Language-Driven 3D Stylization

CSCI-677: Advanced Computer Vision

Pranav Budhwant, Jingmin Wei, Xianshi Ma

Nov 28, 2023

USC

Introduction

Traditional Style Transfer

Given a set of 2D calibrated images, and a 2D style image, generate a 3D stylized radiance field.

Images from Zhang, Kai, et al. "Arf: Artistic radiance fields." ECCV 2022.

Previous Work & Limitations

- Most approaches [1-4], focus on stylizing **entire scenes**
  - Usually 2 stages:
    - Train a photo-realistic radiance field
    - Fine-tune the 3D scene representation

- Object-specific style-transfer methods [3] perform instance based style transfer, and suffer from artifacts.

- These approaches do not incorporate **language**, and don't allow **open-ended queries** for object selection/style selection.

[1Pei-Ze Chiang, et al. Stylizing 3d scene via implicit representation and hypernetwork. WACV 2022
[2]Yuechen Zhang, et al. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization, CVPR 2023
[3]Chong Bao, et al. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. CVPR 2023.
[4] Images from Zhang, Kai, et al. "Arf: Artistic radiance fields." ECCV 2022.

USC

How can language help?

1. **Object Selection**
   - User specifies object(s) to be stylized in the scene
     - Eg: Table, TV, Flower, Fern, …
     - Allows semantic style transfer, instead of instance based

2. **Style Specification**
   - User specifies style(s) using language
     - Eg: “in the style of Vincent Van Gogh”, “floral print”, …

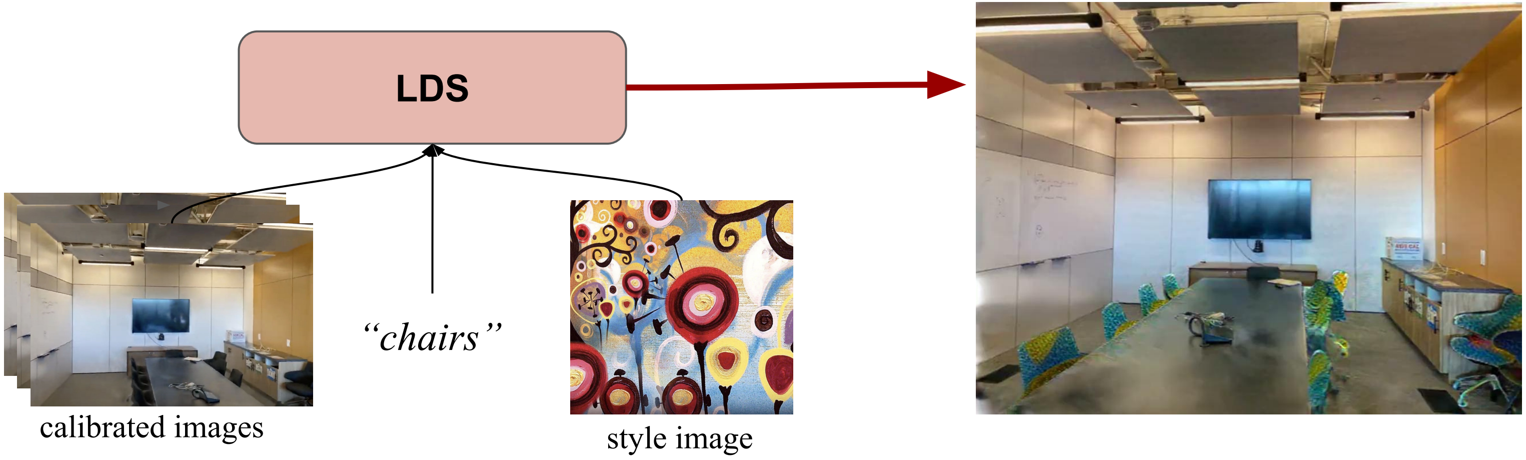
*In this work, **we focus on language driven object selection** and plan to extend our work to allow language based style specification.*

APPROACH

Overview

Inputs: Calibrated images of the 3D Scene, object query specified in natural language, and the style image.

1. Train a **photo-realistic radiance field** using calibrated images
2. Generate **semantic segmentation masks** for given object
3. Fine-tune (style-transfer) only the mask area using **VGG-based NNFM** [4]

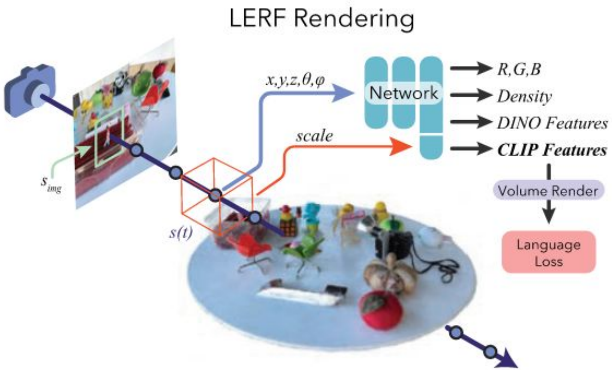


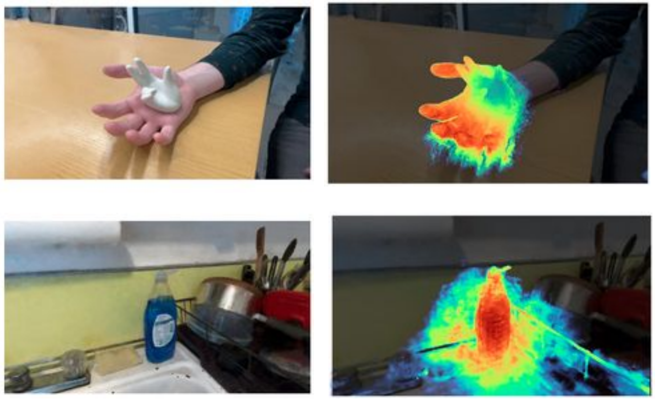
USC

[4]Zhang, Kai, et al. "Arf: Artistic radiance fields." *ECCV* 2022.

Approach 1: LERF (Language Embedded Radiance Fields) [5]

1. Train a LERF model
   a. Jointly optimize a language field along with a radiance field using CLIP+DINO
2. Use the user specified object to query the trained LERF model and obtain the **relevancy map**
3. Convert this relevancy map to a **segmentation mask**, by thresholding
4. Fine-tune the trained LERF model with **NNFM** for style transfer



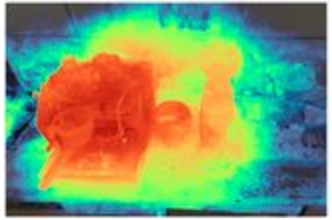


USC

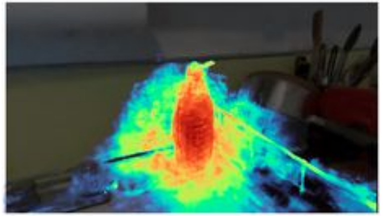
[5]Kerr, Justin, et al. "Lerf: Language embedded radiance fields." Proceedings of the IEEE/CVF ICCV 2023.

LERF Challenges

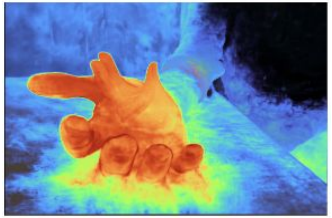
- Noisy Relevancy Maps



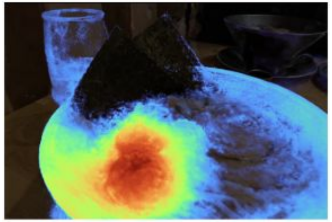
Espresso Machine



Blue Dish Soap



Hand



Eggs

- Expensive Compute
    - Training time: ~20min/epoch
    - Out of memory errors
    - Difficult to setup environment and dependencies

USC

Approach 2: GroundedSAM (*GroundingDINO*[6] + *SAM*[7])

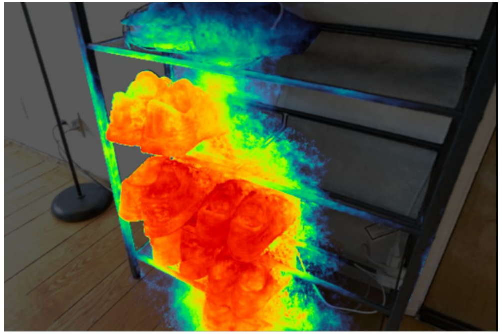
1. Train a radiance field on given calibrated images
2. Generate **object bounding boxes** for given object query using GroundingDINO
3. Pass the bounding boxes to SAM to generate **segmentation masks**
4. Fine-tune the pretrained radiance field with masked NNFM for style transfer

[6]Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." *arXiv preprint arXiv:2303.05499* (2023).
[7]Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).

Advantages over LERF

- Accurate Segmentation Masks

Query: *"shoes"*

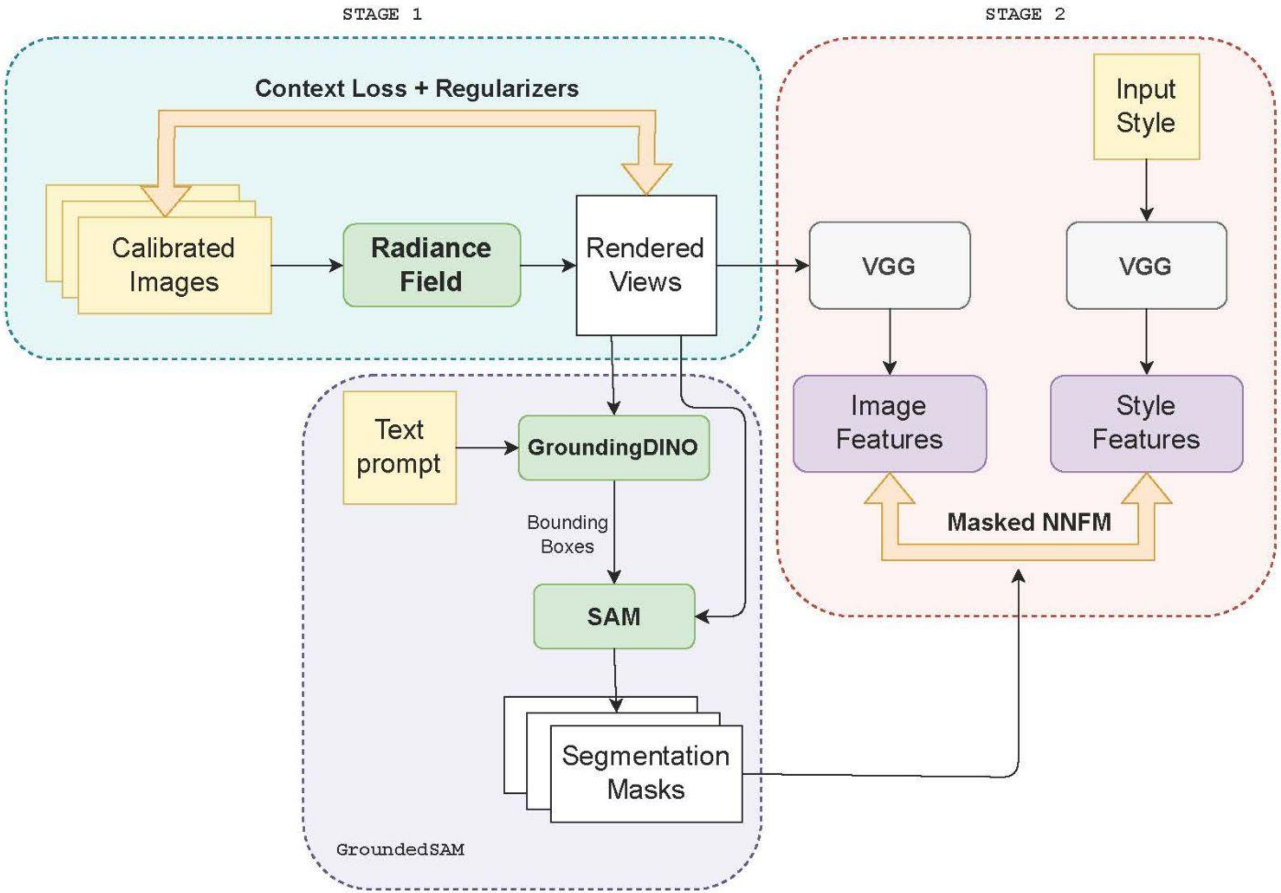


LERF

GroundedSAM

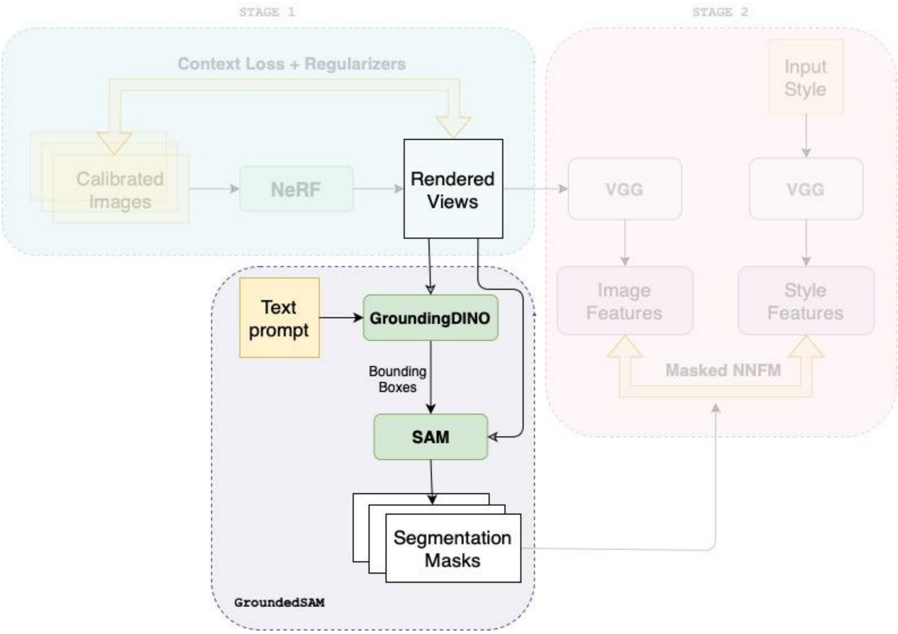
- Reduced compute requirements
  - ~45 minutes/experiment

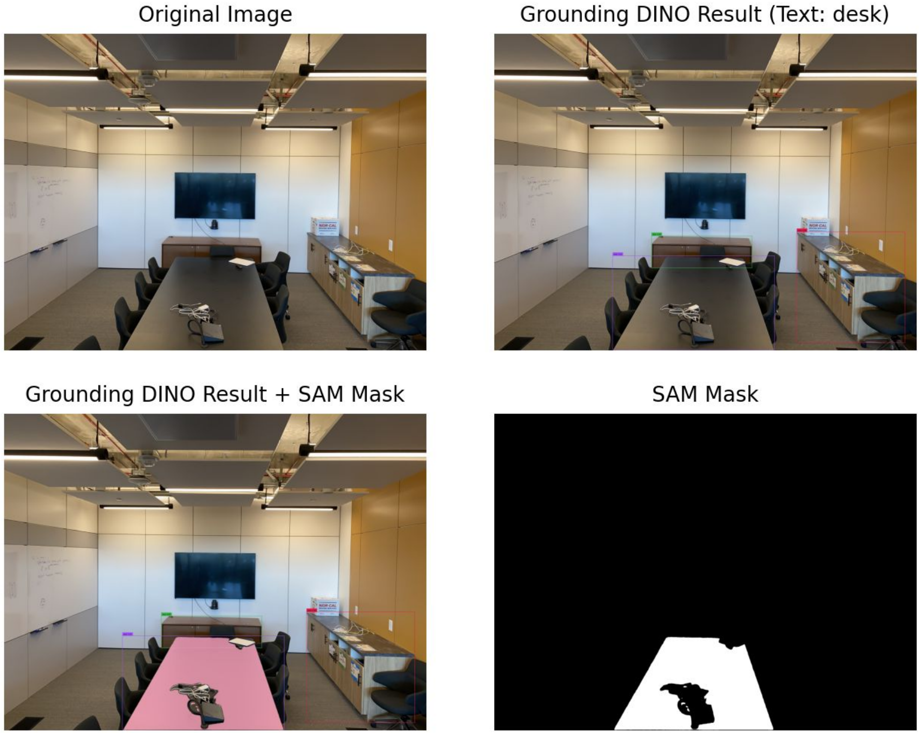
Pipeline



USC

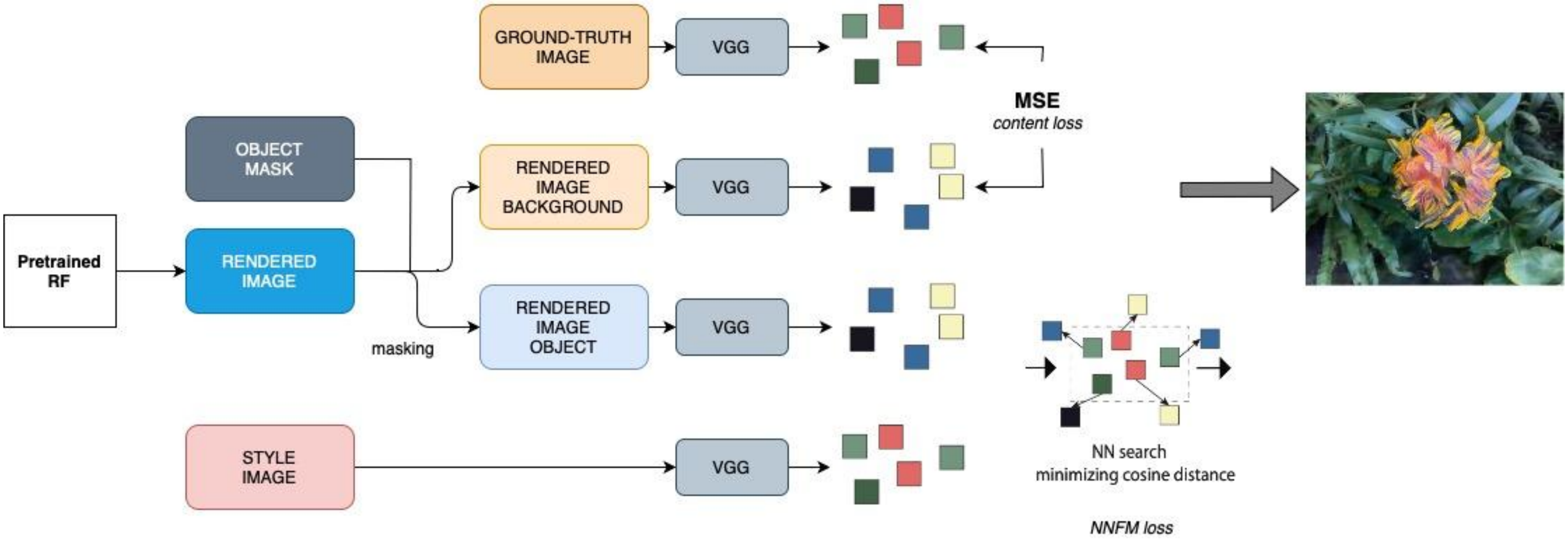
GroundedSAM





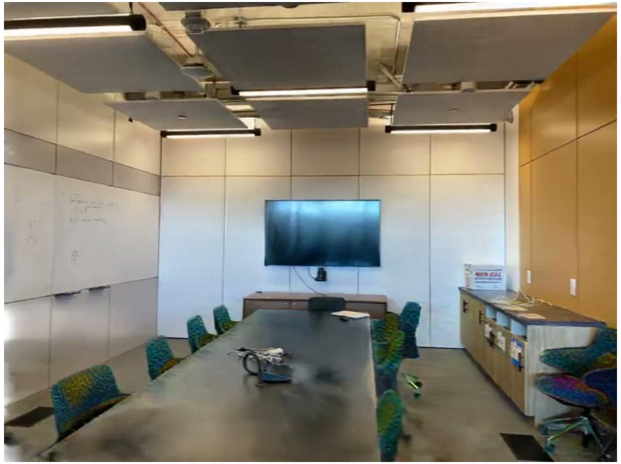
USC

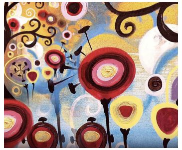
Masked NNFM


GROUND-TRUTH IMAGE
VGG
MSE
content loss
OBJECT MASK
RENDERED IMAGE BACKGROUND
VGG
Pretrained RF
RENDERED IMAGE
RENDERED IMAGE OBJECT
VGG
masking
NN search
minimizing cosine distance
STYLE IMAGE
VGG
NNFM loss

Training

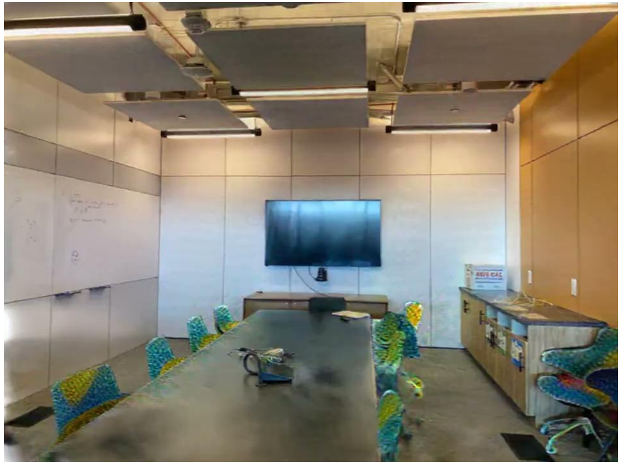
Pretraining

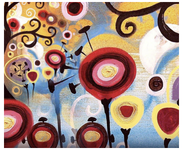




style

2 Epochs
Prompt: "chairs"



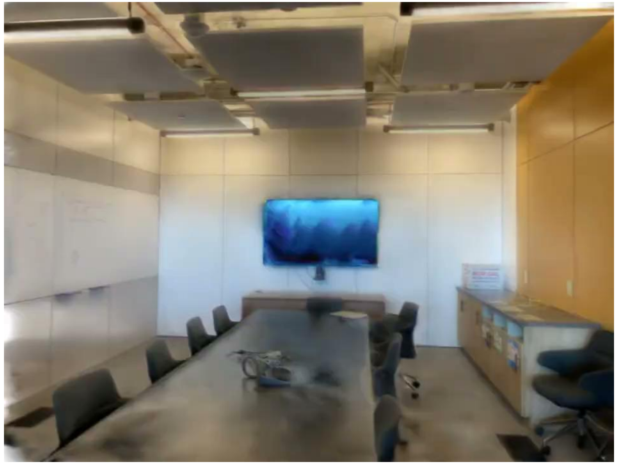


style

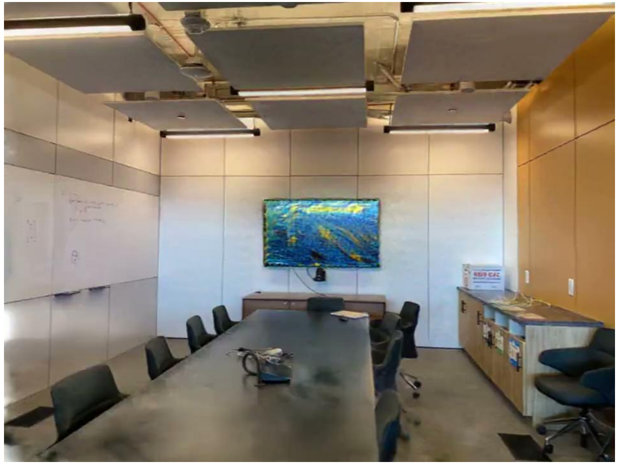
10 Epochs
Prompt: "chairs"

EXPERIMENTS

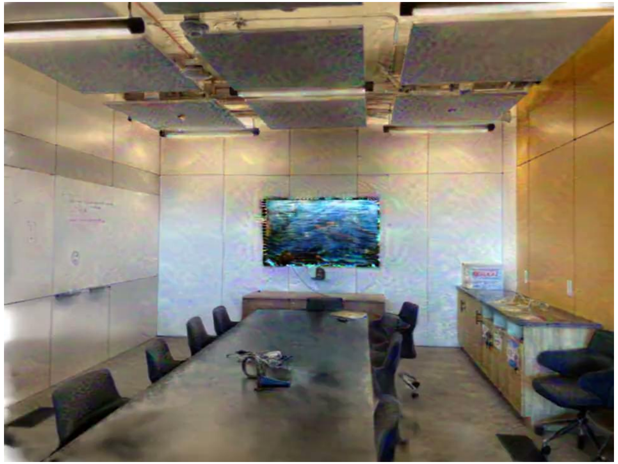
VGG Block



Block 0



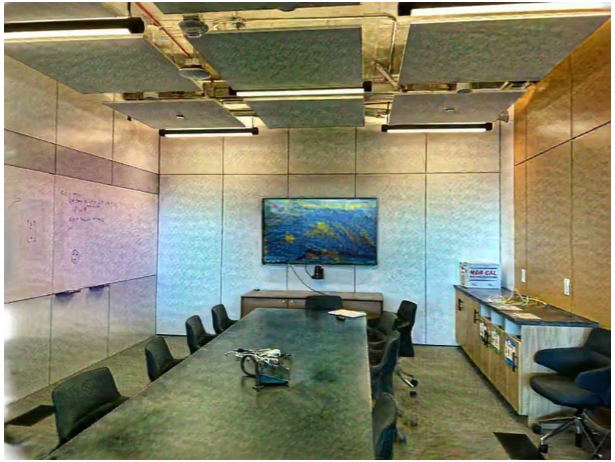
Block 2



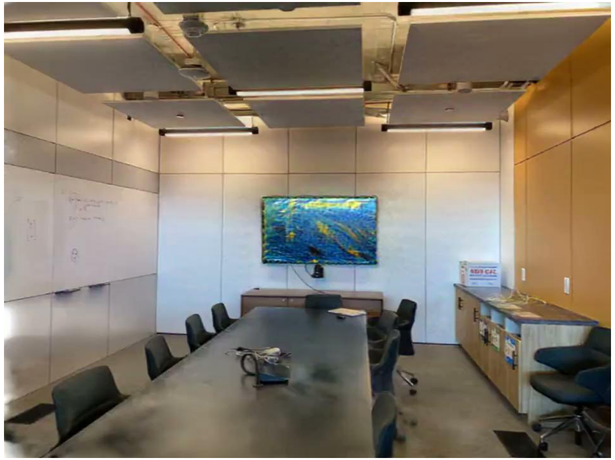
Block 4

Prompt: “tv”
Style: Starry Night

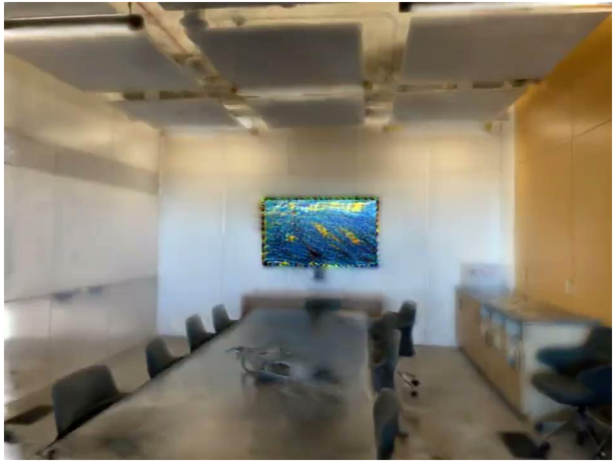
Content Weight



1



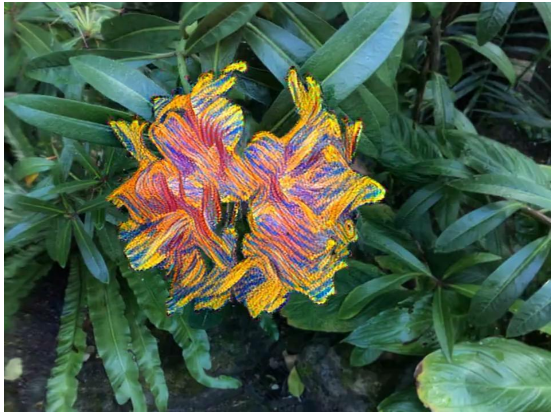
1e-3



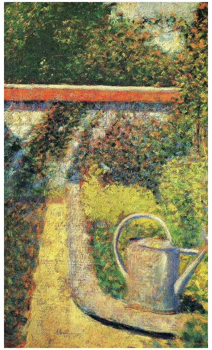
1e-6

Prompt: “tv”
Style: Starry Night

Qualitative Results (Different Styles)



Prompt: “flower”
Style: Starry Night



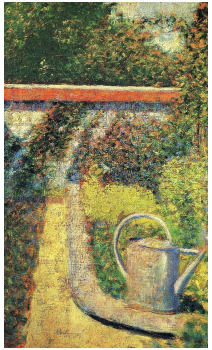
Prompt: “flower”
Style: Abstract painting

Prompt: “flower”
Style: Landscape

USC

Qualitative Results (Different Styles)

Prompt: “fern”
Style: Starry Night



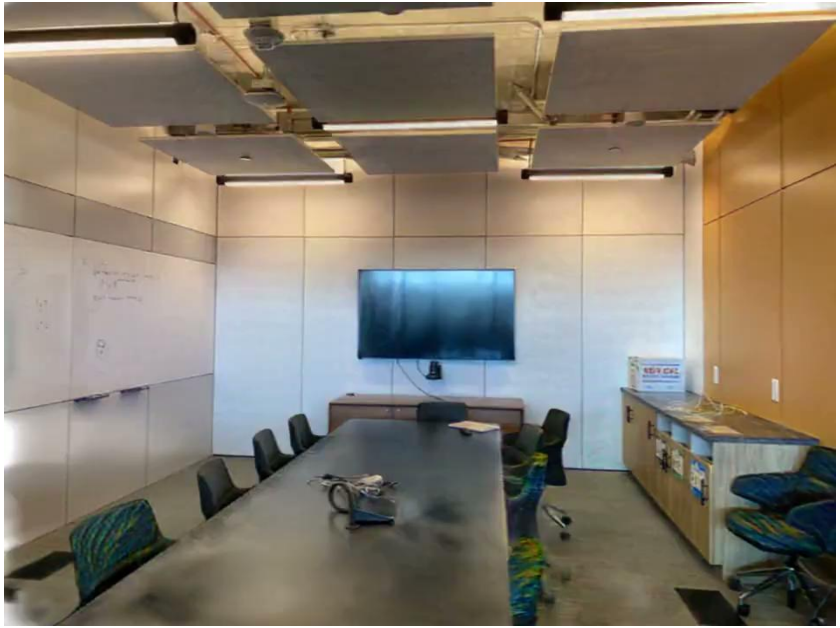
Prompt: “fern”
Style: Abstract painting

Prompt: “fern”
Style: Landscape

USC

Qualitative Results (Different Text Prompts)

Prompt: “desk”
Style: Starry Night



Prompt: “chairs”
Style: Starry Night

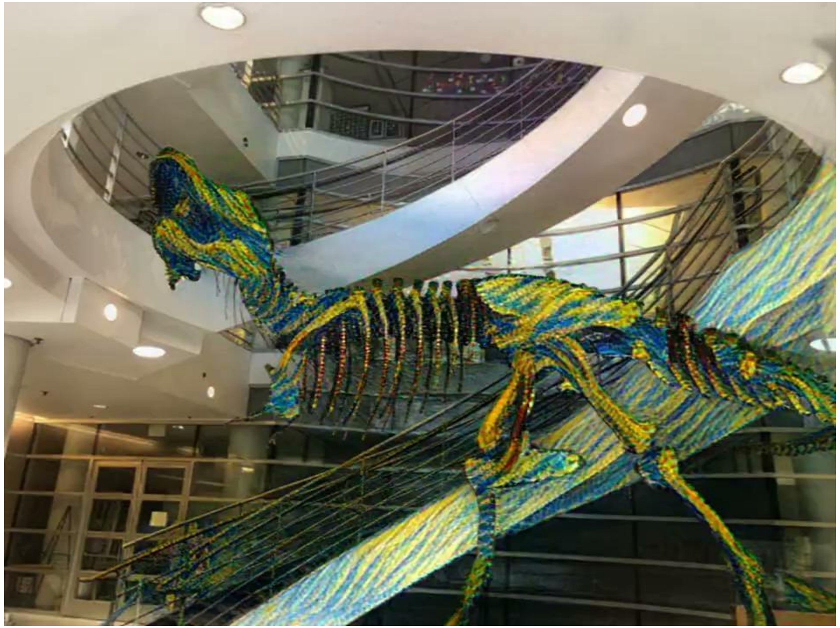
USC

Qualitative Results (Different Prompts, Styles)

Prompt: “castle”
Style: Starry Night

Prompt: “fortress”
Style: Landscape

Qualitative Results (A Failure Exp)



Prompt: “dinosaur”
Style: Starry Night

Prompt: “dinosaur”
Grounded SAM Result

USC

Future work…

Limitations & Future Work

Our work inherits limitations of GroundingDINO, SAM.

- SAM segmentations are guided by DINO bounding boxes, and not the input query
- With different views, bounding boxes have different scores -> different boxes to SAM
  - may lead to noisy mask and style transfer

We plan to extend our work to incorporate

- Stylizing multiple objects
- Stylizing multiple instances of same object with different styles
- **StyleAnything**: Language based style specification
  - Text conditioned style generation using stable diffusion

USC

Q/A

THANK YOU

Outline

- Introduction & Problem Statement
- Previous work & limitations
- Approach
  - Overview
  - Approach 1: LERF
    - Challenges
  - Approach 2: GroundingDINO + SAM
    - Advantages
  - Final Pipeline
    - NeRF Pretraining
    - GroundingDINO + SAM
    - Style Transfer: VGG + NNFM
- Experiments
  - Content Weight
  - VGG-Block
  - Epochs
- Qualitative Results
- Future Work

USC

Conclusions

Our method lies in the application of masked NNFM loss, enabling a more controllable style transfer;

Our method effectively achieves style transfer on both semantic and instance level, successfully applying distinct style(s) to multiple object(s) within a single scene.

(copied from ICCV)

Previous Work

Perform 3D stylization on point clouds or meshes are sensitive to **geometric reconstruction errors** for complex real-world scenes;

Commonly used **Gram matrix-based loss** tends to produce blurry results without faithful brushstrokes; (these two above copied from ARF)

Methods differ in the way they fine-tune or modify the 3D scene representation. Some works utilize a separate hyper networkwhile others alter the implicit representations themselves.

Focus on **whole-scene** stylization, be it through image or text modalities.; (these two above copied from S2RF)

Motivations

Enable language based object selection for stylization

TODO:Why GroundedSAM may performs better in this specified segmentation subtask?

In the realm of 3D scene stylization, we need to address **spatial consistency** challenge(intro to the NNFM Loss);

Constraint the style transfer to a **specified object** is challenging and active for research.

Approach 1:Using LERF

USC

LERF

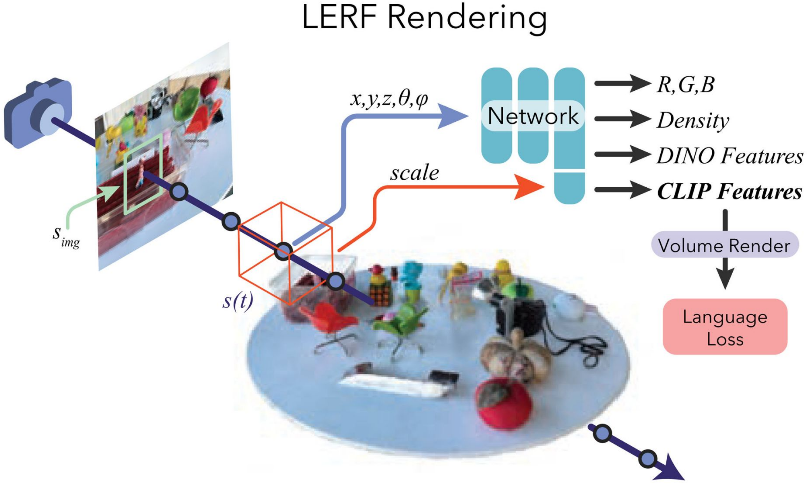
1.Train a LERF model

(I.e. jointly optimize a language field along with a radiance field using CLIP+DINO supervision)

2.Use the **user specified object** to query the trained LERF model and obtain the **relevancy map**

3.Convert this **relevancy map** to a **segmentation mask**, by thresholding

4.Fine-tune the trained LERF model with Nearest Neighbor Feature Matching(NNFM) loss for style transfer



Eg: Relevancy map for text queries

USC

Issues

Generated relevancy maps are very noisy

Example: Like CLIP, language queries from LERF often exhibit “bag-of-words” behavior

(i.e., “not red” is similar to “red”) and struggles to capture spatial relationships between objects.(copied from LERF paper)

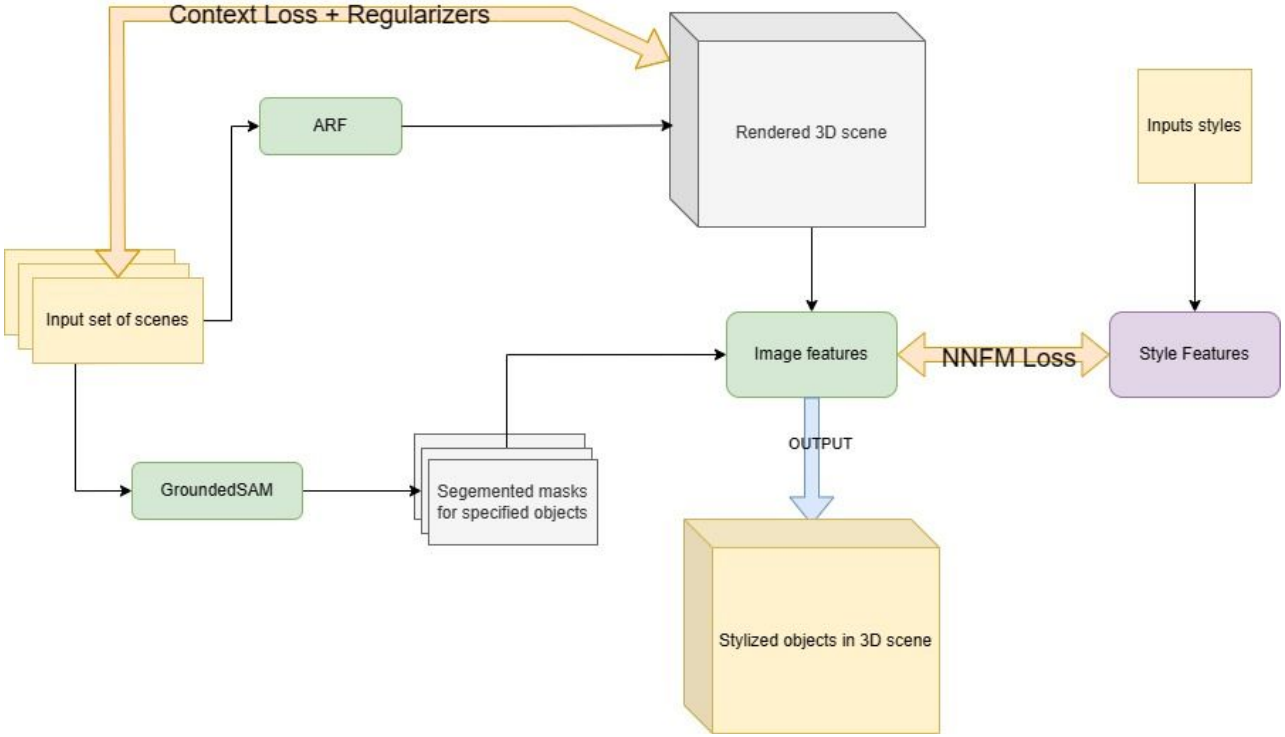
Efficiency on our machines: **Long training times** (2 hours per experiment) + **low GPU memory**(hard to implement the best version “lerf”, can only implement the small-scale version “lerf-lite”)

Approach 2:Using ARF with GroundedSAM (Now we used)

Pipeline

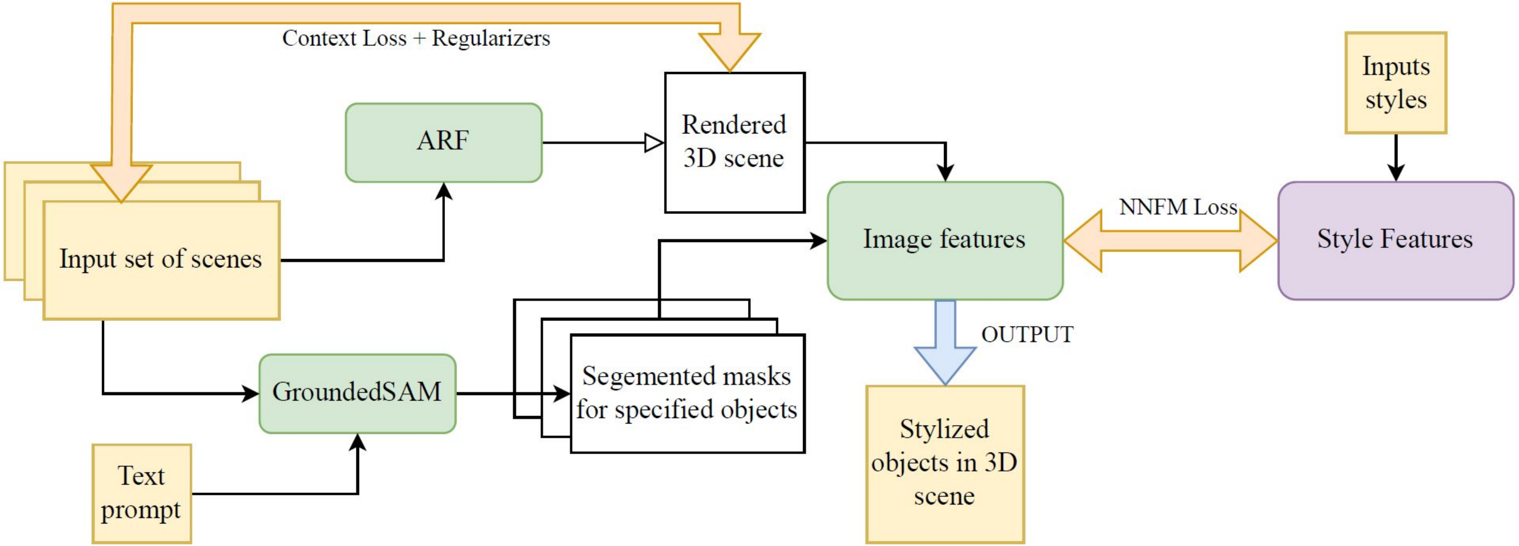
TODO: if any fault , can open/modify through draw.io (drawio.com)(regularizer should be DINO?)

Shared link: Language-based Object Selection for 3D Stylization.drawio



Pipeline

I draw a new pipeline here (Jingmin) . See share drive here



Artistic Radiance Fields(ARF)

Use the user specified object query with GroundedSAM to generate the segmentation mask

Unlike a Gram matrix describing global feature statistics across the entire image, NN feature matching focuses on local image descriptions, better capturing distinctive local details.

VGG feature-based content loss(if used) : balances stylization and content preservation, improves the color match between our final renderings and the input style;

Stylized NeRF

Fine-tune the pretrained NeRF model with NNFM for style transfer

Improvements

Accurate segmentation masks (show comparison between LERF and DINO+SAM)

Memory-friendly model + lower training time (~45 minutes per experiment)

NNFM Loss

Experiments & Qualitative Results

Experiment

VGG Block Ablation

Content weight ablation

Epoch-wise training progress

Experiment

Content weight

Show comparison between different content weight for stylization

1, 1e-1, 1e-2, 1e-3, 1e-5

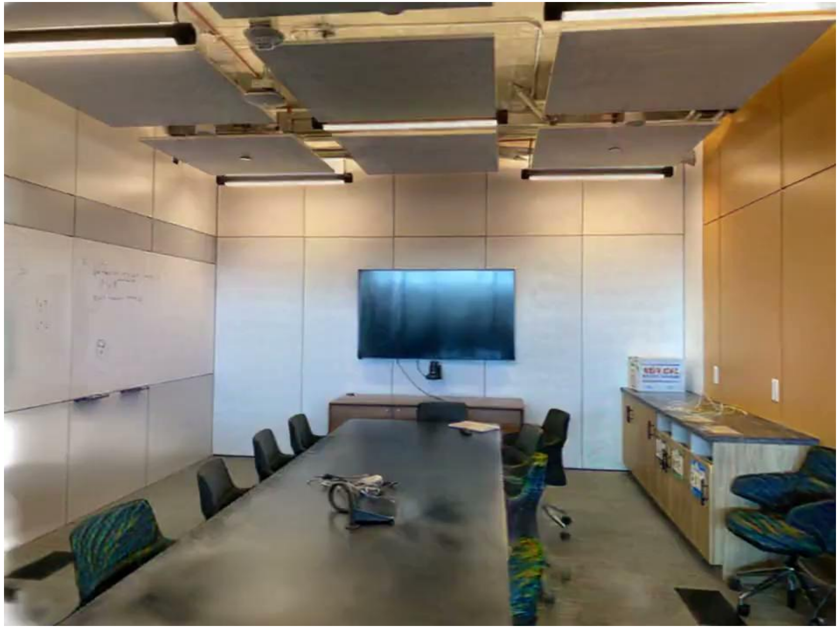
Write some observations

Qualitative Results

Dataset: room, Style image: Starry. Text prompt: “tv” .

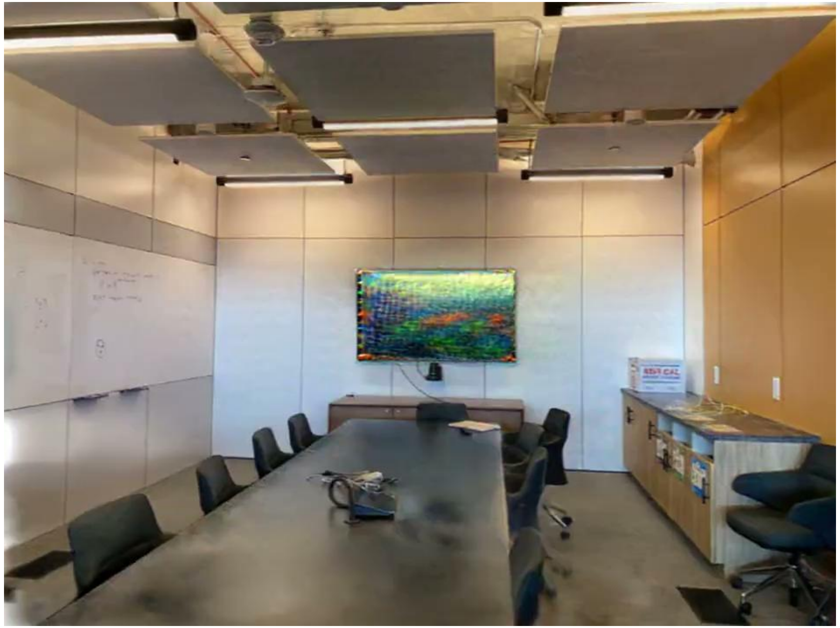
Qualitative Results

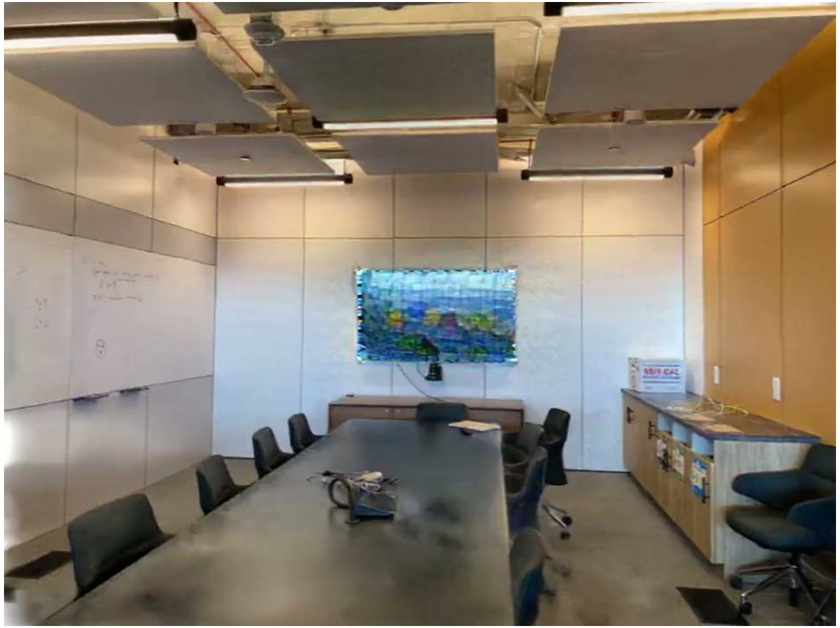
Different text prompt to model: “table” and “chair” .



Qualitative Results

Different style init with same text prompt “tv”.





Qualitative Results

Epochs

MSE_NUM_EPOCHS & NNFM_N_EPOCHS

Conclusions

Future Works

Encompass a broader range of scenes, including 360-degree environments and scenes with an increased number of objects.

Conduct more quantitative evaluations to thoroughly assess the effectiveness of our method.

(copied from ICCV)