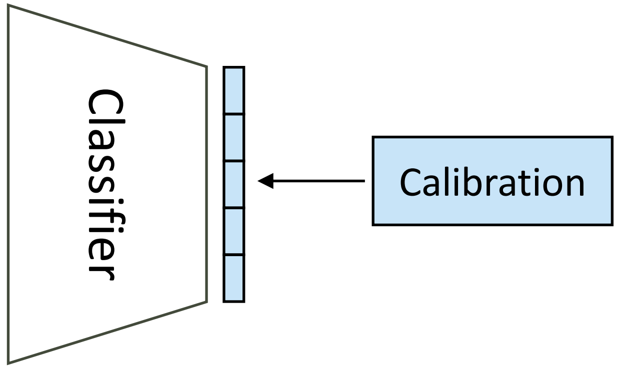


1. Reconstruction-based

Drawback: Fingerprint reconstruction error is too subtle to be thresholded

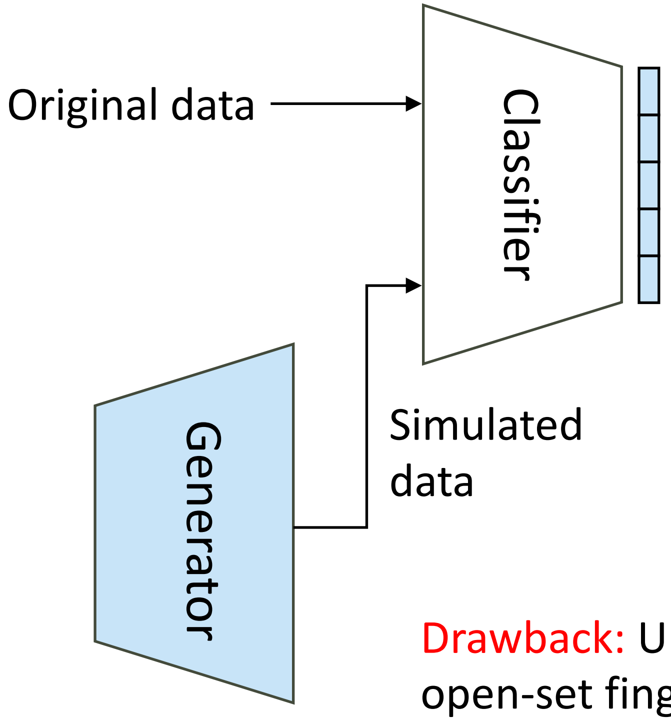
Existing Works on OSR

Discriminative-Based



Drawback: The performance depends heavily on the closed-set classifier

Generative-Based



2. Open data simulation-based

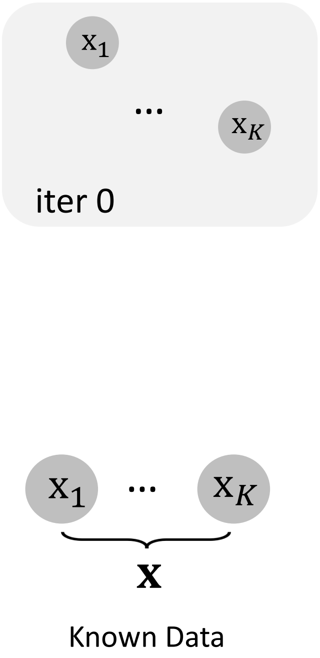
Drawback: Unable to produce diverse open-set fingerprints

POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.

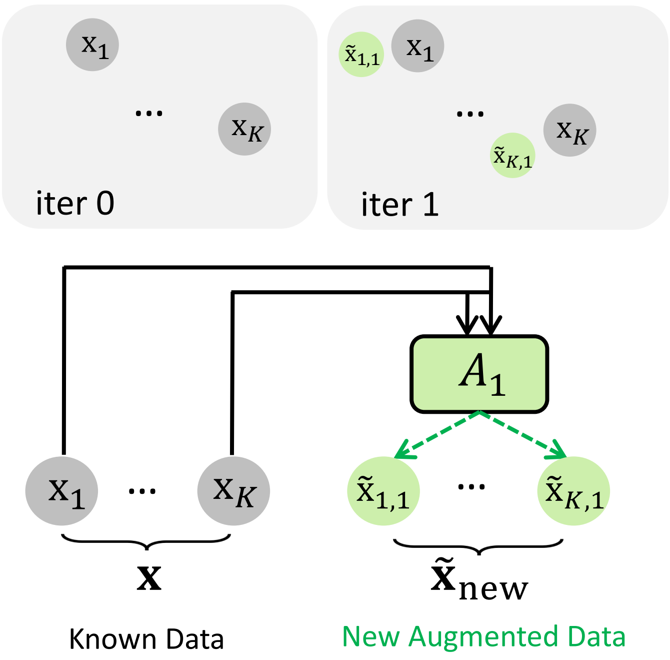
POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



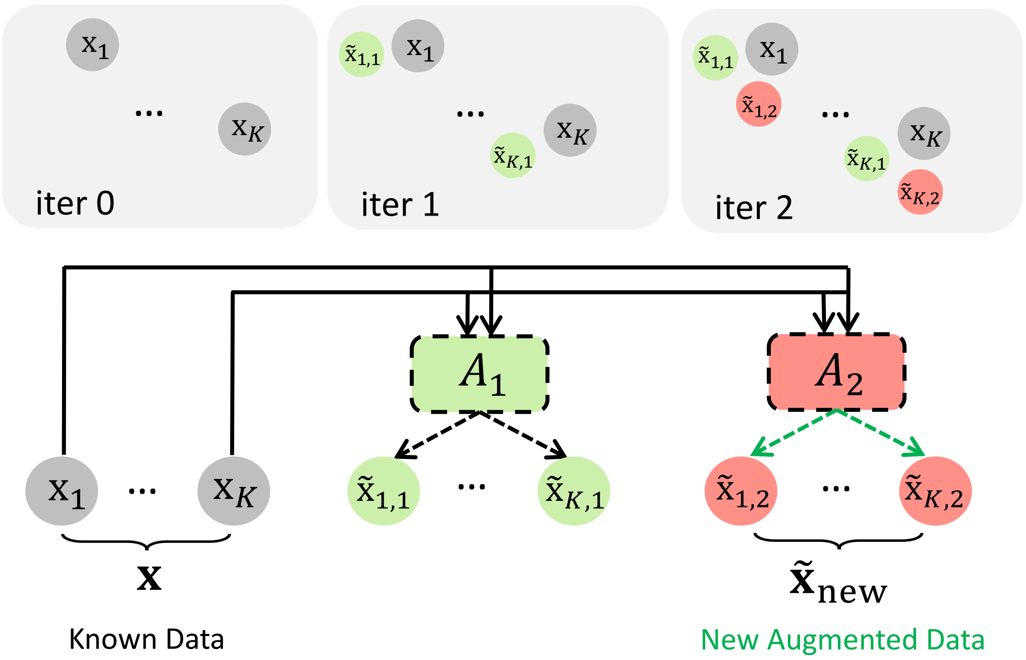
POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



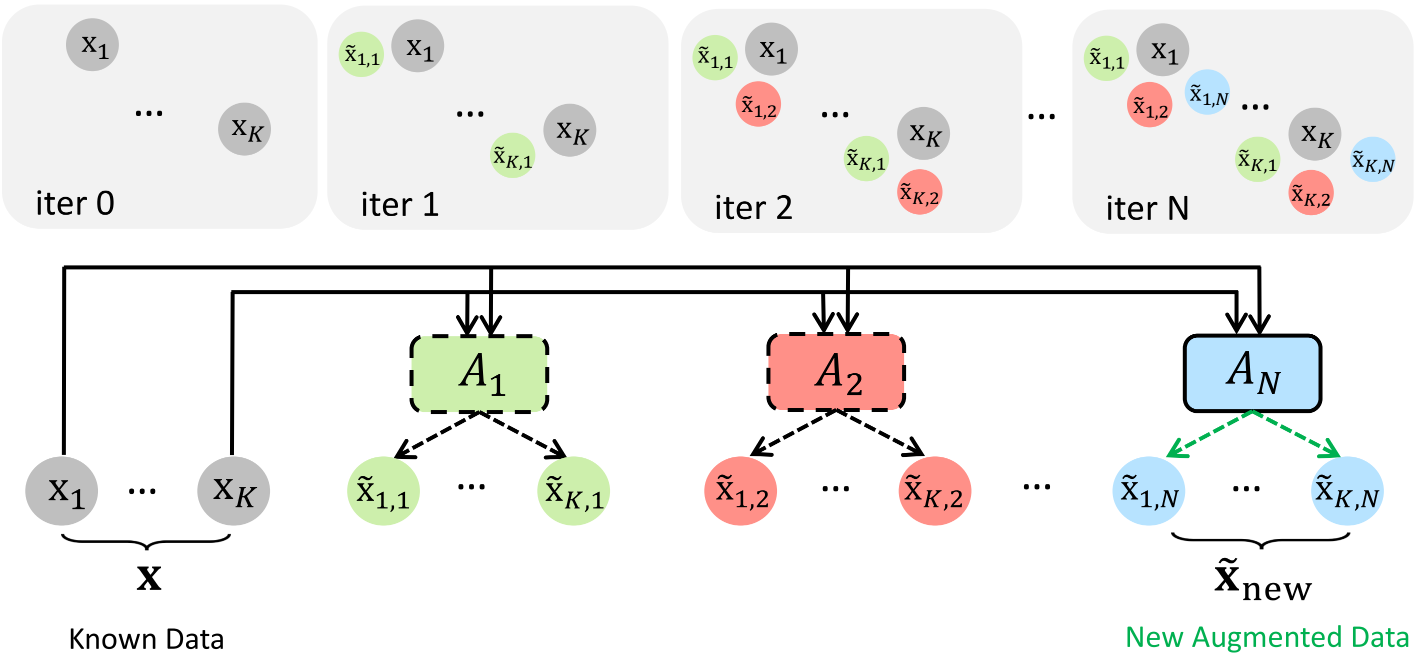
POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



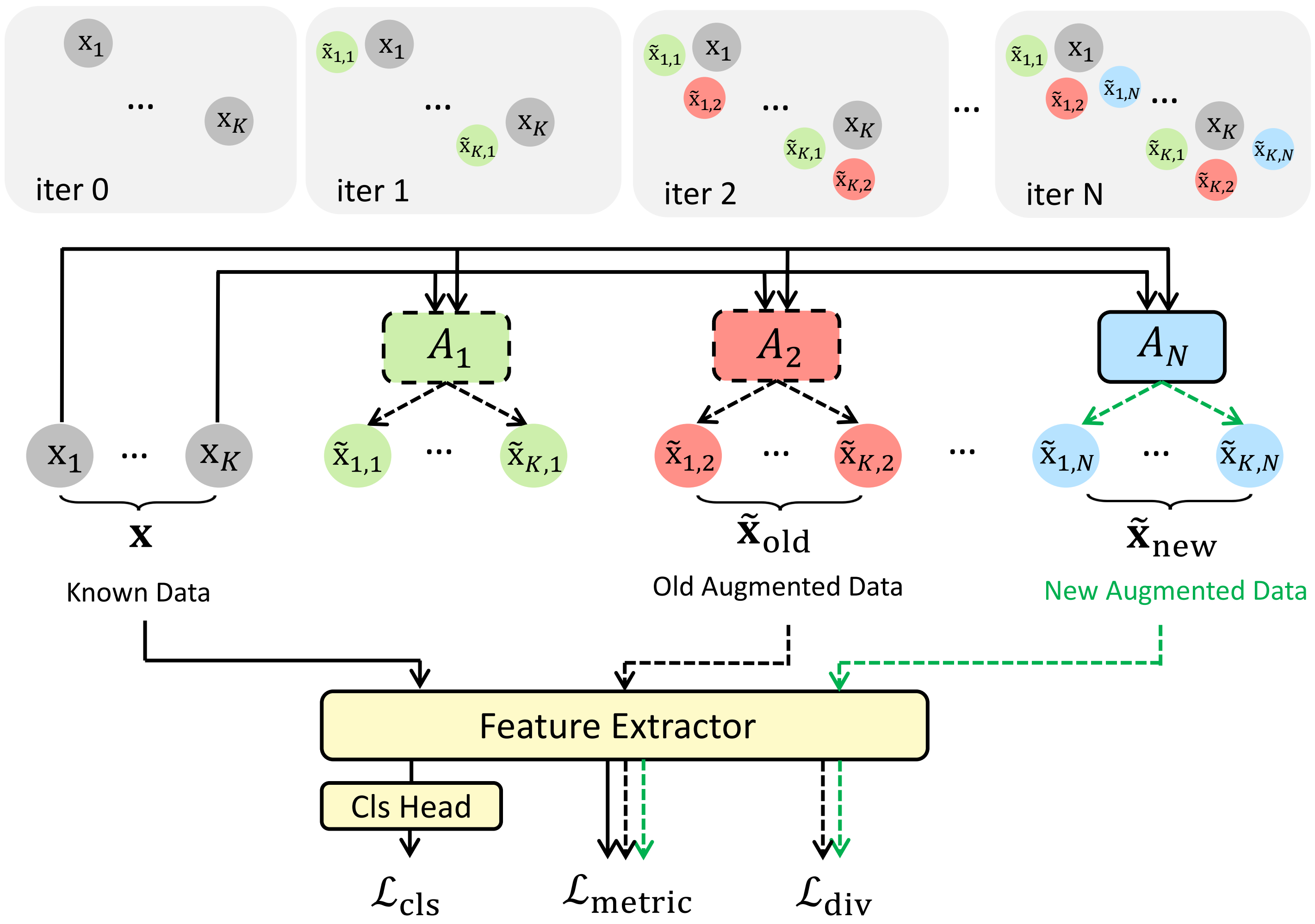
POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



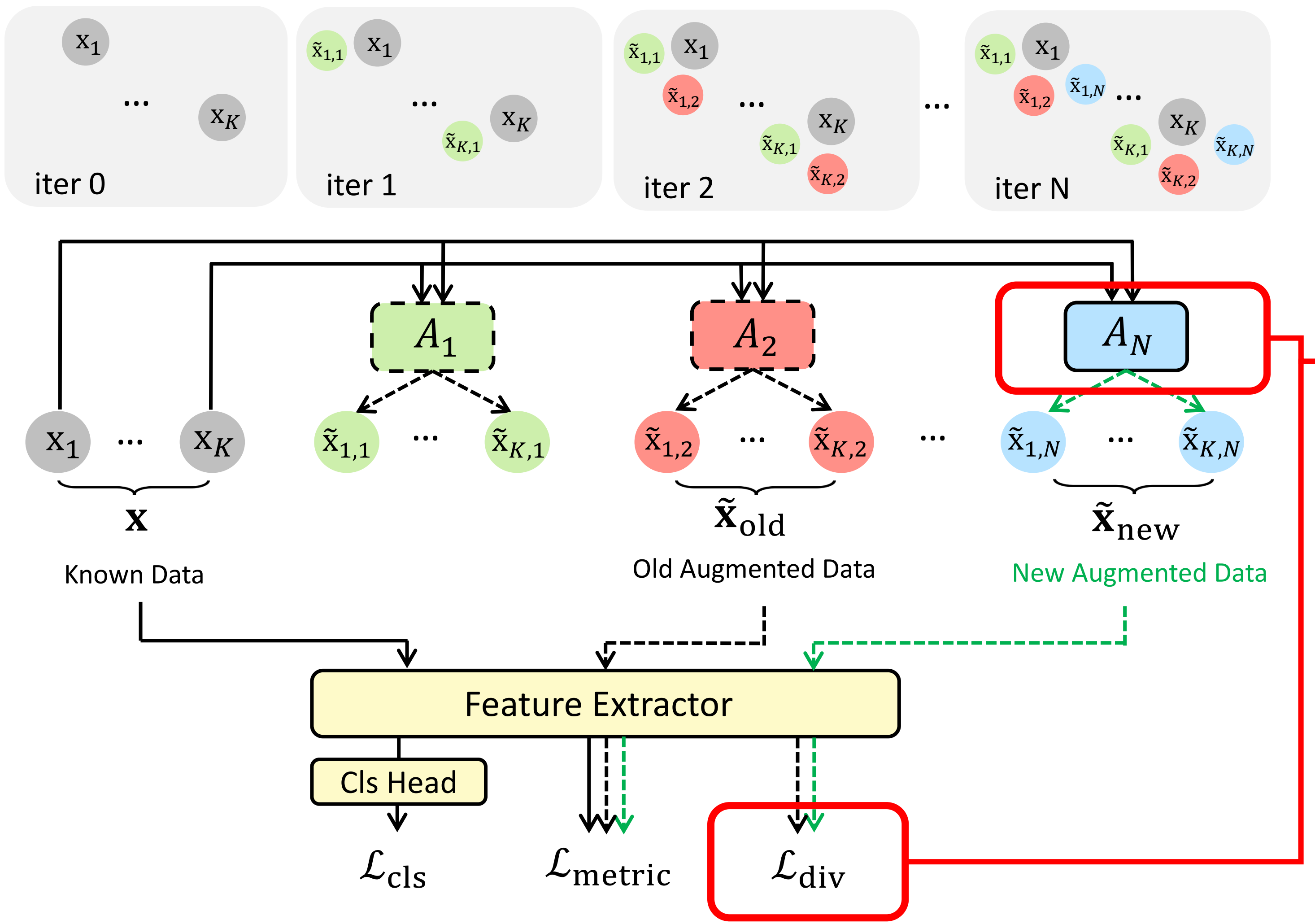
POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



Train augmentation model

- Objective
  - Expand the simulated open space progressively with diversity.
- Loss function

$$\mathcal{L}_{\text{aug}} = \mathcal{L}_{\text{recons}}(\mathbf{x}, \tilde{\mathbf{x}}_{\text{new}}) + \mathcal{L}_{\text{div}}(\tilde{\mathbf{x}}_{\text{old}}, \tilde{\mathbf{x}}_{\text{new}}),$$

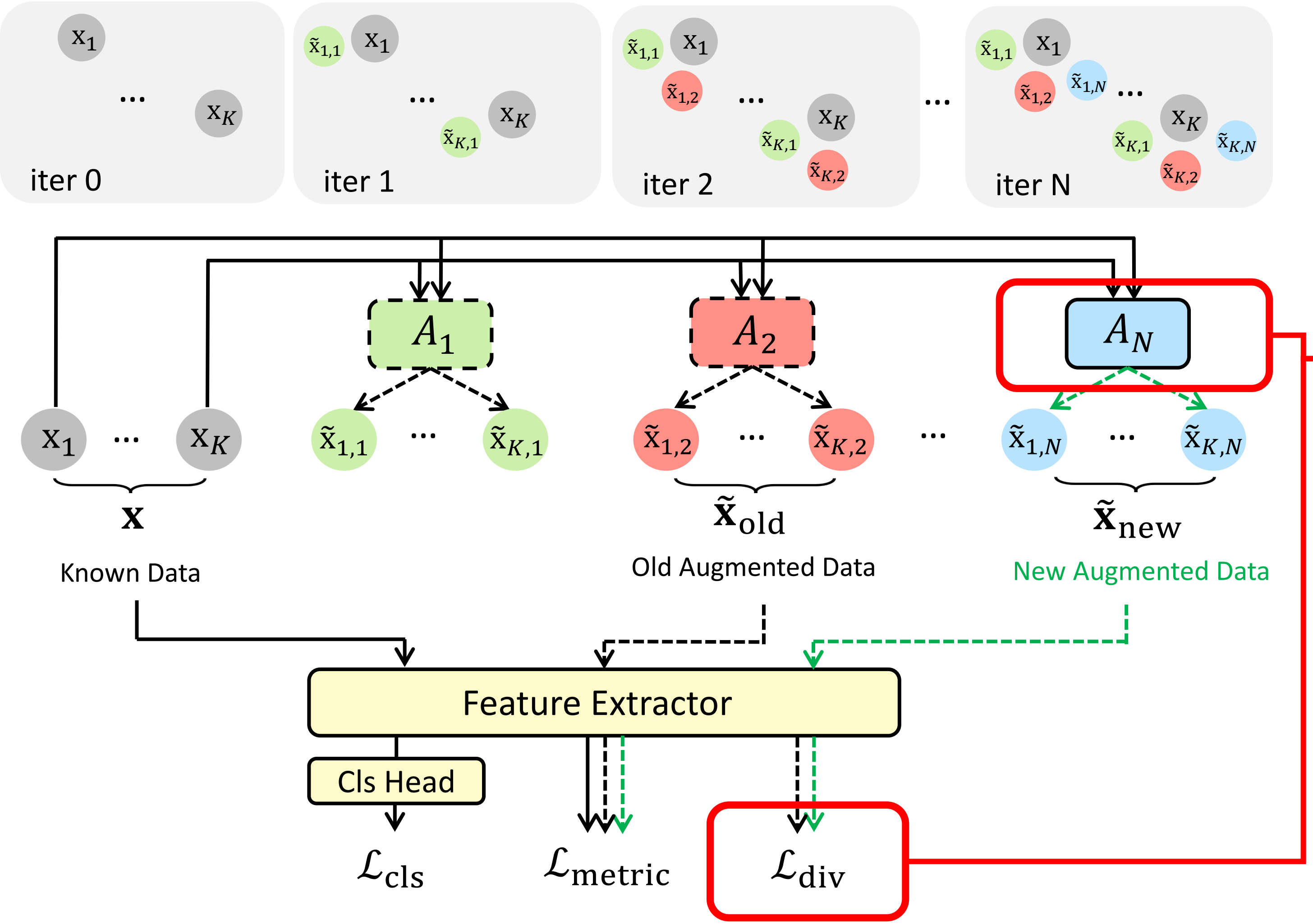
Known Data

New Augmented Data

Old Augmented Data

POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



Train augmentation model

➢ Objective

- Expand the simulated open space progressively with diversity.

➢ Loss function

$$\mathcal{L}_{\text{aug}} = \mathcal{L}_{\text{recons}}(\mathbf{x}, \tilde{\mathbf{x}}_{\text{new}}) + \mathcal{L}_{\text{div}}(\tilde{\mathbf{x}}_{\text{old}}, \tilde{\mathbf{x}}_{\text{new}}),$$

Known Data

New Augmented Data

Old Augmented Data

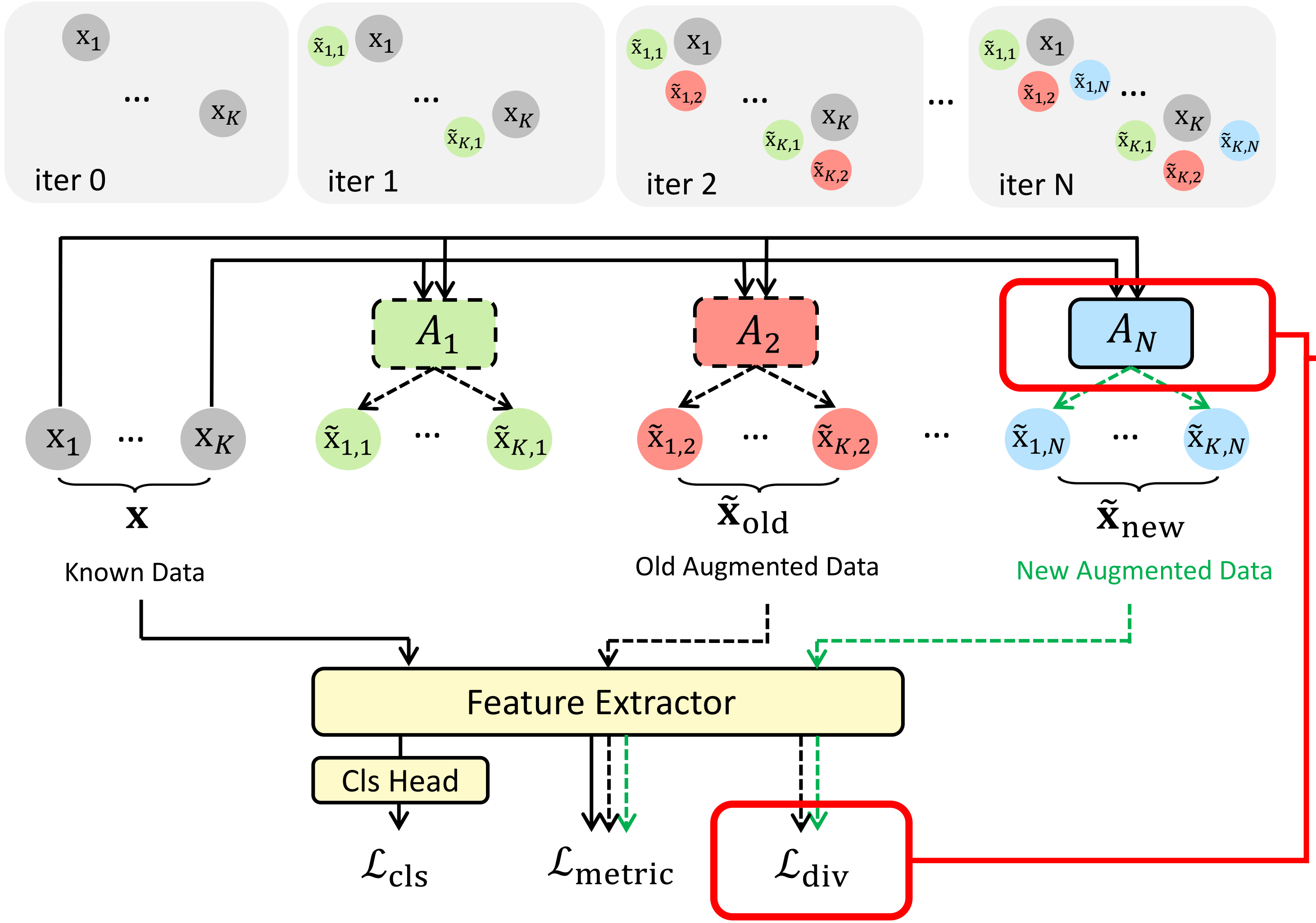
$$\mathcal{L}_{\text{div}} = \alpha F_{cos}(\tilde{\mathbf{z}}_{\text{new}}, \tilde{\mathbf{z}}_{\text{old}}) - \beta \min(F_{cos}(\tilde{\mathbf{z}}_{\text{new}}, \mathbf{z}), d),$$

Enlarge the embedding distance of new and old augmented data

Narrow the embedding distance of new augmented data and known data under a margin *d*

POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



Train augmentation model

- Objective
  - Expand the simulated open space progressively with diversity.
- Loss function

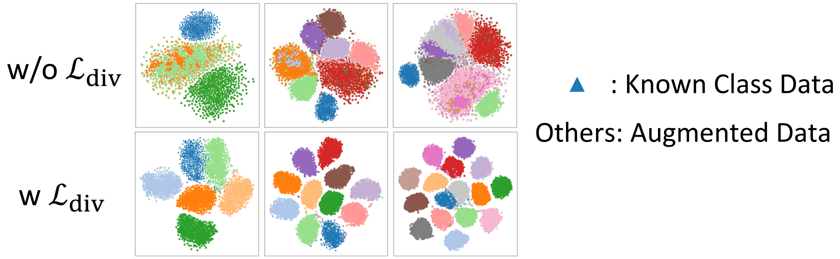
$$\mathcal{L}_{\mathrm{aug}} = \mathcal{L}_{\mathrm{recons}}(\mathbf{x}, \tilde{\mathbf{x}}_{\mathrm{new}}) + \mathcal{L}_{\mathrm{div}}(\tilde{\mathbf{x}}_{\mathrm{old}}, \tilde{\mathbf{x}}_{\mathrm{new}}),$$

Known Data

New Augmented Data

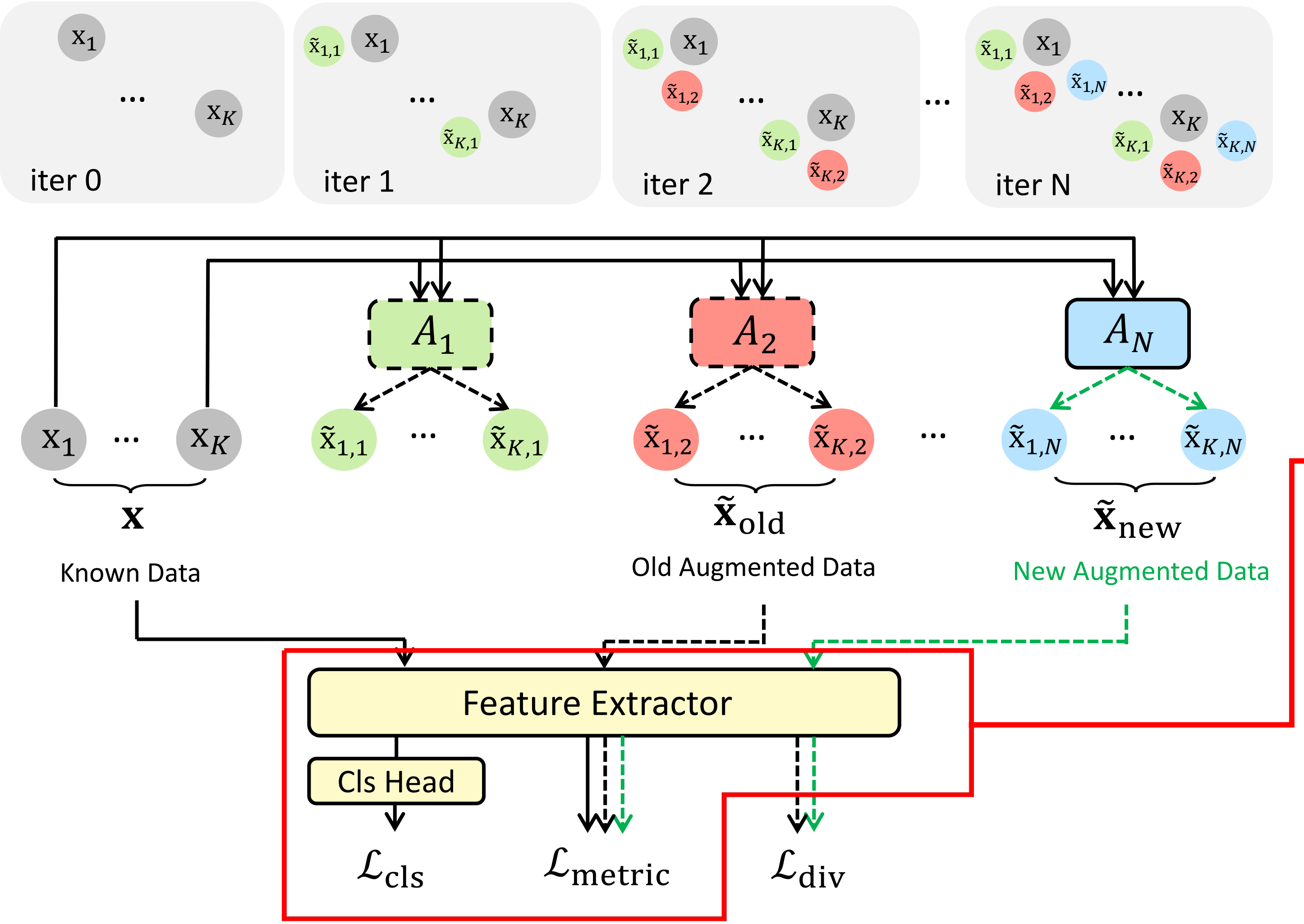
Old Augmented Data

$$\mathcal{L}_{\mathrm{div}} = \alpha F_{cos}(\tilde{\mathbf{z}}_{\mathrm{new}}, \tilde{\mathbf{z}}_{\mathrm{old}}) - \beta \min(F_{cos}(\tilde{\mathbf{z}}_{\mathrm{new}}, \mathbf{z}), d),$$



POSE (Progressive Open Space Expansion)

- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



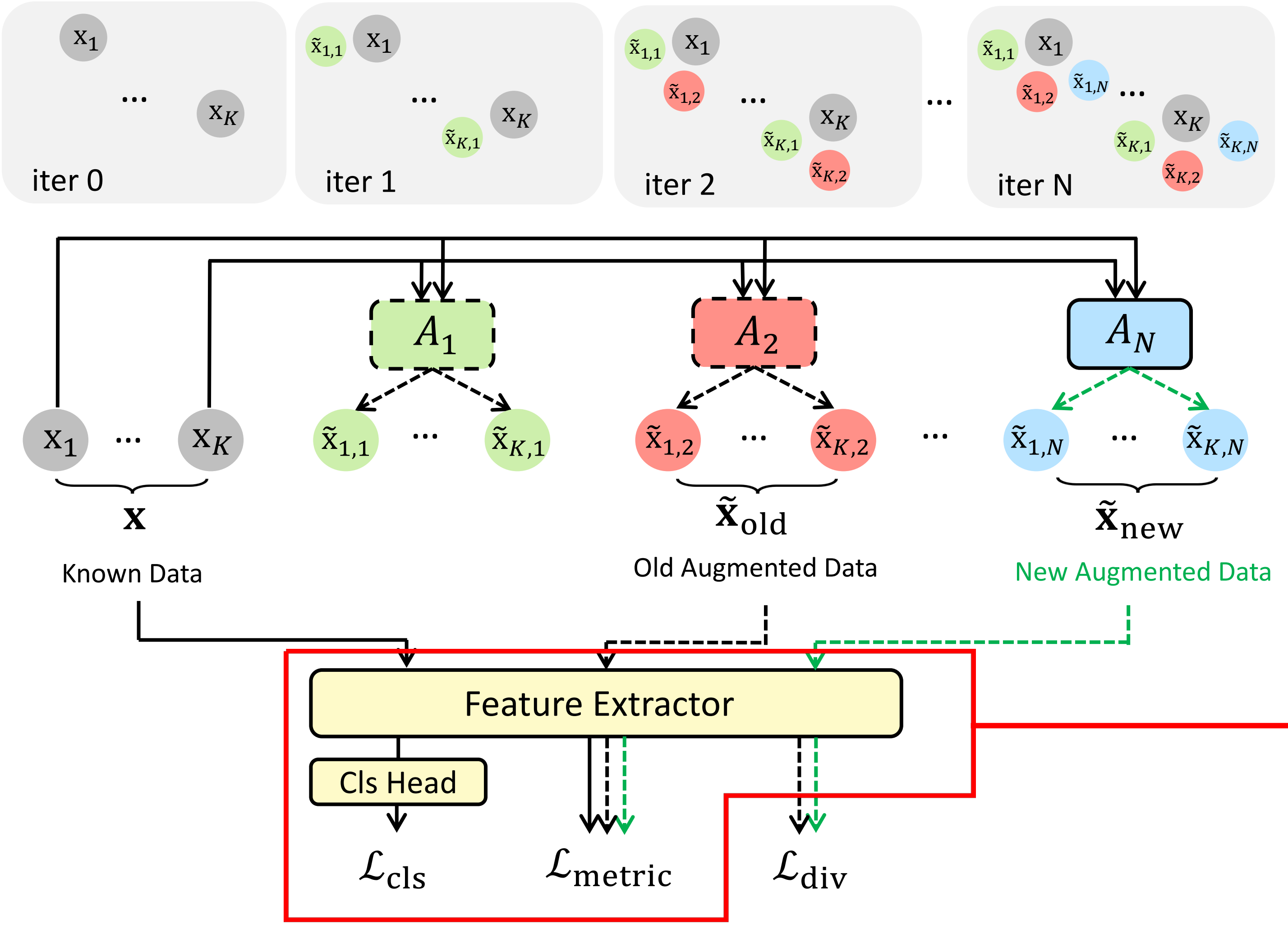
Train feature extractor and classification head

- Objective
  1. Known class classification.
  2. Distinguish augmented data from known data, and separate different known and augmented classes.
- Loss function

$$\mathcal{L}_{\text{task}} = \underbrace{\mathcal{L}_{\text{cls}}(\mathbf{x})}_{\text{Objective 1}} + \underbrace{\mathcal{L}_{\text{metric}}(\mathbf{x}, \tilde{\mathbf{x}}_{\text{old}}) + \mathcal{L}_{\text{metric}}(\mathbf{x}, \tilde{\mathbf{x}}_{\text{new}})}_{\text{Objective 2}}$$

POSE (Progressive Open Space Expansion)

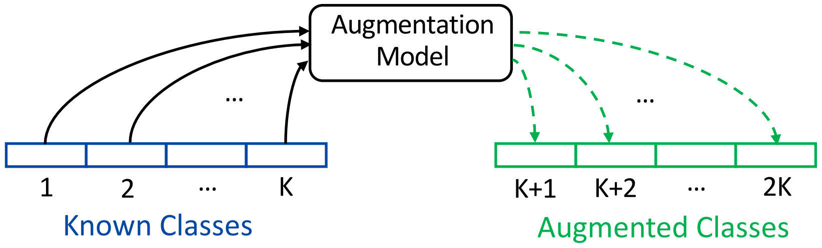
- **Key Idea:** Progressively simulate the potential open space of unknown models via a set of lightweight augmentation models.



Train feature extractor and classification head

- Objective
  1. Known class classification.
  2. Distinguish augmented data from known data, and separate different known and augmented classes.
- Loss function

$$\mathcal{L}_{\text{task}} = \underbrace{\mathcal{L}_{\text{cls}}(\mathbf{x})}_{\text{Objective 1}} + \underbrace{\mathcal{L}_{\text{metric}}(\mathbf{x}, \tilde{\mathbf{x}}_{\text{old}}) + \mathcal{L}_{\text{metric}}(\mathbf{x}, \tilde{\mathbf{x}}_{\text{new}})}_{\text{Objective 2}}$$



Dataset

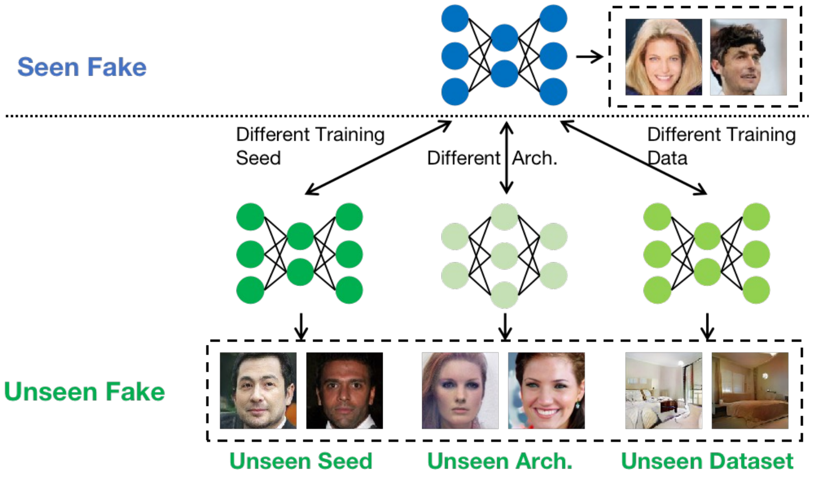
- Four groups of data: Seen Real, Seen Fake, Unseen Real, and three type of Unseen Fake
  - Unseen fake include Unseen Seed, Unseen Architecture, and Unseen Dataset

<table>
<tr><th colspan="2">Seen Real</th><th>CelebA</th><th>Face-HQ</th><th>ImageNet</th><th>Youtube</th><th>LSUN-Bedroom</th><th>LSUN-Cat</th><th>LSUN-Bus</th></tr>
<tr><td colspan="2">Seen Fake</td><td>StarGAN [10],
ProGAN_seed0 [22]</td><td>StyleGAN3-r [23],
StyleGAN3-t</td><td>SAGAN [56],
SNGAN</td><td>FSGAN [37],
FaceSwap [1]</td><td>ProGAN_seed0,
MMDGAN</td><td>StyleGAN,
StyleGAN3</td><td>ProGAN,
StyleGAN</td></tr>
<tr><td rowspan="3">Unseen Fake</td><td>Unseen Seed</td><td>ProGAN
(seed1,2,3,4,5)</td><td>-</td><td>-</td><td>-</td><td>ProGAN
(seed1,2,3,4,5)</td><td>-</td><td>-</td></tr>
<tr><td>Unseen Architec-ture</td><td>SNGAN [34],
AttGAN [19],
MMDGAN [3],
InfoMaxGAN [28]</td><td>StyleGAN2 [25],
ProGAN,
StyleGAN [24]</td><td>S3GAN [32],
BigGAN [4],
ContraGAN [21]</td><td>Wav2Lip [40],
FaceShifter [29]</td><td>SNGAN,
InfoMaxGAN</td><td>SNGAN,
ProGAN,
MMDGAN,
StyleGAN2</td><td>SNGAN,
MMDGAN,
StyleGAN2,
StyleGAN3</td></tr>
<tr><td>Unseen Dataset</td><td colspan="7">ProGAN, StyleGAN, StyleGAN3 (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)</td></tr>
<tr><td colspan="2">Unseen Real</td><td colspan="7">Coco, Summer</td></tr>
</table>

Dataset

- Four groups of data: Seen Real, Seen Fake, Unseen Real, and three type of Unseen Fake
  - Unseen fake include Unseen Seed, Unseen Architecture, and Unseen Dataset

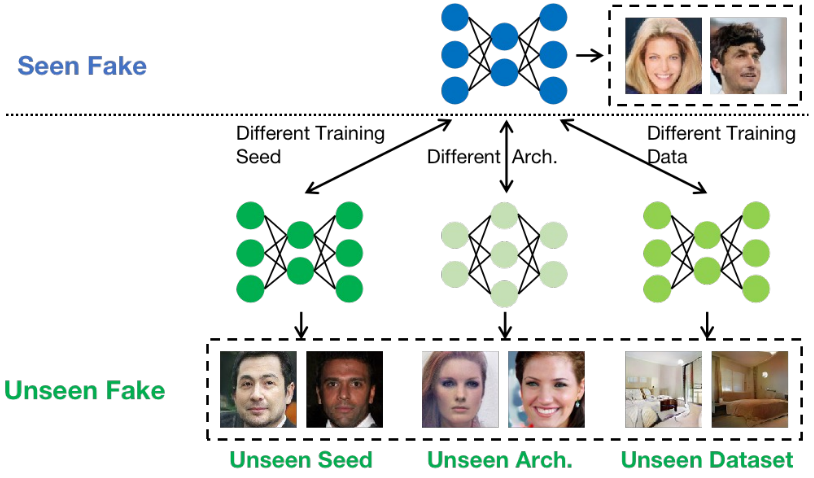
Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		StarGAN [10], ProGAN_seed0 [22]	StyleGAN3-r [23], StyleGAN3-t	SAGAN [56], SNGAN	FSGAN [37], FaceSwap [1]	ProGAN_seed0, MMDGAN	StyleGAN, StyleGAN3	ProGAN, StyleGAN
Unseen Fake	**Unseen Seed**	ProGAN (seed1,2,3,4,5)	-	-	-	ProGAN (seed1,2,3,4,5)	-	-
	Unseen Architec-ture	SNGAN [34], AttGAN [19], MMDGAN [3], InfoMaxGAN [28]	StyleGAN2 [25], ProGAN, StyleGAN [24]	S3GAN [32], BigGAN [4], ContraGAN [21]	Wav2Lip [40], FaceShifter [29]	SNGAN, InfoMaxGAN	SNGAN, ProGAN, MMDGAN, StyleGAN2	SNGAN, MMDGAN, StyleGAN2, StyleGAN3
	Unseen Dataset	ProGAN, StyleGAN, StyleGAN3 (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						

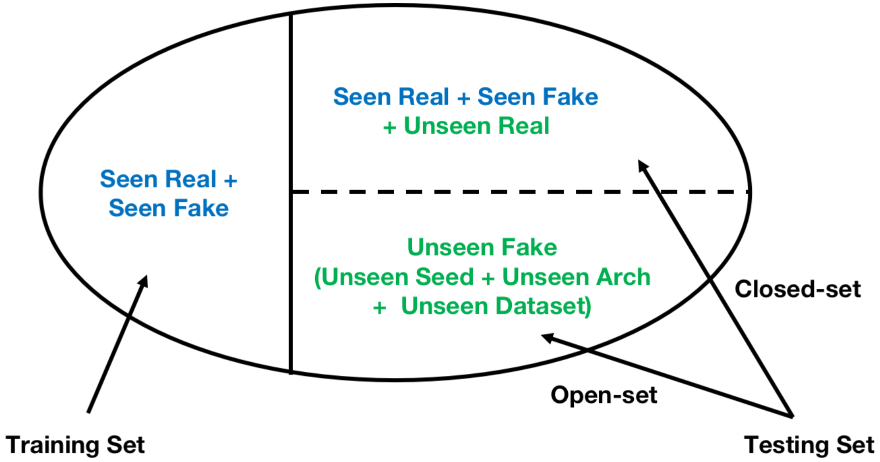


Dataset

- Four groups of data: Seen Real, Seen Fake, Unseen Real, and three type of Unseen Fake
  - Unseen fake include Unseen Seed, Unseen Architecture, and Unseen Dataset

Seen Real		CelebA	Face-HQ	ImageNet	Youtube	LSUN-Bedroom	LSUN-Cat	LSUN-Bus
Seen Fake		StarGAN [10], ProGAN_seed0 [22]	StyleGAN3-r [23], StyleGAN3-t	SAGAN [56], SNGAN	FSGAN [37], FaceSwap [1]	ProGAN_seed0, MMDGAN	StyleGAN, StyleGAN3	ProGAN, StyleGAN
Unseen Fake	**Unseen Seed**	ProGAN (seed1,2,3,4,5)	-	-	-	ProGAN (seed1,2,3,4,5)	-	-
	Unseen Architec-ture	SNGAN [34], AttGAN [19], MMDGAN [3], InfoMaxGAN [28]	StyleGAN2 [25], ProGAN, StyleGAN [24]	S3GAN [32], BigGAN [4], ContraGAN [21]	Wav2Lip [40], FaceShifter [29]	SNGAN, InfoMaxGAN	SNGAN, ProGAN, MMDGAN, StyleGAN2	SNGAN, MMDGAN, StyleGAN2, StyleGAN3
	Unseen Dataset	ProGAN, StyleGAN, StyleGAN3 (Cow, Sheep, Classroom, Bridge, Kitchen, Airplane, Church)						
Unseen Real		Coco, Summer						





Experimental Setup

- **Compared Methods**
  - GAN attribution: PRNU [1], Yu *et al*. [2], DCT CNN [3], DNA-Det [4], and RepMix [5]
  - GAN discovery: Girish *et al*. [6]
  - Open-set recognition: OpenMax [7], PROSER [8], ARPL+CS [9], and DIAS [10]

[1] Do gans leave artificial fingerprints? In *MIPR*, 2019

[2] Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019.

[3] Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020.

[4] Deepfake network architecture attribution. In *AAAI*, 2022.

[5] Repmix: Representation mixing for robust attribution of synthesized images. In *ECCV*, 2022.

[6] Towards discovery and attribution of open-world gan generated images. In *ICCV*, 2021

[7] Towards open set deep networks. In *CVPR*, 2016.

[8] Learning placeholders for open-set recognition. In *CVPR*, 2021.

[9] Adversarial reciprocal points learning for open set recognition. In *TPAMI, 2021*

[10] Difficulty-aware simulator for open set recognition. In *ECCV*, 2022

Experimental Setup

- **Compared Methods**
  - GAN attribution: PRNU [1], Yu *et al*. [2], DCT CNN [3], DNA-Det [4], and RepMix [5]
  - GAN discovery: Girish *et al*. [6]
  - Open-set recognition: OpenMax [7], PROSER [8], ARPL+CS [9], and DIAS [10]

- **Testing**
  - Test image → Feature extractor F → Classification head H → *Softmax* → Confidence scores
    - If the max confidence score is larger than a threshold → Known category of the index
    - Otherwise → Unknown

[1] Do gans leave artificial fingerprints? In *MIPR*, 2019
[2] Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019.
[3] Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020.
[4] Deepfake network architecture attribution. In *AAAI*, 2022.
[5] Repmix: Representation mixing for robust attribution of synthesized images. In *ECCV*, 2022.
[6] Towards discovery and attribution of open-world gan generated images. In *ICCV*, 2021
[7] Towards open set deep networks. In *CVPR*, 2016.
[8] Learning placeholders for open-set recognition. In *CVPR*, 2021.
[9] Adversarial reciprocal points learning for open set recognition. In *TPAMI, 2021*
[10] Difficulty-aware simulator for open set recognition. In *ECCV*, 2022

Experimental Setup

- **Compared Methods**
  - GAN attribution: PRNU [1], Yu *et al*. [2], DCT CNN [3], DNA-Det [4], and RepMix [5]
  - GAN discovery: Girish *et al*. [6]
  - Open-set recognition: OpenMax [7], PROSER [8], ARPL+CS [9], and DIAS [10]
- **Testing**
  - Test image → Feature extractor F → Classification head H → *Softmax* → Confidence scores
    - If the max confidence score is larger than a threshold → Known category of the index
    - Otherwise → Unknown
- **Evaluation**
  - Accuracy: closed-set classification
  - AUC: closed/open discrimination
  - OSCR: trade-off between the two aspects

[1] Do gans leave artificial fingerprints? In *MIPR*, 2019
[2] Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019.
[3] Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020.
[4] Deepfake network architecture attribution. In *AAAI*, 2022.
[5] Repmix: Representation mixing for robust attribution of synthesized images. In *ECCV*, 2022.
[6] Towards discovery and attribution of open-world gan generated images. In *ICCV*, 2021
[7] Towards open set deep networks. In *CVPR*, 2016.
[8] Learning placeholders for open-set recognition. In *CVPR*, 2021.
[9] Adversarial reciprocal points learning for open set recognition. In *TPAMI, 2021*
[10] Difficulty-aware simulator for open set recognition. In *ECCV*, 2022

Experimental Result

- **Compare with GAN attribution methods**

POSE outperforms existing fake image attribution methods in terms of closed-set classification and closed/open discrimination.

Method	Closed-Set	Unseen Seed		Unseen Architecture		Unseen Dataset		Unseen All	
	Accuracy	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR
PRNU [33]	55.27	**69.20**	49.16	70.02	49.49	67.68	48.57	68.94	49.06
Yu *et al.* [53]	85.71	53.14	50.99	69.04	64.17	<u>78.79</u>	72.20	69.90	64.86
DCT-CNN [14]	86.16	55.46	52.68	72.56	67.43	72.87	67.57	69.46	64.70
DNA-Det [50]	93.56	61.46	<u>59.34</u>	<u>80.93</u>	<u>76.45</u>	66.14	63.27	71.40	68.00
RepMix [5]	<u>93.69</u>	54.70	53.26	72.86	70.49	78.69	<u>76.02</u>	<u>71.74</u>	<u>69.43</u>
POSE	**94.81**	<u>68.15</u>	**67.25**	**84.17**	**81.62**	**88.24**	**85.64**	**82.76**	**80.50**

- **Compare with OSR methods**

The simulated open space by POSE is more suitable for OSMA than off-the-shelf OSR methods.

Method	Closed-Set	Unseen Seed		Unseen Architecture		Unseen Dataset		Unseen All	
	Accuracy	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR
Base	90.68	62.02	60.58	76.03	72.92	77.01	73.88	73.78	70.97
Base+OpenMax [2]	91.11	63.27	61.60	76.40	73.29	75.33	72.32	73.50	70.70
Base+PROSER [58]	92.12	63.32	62.19	79.55	76.57	81.43	78.64	77.22	74.66
Base+ARPL+CS [7]	91.77	54.94	54.17	79.09	75.97	80.48	77.52	75.08	72.47
Base+DIAS [35]	92.77	62.15	61.02	79.34	76.49	84.14	81.13	78.00	75.41
Base+AM	<u>93.41</u>	<u>66.17</u>	<u>65.04</u>	<u>82.21</u>	<u>79.42</u>	<u>85.04</u>	<u>82.20</u>	<u>80.31</u>	<u>77.80</u>
Base+AM+$\mathcal{L}_{div}$(**POSE**)	**94.81**	**68.15**	**67.25**	**84.17**	**81.62**	**88.24**	**85.64**	**82.76**	**80.50**

- **Compare with GAN discovery method**

POSE is better in unknown model clustering.

Method	Avg. Purity	NMI	ARI
Girish *et al.* [16] (k=49)	32.89	61.89	21.05
POSE (k=49)	39.16	**61.91**	**27.48**
POSE (k=68)	**41.04**	60.59	26.39

- *k* = 68: the true number of classes for seen and unseen data
- *k* = 49: the number of clusters that Girish *et al.* returns after four iterations.

Ablation Study

- **The diversity loss** increase the diversity of open space simulated by different augmentation models, and reduces the open space risk better.

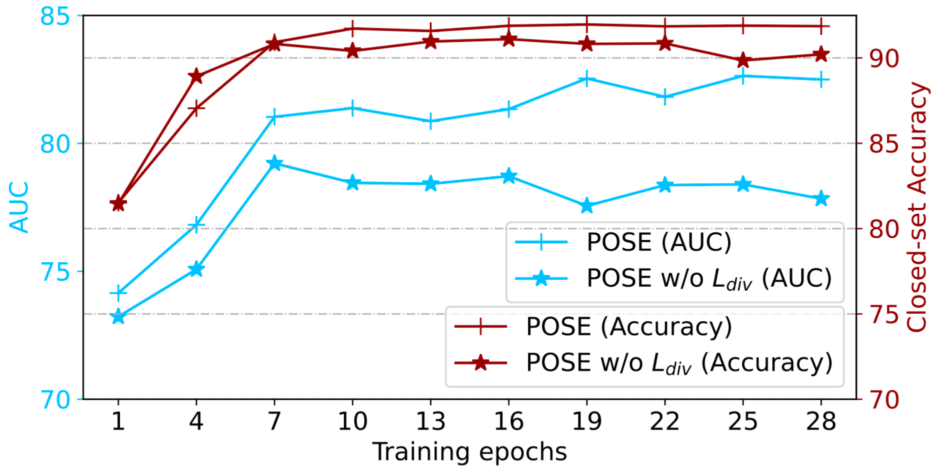
	Closed-set	Open-set	
	Acc	AUC	OSCR
Base	90.68	73.78	70.97
Base+AM	93.41	80.31	77.80
Base+AM+Ldiv	94.81	82.76	80.50

Ablation Study

- **The diversity loss** increase the diversity of open space simulated by different augmentation models, and reduces the open space risk better.

	Closed-set	Open-set	
	Acc	AUC	OSCR
Base	90.68	73.78	70.97
Base+AM	93.41	80.31	77.80
Base+AM+Ldiv	94.81	82.76	80.50

With L_{div}, the AUC increases continually until about 19 epochs.

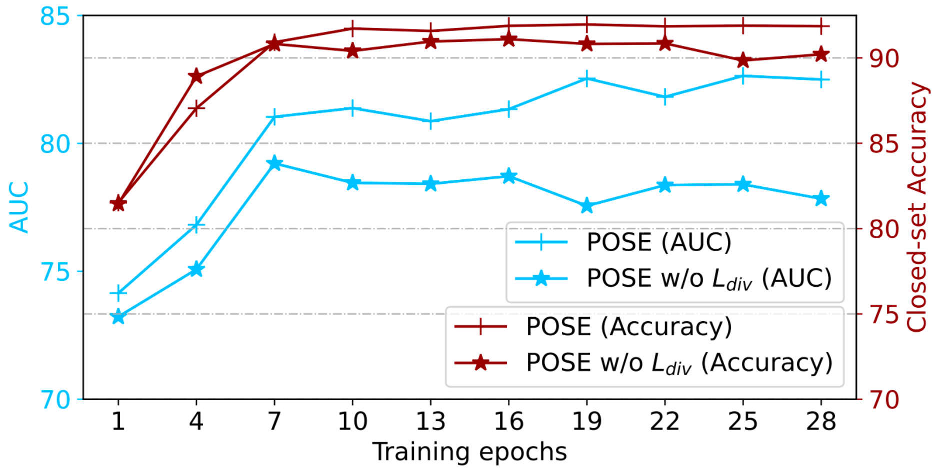


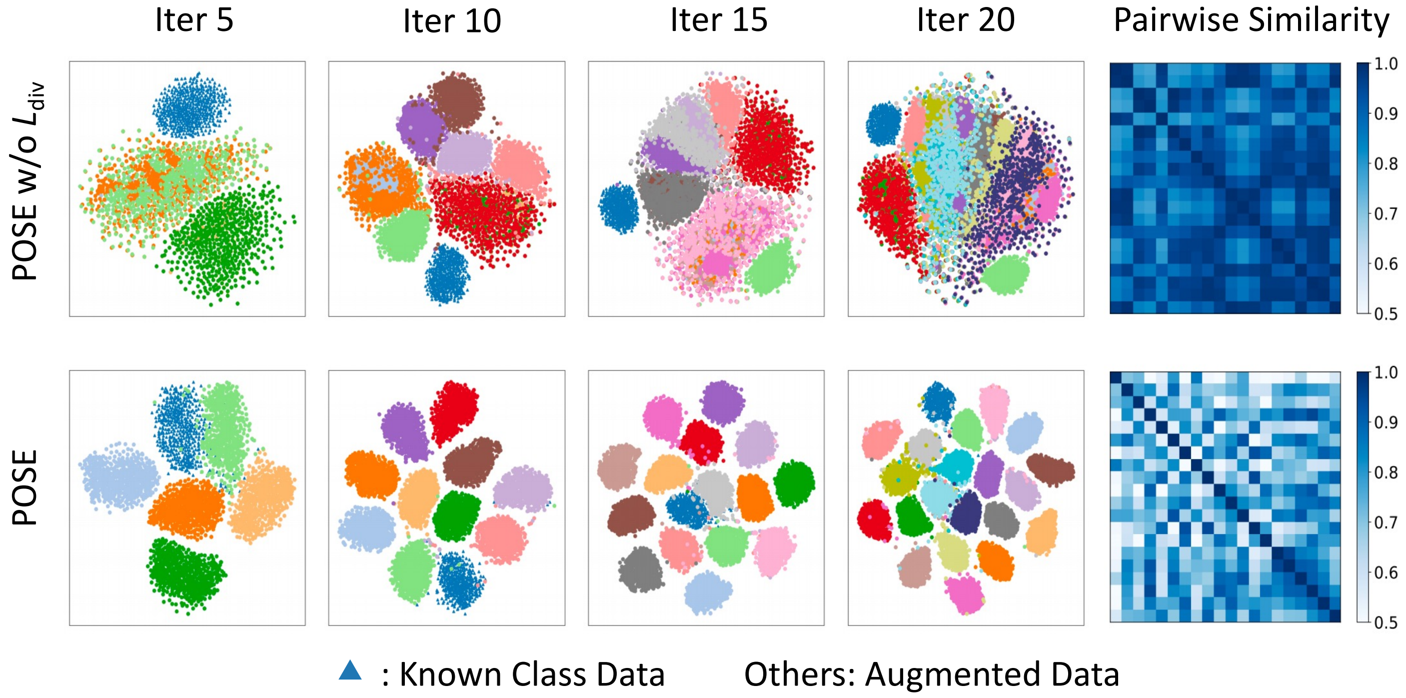
Ablation Study

- **The diversity loss** increase the diversity of open space simulated by different augmentation models, and reduces the open space risk better.

	Closed-set	Open-set	
	Acc	AUC	OSCR
Base	90.68	73.78	70.97
Base+AM	93.41	80.31	77.80
Base+AM+Ldiv	94.81	82.76	80.50

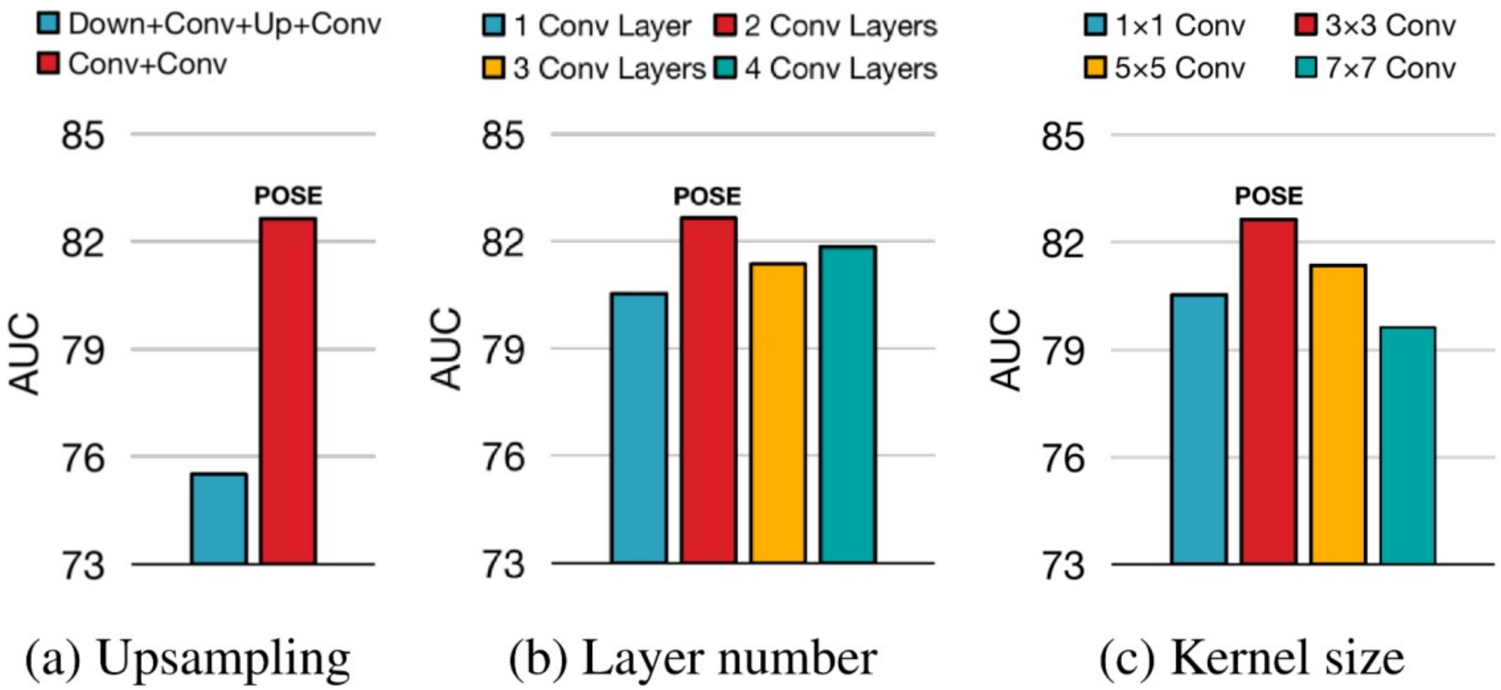
With L_{div}, the AUC increases continually until about 19 epochs.





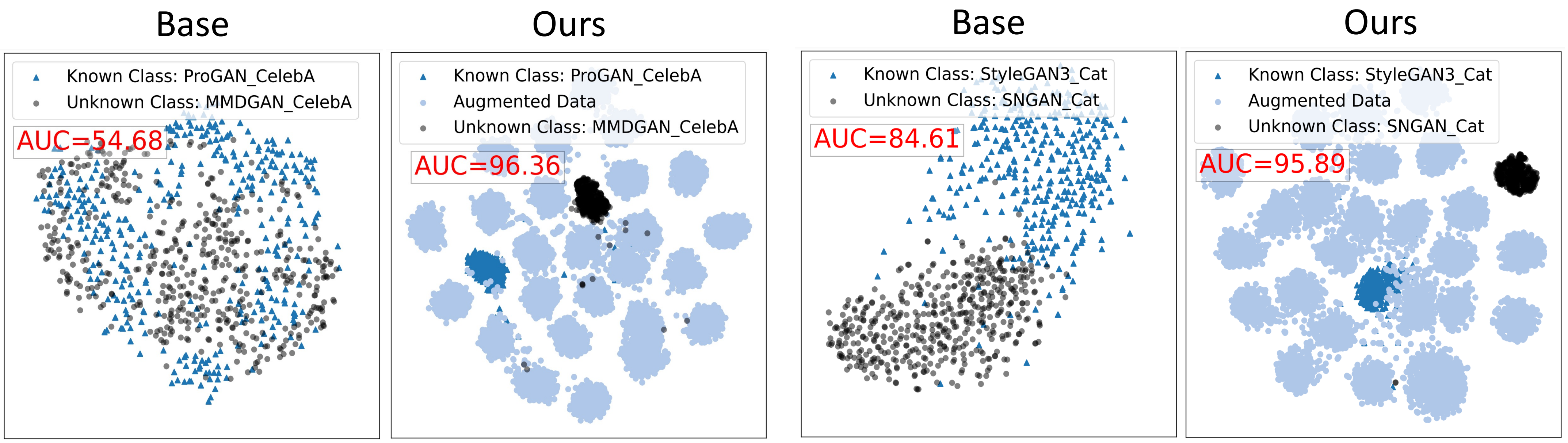
Ablation Study

- **Ablation study on the architecture of augmentation models.**
  - Best option: only convolution layer, Layer number = 2, Kernel size = 3



Visualization Examples

- The augmented data simulates a rich open space enclosing the known data points, resulting in a clear better close/open discrimination.



▲ Known Class Data ● Augmented Data for the Known Class ● Unknown Class Data（Easiest to be confused）

Summary

- Highlights
  - **Problem:** A new task named open-set model attribution.
  - **Method:** Simulate the potential open space progressively via lightweight augmentation models.
  - **Dataset:** A dataset considering Seen Real, Seen Fake, Unseen Real, and three types of Unseen Fake.
  - **Evaluation:** Superior than model attribution methods and off-the-shelf OSR methods.
  - Code, dataset, and models are at https://github.com/ICTMCG/POSE
- Future Work
  - Unified framework for architecture-level and model-level attribution.
  - Model retrieval, model lineage analysis.

Thanks

Feel free to contact :

yangtianyun19z@ict.ac.cn
wangdanding@ict.ac.cn