Review for Workshop & Tutorial

YeongHyeon Park

Dept. of ECE, SungKyunKwan University

Short Review

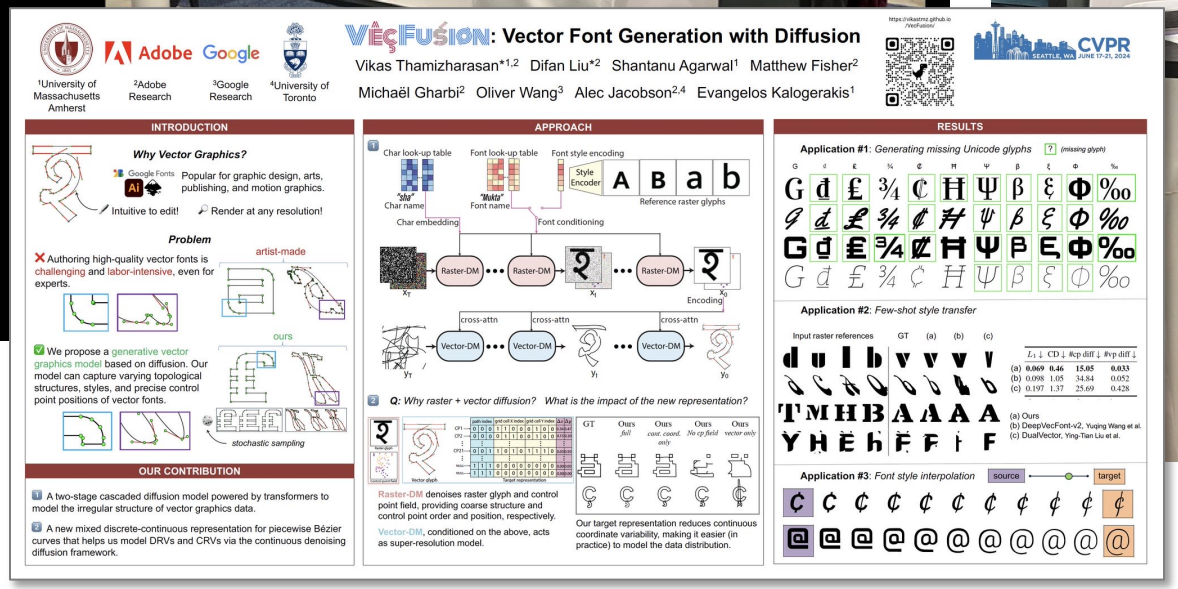
Good

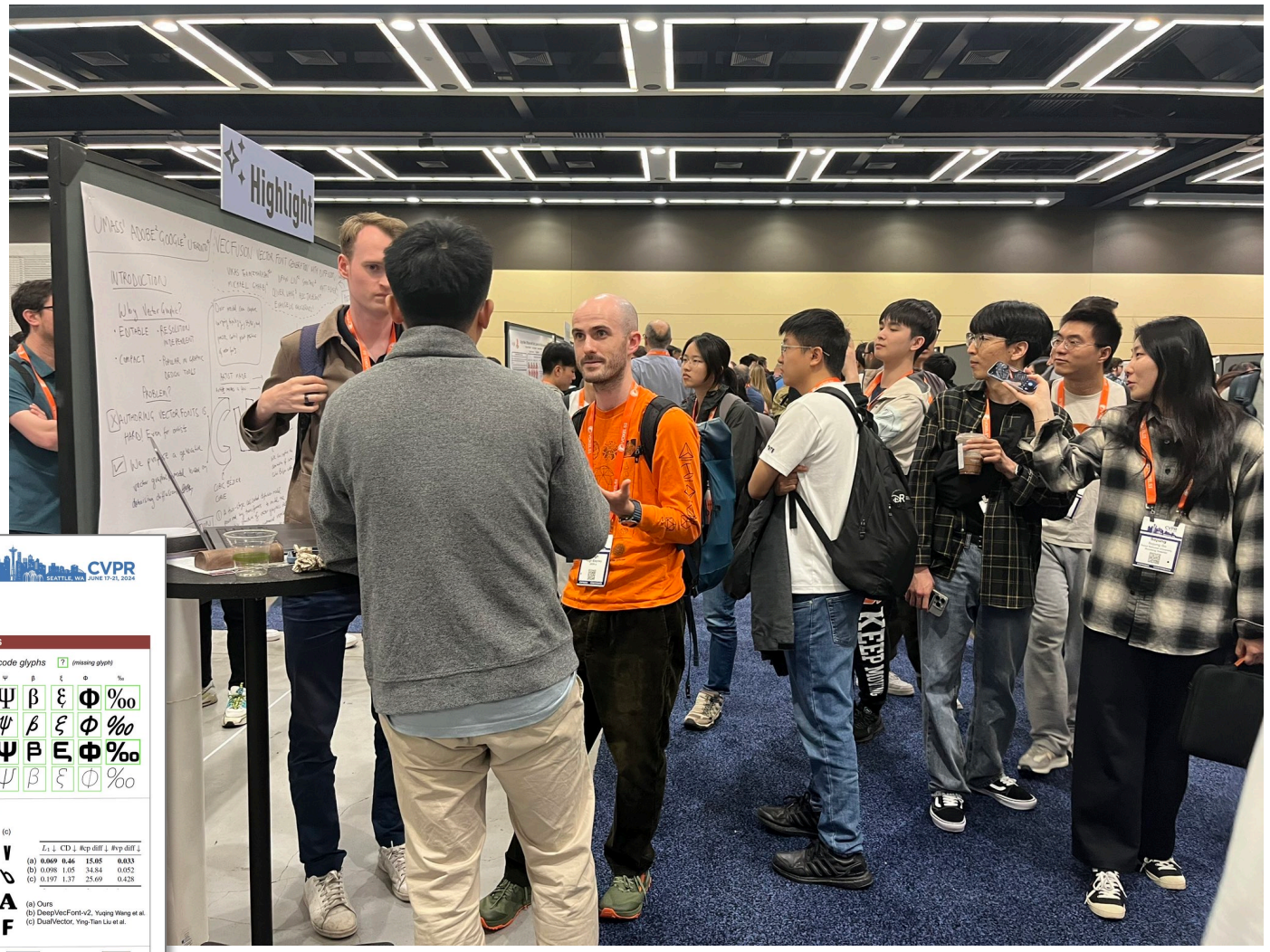
- Tutorial/workshop planning worked well
- Reliable streaming service
  - Well-prepared presentation
  - But only for available virtual streaming
- Interesting new topics
  - Focusing on research topics rather than NN structure
  - Makes me think about research in other fields than AD
- Daily news letter

Short

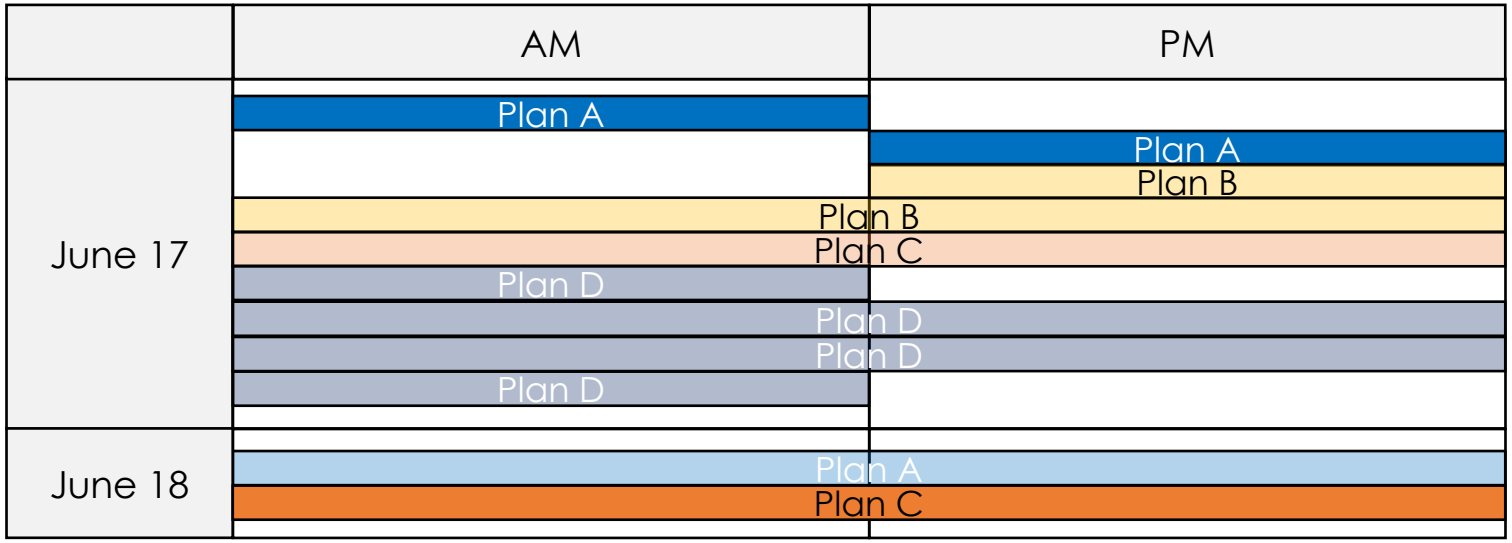
- Late notice
- My job schedule (Meeting @ 10 AM every Wed.)

Improvisation





Overview of plans



Plan-A (Toturial)

- 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

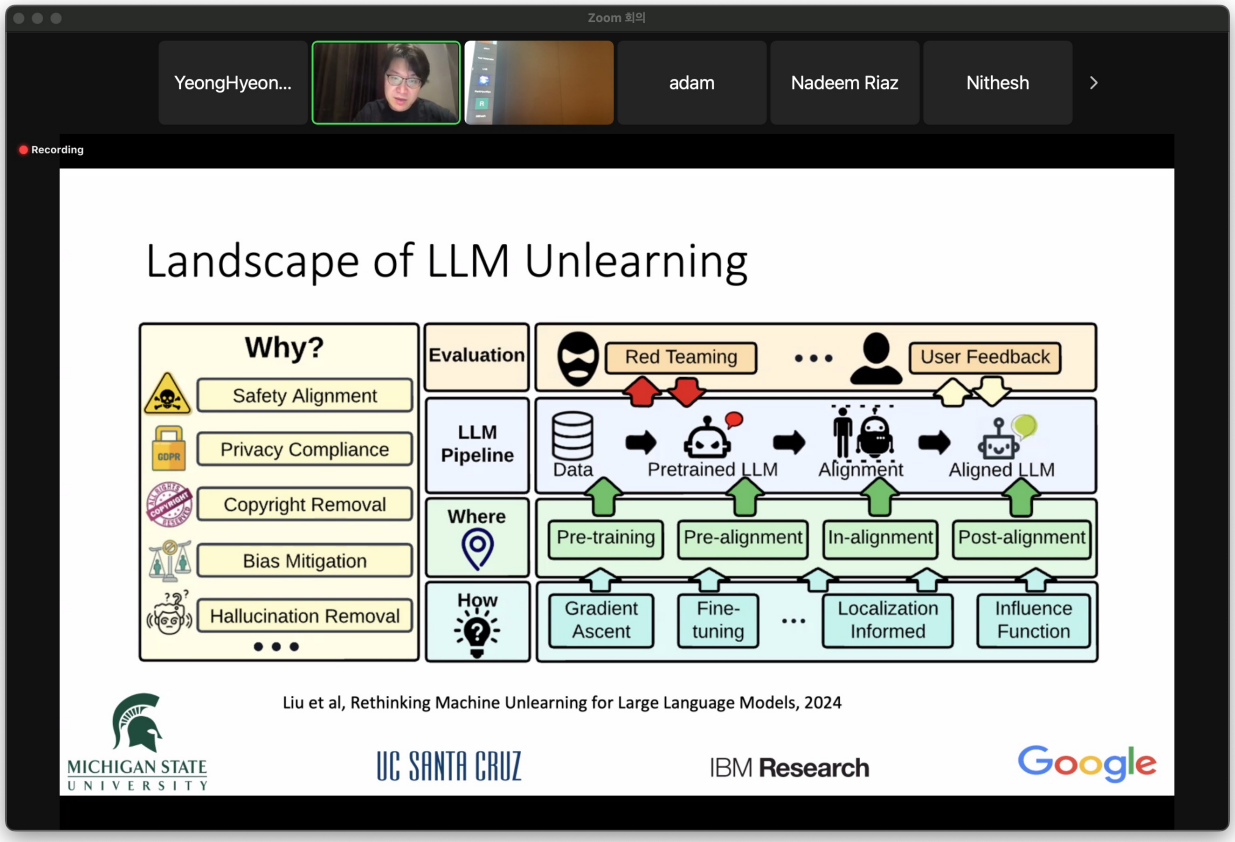
Date: Monday, June 17

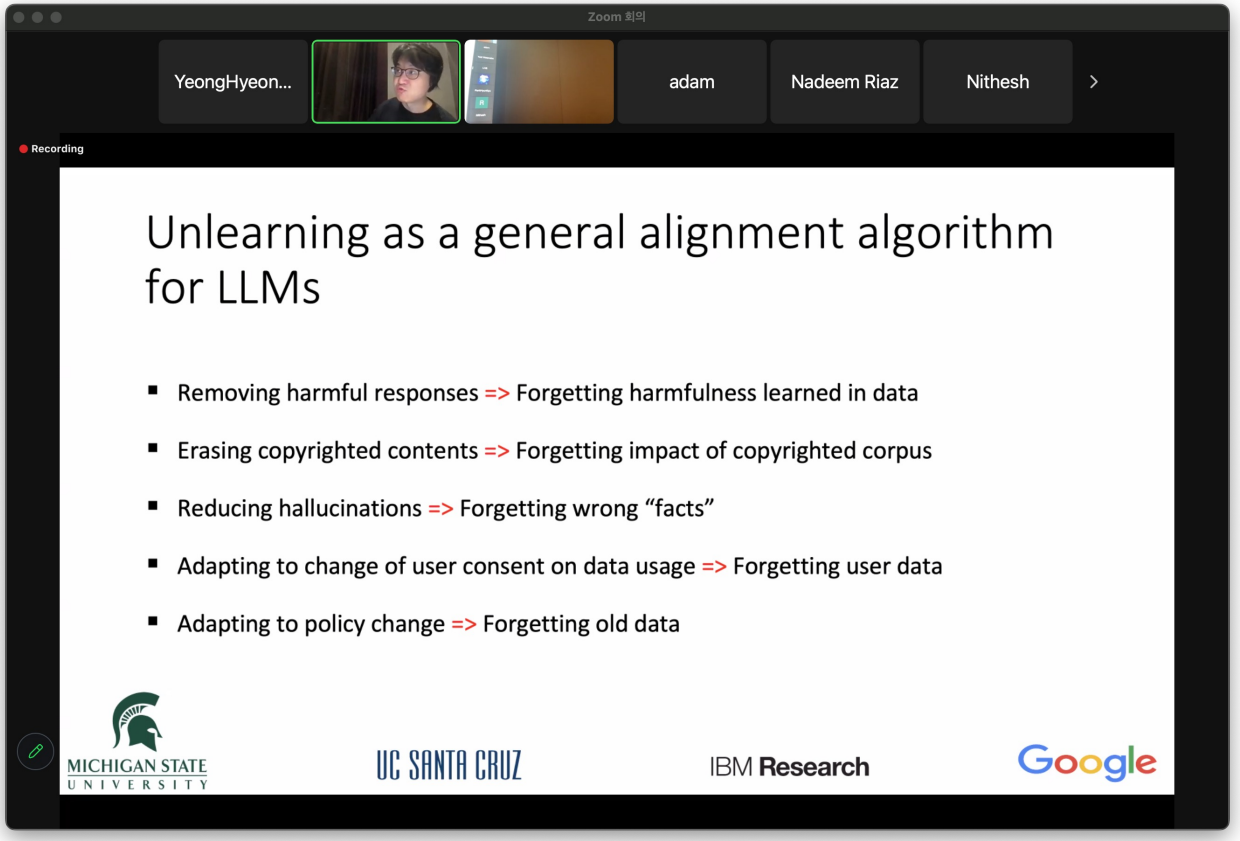
Time: 9:00 AM–12:00 AM

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

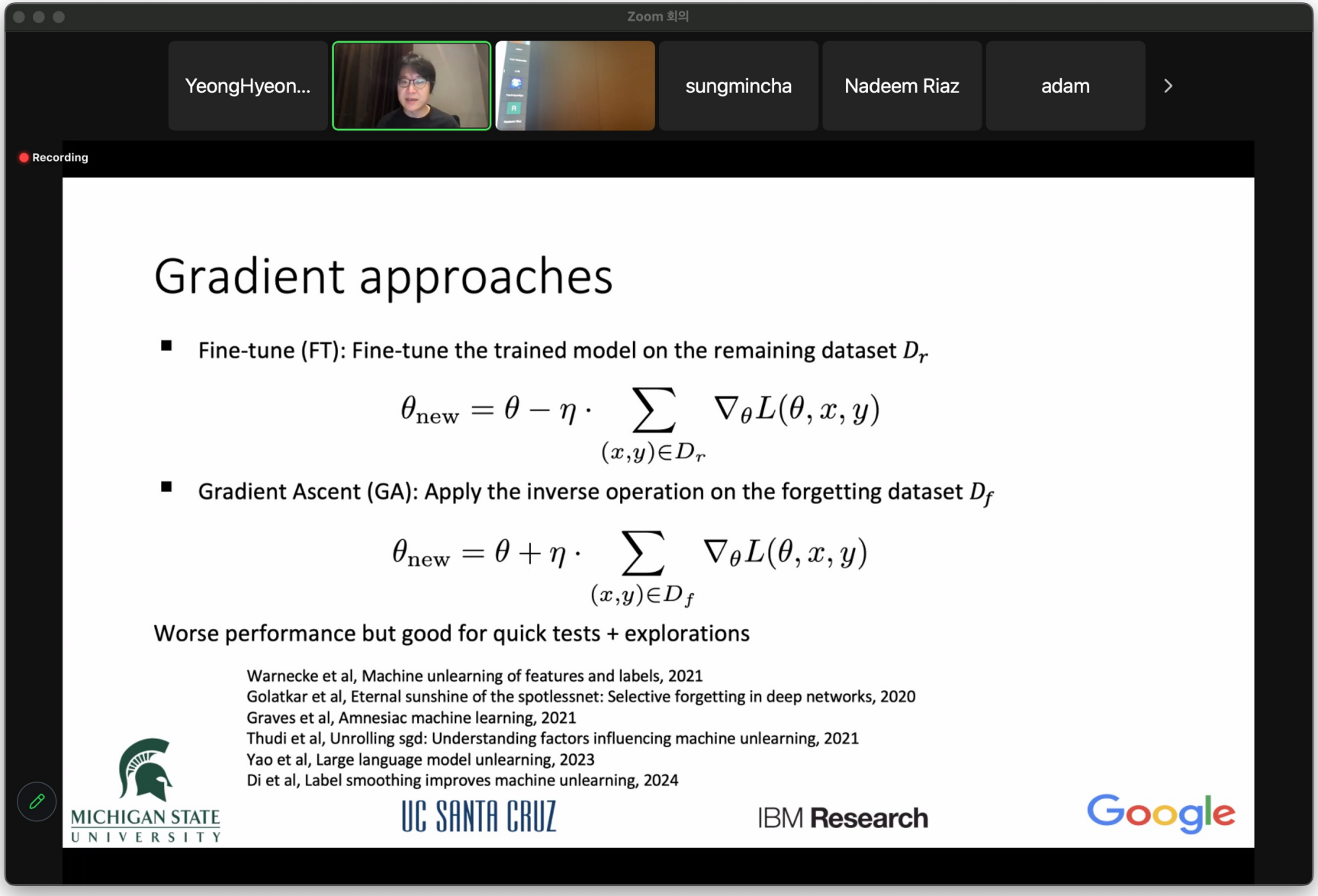




Rewinding of training process
to prevent unintentional information leakage

Machine Unlearning in Computer Vision

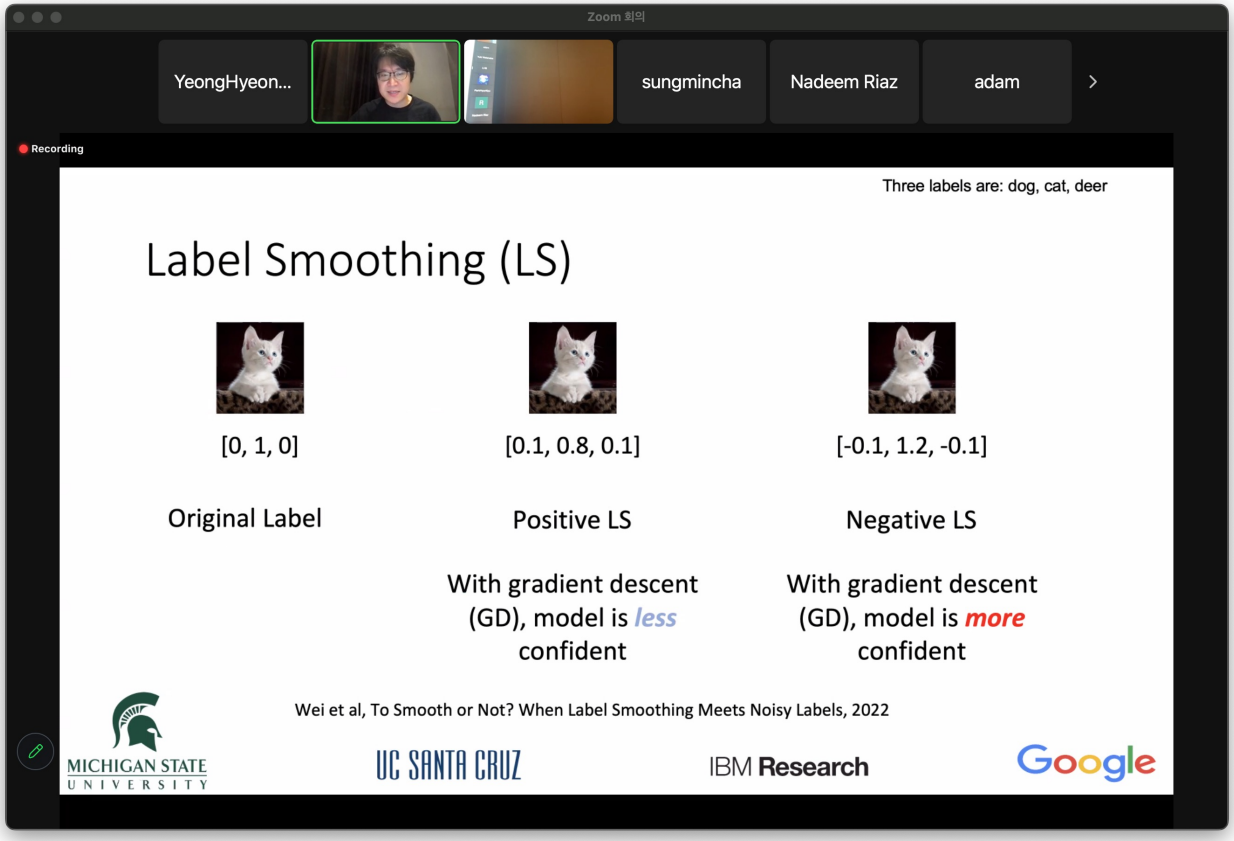
Gradient approach

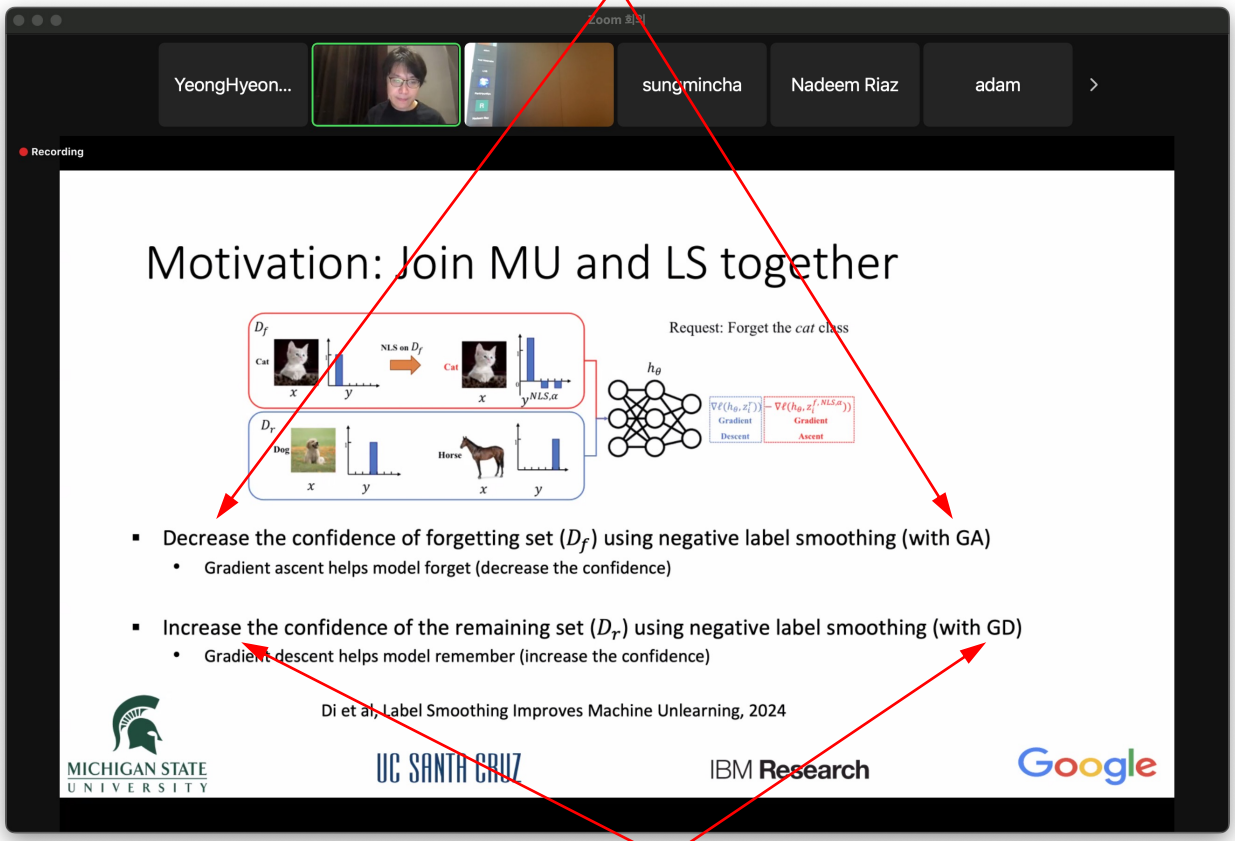


Machine Unlearning in Computer Vision

Gradient approach

Soft knowledge forgetting





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** $\boldsymbol{m}$, the unlearning error, $e(\boldsymbol{m})$, characterized by weight distance between **an approximate unlearn** and the **exact unlearn (retrain)** yields [Jia. Liu, et al., NeurIPS'23]

  $$e(\boldsymbol{m}) = \mathcal{O}(|\boldsymbol{m} \odot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)|_2)$$

  $\odot$ is entry-wise product, $\boldsymbol{\theta}_t$ 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

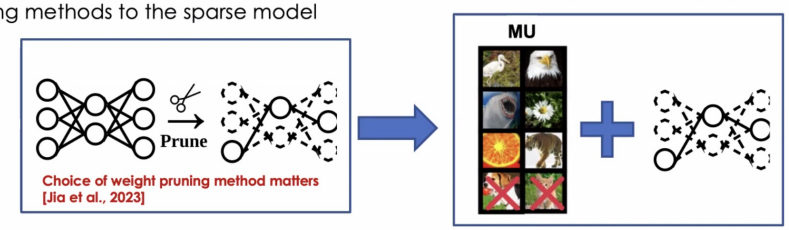
Zoom 회의

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



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

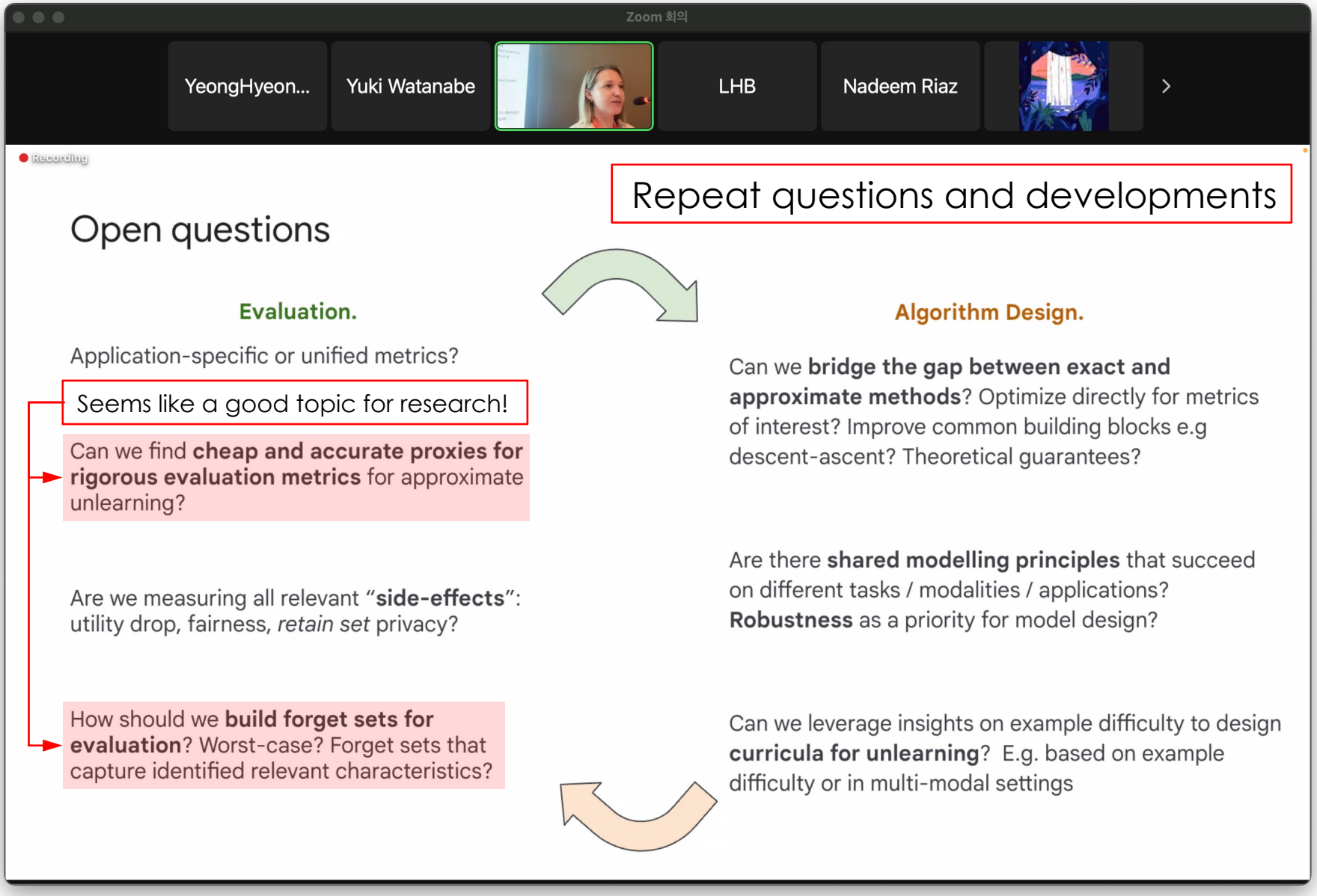
$$\boldsymbol{\theta}_u = \text{argmin}_{\boldsymbol{\theta}} L_{MU}(\boldsymbol{\theta};\ \mathcal{D}_r) + \gamma\|\boldsymbol{\theta}\|_1$$

MU objective function on retain dataset $\mathcal{D}_r$ — ℓ_1 sparse regularization

MICHIGAN STATE UNIVERSITY UC SANTA CRUZ IBM Research Google

Machine Unlearning in Computer Vision

Better way to research



Day 1

Robustness at Inference:
Towards Explainability, Uncertainty, and Intervenability

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

Organizers: Ghassan AlRegib, Mohit Prabhushankar

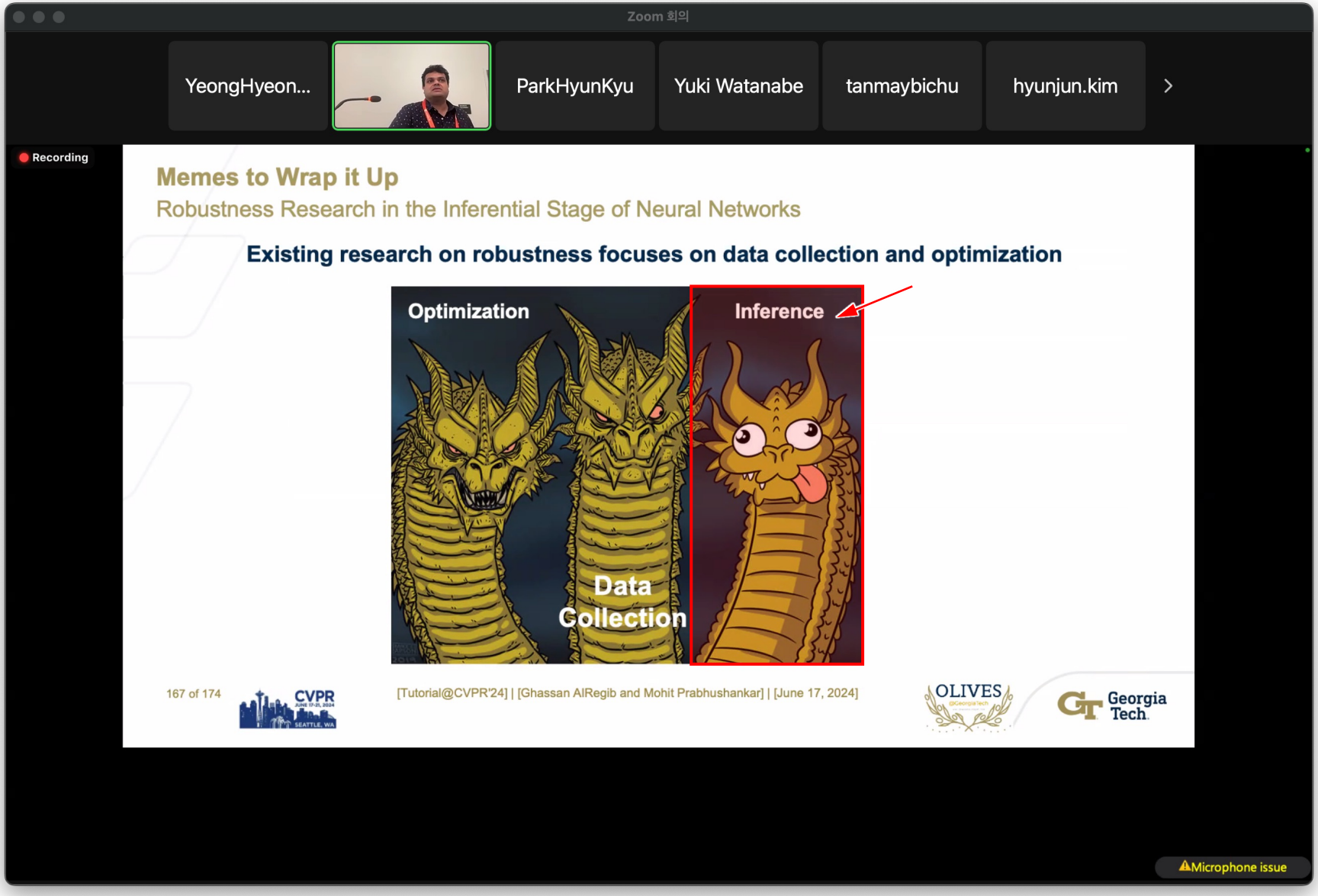
Date: Monday, June 17

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

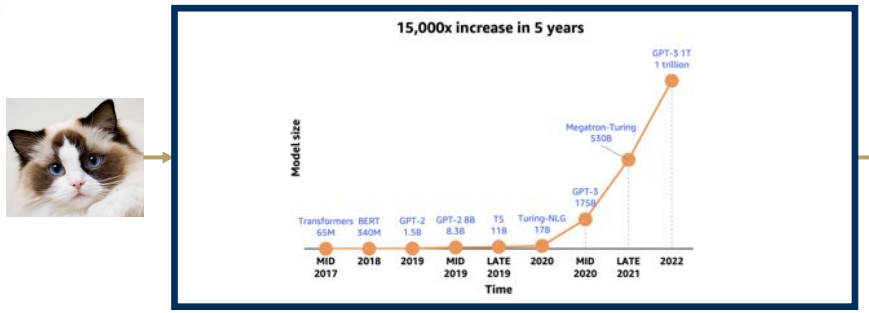


Robustness at inference

Deep Deep Deep Deep Deep … Learning
Recent Advancements

Development of DLs

Transformers, Large Language Models and Foundation Models

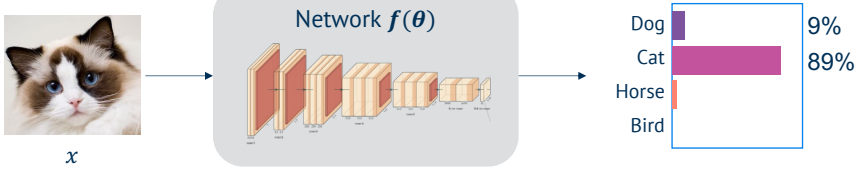


Primary reasons for advancements:
1. Expanded interests from the research community
2. Computational resources availability
3. **Big data availability**

Deep Learning at Inference
Classification

Seen sample

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



$\hat{y} = f(x)$
$y = argmax_i\ \hat{y}$
$p(\hat{y}) = T(f(x))$

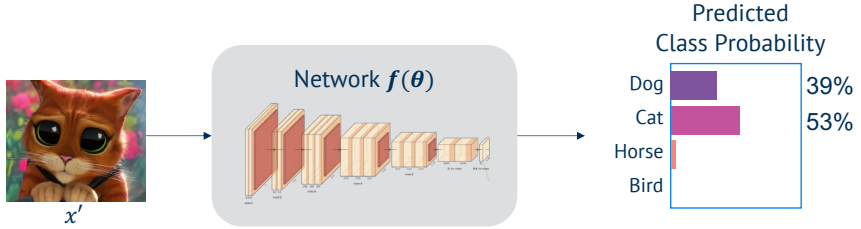
$\hat{y}$ = Logits
y = Predicted Class
$p(\hat{y})$ = Probabilities
$f(\cdot)$ = Trained Network
χ = Training data

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

Deep Learning at Inference
Robust Classification in Deep Networks

Unseen sample

Deep learning robustness: Correctly predict class even when data is novel



To achieve robustness at Inference, we need the following:

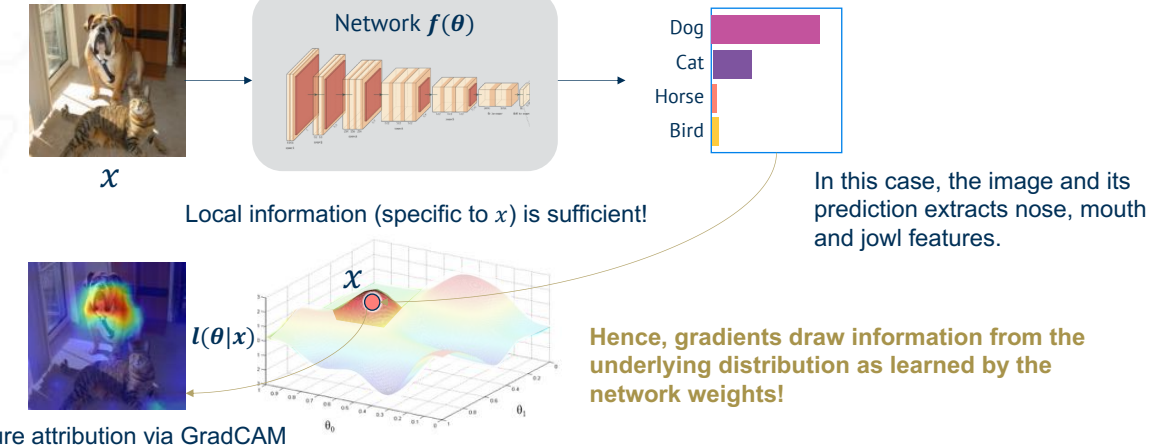
- **Information** provided by the novel data as **a function of training distribution**
- Methodology to **extract information** from novel data
- **Techniques** that utilize the information from novel data

Why is this Challenging?

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

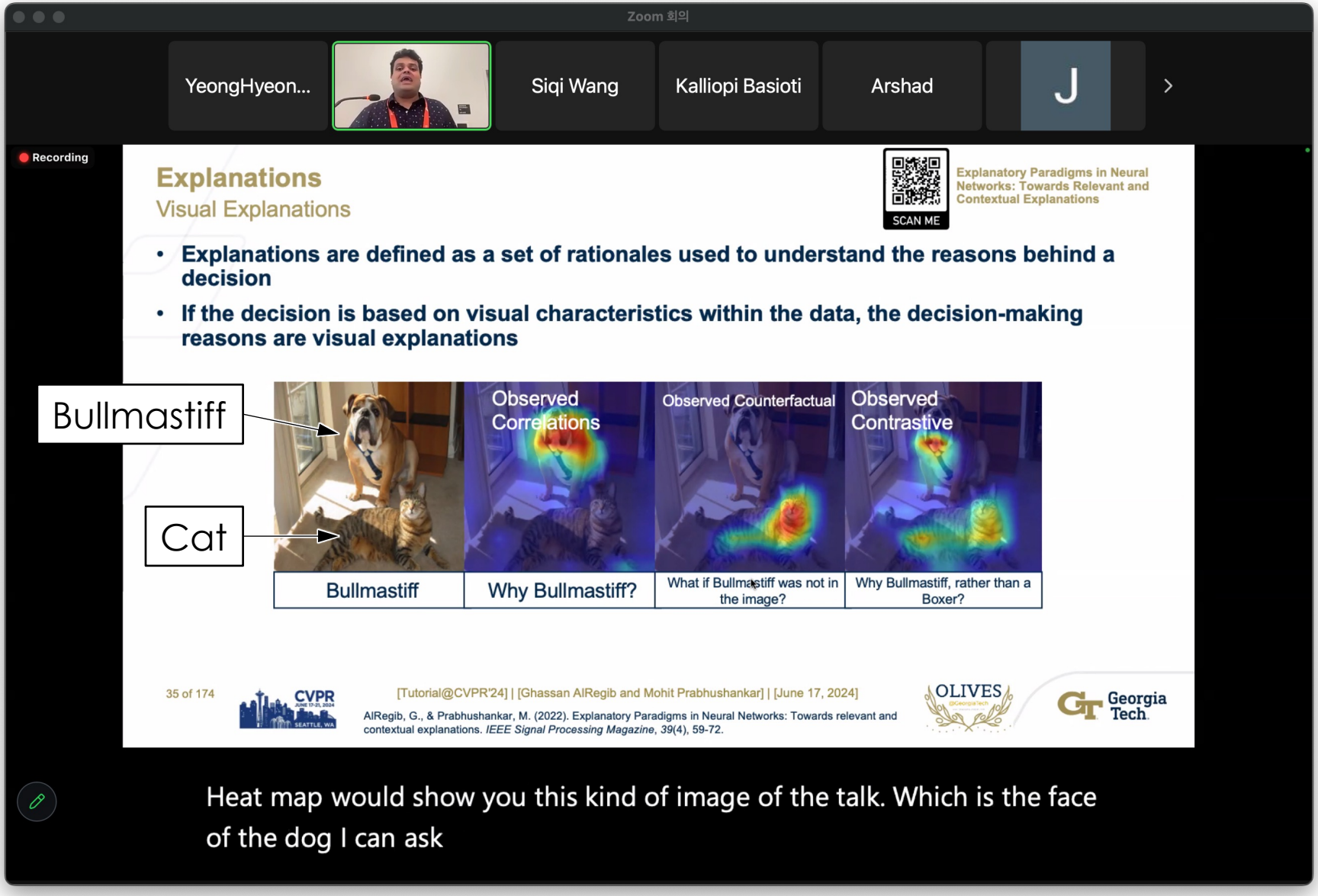
Visual reasoning

Gradients infer information about the statistics of underlying manifolds



Robustness at inference

Visual reasoning



Robustness at inference

Visual reasoning

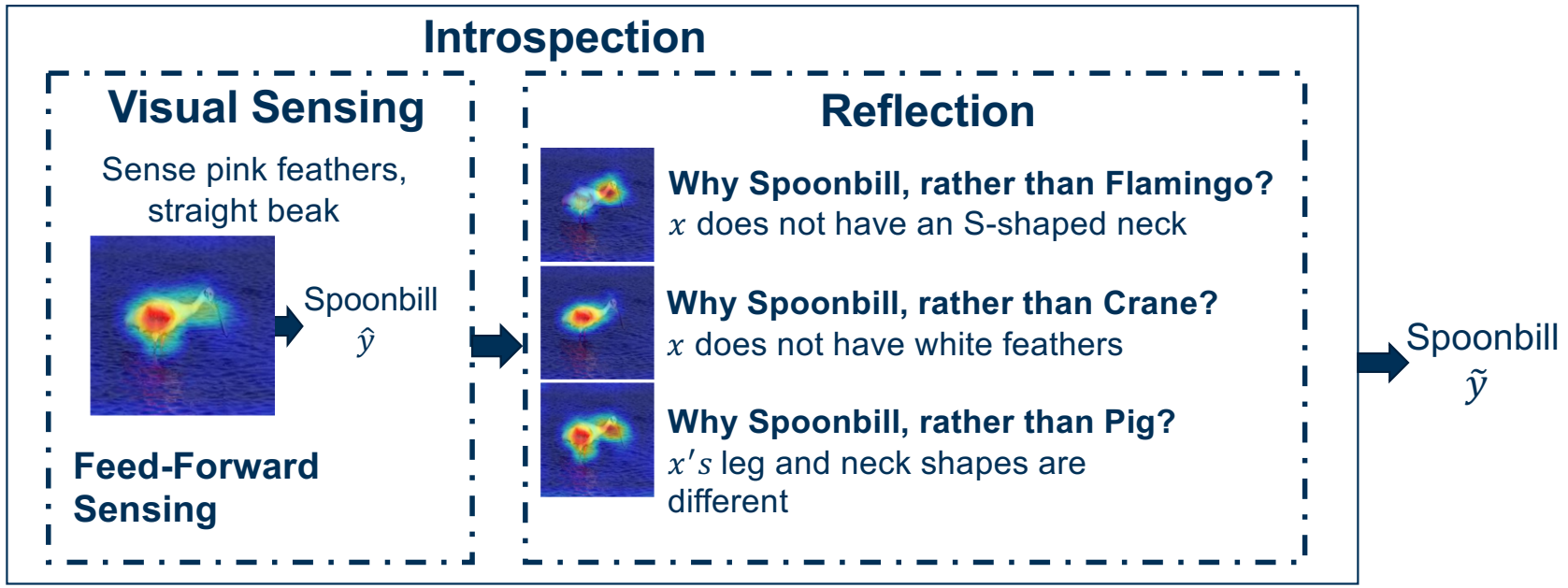
Introspection

What is Introspection?

SCAN ME

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



M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

Robustness at inference
Uncertainty of inference (Inference by multiple models)

Uncertainty
What is Uncertainty?

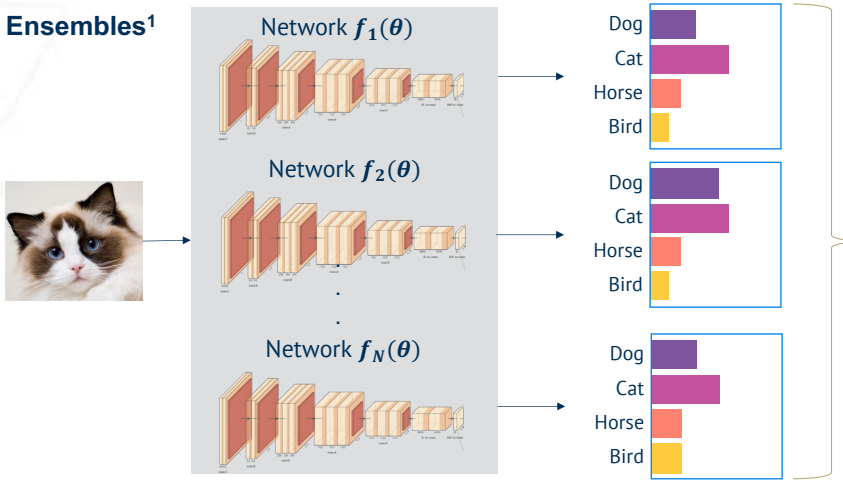
Uncertainty is a model knowing that it does not know

Input Image | **Neural Network Output** | **Uncertainty Heatmap**

Kendall, Gal "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision." *NIPS 2017*

Uncertainty
Uncertainty Quantification in Neural Networks

Via Ensembles[1]



Variation within outputs is the uncertainty.

Commonly referred to as **Prediction Uncertainty.**

Requires multiple trained models – not exactly an inferential method

[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty measuring by multiple models
But not effective approach

Robustness at inference

Uncertainty of inference (Iterative inference by dropout)

Uncertainty

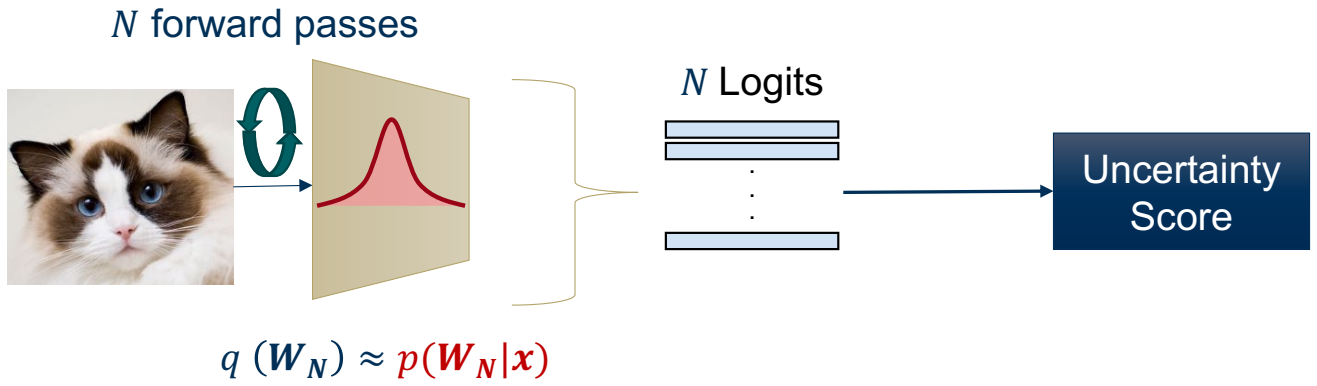
Iterative Uncertainty Quantification

Via Monte-Carlo Dropout[1]: 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

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



$$q(\boldsymbol{W_N}) \approx p(\boldsymbol{W_N}|\boldsymbol{x})$$

Final prediction is the mean of the outputs

Variation or entropy of logits is the uncertainty

[1] Y Gal, Z Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", ICML 2016

Robustness at inference

Uncertainty of inference (Multiple inference by multiple masking)

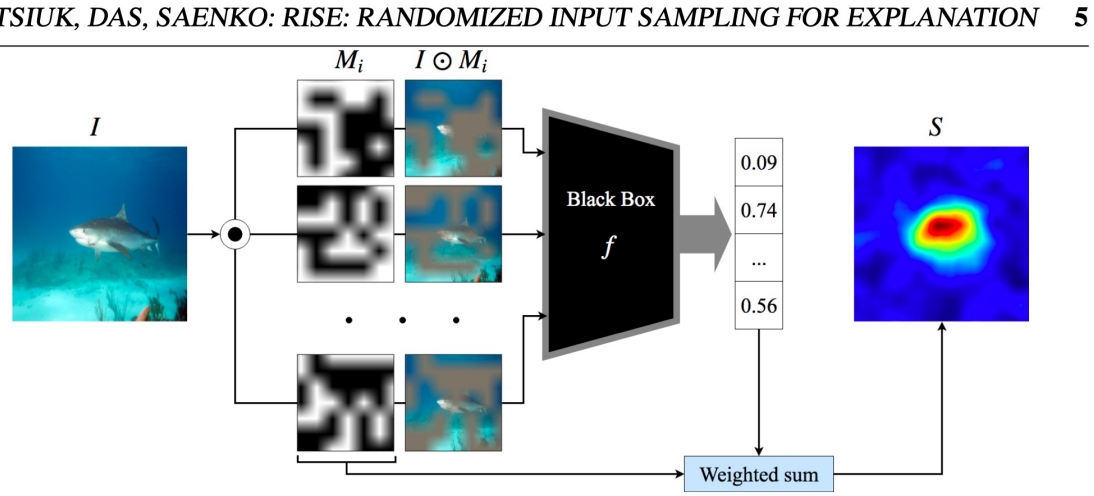


Figure 3: Overview of RISE: Input image I is element-wise multiplied with random masks M_i and the masked images are fed to the base model. The saliency map is a linear combination of the masks where the weights come from the score of the target class corresponding to the respective masked inputs.

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

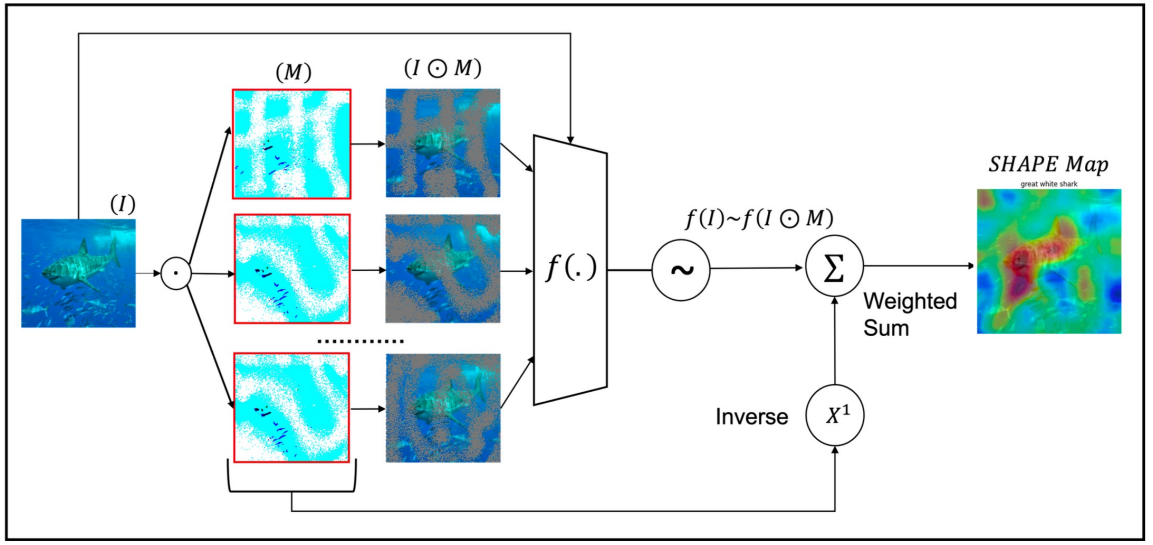


Fig. 2. Overview of **SHAPE** to generate adversarial explanations: Input image I is element-wise multiplied by the random masks M_i and are fed into the model f along with the original image to calculate the change in prediction scores.The importance map is a weighted sum of masks where the weights 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, Shobhita Sundaram, Wei-Chiu Ma, Phillip Isola, Ranjay Krishna

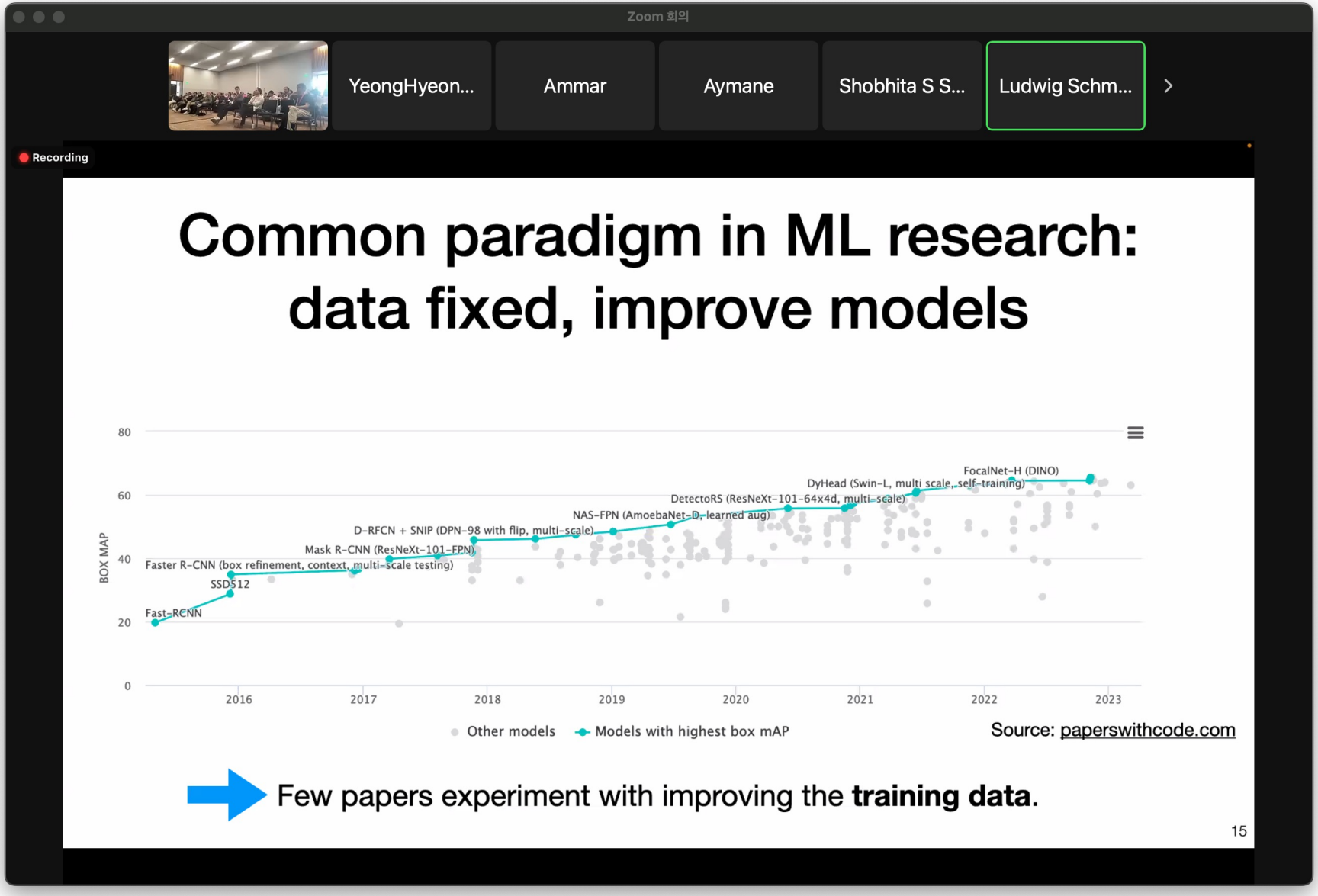
Date: Tuesday, June 18

Time: 8:30 AM–5:30 PM

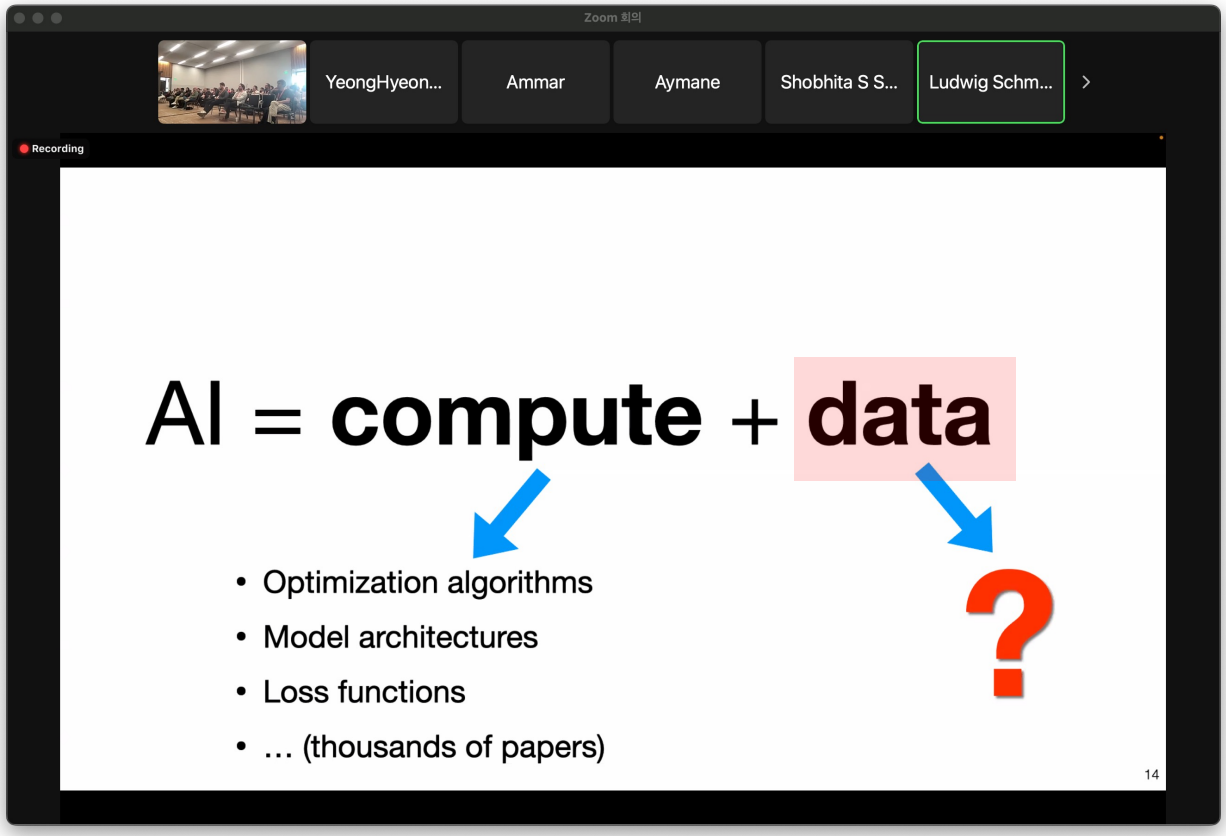
Location: Summit 423-425

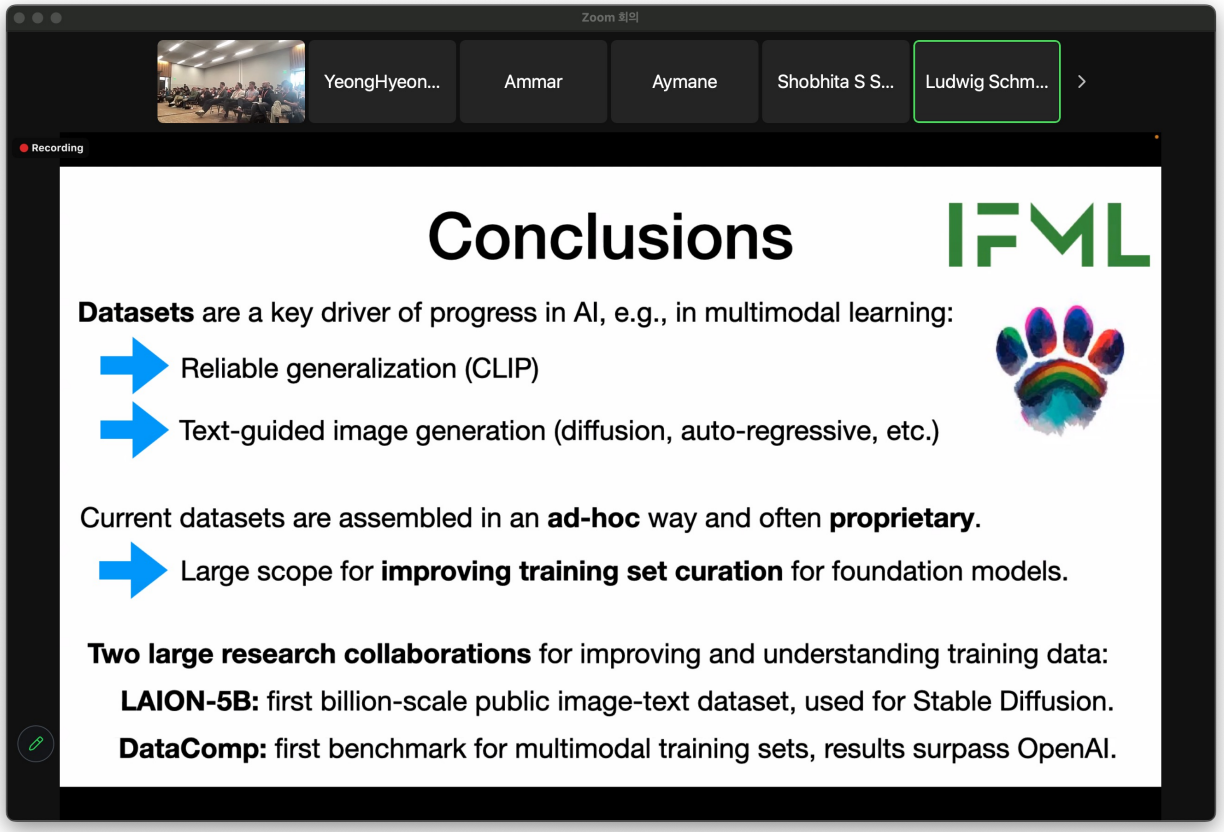
Summary: The workshop aims to explore the use of synthetic data in training and evaluating computer vision models, as well as in other related domains. During the last decade, advancements in computer vision were catalyzed by the release of painstakingly curated human-labeled datasets. Recently, people 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



Synthetic Data for Computer Vision

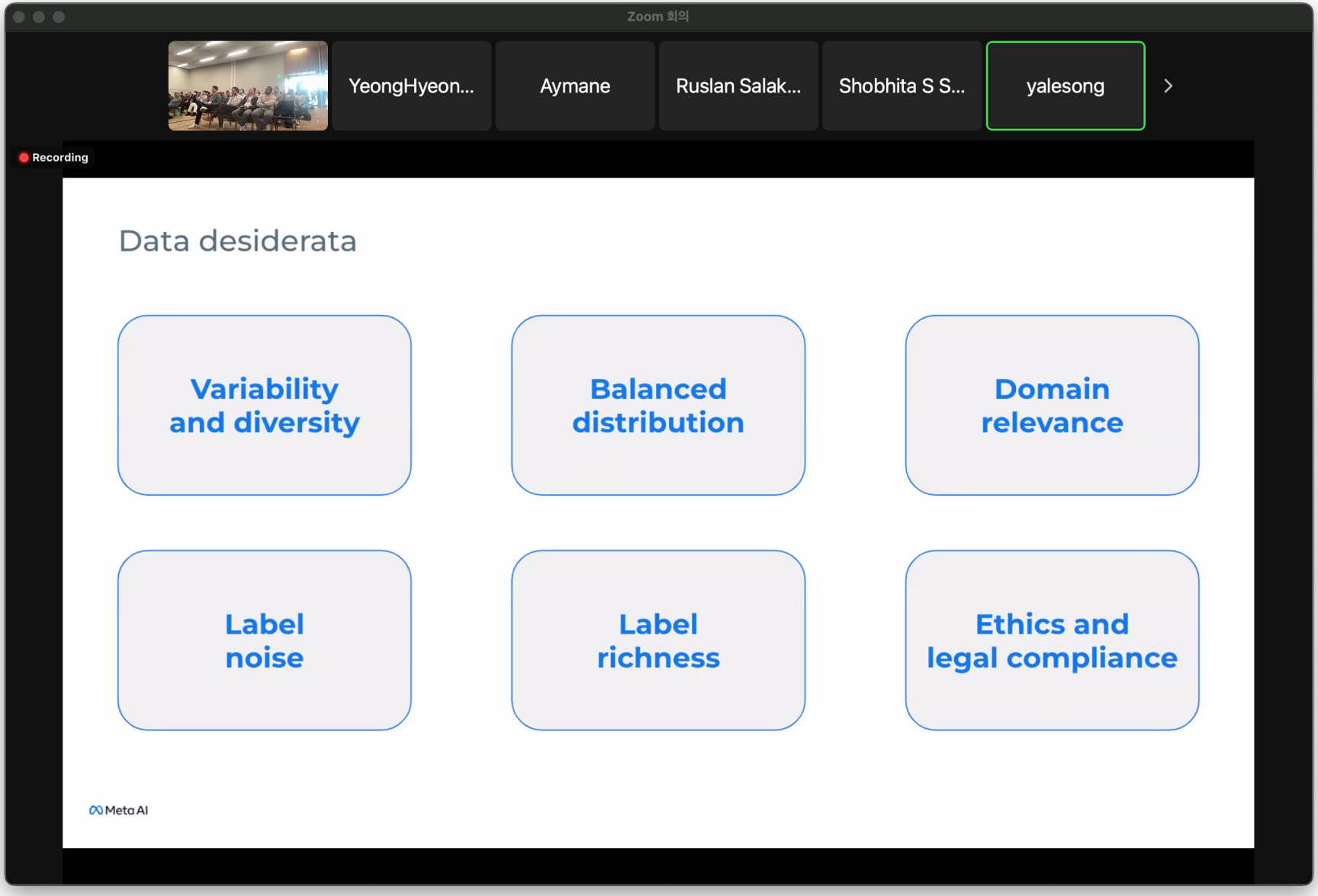




Data! Data! Data!

Synthetic Data for Computer Vision

Why synthetic data?



Synthetic Data for Computer Vision

Why synthetic data?

Zoom 회의

YeongHyeon... Aymane Ruslan Salak... Shobhita S S... yalesong

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

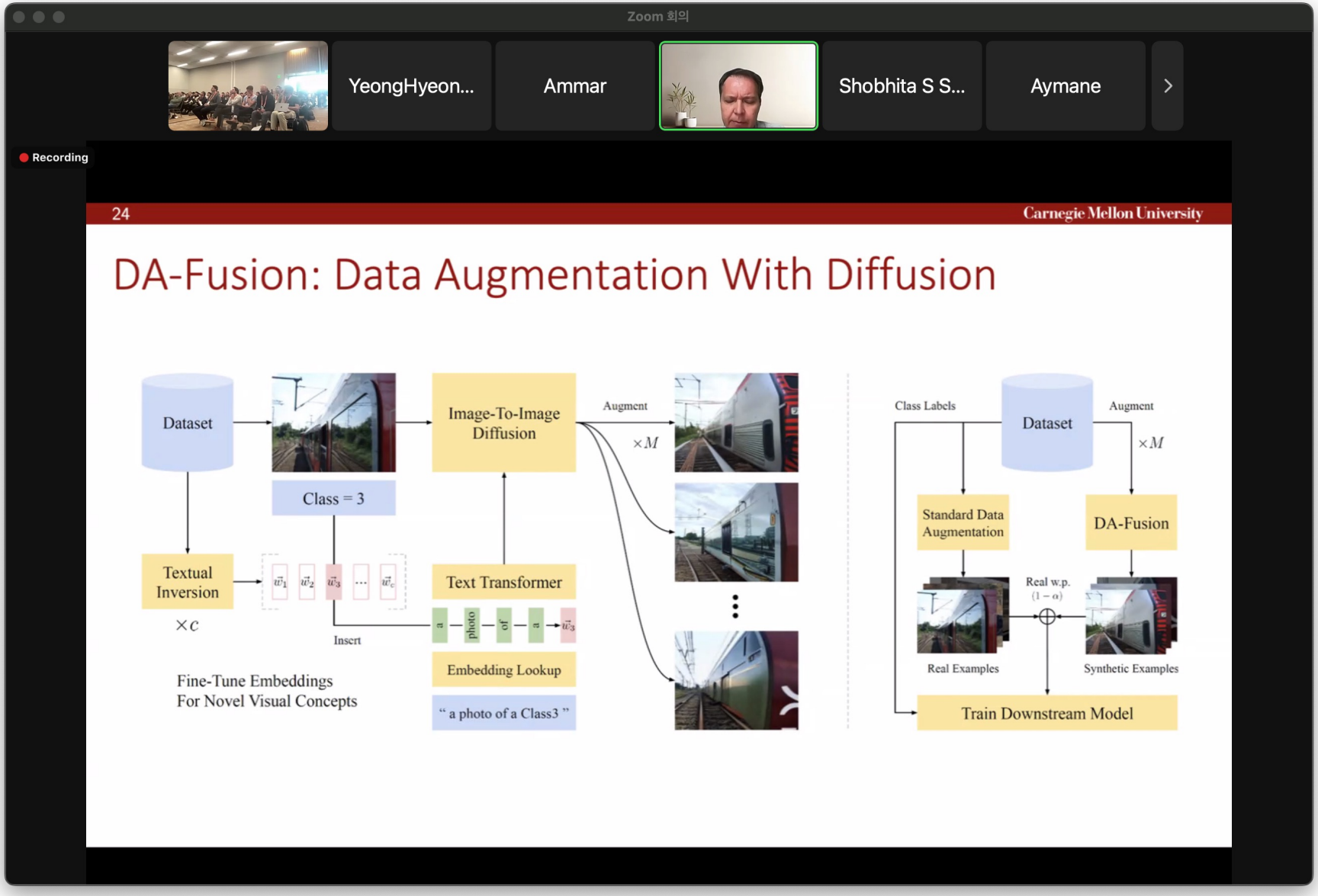
Fake it till you make it [Wood et al., ICCV 2021]

Meta AI

Synthetic data takes all the advantages

Synthetic Data for Computer Vision

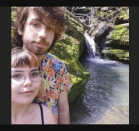
Diffusion usage



Synthetic Data for Computer Vision

Simulator usage (1)

Zoom 회의

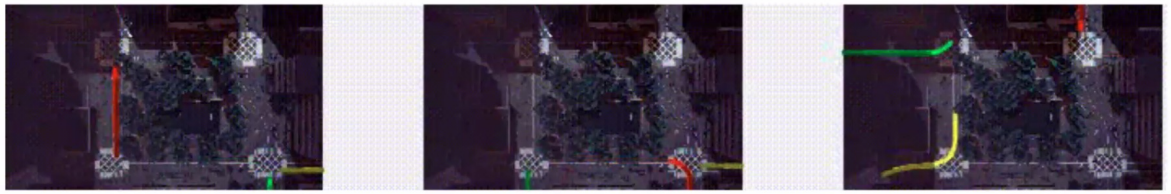

YeongHyeon...

Ruslan Salak...
Shobhita S S...
yalesong

Recording

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

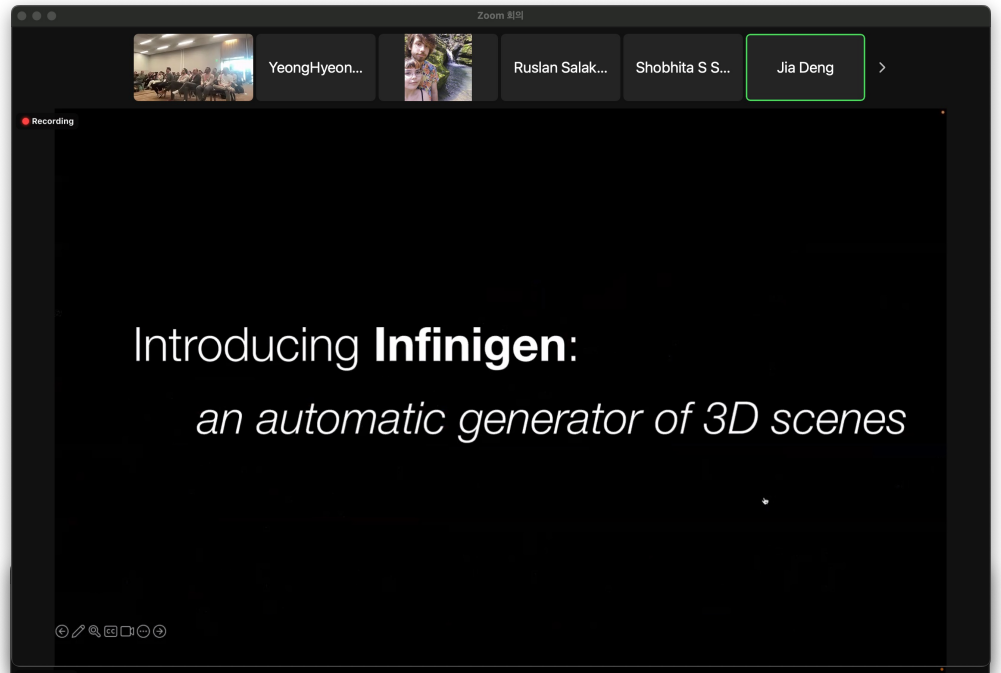


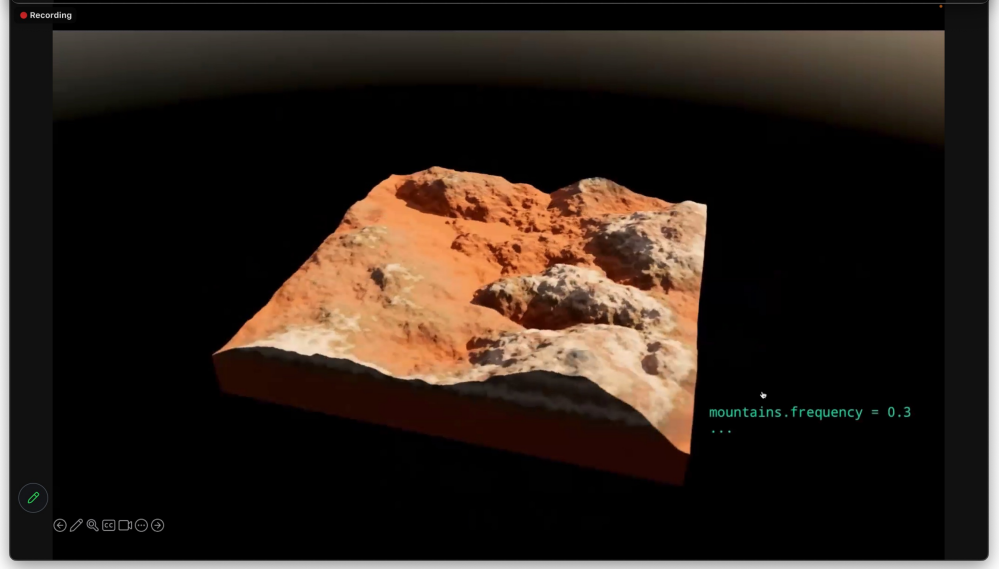
No "leader-follower" relationships, no confounders:

Meta AI

Synthetic Data for Computer Vision

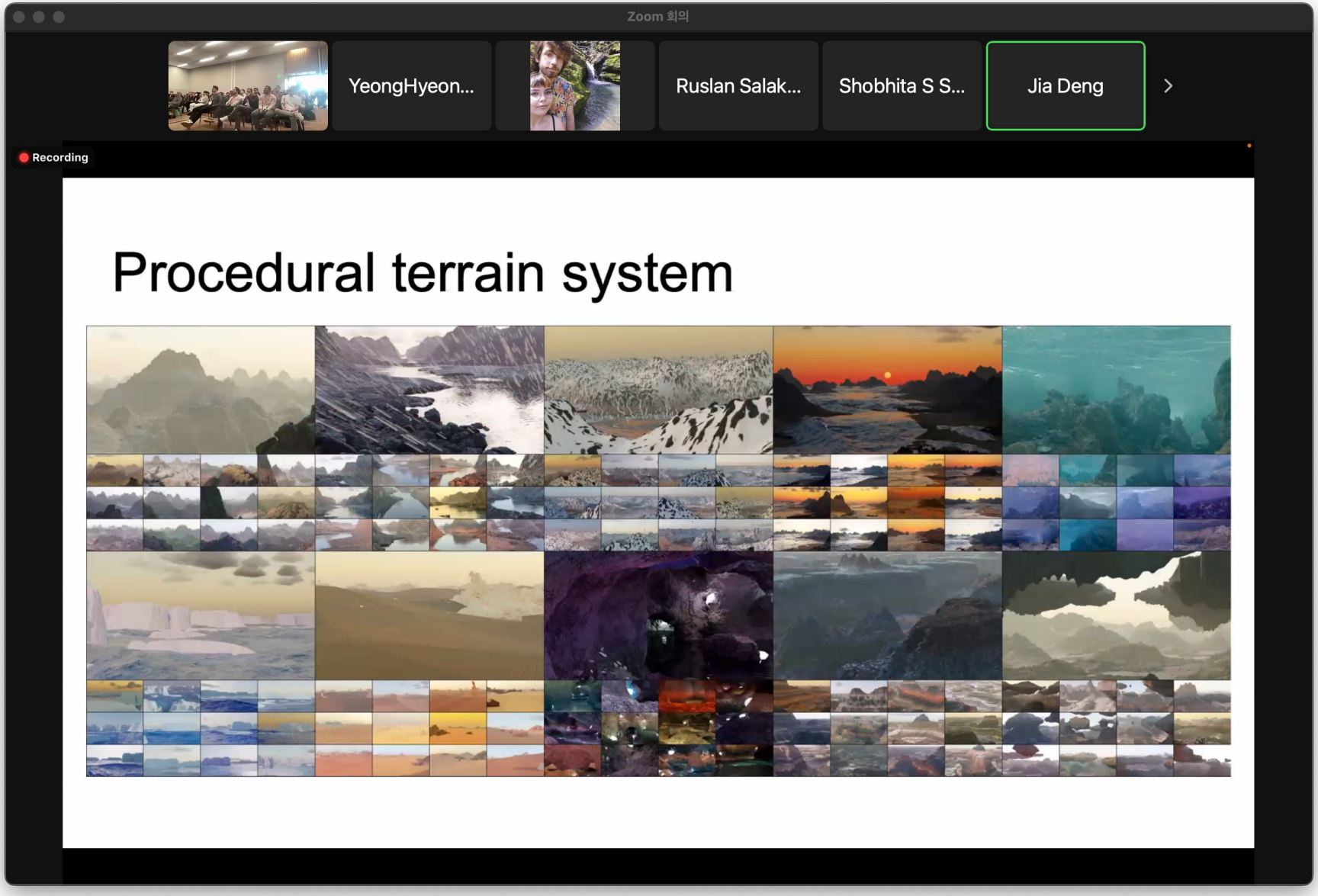
Simulator usage (2)





Synthetic Data for Computer Vision

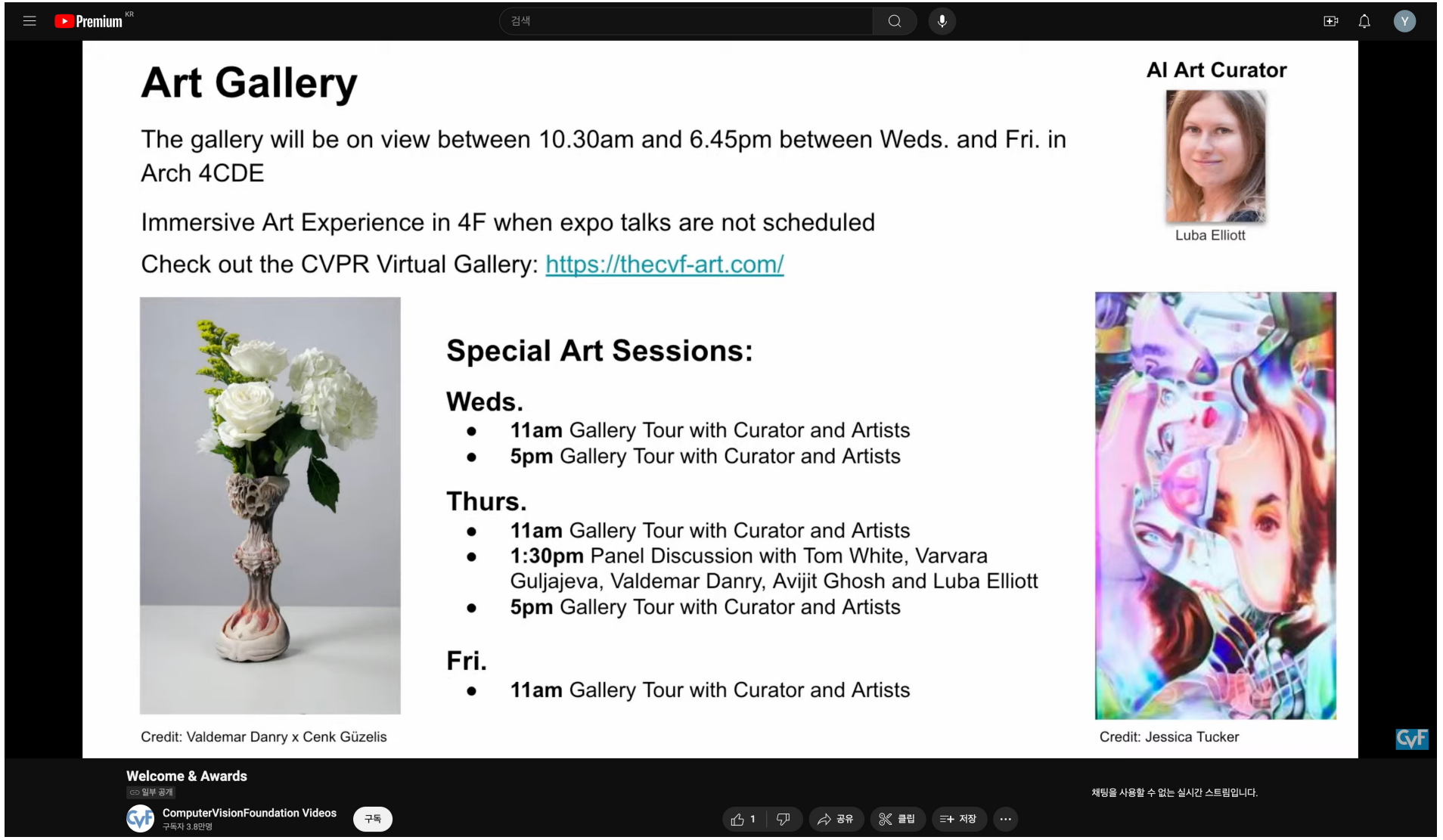
Simulator usage (2)



Day 3

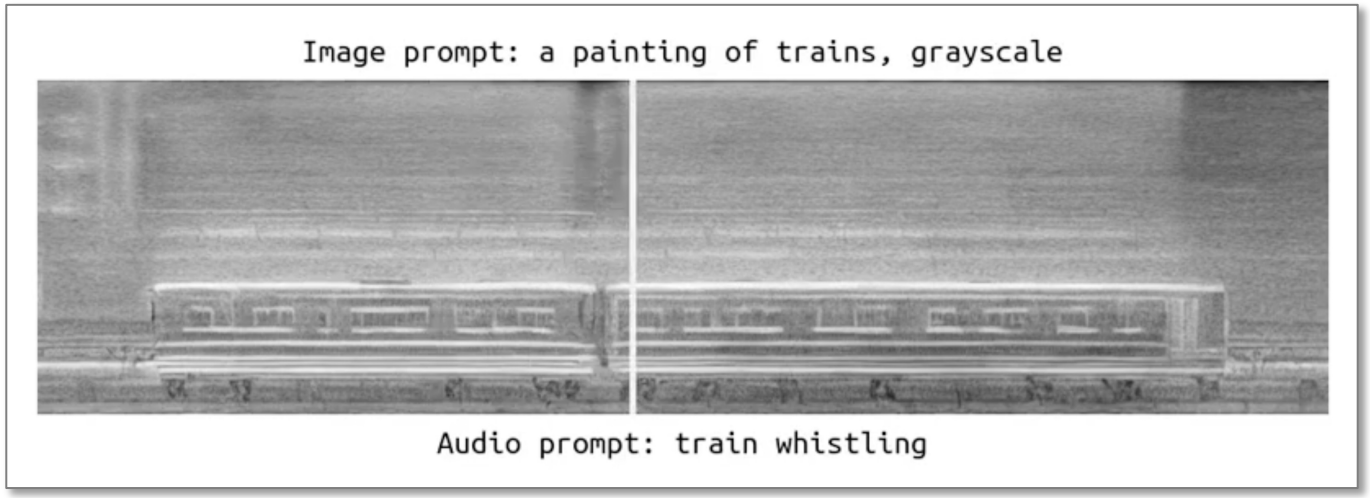
Art Gallery

Art Gallery



https://thecvf-art.com/

Art Gallery



https://thecvf-art.com/project/images-that-sound/

Interesting Paper

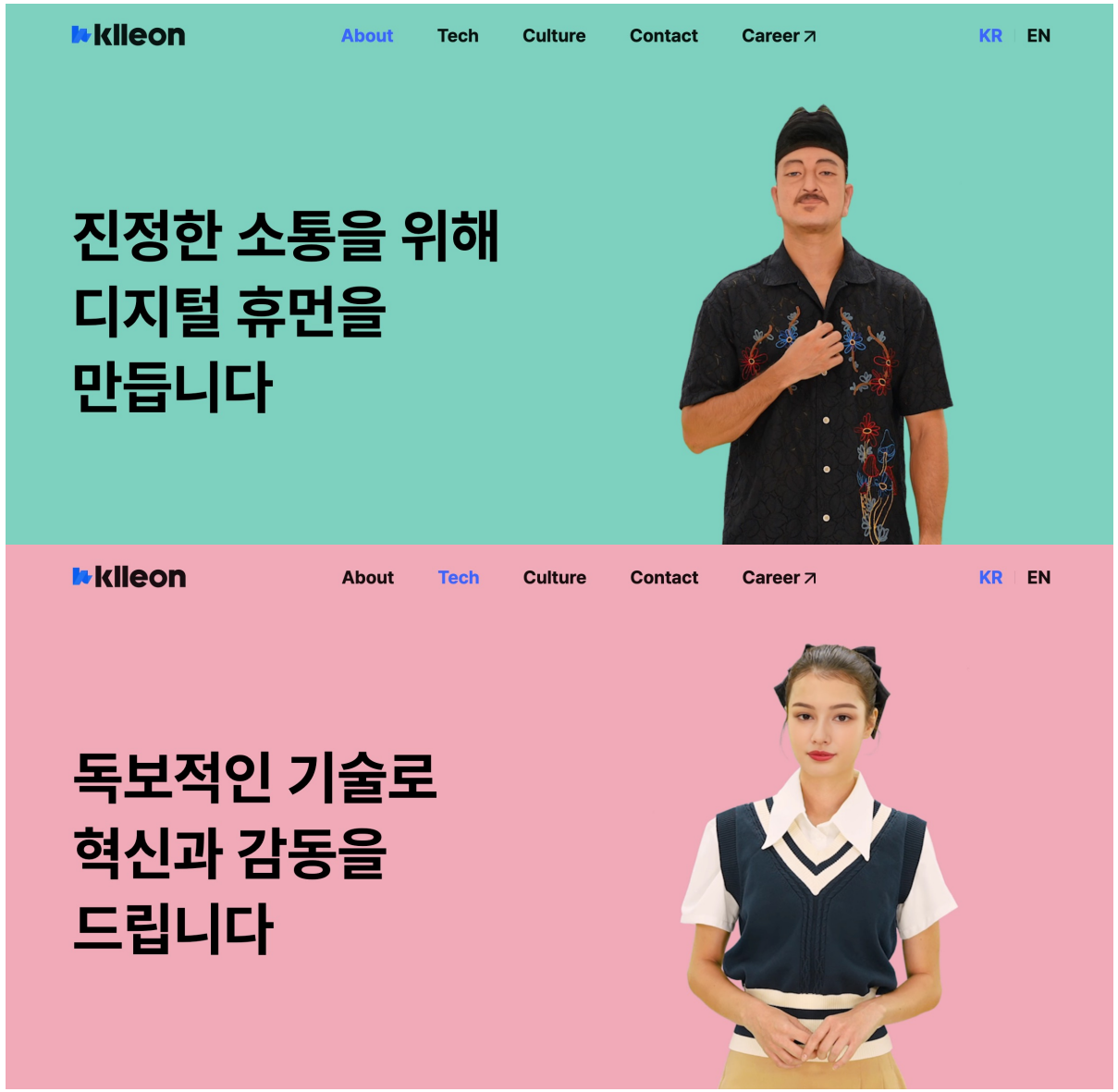
Generative Unlearning for Any Identity

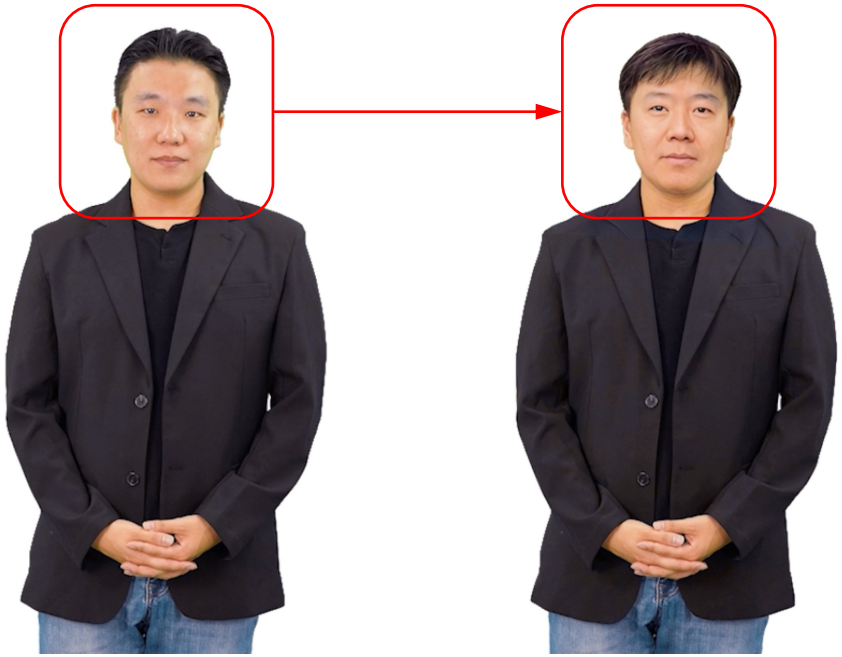
Juwon Seo[1*] Sung-Hoon Lee[1*] Tae-Young Lee[1*] Seungjun Moon[2]
Gyeong-Moon Park[1†]

[1]Kyung Hee University, Yongin, Republic of Korea
[2]KLleon Tech., Seoul, Republic of Korea

{jwseo001, sunghoonlee961, slcks1, gmpark}@khu.ac.kr
seungjun.moon@klleon.io

Kleon Tech


kleon
About
Tech
Culture
Contact
Career
KR
EN
진정한 소통을 위해
디지털 휴먼을
만듭니다
kleon
About
Tech
Culture
Contact
Career
KR
EN
독보적인 기술로
혁신과 감동을
드립니다



Generative Unlearning for Any Identity

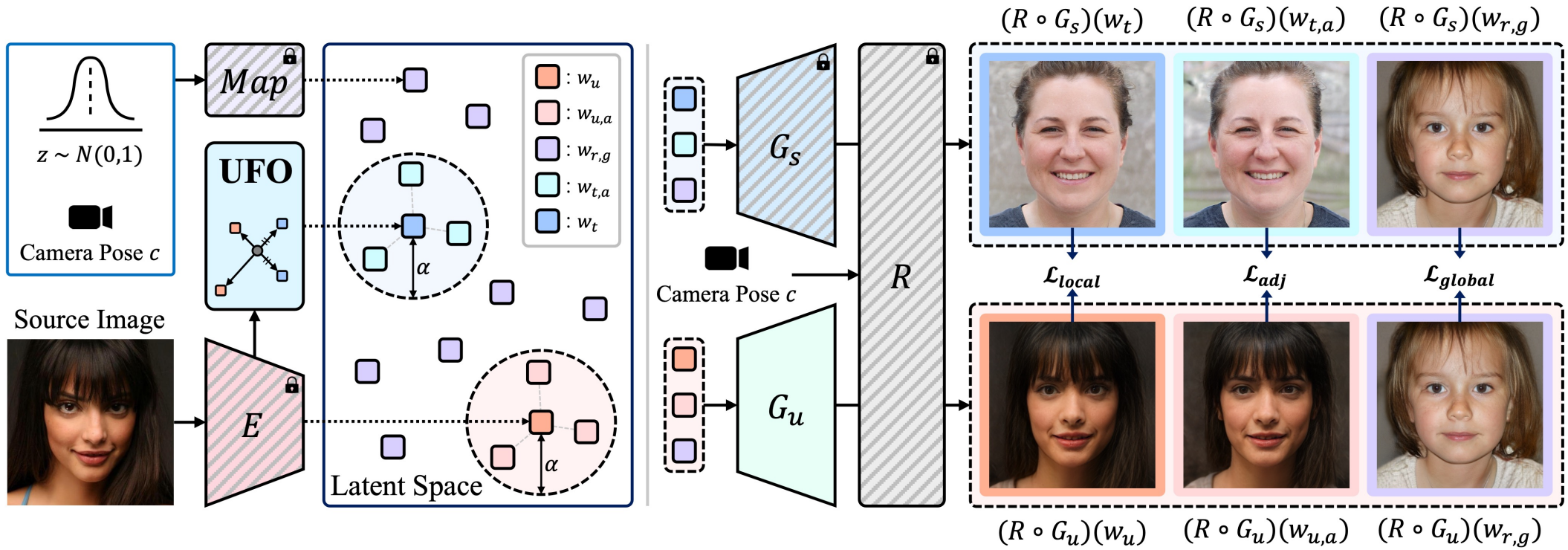


Figure 3. An overview of GUIDE. Starting with a source image, we employ a GAN inversion network E, specifically GOAE [52], to embed this image into the latent space of a pre-trained generative model, namely EG3D [4], obtaining the source latent code w_u. The target latent code w_t is designated through 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 $w_{r,g}$.

Generative Unlearning for Any Identity

Juwon Seo[1*], Sung-Hoon Lee[1*], Tae-Young Lee[1*], Seungjun Moon[2], Gyeong-Moon Park[1†]
[1]Kyung Hee University, Yongin, Republic of Korea [2]KLleon Tech, Seoul, Republic of Korea

CVPR SEATTLE, WA JUNE 17-21, 2024

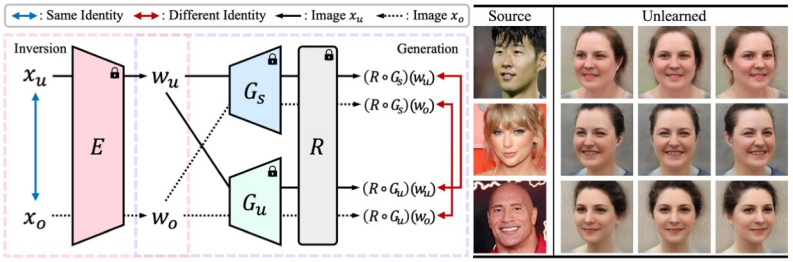
TL;DR

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).

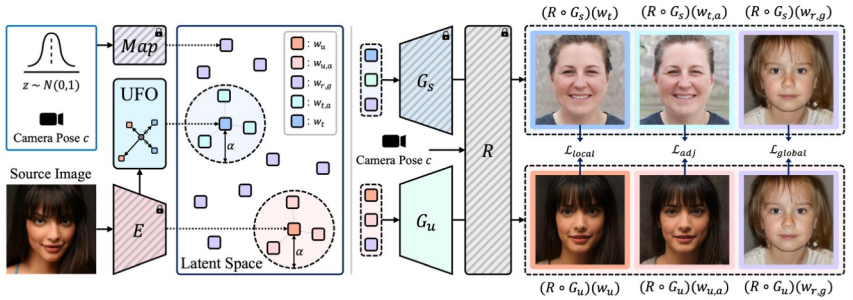
- **Task: Generative Identity Unlearning**



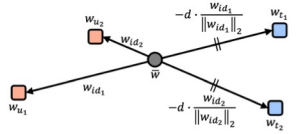
- **Goals**
  - **Prevent** pre-trained 3D GANs from generating a certain identity.
  - Erase the given identity effectively by utilizing a **single image**.
  - **Preserve** the performance of pre-trained models.

Method

- **Framework: G**enerative **U**nlearning for Any **IDE**ntity (**GUIDE**)



- **Un-Identifying Face On latent space (UFO)**
  - Goal: Determine the **effective target identity**.
  - Approach: Extrapolation between w_u and $\overline{w}$.



- **Latent Target Unlearning (LTU)**

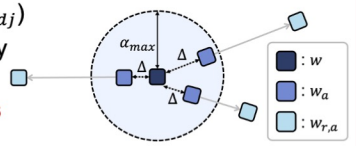
1. Local Unlearning Loss ($\mathcal{L}_{local}$)

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

$$\mathcal{L}_{local}(\hat{x}_u, \hat{x}_t) = \lambda_{L2}\mathcal{L}_{L2}(F_u, F_t) + \lambda_{per}\mathcal{L}_{per}(\hat{x}_u, \hat{x}_t) + \lambda_{id}\mathcal{L}_{id}(\hat{x}_u, \hat{x}_t).$$

2. Adjacency-Aware Unlearning Loss ($\mathcal{L}_{adj}$)

- Goal: **Effectively erase** identity with only utilizing single image.
- Approach: Approximate **neighborhoods** of both source and target. $\mathcal{L}_2, \mathcal{L}_{per}$, and $\mathcal{L}_{id}$, between neighborhoods.

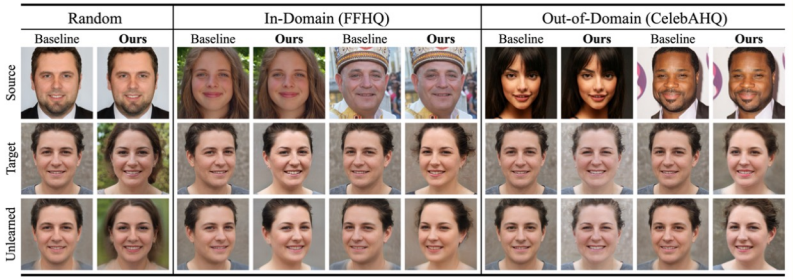


3. Global Preservation Loss ($\mathcal{L}_{global}$)

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

Experiments

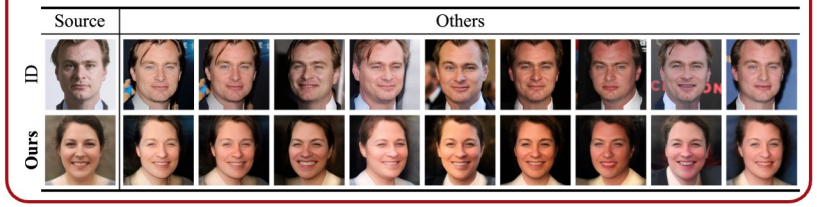
- **Qualitative** Results of GUIDE



- **Quantitative** Results of GUIDE

Method	Random			In-Domain (FFHQ)			Out-of-Domain (CelebAHQ)		
	ID (↓)	FID_{pre} (↓)	ΔFID_{real} (↓)	ID (↓)	FID_{pre} (↓)	ΔFID_{real} (↓)	ID (↓)	FID_{pre} (↓)	ΔFID_{real} (↓)
Baseline	0.19 ±0.09	11.73 ±2.74	7.46 ±2.20	0.16 ±0.07	9.00 ±1.15	4.15 ±1.18	0.12 ±0.06	9.52 ±1.53	4.75 ±0.89
+ extrapolated w_t	**0.12** ±0.06	14.28 ±3.34	9.63 ±2.53	0.05 ±0.06	12.78 ±1.82	6.76 ±1.41	0.02 ±0.05	13.02 ±3.20	7.31 ±1.98
+ $\mathcal{L}_{adj}$	0.14 ±0.07	19.65 ±4.90	13.94 ±3.59	**0.04** ±0.06	13.53 ±2.08	7.35 ±1.70	**0.01** ±0.05	13.63 ±3.52	7.83 ±2.19
+ $\mathcal{L}_{global}$ **(GUIDE)**	0.14 ±0.06	**10.80** ±2.70	**6.64** ±1.60	0.06 ±0.06	**8.00** ±1.20	**3.05** ±0.81	0.03 ±0.05	**7.88** ±1.96	**3.34** ±1.10

- **Erasing the Entire Identity** using GUIDE



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.