



# Review for Workshop & Tutorial

**YeongHyeon Park**

Dept. of ECE, SungKyunKwan University

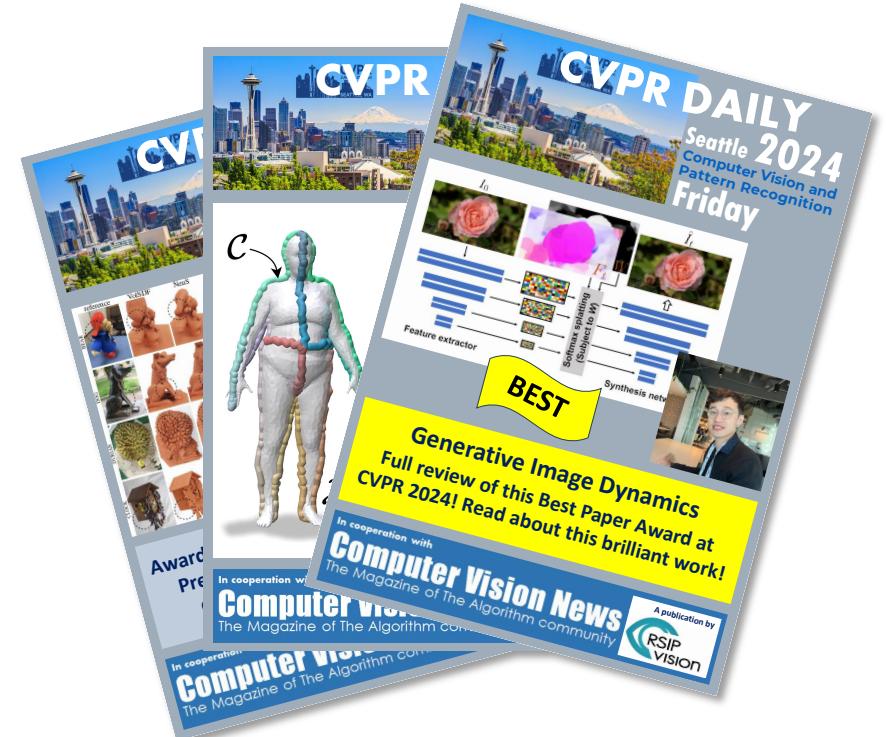
# Short Review

## Good

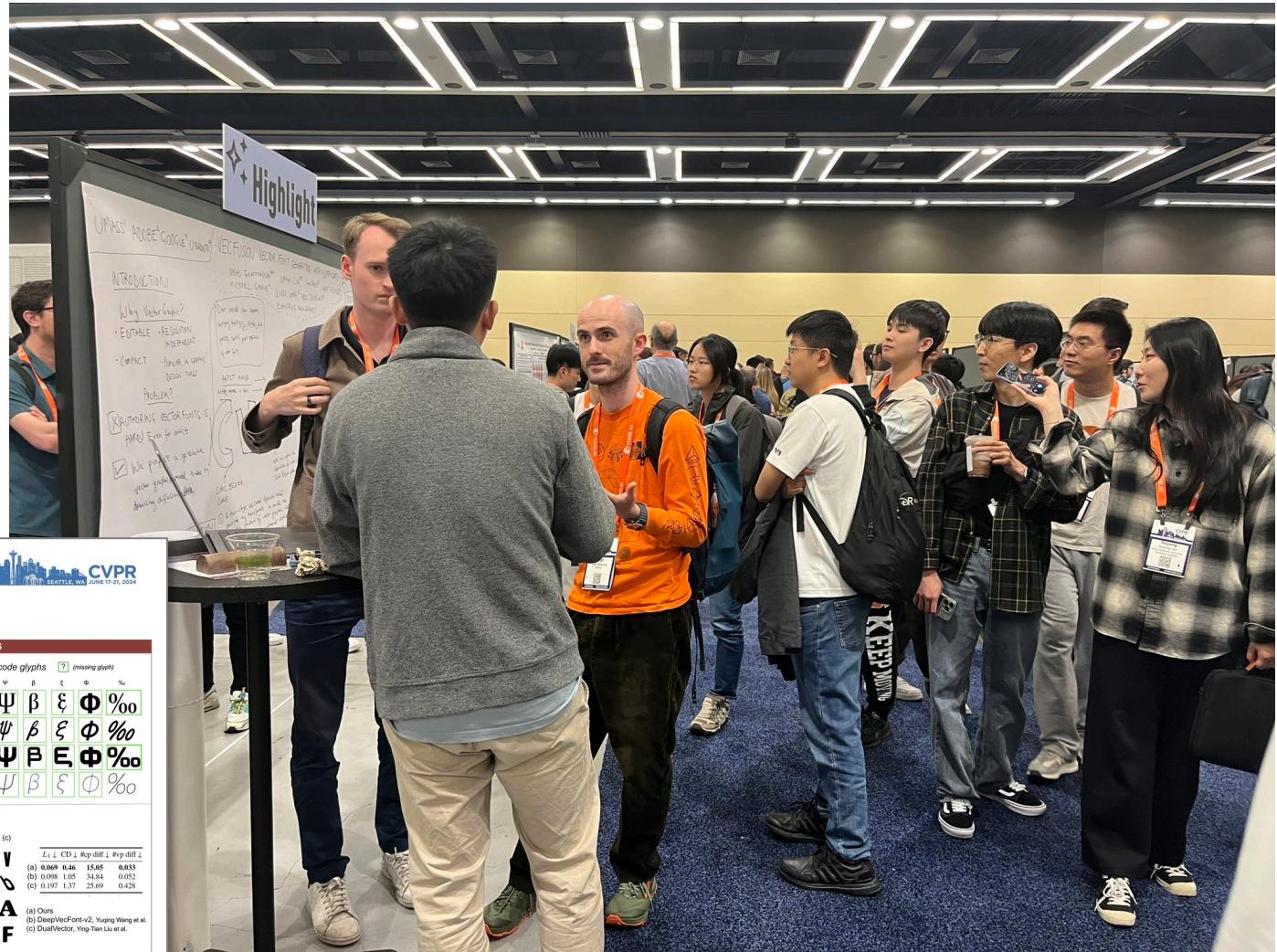
- Tutorial/workshop planning worked well
- Reliable streaming service
  - Well-prepared presentation
  - But only for available virtual streaming
- Interesting new topics
  - Focusing on research topics rather than NN structure
  - Makes me think about research in other fields than AD
- Daily news letter

## Short

- Late notice
- My job schedule (Meeting @ 10 AM every Wed.)



# Improvisation



# Overview of plans

	AM	PM
June 17	Plan A	
		Plan A
		Plan B
	Plan B	
	Plan C	
	Plan D	
		Plan D
		Plan D
	Plan D	
June 18		Plan A
		Plan C

## Plan-A (Tutorial)

- June 17
  - AM) Machine Unlearning in Computer Vision: Foundations and Applications
  - PM) Robustness at Inference: Towards Explainability, Uncertainty, and Intervenability
- June 18
  - Learning Deep Low-dimensional Models from High-Dimensional Data: From Theory to Practice

## Plan-C (Workshop)

- June 17
  - Full day) Prompting in Vision → Not provided
- June 18
  - Full day) Synthetic Data for Computer Vision

# Day 1

## Machine Unlearning in Computer Vision: Foundations and Applications

### Tutorial: Machine Unlearning in Computer Vision: Foundations and Applications

**Organizers:** Sijia Liu , Yang Liu,  
Nathalie Baracaldo,  
Eleni Triantafillou

**Date:** Monday, June 17

**Time:** 9:00 AM-12:00 AM

**Location:** Arch 305



**Summary:** This tutorial aims to offer a comprehensive understanding of emerging machine unlearning (MU) techniques. These techniques are designed to accurately assess the impact of specific data points, classes, or concepts on model performance and efficiently eliminate their potentially harmful influence within a pre-trained model, in response to users' unlearning requests. With the recent shift to foundation models, MU has become indispensable, as re-training from scratch is prohibitively costly in terms of time, computational resources, and finances. Consequently, the field has expanded beyond the realm of security and privacy (SP) to include the removal of toxic content, copyright material, harmful information, and personally identifying data. Despite increasing research interest, MU for vision tasks remains significantly underexplored compared to its prominence in the SP field. Therefore, it is crucial to meticulously review, thoroughly explore, and comprehensively survey MU for computer vision (CV) through this tutorial. Within this tutorial, we will delve into the algorithmic foundations of MU methods, including techniques such as localization-informed unlearning, unlearning-focused finetuning, and vision model-specific optimizers. We will provide a comprehensive and clear overview of the diverse range of applications for MU in CV. Furthermore, we will emphasize the importance of unlearning from an industry perspective, where modifying the model during its life-cycle is preferable to re-training it entirely, and where metrics to verify the unlearning process become paramount. Our tutorial will furnish the general audience with sufficient background information to grasp the motivation, research progress, opportunities, and ongoing challenges in MU.

# Machine Unlearning in Computer Vision

Zoom 회의

YeongHyeon... adam Nadeem Riaz Nithesh >

Recording

## Landscape of LLM Unlearning

**Why?**

- Safety Alignment
- Privacy Compliance
- Copyright Removal
- Bias Mitigation
- Hallucination Removal
- ...

**Evaluation**

Red Teaming, User Feedback

**LLM Pipeline**

Data → Pretrained LLM → Alignment → Aligned LLM

**Where**

Pre-training, Pre-alignment, In-alignment, Post-alignment

**How**

Gradient Ascent, Fine-tuning, Localization Informed, Influence Function

Liu et al, Rethinking Machine Unlearning for Large Language Models, 2024

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

Zoom 회의

YeongHyeon... adam Nadeem Riaz Nithesh >

Recording

## Unlearning as a general alignment algorithm for LLMs

- Removing harmful responses => Forgetting harmfulness learned in data
- Erasing copyrighted contents => Forgetting impact of copyrighted corpus
- Reducing hallucinations => Forgetting wrong “facts”
- Adapting to change of user consent on data usage => Forgetting user data
- Adapting to policy change => Forgetting old data

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

Rewinding of training process  
to prevent unintentional information leakage

# Machine Unlearning in Computer Vision

## Gradient approach

The image shows a Zoom meeting interface. At the top, there are five participant thumbnails: 'YeongHyeon...', 'sungmincha' (highlighted with a green border), 'Nadeem Riaz', and 'adam'. Below the thumbnails, a red dot indicates 'Recording'. The main content area displays a slide with the following text and equations:

## Gradient approaches

- Fine-tune (FT): Fine-tune the trained model on the remaining dataset  $D_r$ 
$$\theta_{\text{new}} = \theta - \eta \cdot \sum_{(x,y) \in D_r} \nabla_{\theta} L(\theta, x, y)$$
- Gradient Ascent (GA): Apply the inverse operation on the forgetting dataset  $D_f$ 
$$\theta_{\text{new}} = \theta + \eta \cdot \sum_{(x,y) \in D_f} \nabla_{\theta} L(\theta, x, y)$$

Worse performance but good for quick tests + explorations

Warnecke et al, Machine unlearning of features and labels, 2021  
Golatkar et al, Eternal sunshine of the spotlessnet: Selective forgetting in deep networks, 2020  
Graves et al, Amnesiac machine learning, 2021  
Thudi et al, Unrolling sgd: Understanding factors influencing machine unlearning, 2021  
Yao et al, Large language model unlearning, 2023  
Di et al, Label smoothing improves machine unlearning, 2024

 MICHIGAN STATE UNIVERSITY

UC SANTA CRUZ

IBM Research

Google

# Machine Unlearning in Computer Vision

## Gradient approach

Zoom 회의

Recording

YeongHyeon...  sungmincha Nadeem Riaz adam >

Three labels are: dog, cat, deer

**Label Smoothing (LS)**

Original Label	Positive LS	Negative LS
 [0, 1, 0]	 [0.1, 0.8, 0.1]	 [-0.1, 1.2, -0.1]
With gradient descent (GD), model is <i>less</i> confident	With gradient descent (GD), model is <i>more</i> confident	

Wei et al, To Smooth or Not? When Label Smoothing Meets Noisy Labels, 2022

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

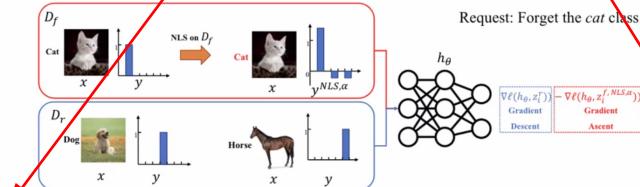
Zoom 회의

Recording

YeongHyeon...  sungmincha Nadeem Riaz adam >

**Soft knowledge forgetting**

**Motivation: Join MU and LS together**



- Decrease the confidence of forgetting set ( $D_f$ ) using negative label smoothing (with GA)
  - Gradient ascent helps model forget (decrease the confidence)
- Increase the confidence of the remaining set ( $D_r$ ) using negative label smoothing (with GD)
  - Gradient descent helps model remember (increase the confidence)

Di et al, Label Smoothing Improves Machine Unlearning, 2024

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

**Hard knowledge preservation**

# Machine Unlearning in Computer Vision

## Pruning approach

Zoom 회의

YeongHyeon... adam Nadeem Riaz Nithesh >

Recording

### Pruning Helps Unlearning

- Pruning introduces sparsity, thus reduces unlearning dimension and simplifies weight optimization for MU
- **Provable guarantee:**

**Theorem:** Given SGD-based training and model **pruning mask  $m$** , the unlearning error,  $e(m)$ , characterized by weight distance between **an approximate unlearn** and the **exact unlearn (retrain)** yields [Jia, Liu, et al., NeurIPS'23]

$$e(m) = \mathcal{O}(|m \odot (\theta_t - \theta_0)|_2)$$

○ is entry-wise product,  $\theta_t$  is model trained after t SGD iterations
- **Sparsity:** Helps reduce unlearning error compared to Retrain, but causing tradeoff with generalization

Jia, Liu, et al. "Model sparsity can simplify machine unlearning." NeurIPS'23

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

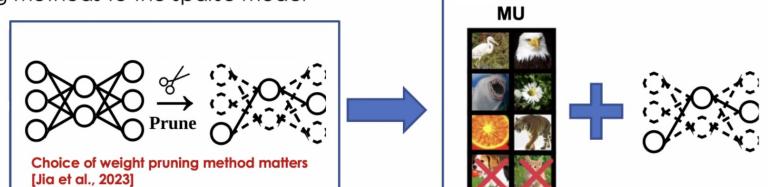
Zoom 회의

YeongHyeon... adam Nadeem Riaz Nithesh >

Recording

### Integrating weight pruning into MU

- **(Strategy 1) Prune first, then unlearn:** Find sparse model first, then applies existing approximate unlearning methods to the sparse model

  
Choice of weight pruning method matters [Jia et al., 2023]

**(Strategy 2) Sparsity-regularized unlearning:** Promoting weight sparsity as a regularization

$$\theta_u = \operatorname{argmin}_{\theta} L_{MU}(\theta; \mathcal{D}_r) + \gamma \|\theta\|_1$$

MU objective function on retain dataset  $\mathcal{D}_r$        $\ell_1$  sparse regularization

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

# Machine Unlearning in Computer Vision

Better way to research

The screenshot shows a Zoom video conference with five participants: YeongHyeon..., Yuki Watanabe, LHB, Nadeem Riaz, and a participant whose video feed is not visible. The video feed for Yuki Watanabe is highlighted with a green border. The interface includes a recording indicator and a 'Zoom 회의' label.

**Open questions**

**Evaluation.**

- Application-specific or unified metrics?
- Seems like a good topic for research!
- Can we find **cheap and accurate proxies for rigorous evaluation metrics** for approximate unlearning?

Are we measuring all relevant "**side-effects**": utility drop, fairness, *retain* set privacy?

How should we **build forget sets for evaluation**? Worst-case? Forget sets that capture identified relevant characteristics?

**Repeat questions and developments**

**Algorithm Design.**

- Can we **bridge the gap between exact and approximate methods**? Optimize directly for metrics of interest? Improve common building blocks e.g descent-ascent? Theoretical guarantees?

Are there **shared modelling principles** that succeed on different tasks / modalities / applications? **Robustness** as a priority for model design?

Can we leverage insights on example difficulty to design **curricula for unlearning**? E.g. based on example difficulty or in multi-modal settings

# Day 1

## Robustness at Inference: Towards Explainability, Uncertainty, and Intervenability

### Tutorial: Robustness at Inference: Towards Explainability, Uncertainty, and Intervenability

**Organizers:** Ghassan AlRegib,  
Mohit Prabhushankar

**Date:** Monday, June 17

**Time:** 1:30 PM–5:30 PM

**Location:** Summit 440-441



**Summary:** Neural networks provide generalizable and task independent representation spaces that have garnered widespread applicability in image understanding applications. The complicated semantics of feature interactions within image data has been broken down into a set of non-linear functions, convolution parameters, attention, as well as multi-modal inputs among others. The complexity of these operations has introduced multiple vulnerabilities within neural network architectures. These vulnerabilities include adversarial samples, confidence calibration issues, and catastrophic forgetting among others. Given that AI promises to herald the fourth industrial revolution, it is critical to understand and overcome these vulnerabilities. Doing so requires creating robust neural networks that drive the AI systems. Defining robustness, however, is not trivial. Simple measurements of invariance to noise and perturbations are not applicable in real life settings. In this tutorial, we provide a human-centric approach to understanding robustness in neural networks that allow AI to function in society. Doing so allows us to state the following: 1) All neural networks must provide contextual and relevant explanations to humans, 2) Neural networks must know when and what they don't know, 3) Neural Networks must be amenable to being intervened upon by humans at decision-making stage. These three statements call for robust neural networks to be explainable, equipped with uncertainty quantification, and be intervenable.

# Robustness at inference

Zoom 회의

YeongHyeon... ParkHyunKyu Yuki Watanabe tanmaybichu hyunjun.kim >

Recording

Memes to Wrap it Up

Robustness Research in the Inferential Stage of Neural Networks

Existing research on robustness focuses on data collection and optimization

Optimization

Data Collection

Inference

167 of 174

CVPR  
Seattle, WA  
June 17-21, 2024

[Tutorial@CVPR'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [June 17, 2024]

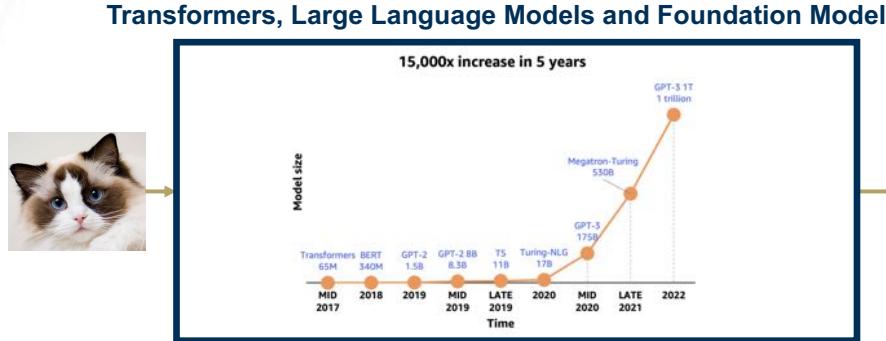
OLIVES Georgia Tech

▲Microphone issue

# Robustness at inference

Deep Deep Deep Deep Deep ... Learning  
Recent Advancements

## Development of DLs



- Primary reasons for advancements:
1. Expanded interests from the research community
  2. Computational resources availability
  3. **Big data availability**

19 of 172

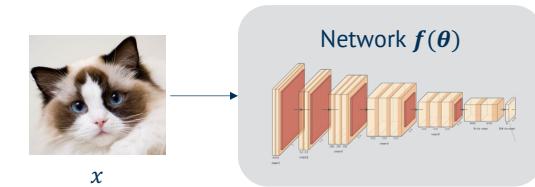


[Tutorial@CVPR'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [June 17, 2024]



Deep Learning at Inference  
Classification

Given : One network, One image. Required: Class Prediction



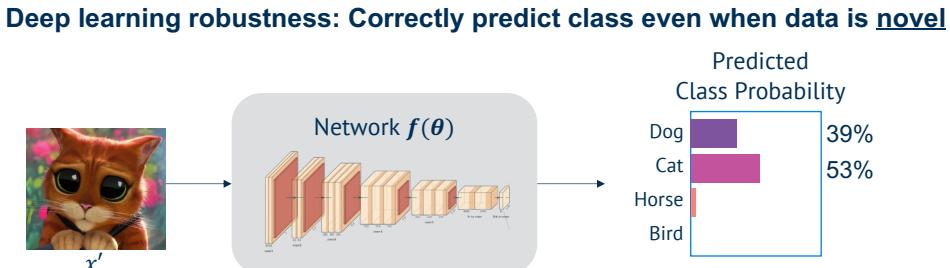
## Seen sample



If  $x \in \chi$ , the data is **not novel**

Deep Learning at Inference  
Robust Classification in Deep Networks

## Unseen sample



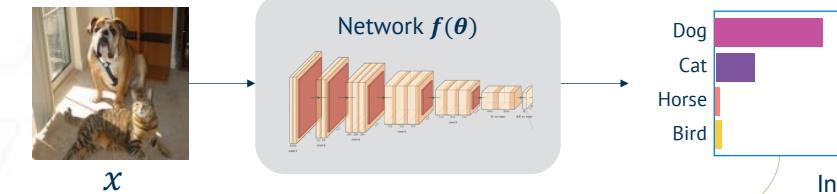
To achieve robustness at Inference, we need the following:

- **Information** provided by the novel data as a **function of training distribution**
- Methodology to **extract information** from novel data
- **Techniques** that utilize the information from novel data

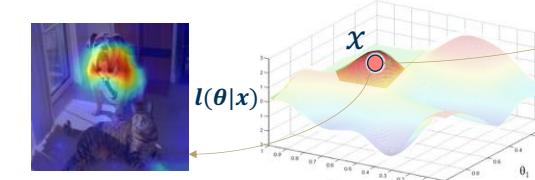
Why is this Challenging?

Information at Inference  
Case Study: Gradients as Fisher Information in Explainability

Gradients infer information about the statistics of underlying manifolds



Local information (specific to  $x$ ) is sufficient!



Feature attribution via GradCAM

## Visual reasoning

In this case, the image and its prediction extracts nose, mouth and jowl features.

Hence, gradients draw information from the underlying distribution as learned by the network weights!

# Robustness at inference

## Visual reasoning

Zoom 회의

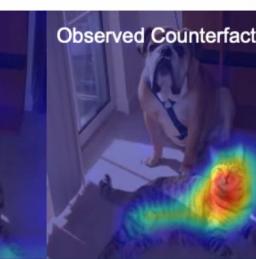
YeongHyeon... Siqi Wang Kalliopi Basioti Arshad J >

● Recording

### Explanations

#### Visual Explanations

- Explanations are defined as a set of rationales used to understand the reasons behind a decision
- If the decision is based on visual characteristics within the data, the decision-making reasons are visual explanations

**Bullmastiff**  Observed Correlations  Observed Counterfactual  Observed Contrastive

**Cat**  Observed Correlations  Observed Counterfactual  Observed Contrastive

Bullmastiff	Why Bullmastiff?	What if Bullmastiff was not in the image?	Why Bullmastiff, rather than a Boxer?
-------------	------------------	---	---------------------------------------

35 of 174 CVPR JUNE 17-21, 2024 [Tutorial@CVPR'24] | [Ghassan AlRegib and Mohit Prabhushankar] | [June 17, 2024]  
AlRegib, G., & Prabhushankar, M. (2022). Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations. *IEEE Signal Processing Magazine*, 39(4), 59-72.

OLIVES Georgia Tech



Heat map would show you this kind of image of the talk. Which is the face of the dog I can ask

# Robustness at inference

## Visual reasoning

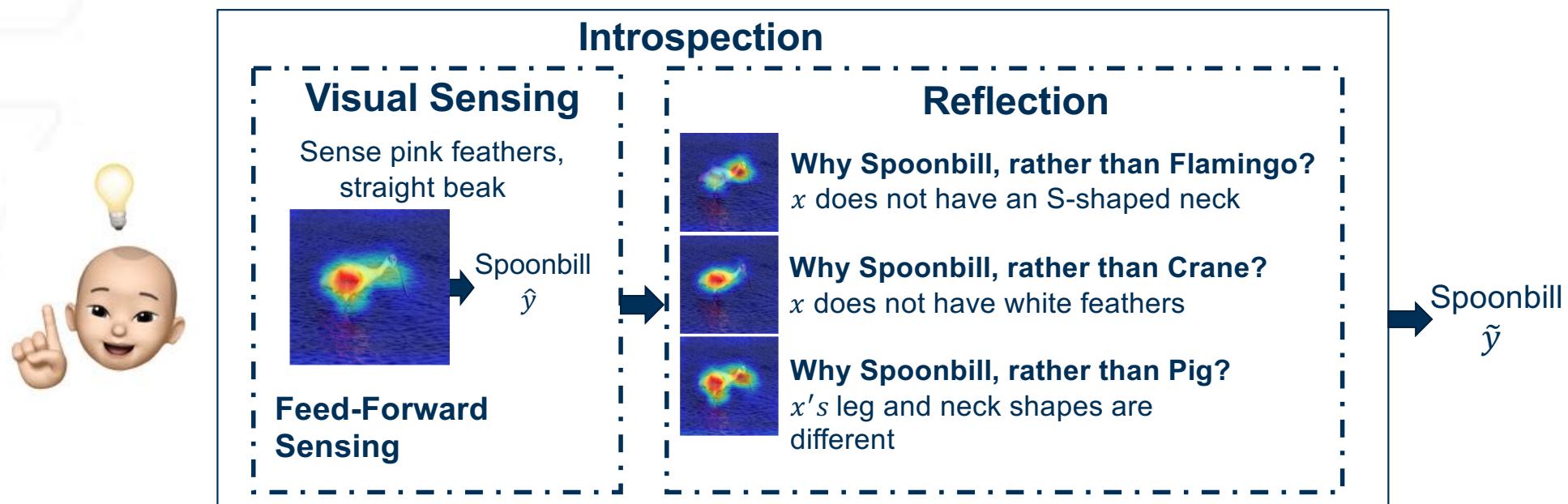
### Introspection

What is Introspection?



Introspective Learning: A Two-stage Approach for Inference in Neural Networks

Introspection Learning is a two-stage approach for Inference that combines visual sensing and reflection



# Robustness at inference

## Uncertainty of inference (Inference by multiple models)

### Uncertainty

What is Uncertainty?

Uncertainty is a model knowing that it does not know

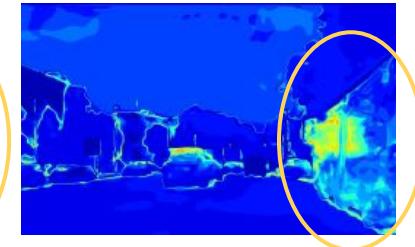
Input Image



Neural Network Output



Uncertainty Heatmap



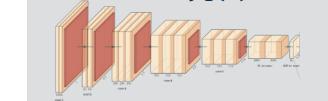
### Uncertainty

Uncertainty Quantification in Neural Networks

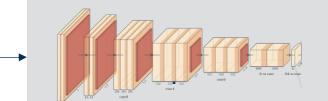
Via Ensembles<sup>1</sup>



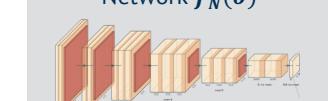
Network  $f_1(\theta)$



Network  $f_2(\theta)$



Network  $f_N(\theta)$



Variation within outputs  
is the uncertainty.

Commonly referred to  
as **Prediction  
Uncertainty**.

**Requires multiple  
trained models – not  
exactly an inferential  
method**

Uncertainty measuring by multiple models  
But not effective approach

# Robustness at inference

## Uncertainty of inference (Iterative inference by dropout)

### Uncertainty

#### Iterative Uncertainty Quantification

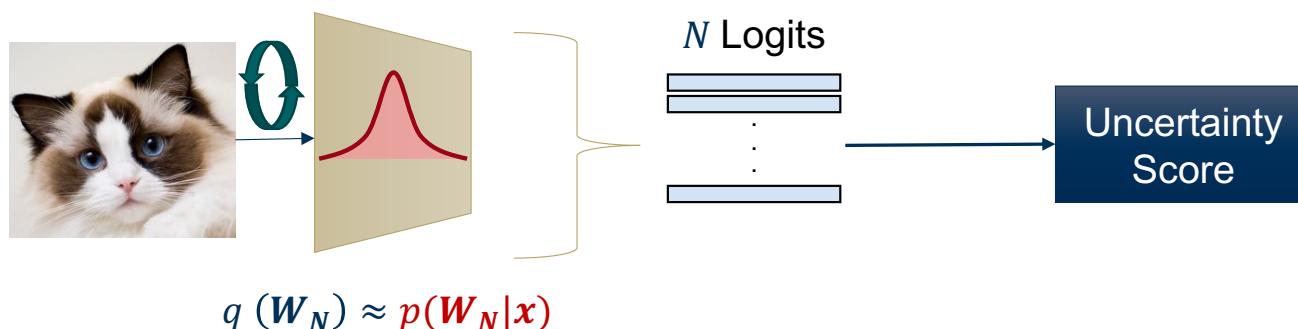
**Via Monte-Carlo Dropout<sup>1</sup>:** During inference repeated evaluations with the same input give different results

Different forward passes with dropout simulate  $f_1(\cdot), f_2(\cdot), f_3(\cdot)$ .

Challenge: intractable denominator

$$p(\mathbf{W}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{W})p(\mathbf{W})}{\int p(\mathbf{x}|\mathbf{W})p(\mathbf{W})d\mathbf{W}}$$

$N$  forward passes



Final prediction is the mean of the outputs

Variation or entropy of logits is the uncertainty

# Robustness at inference

## Uncertainty of inference (Multiple inference by multiple masking)

PETSIUK, DAS, SAENKO: RISE: RANDOMIZED INPUT SAMPLING FOR EXPLANATION 5

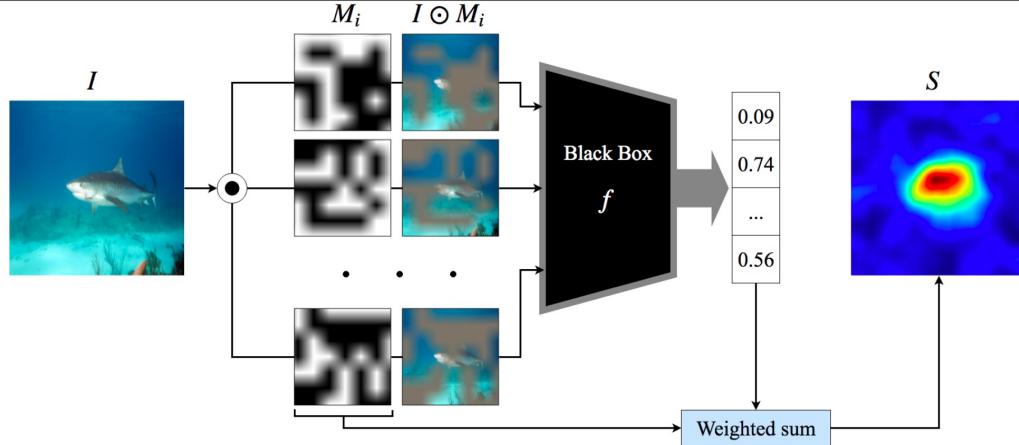


Figure 3: Overview of RISE: Input image  $I$  is element-wise multiplied with random masks  $M_i$  and the masked images are fed to the base model. The saliency map is a linear combination of the masks where the weights come from the score of the target class corresponding to the respective masked inputs.

Vitali Petsiuk, et al. "RISE: Randomized input sampling for explanation of black-box models." arXiv. 2018.

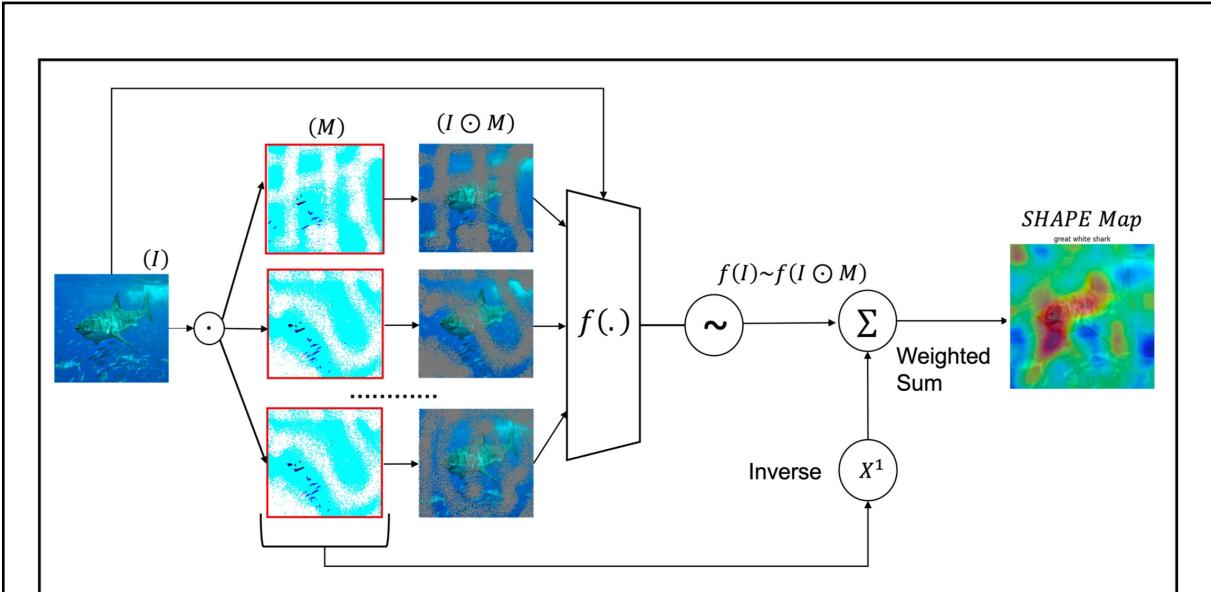


Fig. 2. Overview of SHAPE to generate adversarial explanations: Input image  $I$  is element-wise multiplied by the random masks  $M_i$  and are fed into the model  $f$  along with the original image to calculate the change in prediction scores. The importance map is a weighted sum of masks where the weights for each mask is its corresponding change in probability scores.

Prithwijit Chowdhury, et al. "Are Objective Explanatory Evaluation metrics Trustworthy? An Adversarial Analysis." arXiv. 2024

## Day 2

# Synthetic Data for Computer Vision

### Synthetic Data for Computer Vision

**Organizers:** Jieyu Zhang, Cheng-Yu Hsieh, Zixian Ma, Shobhit Sundaram, Wei-Chiu Ma, Phillip Isola, Ranjay Krishna

**Date:** Tuesday, June 18

**Time:** 8:30 AM-5:30 PM

**Location:** Summit 423-425

**Summary:** The workshop aims to explore the use of synthetic data in training and evaluating computer vision models, as well as in other related domains. During the last decade, advancements in computer vision were catalyzed by the release of painstakingly curated human-labeled datasets. Recently, people have increasingly resorted to synthetic data as an alternative to laborintensive human-labeled datasets for its scalability, customizability, and costeffectiveness. Synthetic data offers the potential to generate large volumes of diverse and high-quality vision data, tailored to specific scenarios and edge cases that are hard to capture in real-world data. However, challenges such as the domain gap between synthetic and real-world data, potential biases in synthetic generation, and ensuring the generalizability of models trained on synthetic data remain. We hope the workshop can provide a forum to discuss and encourage further exploration in these areas.



# Synthetic Data for Computer Vision

The image shows a Zoom video call interface. In the top left, there's a thumbnail of a room full of people. To the right are five participant names: YeongHyeon..., Ammar, Aymane, Shobhita S S..., and Ludwig Schm... (the last one is highlighted with a green border). Below the names, a red dot indicates 'Recording'. The main content area displays a presentation slide with the following text:

## Common paradigm in ML research: data fixed, improve models

Below the text is a scatter plot titled 'BOX MAP' on the y-axis (ranging from 0 to 80) and 'Year' on the x-axis (ranging from 2016 to 2023). The plot shows numerous small grey dots representing 'Other models' scattered across the chart. Overlaid on these are several teal-colored diamond markers representing 'Models with highest box mAP', connected by a teal line. The labeled models and their descriptions are:

- Fast-RCNN (approx. 20, 2015)
- SSD512 (approx. 30, 2016)
- Faster R-CNN (box refinement, context, multi-scale testing) (approx. 35, 2016)
- D-RFCN + SNIP (DPN-98 with flip, multi-scale) (approx. 40, 2017)
- Mask R-CNN (ResNeXt-101-FPN) (approx. 45, 2017)
- NAS-FPN (AmoebaNet-D, learned aug) (approx. 50, 2018)
- Detector3D (ResNeXt-101-64x4d, multi-scale) (approx. 55, 2019)
- DyHead (Swin-L, multi scale, self-training) (approx. 60, 2021)
- NAS-FPN (AmoebaNet-D, learned aug) (approx. 60, 2021)
- FocalNet-H (DINO) (approx. 65, 2022)
- DyHead (Swin-L, multi scale, self-training) (approx. 65, 2022)

A legend at the bottom left identifies the symbols: a grey circle for 'Other models' and a teal diamond for 'Models with highest box mAP'. At the bottom right, the source is cited: 'Source: [paperswithcode.com](#)'. A large blue arrow points downwards to the text 'Few papers experiment with improving the training data.'.

Few papers experiment with improving the **training data**.

15

# Synthetic Data for Computer Vision

A screenshot of a Zoom video conference interface. The top bar shows the title "Zoom 회의". Below it, participant names are listed: YeongHyeon..., Ammar, Aymane, Shobhita S S..., and Ludwig Schm... (with the last name highlighted in green). The status "Recording" is visible. The main content area contains the text "AI = compute + data" where "data" is highlighted in a pink box and has a red question mark icon to its right. Two blue arrows point from a bulleted list below to the words "compute" and "data". The list includes: Optimization algorithms, Model architectures, Loss functions, and "... (thousands of papers)". The slide is numbered 14 at the bottom right.

A screenshot of a Zoom video conference interface. The top bar shows the title "Zoom 회의". Below it, participant names are listed: YeongHyeon..., Ammar, Aymane, Shobhita S S..., and Ludwig Schm... (with the last name highlighted in green). The status "Recording" is visible. The main content area features the word "Conclusions" in large black font. To the right is the IFML logo, which is a colorful paw print with a rainbow inside. Below "Conclusions", the text states: "Datasets are a key driver of progress in AI, e.g., in multimodal learning:" followed by two blue arrow points: "Reliable generalization (CLIP)" and "Text-guided image generation (diffusion, auto-regressive, etc.)". Further down, it says: "Current datasets are assembled in an **ad-hoc** way and often **proprietary**." Another blue arrow point leads to: "Large scope for **improving training set curation** for foundation models." At the bottom, it mentions: "Two large research collaborations for improving and understanding training data: LAION-5B: first billion-scale public image-text dataset, used for Stable Diffusion. DataComp: first benchmark for multimodal training sets, results surpass OpenAI."

Data! Data! Data!

# Synthetic Data for Computer Vision

Why synthetic data?

Zoom 회의

Recording

Data desiderata

- Variability and diversity
- Balanced distribution
- Domain relevance
- Label noise
- Label richness
- Ethics and legal compliance

Meta AI

# Synthetic Data for Computer Vision

Why synthetic data?

Recording

## Why use synthetic data?

- ❑ Easy to control variation and diversity
- ❑ Balanced distribution comes for free
- ❑ Clean labels without annotation error
- ❑ Superhuman-level label richness
- ❑ No personal identifiable information
- ❑ Possible to match your target domain

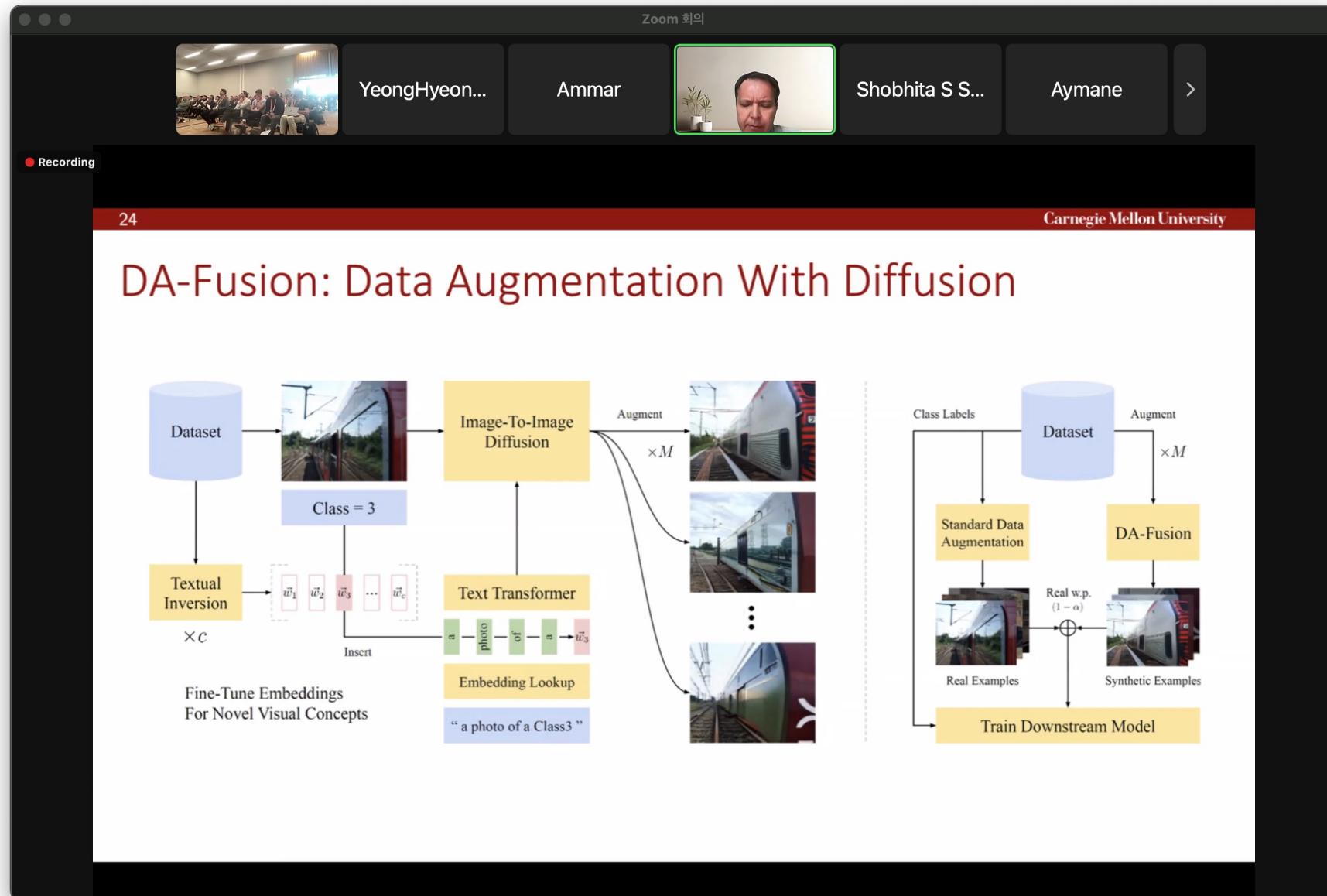
Fake it till you make it [Wood et al., ICCV 2021]

Meta AI

Synthetic data takes all the advantages

# Synthetic Data for Computer Vision

## Diffusion usage



# Synthetic Data for Computer Vision

## Simulator usage (1)

Zoom 회의

Recording

YeongHyeon... Ruslan Salak... Shobhita S S... yalesong >

### Leader-Follower Dataset

Controlled causal relationships:  
“leaders” (intervention), “followers” (outcomes), traffic signals (confounder)

Causally connected by "leader-follower" routes: Car 1 ⇒ Car 2 , Car 3 ⇒ Car 4 , Car 5 ⇒ Car 6



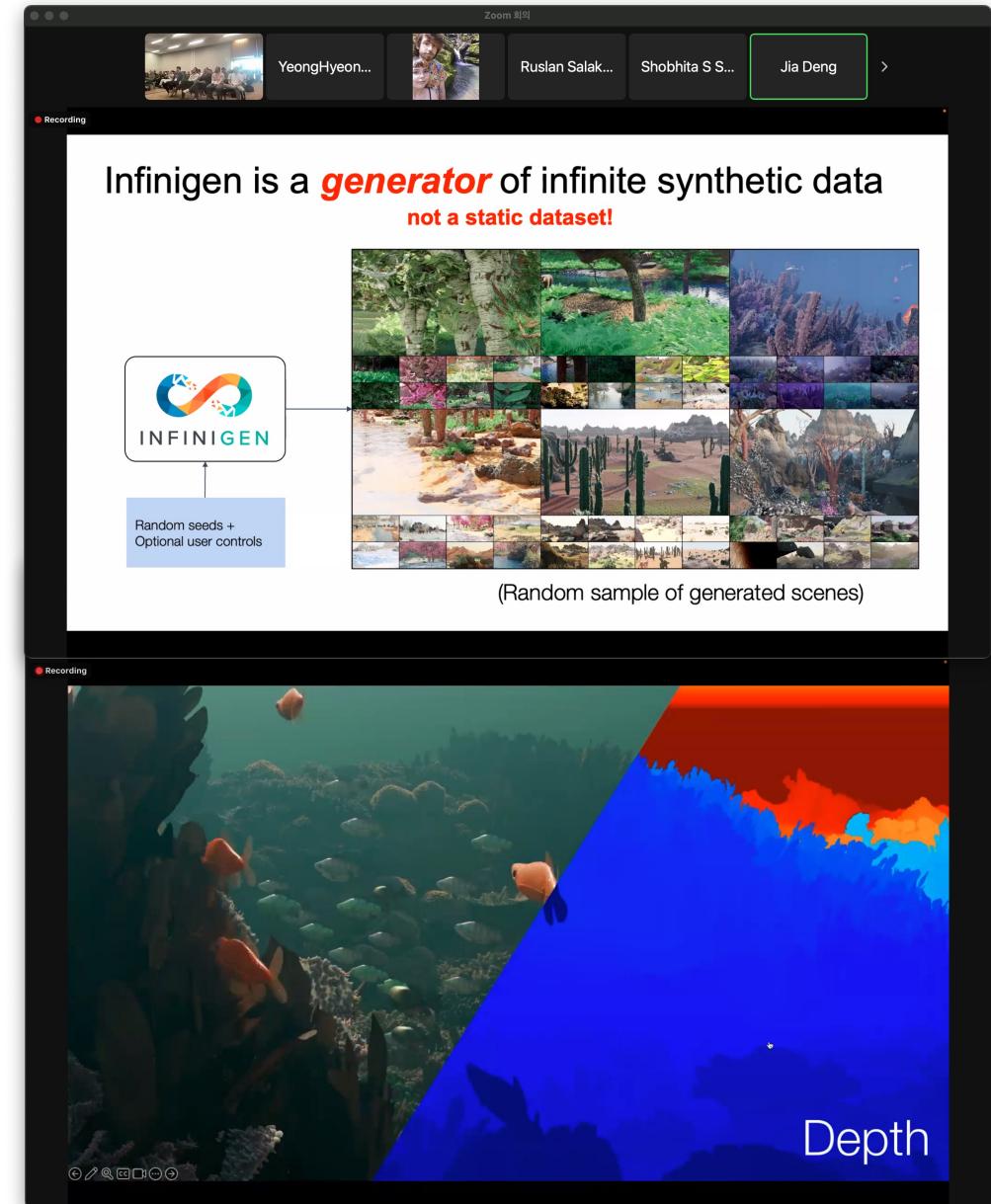
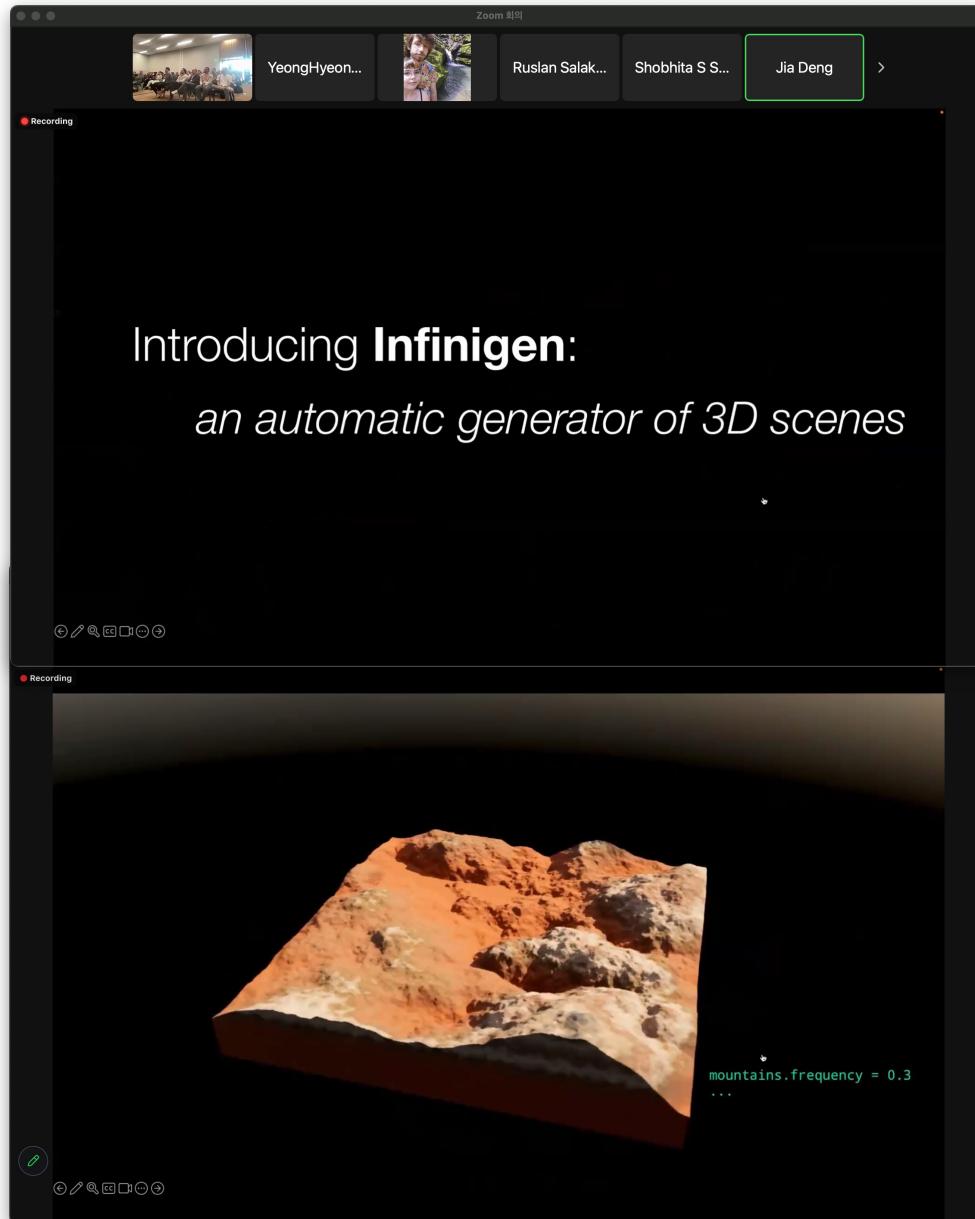
No "leader-follower" relationships, no confounders:



Meta AI

# Synthetic Data for Computer Vision

## Simulator usage (2)



# Synthetic Data for Computer Vision

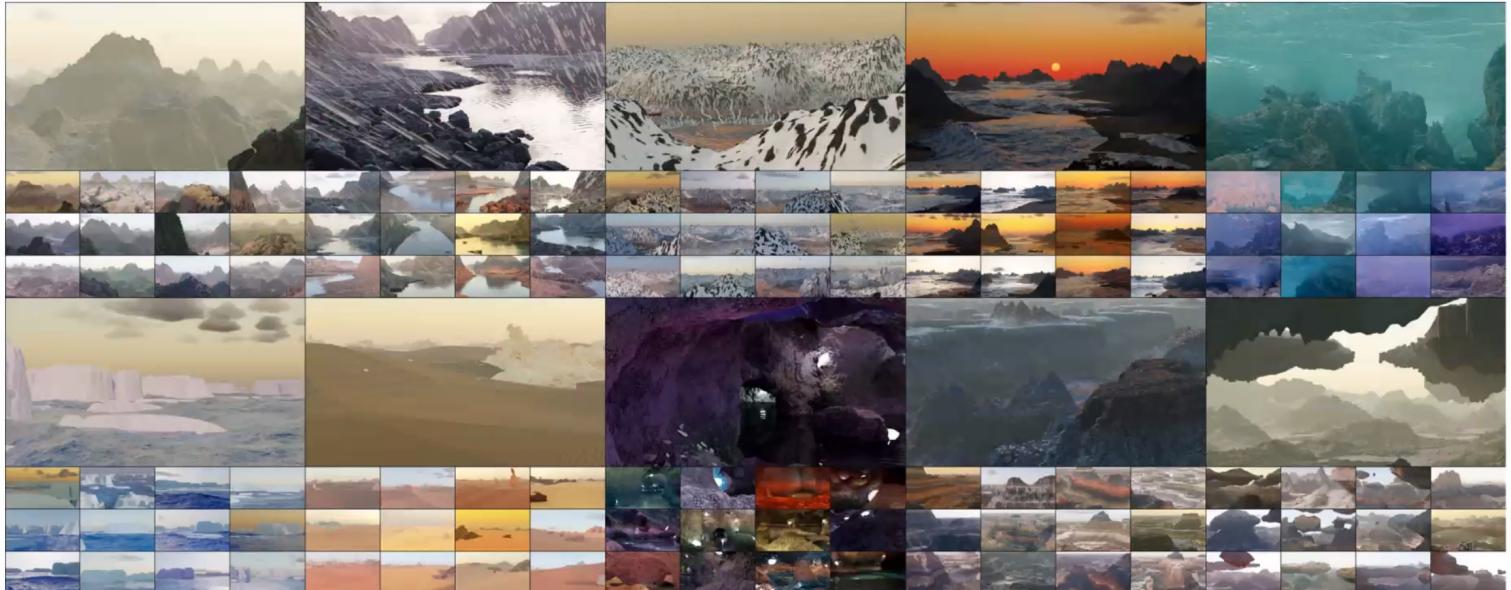
## Simulator usage (2)

Zoom 회의

Recording

YeongHyeon... Ruslan Salak... Shobhita S S... Jia Deng >

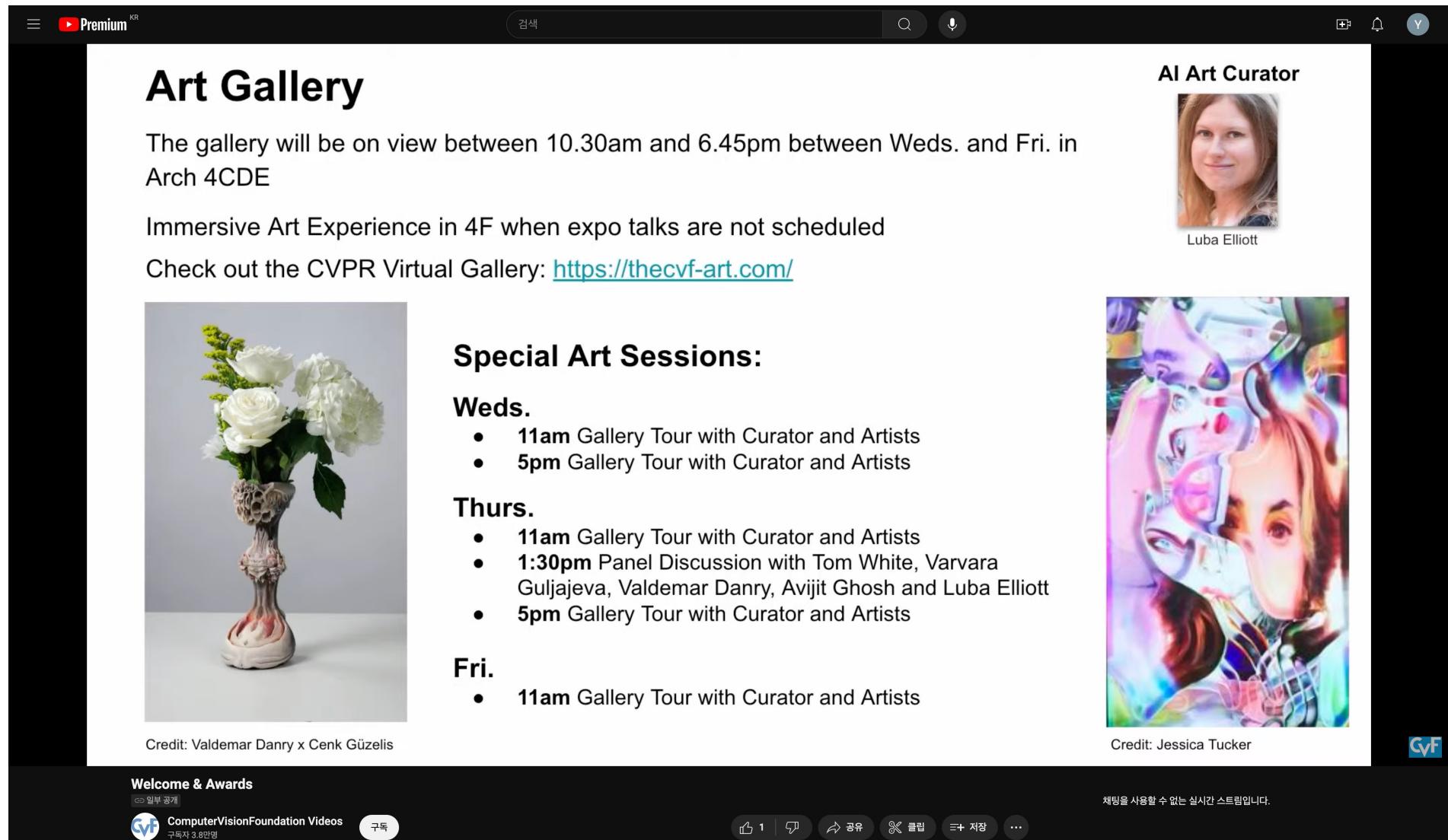
### Procedural terrain system

A large grid of procedural terrain images generated by a computer vision simulator. The grid is composed of numerous smaller images showing various landscapes, including mountains, deserts, forests, and underwater scenes, demonstrating the diversity of terrain generated by the system.

# Day 3

## Art Gallery

# Art Gallery



The gallery will be on view between 10.30am and 6.45pm between Weds. and Fri. in Arch 4CDE

Immersive Art Experience in 4F when expo talks are not scheduled

Check out the CVPR Virtual Gallery: <https://thecvf-art.com/>

**AI Art Curator**



Luba Elliott

**Special Art Sessions:**

**Weds.**

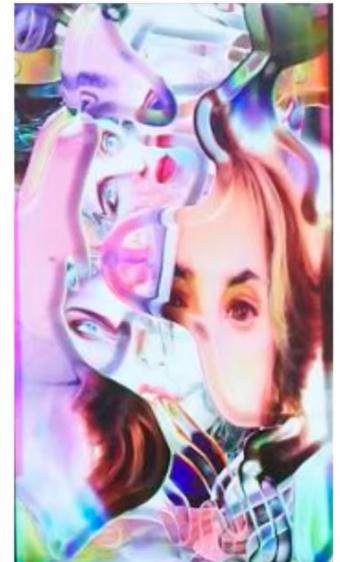
- **11am** Gallery Tour with Curator and Artists
- **5pm** Gallery Tour with Curator and Artists

**Thurs.**

- **11am** Gallery Tour with Curator and Artists
- **1:30pm** Panel Discussion with Tom White, Varvara Guljajeva, Valdemar Danry, Avijit Ghosh and Luba Elliott
- **5pm** Gallery Tour with Curator and Artists

**Fri.**

- **11am** Gallery Tour with Curator and Artists



Credit: Jessica Tucker

**Welcome & Awards**

© 일부 공개

ComputerVisionFoundation Videos 구독자 3.8만명

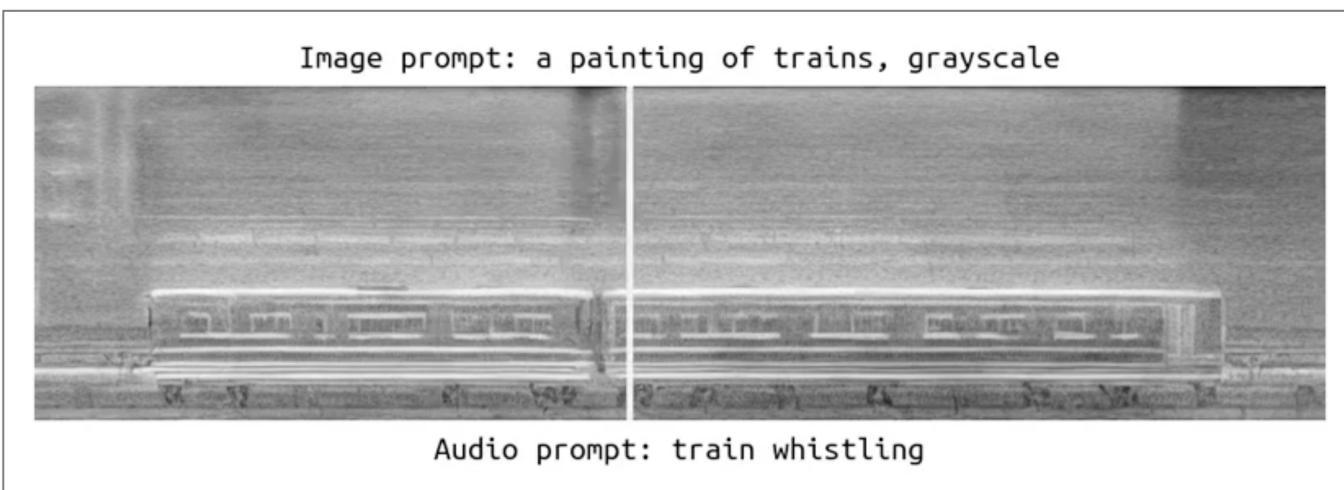
구독

채팅을 사용할 수 있는 실시간 스트리밍입니다.

1 | ⏪ ⏴ 공유 ☰ 클립 ⌂ 저장 ...

<https://thecvf-art.com/>

# Art Gallery



# Interesting Paper

# Generative Unlearning for Any Identity

Juwon Seo<sup>1\*</sup>

Sung-Hoon Lee<sup>1\*</sup>

Tae-Young Lee<sup>1\*</sup>

Seungjun Moon<sup>2</sup>

Gyeong-Moon Park<sup>1†</sup>

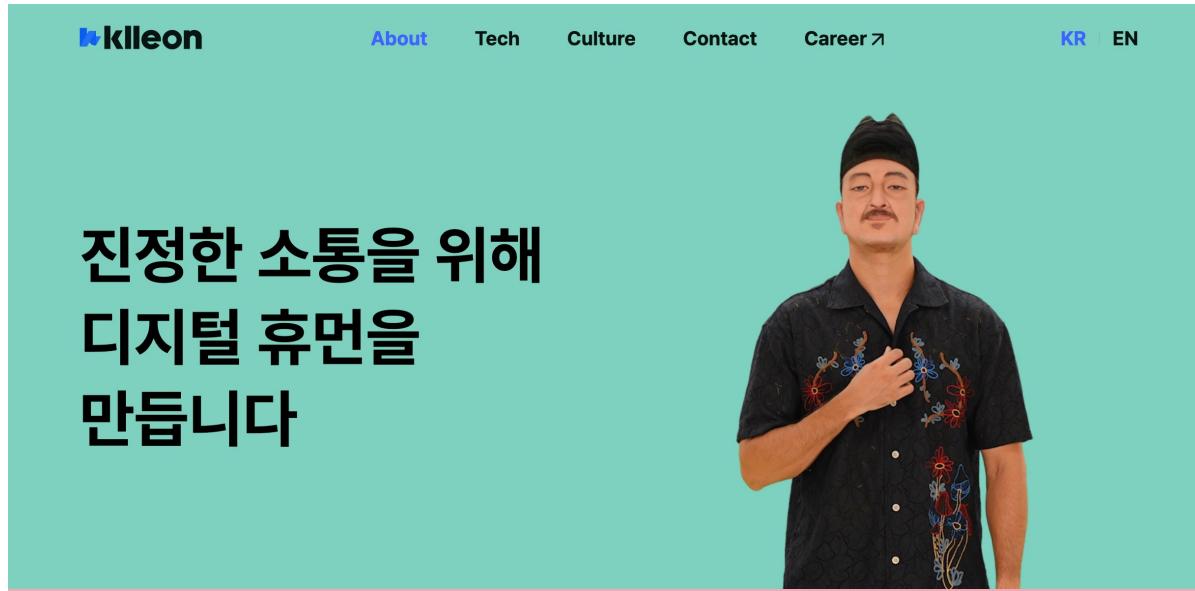
<sup>1</sup>Kyung Hee University, Yongin, Republic of Korea

<sup>2</sup>KLleon Tech., Seoul, Republic of Korea

{jwseo001, sunghoonlee961, slcks1, gmpark}@knu.ac.kr

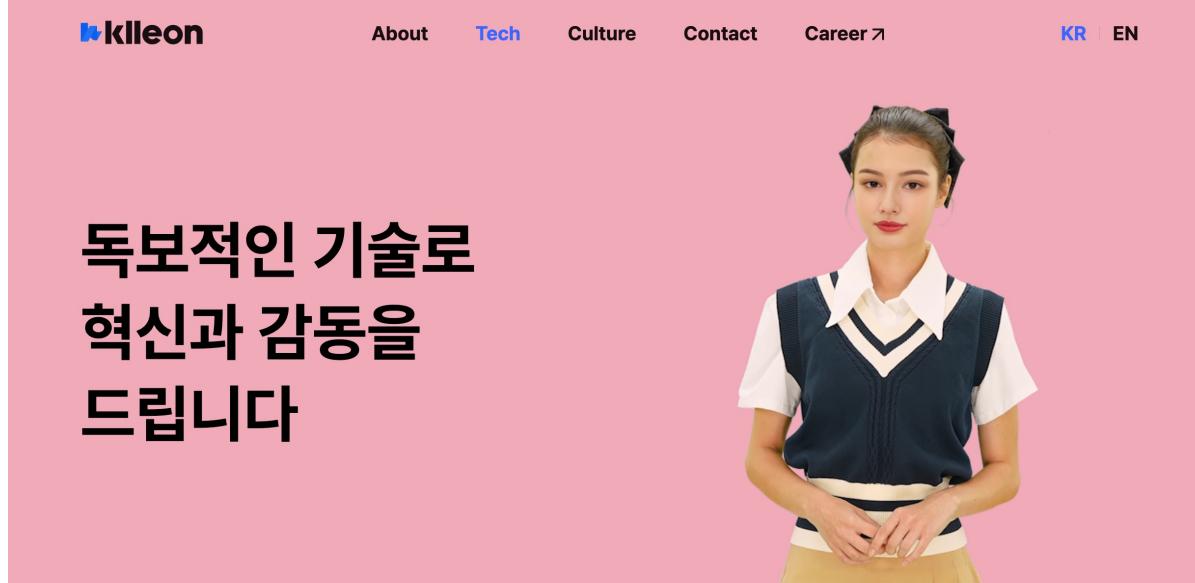
seungjun.moon@klleon.io

# Kleon Tech



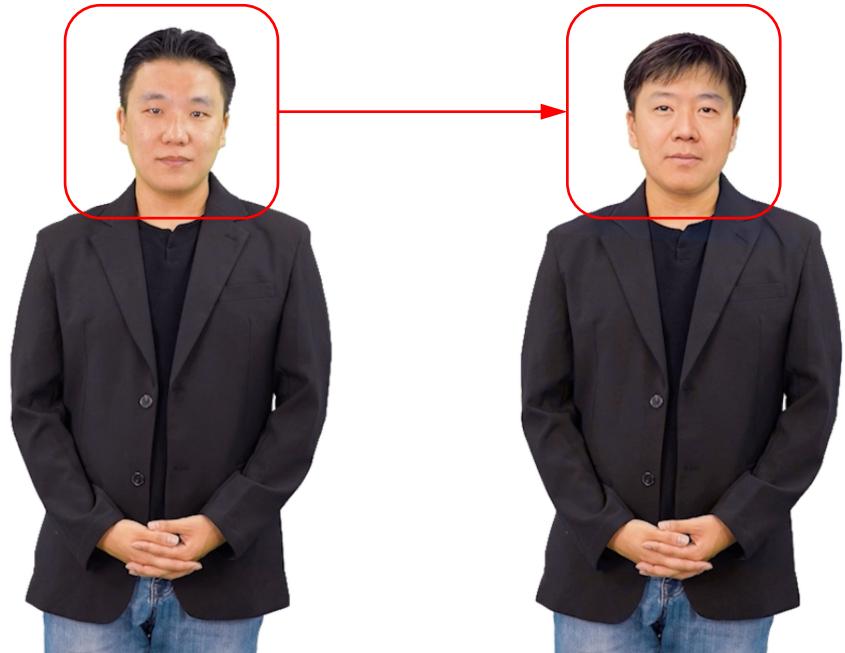
**klleon** About Tech Culture Contact Career KR EN

진정한 소통을 위해  
디지털 휴먼을  
만듭니다



**klleon** About Tech Culture Contact Career KR EN

독보적인 기술로  
혁신과 감동을  
드립니다



# Generative Unlearning for Any Identity

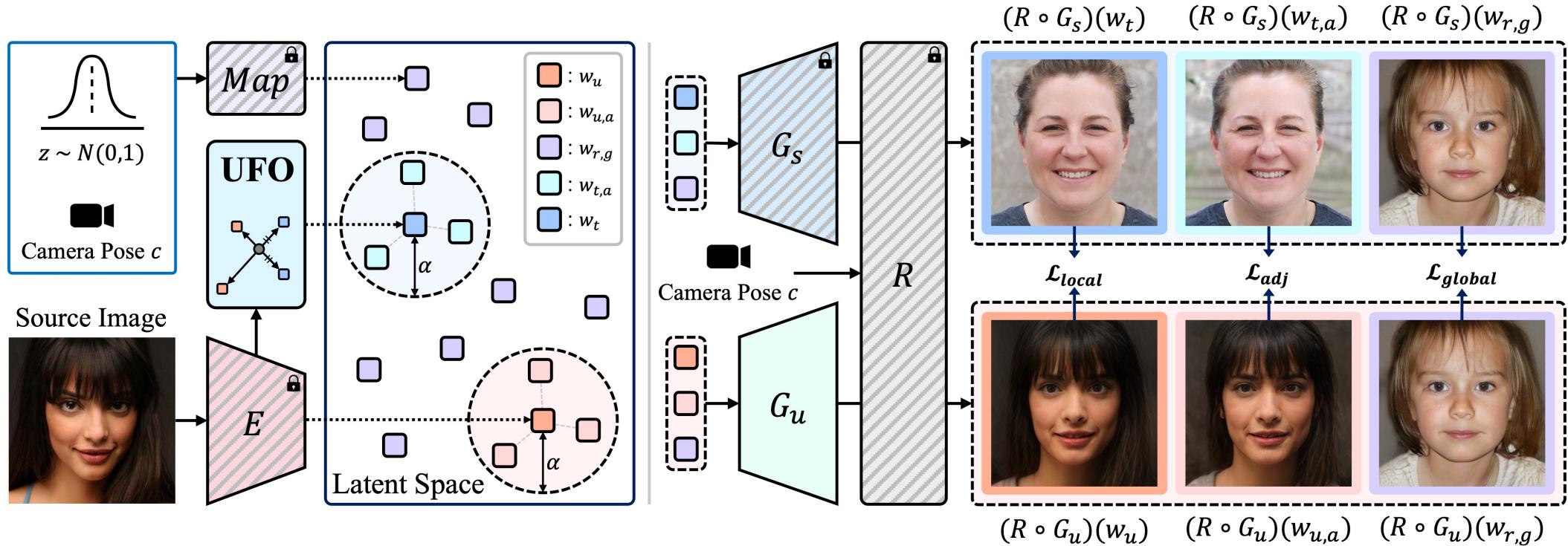


Figure 3. An overview of GUIDE. Starting with a source image, we employ a GAN inversion network  $E$ , specifically GOAE [52], to embed this image into the latent space of a pre-trained generative model, namely EG3D [4], obtaining the source latent code  $w_u$ . The target latent code  $w_t$  is designated through the UFO process. To facilitate identity removal in  $w_u$ , we shift its identity to match that of  $w_t$  with our Latent Target Unlearning (LTU) process. Three loss functions of LTU are designed for this purpose: (i) The generator is optimized to produce an image from the source latent code, denoted as  $(R \circ G_u)(w_u)$ , that is similar to the image from the target latent code, represented as  $(R \circ G_s)(w_t)$ . (ii) To achieve unlearning across the entire identity, we consider latent codes near both the source and target latent codes, denoted as  $w_{u,a}$  and  $w_{t,a}$ , respectively. (iii) To prevent model corruption during the unlearning process, we additionally sample latent codes from a random noise vector, represented as  $w_{r,g}$ , and optimize  $G_u$  to preserve its generation ability on  $w_{r,g}$ .



## TL;DR

**Prevent 3D GANs from generating you!**

## Problem Definition

## • Machine Unlearning

- Address privacy issues in deep neural networks.
- Erase or reduce the effect of certain dataset or knowledge.

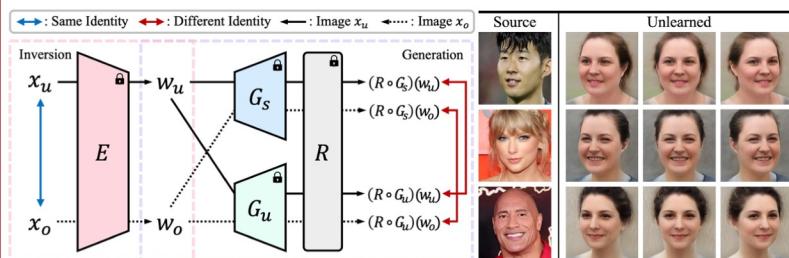
## • Previous Work

- Erase high-level concept (e.g. NSFW) attributes rather than identity.
- Unlearning on identity is unexplored.

## • Necessity

- Even if not used in training, anyone can be easily generated (GAN inversion) and manipulated (editing).

## • Task: Generative Identity Unlearning



## • Goals

- Prevent** pre-trained 3D GANs from generating a certain identity.
- Erase the given identity effectively by utilizing a **single image**.
- Preserve** the performance of pre-trained models.

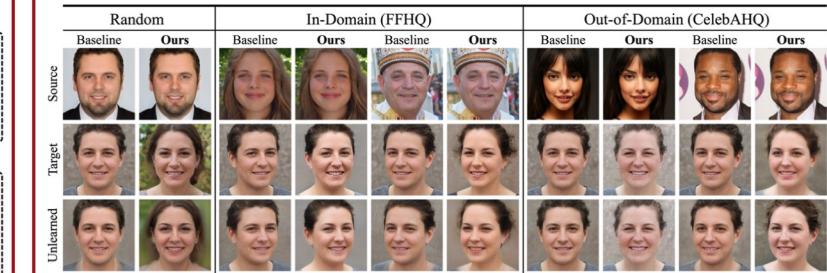
## Generative Unlearning for Any Identity

Juwon Seo<sup>1\*</sup>, Sung-Hoon Lee<sup>1\*</sup>, Tae-Young Lee<sup>1\*</sup>, Seungjun Moon<sup>2</sup>, Gyeong-Moon Park<sup>1†</sup>  
<sup>1</sup>Kyung Hee University, Yongin, Republic of Korea <sup>2</sup>KLLeon Tech, Seoul, Republic of Korea



## Experiments

## • Qualitative Results of GUIDE



## • Quantitative Results of GUIDE

Method	Random			In-Domain (FFHQ)			Out-of-Domain (CelebAHQ)		
	ID (I)	FID <sub>pre</sub> (I)	ΔFID <sub>real</sub> (I)	ID (I)	FID <sub>pre</sub> (I)	ΔFID <sub>real</sub> (I)	ID (I)	FID <sub>pre</sub> (I)	ΔFID <sub>real</sub> (I)
Baseline	0.19 ± .009	11.73 ± 2.74	746 ± 220	0.16 ± .007	9.00 ± 1.15	415 ± 138	0.12 ± .006	9.52 ± 1.53	475 ± .089
+ extrapolated $w_t$	<b>0.12 ± .006</b>	14.28 ± 3.38	963 ± 255	0.05 ± .006	12.78 ± 1.82	676 ± 141	0.02 ± .005	13.02 ± 3.20	731 ± 198
+ $L_{adj}$	0.14 ± .007	19.65 ± .49	1394 ± 359	<b>0.04 ± .006</b>	13.53 ± 2.08	735 ± 170	<b>0.01 ± .005</b>	13.63 ± 3.32	783 ± 219
+ $L_{global}$ (GUIDE)	0.14 ± .006	<b>10.80 ± 2.70</b>	<b>6.64 ± 1.60</b>	0.06 ± .006	<b>8.00 ± 1.20</b>	<b>3.05 ± .881</b>	0.03 ± .005	<b>7.88 ± 1.96</b>	<b>3.34 ± 1.10</b>

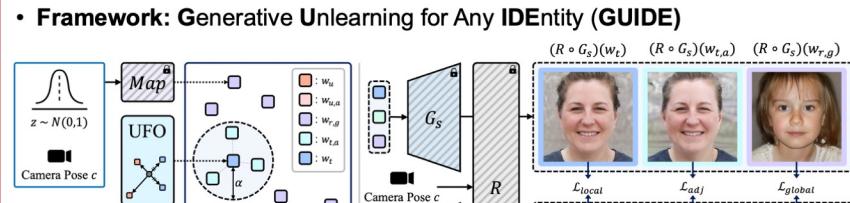
## • Erasing the Entire Identity using GUIDE

ID	Others									
	Source	Others								
ID										
Ours										

## Summary

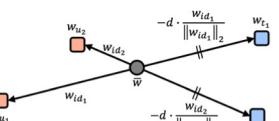
- Tackle a necessary but under-explored unlearning task in generative models, generative identity unlearning.
- Propose a novel framework GUIDE to erase given identity in the pre-trained generative models.
- Shed light on a new direction in alleviating privacy issues.

## Method



## • Un-Identifying Face On latent space (UFO)

- Goal: Determine the **effective target identity**.
- Approach: Extrapolation between  $w_u$  and  $\bar{w}$ .



## • Latent Target Unlearning (LTU)

1. Local Unlearning Loss ( $L_{local}$ )

- Goal: **Shift** source identity to target identity

$$L_{local}(\hat{x}_w, \hat{x}_t) = \lambda_{L2} L_{L2}(F_w, F_t) + \lambda_{per} L_{per}(\hat{x}_w, \hat{x}_t) + \lambda_{id} L_{id}(\hat{x}_w, \hat{x}_t).$$

2. Adjacency-Aware Unlearning Loss ( $L_{adj}$ )

- Goal: **Effectively erase** identity with only utilizing single image.
- Approach: Approximate **neighborhoods** of both source and target.  $L_2, L_{per},$  and  $L_{id}$ , between neighborhoods.

3. Global Preservation Loss ( $L_{global}$ )

- Goal: **Preserve** the pre-trained models' performance.
- Approach:  $L_{per}$  between randomly sampled latent codes from current generator and pre-trained generator.