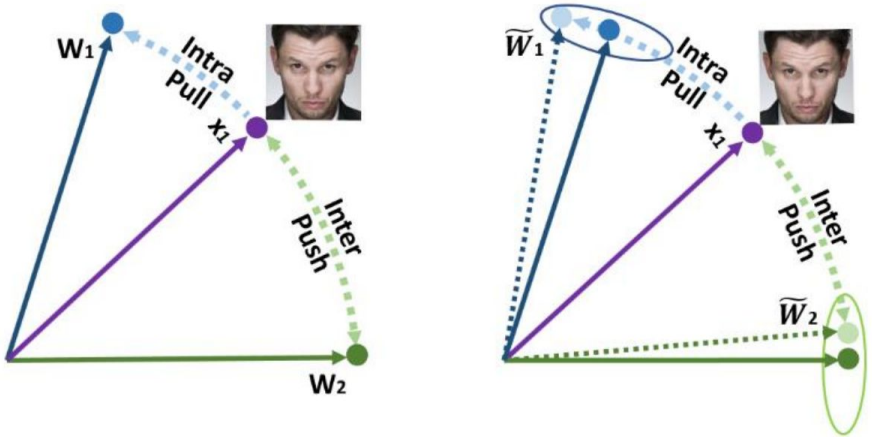
Review

- Variational Prototype Learning for Deep Face Recognition, CVPR 2021

  : Propose a novel **Variational Prototype Learning** method which represents each class as a <u>distribution</u> instead of a <u>point</u> by using the margin-based softmax loss.



(a) Prototype Learning (b) Variational Prototype Learning

CS588 Paper Presentation 1

Towards Universal Fake Image Detectors that Generalize Across Generative Models

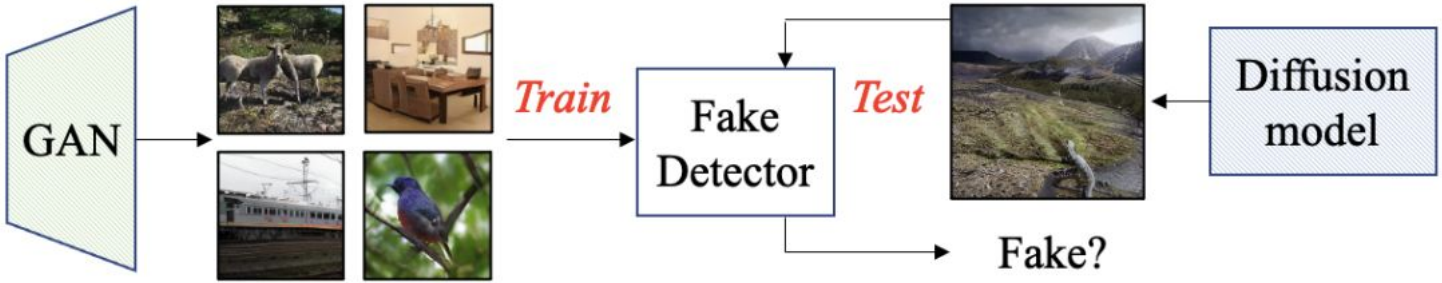
CVPR 2023

Jumin Lee

2024. 04. 29.

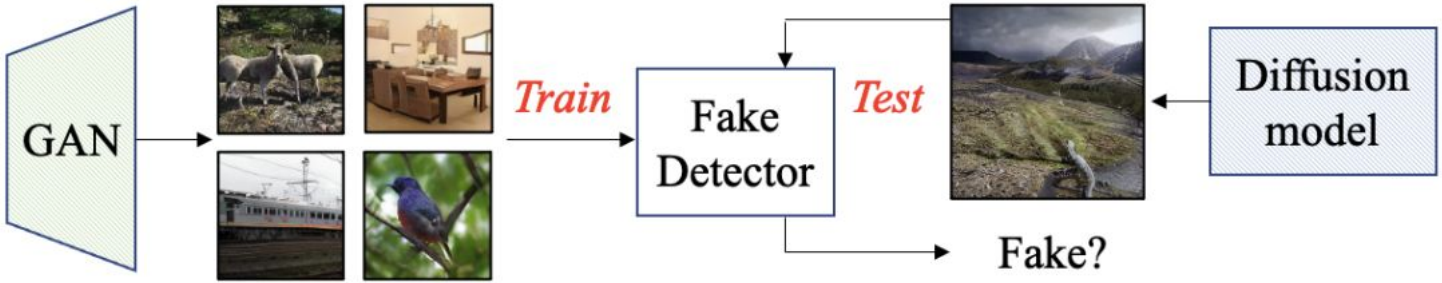
SGVR Lab
KAIST

Goal



- **Fake Image Detection with Generalizability**
  - Train only on real/fake images associated with GAN
  - Achieve high performance on <u>unseen Generative model images</u>

Motivation

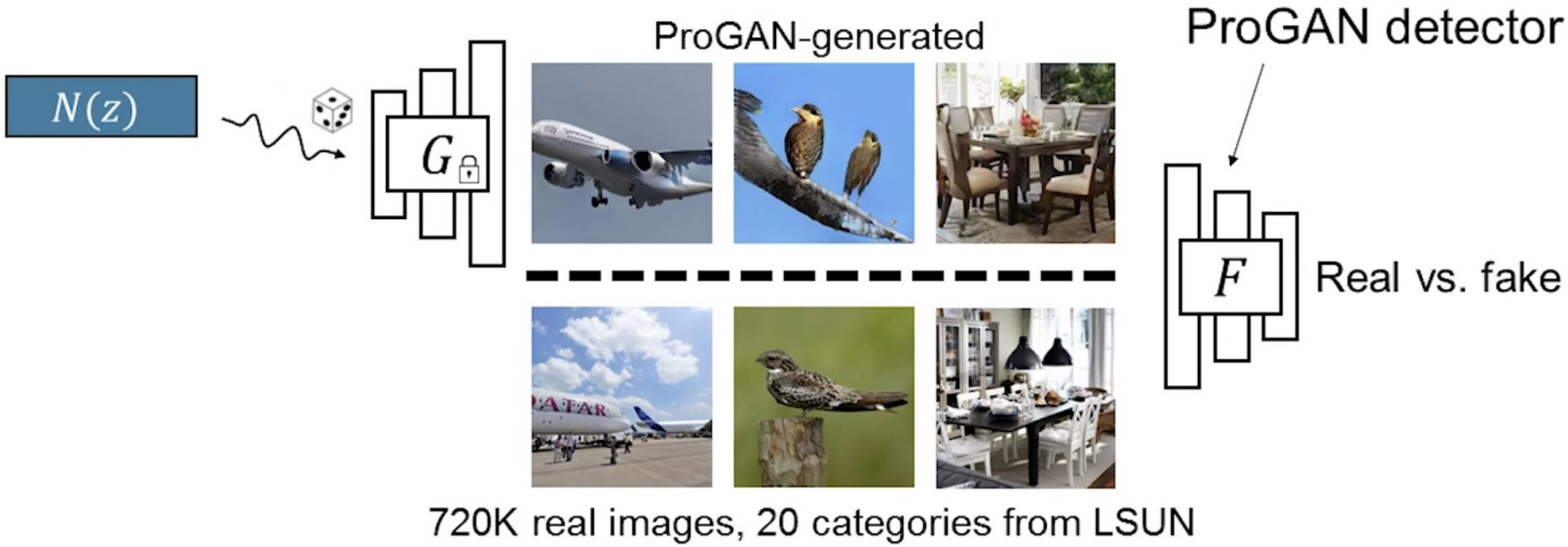


- Generative models are spreading quickly,

  and there are **growing concerns** about using generated images to cause harm.
- However, the **existing method**(real-vs-fake classifier) to distinguish between real and fake images **doesn't work with new generative models.**

Motivation

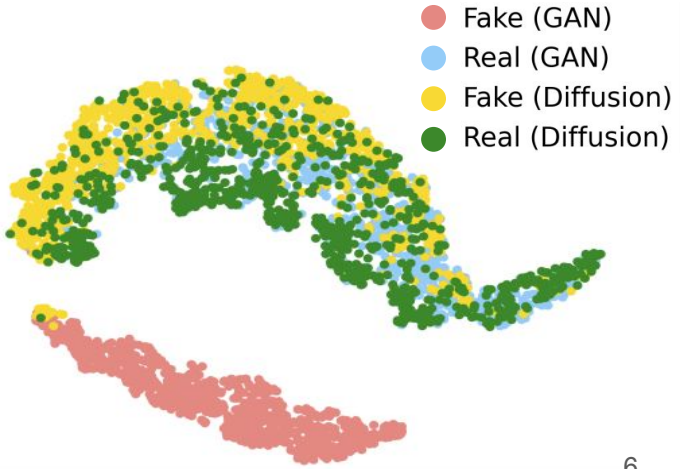
- Training method of existing model[1]



Discovery

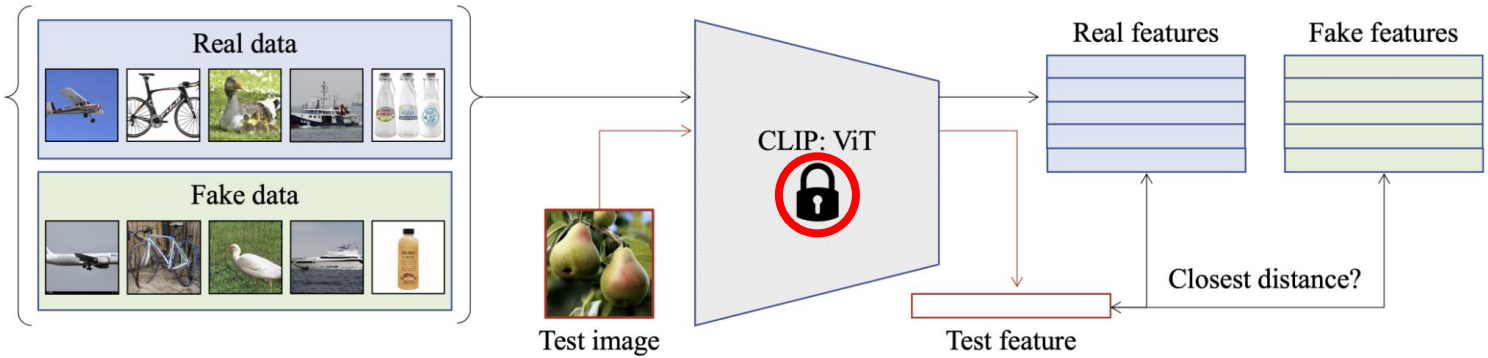
- Why does this happen?
  - Real-vs-fake classifiers learn to identify fake images by using the **fingerprint** of the model, rather than learning all the ways an image could be real.

	CycleGAN	GauGAN	LDM	Guided	DALL-E
Real acc.	98.64	99.4	99.61	99.14	99.61
Fake acc.	62.91	59.1	3.05	4.67	4.9
Average	80.77	79.25	51.33	51.9	52.26
Chance performance	50.00	50.00	50.00	50.00	50.00



t-SNE visualization

Proposed Method



- No training of real vs. fake classifiers

  : The classification process should happen in a **feature space** which has not been trained to separate images from the two classes.

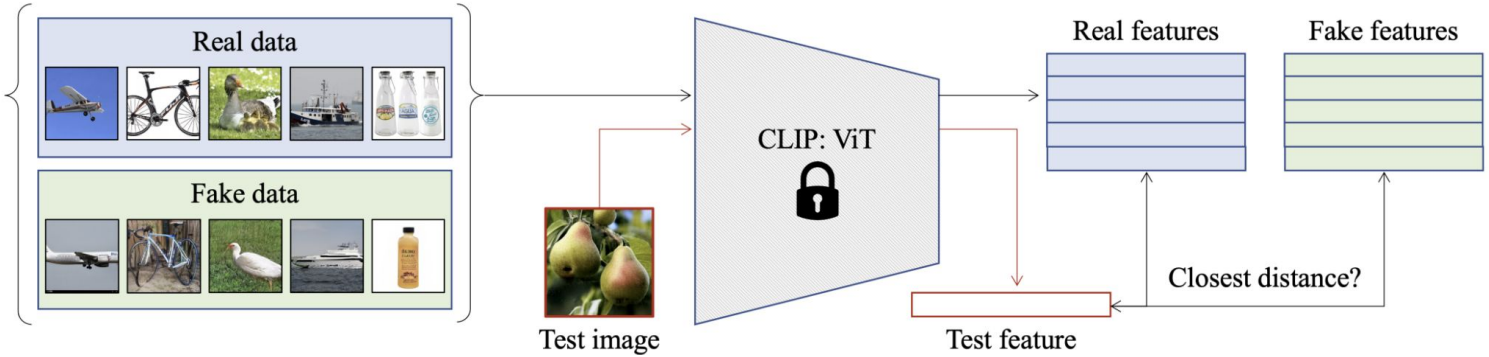
Contribution

- Analyze the limitations of existing methods in detecting fake images from unseen breeds of generative models.
- Present two very simple method(nearest neighbor, linear classification) which **utilize a feature space that is entirely untrained for real/fake classification.**
- Show state-of-the-art generalization performance.

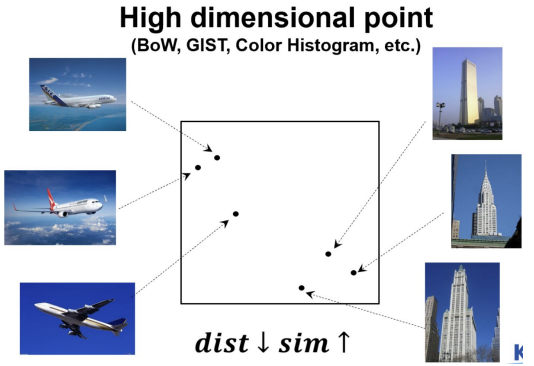
Method

Similar to Image Retrieval

- Extract feature and measure distance to find similar.

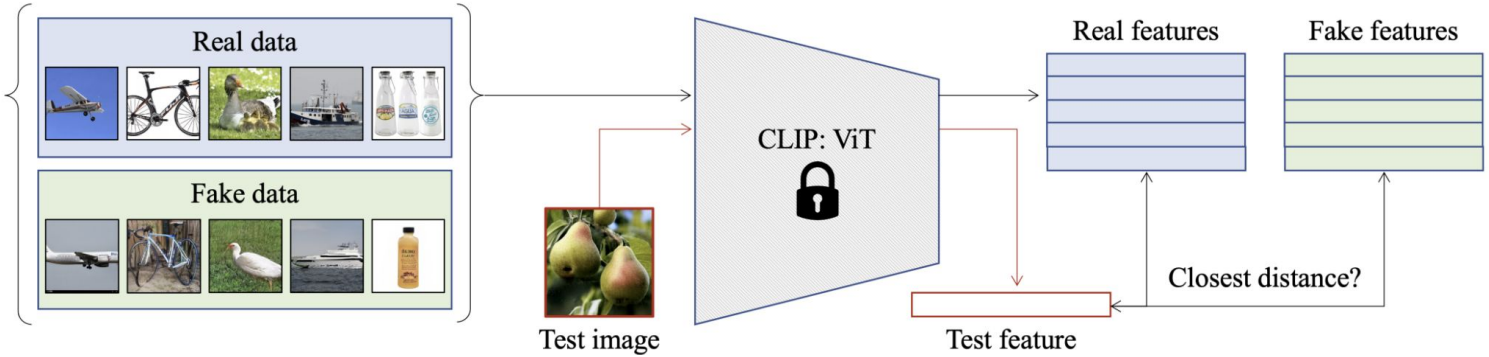


- **Fake image detection** identify feature that signify whether an image is generated.
- **Image retrieval** identify and match features across a database to retrieve relevant images.

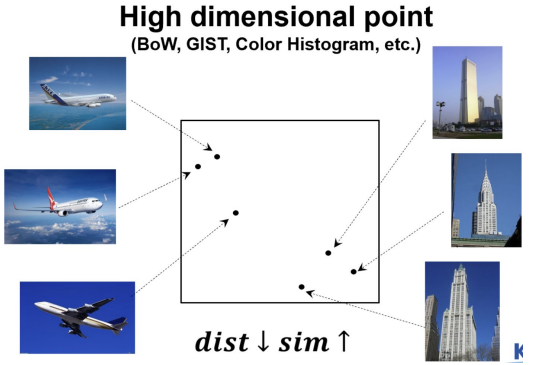


Similar to Image Retrieval

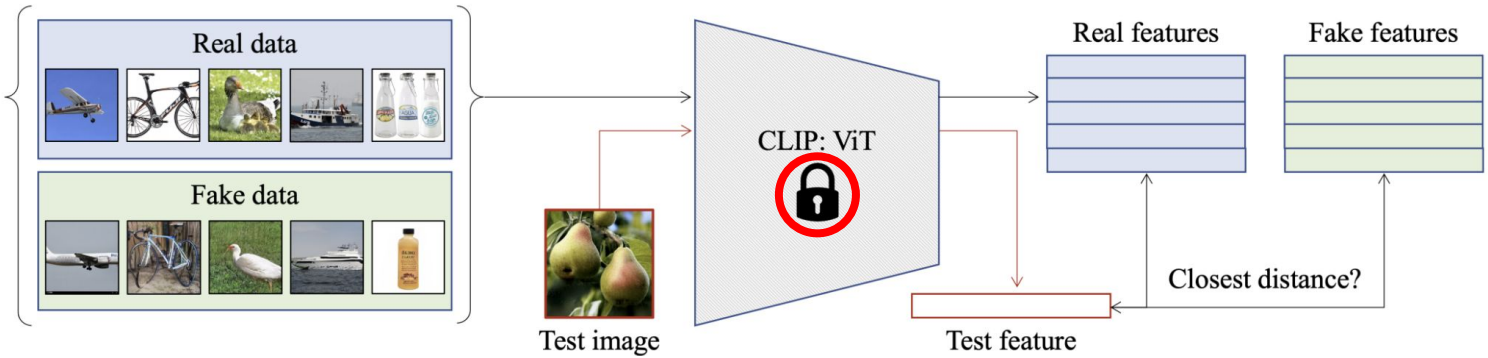
- Generalization



- **Fake image detection** should generalize well across generative models that were not trained on it.
- **Image retrieval** should perform across image sets with a different conditions.



Method



- Choice of feature extractor = CLIP:ViT visual encoder

- Exposed to a **large number** of images.

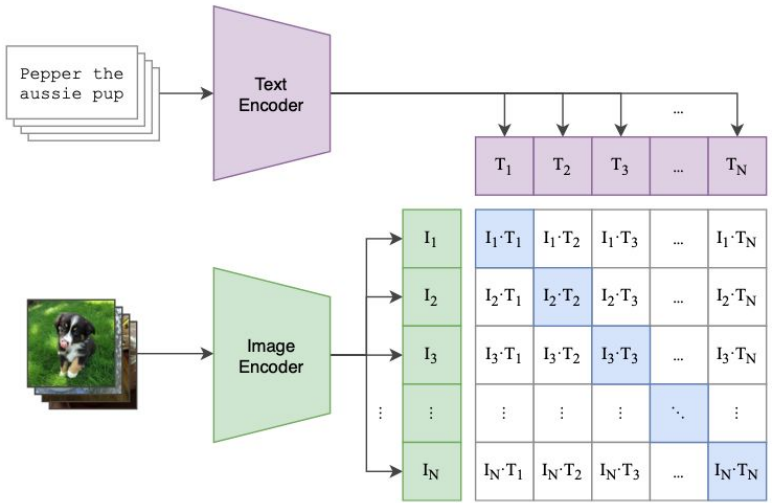
  : Consistent for a wide variety of real/fake images for **generalizability**.

- Capture **low-level details** of an image.

  : **Differences** between real and fake images arise at low-level details.

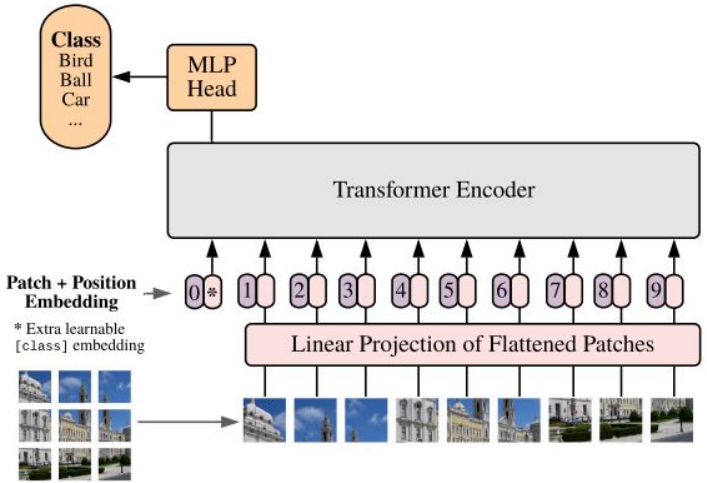
CLIP:ViT

- Motivation
  - Language models have made significant progress with **large-scale models.**
  - Similar advancements are **anticipated in vision models with the large models.**
- Method
  - Assemble a dataset of 400 million image-text pairs from the Internet.
  - Implement contrastive learning.

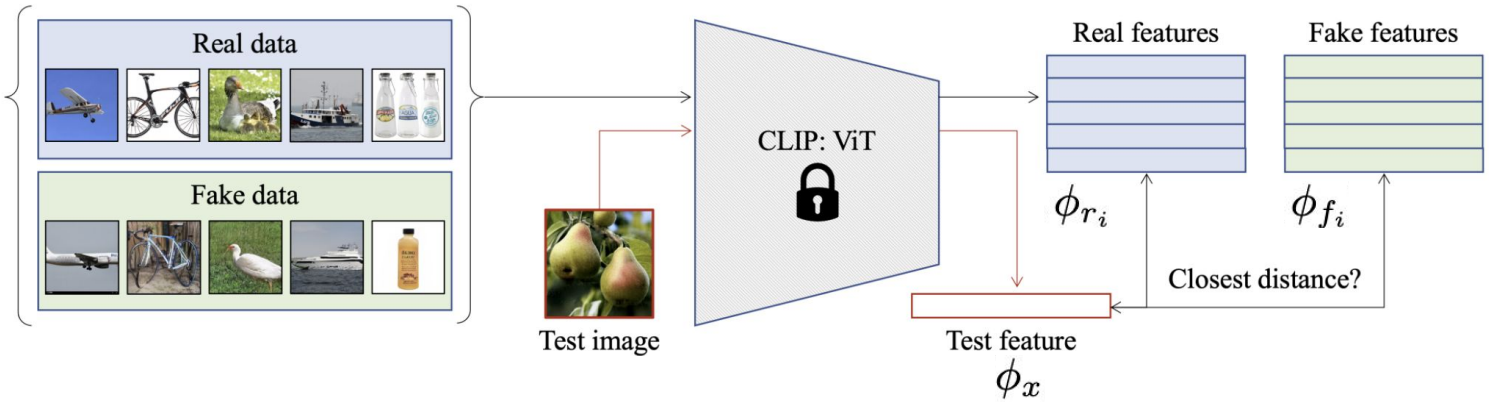


CLIP:**ViT**

- Employs a **transformer architecture** over patches of the image.
  - An image is split into fixed-size patches,
  - each of them are then linearly embedded, position embeddings are added,
  - and the resulting sequence of vectors is fed to a standard Transformer encoder.



Method #1. Nearest Neighbors

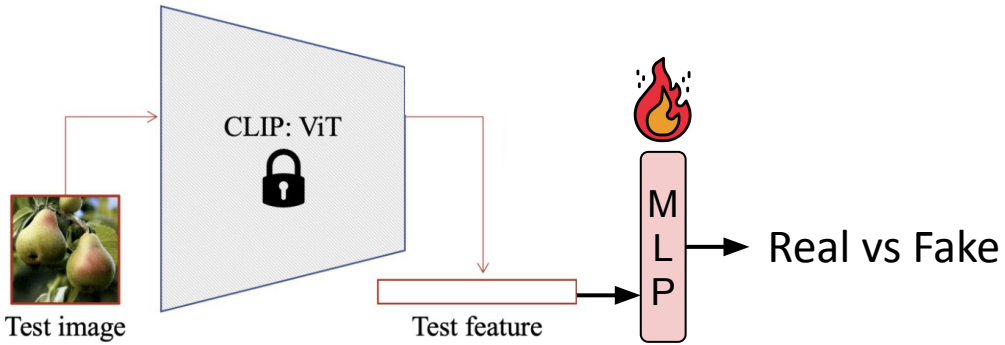


- Using cosine distance as the metric d,

  find the nearest neighbors in both the real and fake feature banks.

$$\text{pred}(x) = \begin{cases} 1, & \text{if } \min_i \left(d(\phi_x, \phi_{f_i})\right) < \min_i \left(d(\phi_x, \phi_{r_i})\right) \quad \text{: Fake} \\ 0, & \text{otherwise.} \quad \text{: Real} \end{cases}$$

Method #2. Linear Classification



- **Add a single linear layer and train only this new classification layer.**
- Since only training a few hundred parameters, perform similarly to the nearest neighbor.
- Has the advantage of being computationally and memory friendly.

Results

Evaluation Metrics

- **Average precision (AP)**
  - Measures the area under the Precision-Recall curve, which plots precision and recall at various threshold levels.
  - How sensitively the model detects fake images.

- **Classification Accuracy**
  - Accuracy = # of correct predictions / Total # of prediction
  - Indicate the overall error rate.

Generalization Results

- Average precision (AP)

Variant	Generative Adversarial Networks						Guided	LDM			Glide			DALL-E	Total
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		mAP
Blur+JPEG (0.1)	**100.0**	93.47	84.5	**99.54**	89.49	98.15	73.72	70.62	71.0	70.54	80.65	84.91	82.07	70.59	83.58
Blur+JPEG (0.5)	**100.0**	96.83	88.24	98.29	98.09	95.44	68.57	66.0	66.68	65.39	73.29	78.02	76.23	65.93	81.52
ViT:CLIP (B+J 0.5)	99.98	93.32	83.63	88.14	92.81	84.62	55.74	52.52	54.51	52.2	56.64	61.13	56.64	62.74	73.44
NN, $k = 1$	**100.0**	98.14	94.49	86.68	99.26	99.53	79.31	95.84	79.84	95.97	93.98	95.17	**96.05**	88.51	90.32
NN, $k = 3$	**100.0**	98.13	94.46	86.67	99.25	99.53	79.26	95.81	79.78	95.94	93.94	95.13	94.60	88.47	90.22
NN, $k = 5$	**100.0**	98.13	94.46	86.66	99.25	99.53	79.25	95.81	79.78	95.94	93.94	95.13	94.60	88.46	90.22
NN, $k = 9$	**100.0**	98.13	94.46	86.66	99.25	99.53	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	90.14
LC	**100.0**	99.46	**99.59**	97.24	**99.98**	99.60	**87.77**	**99.14**	**92.15**	**99.17**	**94.74**	**95.34**	94.57	**97.15**	**93.38**

- **Existing method** distinguishes with good accuracy for other GAN variants. However, the accuracy drops drastically from unseen generative models.
- **Propose method** show a drastically better generalization performance.

Generalization Results

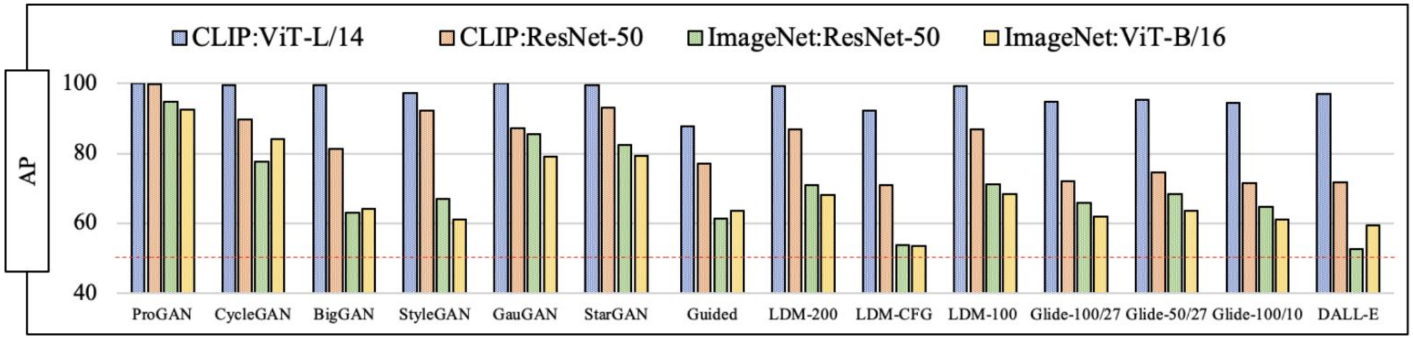
- Classification Accuracy

Variant	Generative Adversarial Networks						Guided	LDM			Glide			DALL-E	Total
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		Avg. acc
Blur+JPEG (0.1)	99.99	85.20	70.20	85.7	78.95	91.7	60.07	54.03	54.96	54.14	60.78	63.8	65.66	55.58	69.58
Blur+JPEG (0.5)	**100.0**	80.77	58.98	69.24	79.25	80.94	51.90	51.33	51.93	51.28	54.43	55.97	54.36	52.26	64.73
ViT:CLIP (B+J 0.5)	98.94	78.80	60.62	60.56	66.82	62.31	50.66	50.74	51.04	50.76	52.15	53.07	52.06	53.18	60.57
NN, $k = 1$	99.58	94.70	86.95	80.24	96.67	98.84	68.76	89.56	68.99	89.51	86.44	88.02	87.27	77.52	82.30
NN, $k = 3$	99.58	95.04	87.63	80.55	96.94	98.77	70.02	90.37	70.17	90.57	87.84	89.34	88.78	79.29	83.28
NN, $k = 5$	99.60	94.32	88.23	80.60	97.00	98.90	70.55	90.89	70.97	91.01	88.42	90.07	89.60	80.19	83.72
NN, $k = 9$	99.54	93.49	88.63	80.75	97.11	98.97	**71.06**	91.29	72.02	91.29	**89.05**	**90.67**	**90.08**	81.47	**84.25**
LC	**100.0**	98.50	**94.50**	82.00	**99.50**	97.00	70.03	**94.19**	73.76	**94.36**	79.07	79.85	78.14	**86.78**	81.38

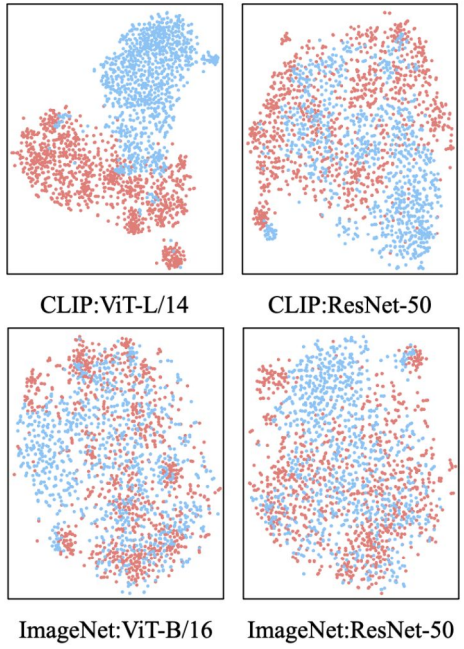
- Results clearly demonstrate the advantage of using the **feature space of a frozen**, pre-trained network that is blind to the downstream real/fake classification task.

Ablation Study about Feature Extractor

- **Networks trained on CLIP tasks** are better able to distinguish between real and fake images, compared to **networks trained on imagenet classification,** even when using the same model architecture.



<Ablation on the network architecture and pre-training dataset>



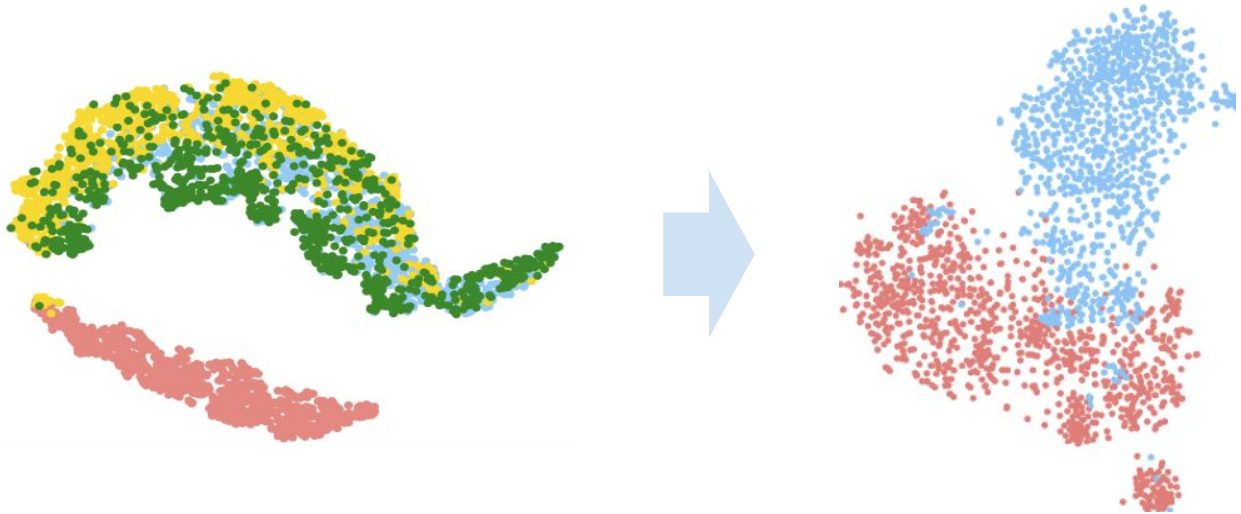
Conclusion

Conclusion

- **Analyze the** limitations **of existing methods** in generalizability of detecting fake images.
- Performing nearest neighbor / linear probing in informative **feature space** not trained for real-vs-fake classification results in a significantly better **generalization** ability of detecting fake images.
- Show state-of-the-art performance.

Limitation

- **The question remains** about the similarity of images generated with different kinds of generative models.



< t-SNE visualization >