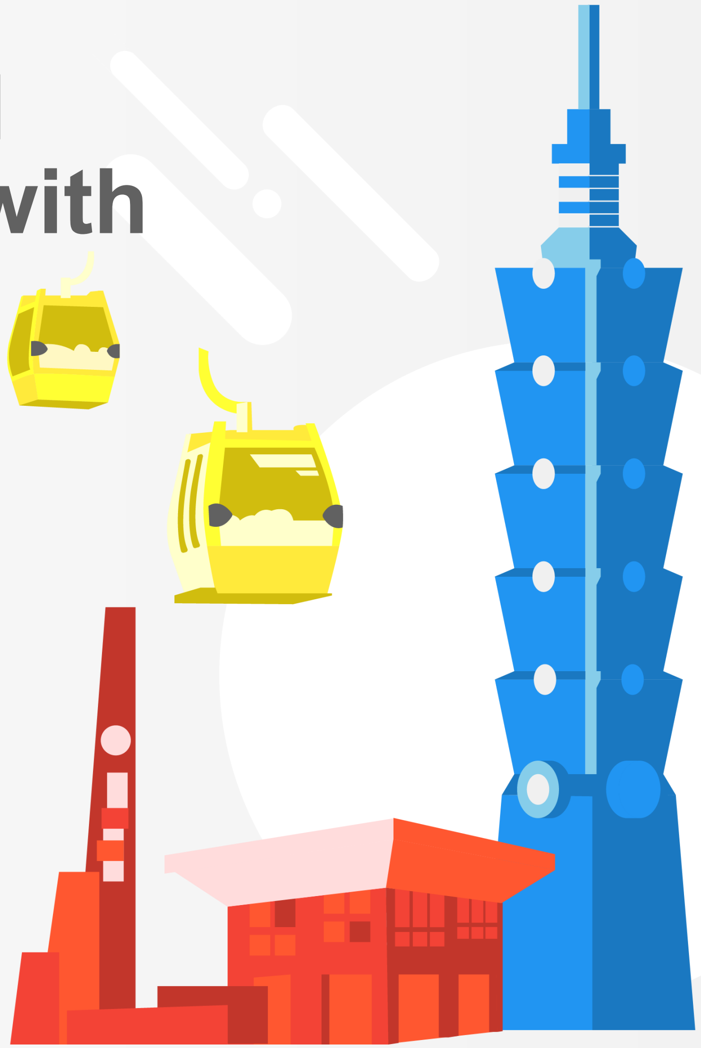
Improved 3D Scene Stylization via Text-Guided Generative Image Editing with Region-Based Control

Haruo Fujiwara, Yusuke Mukuta, Tatsuya Harada

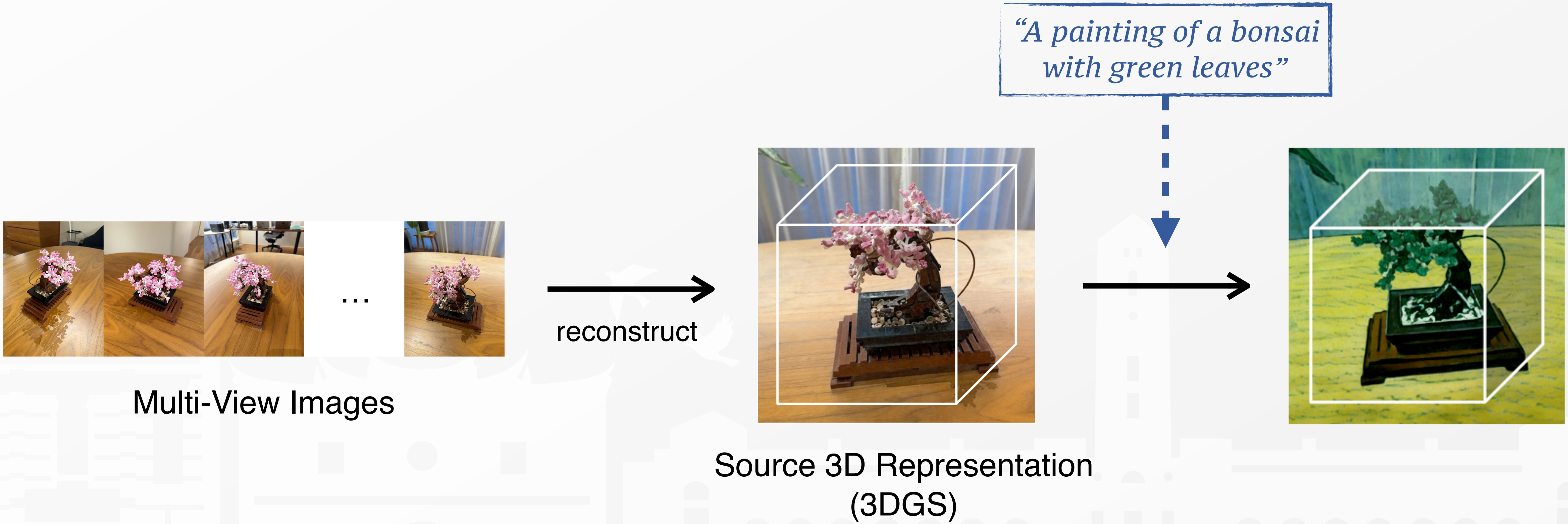
The University of Tokyo

2025 TAIPEI



Task Objective

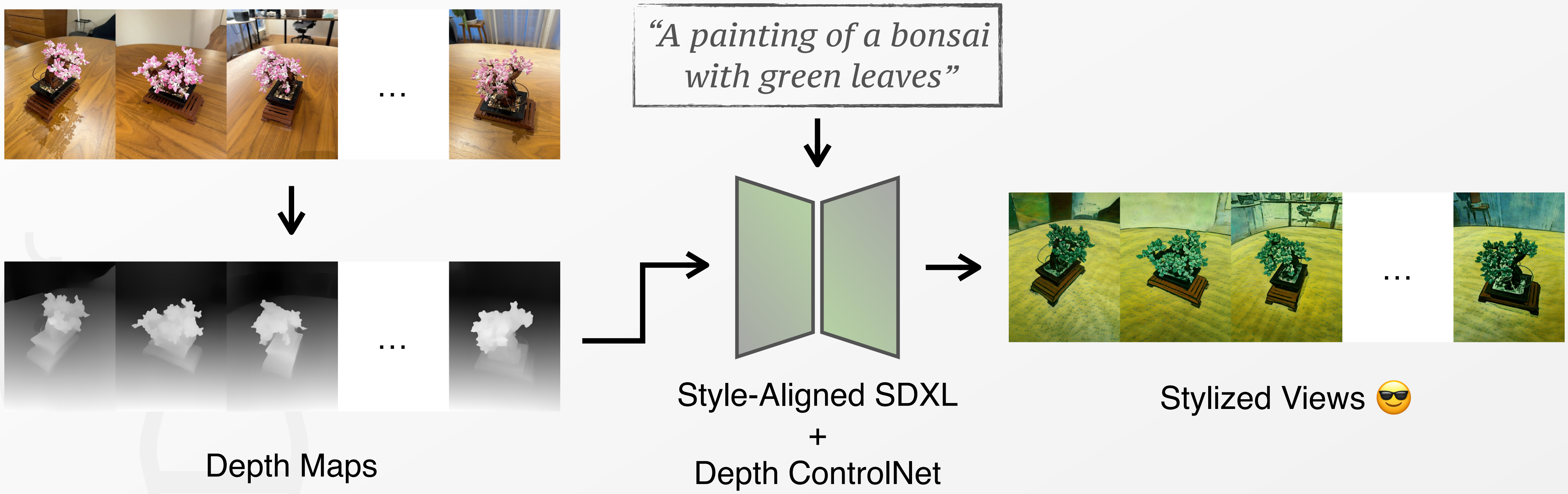
Text-driven stylization of a reconstructed 3D scene



Previous Work

"Style-NeRF2NeRF" [Fujiwara+ SIGGRAPH Asia 2024]

Make stylized multi-views then finetine

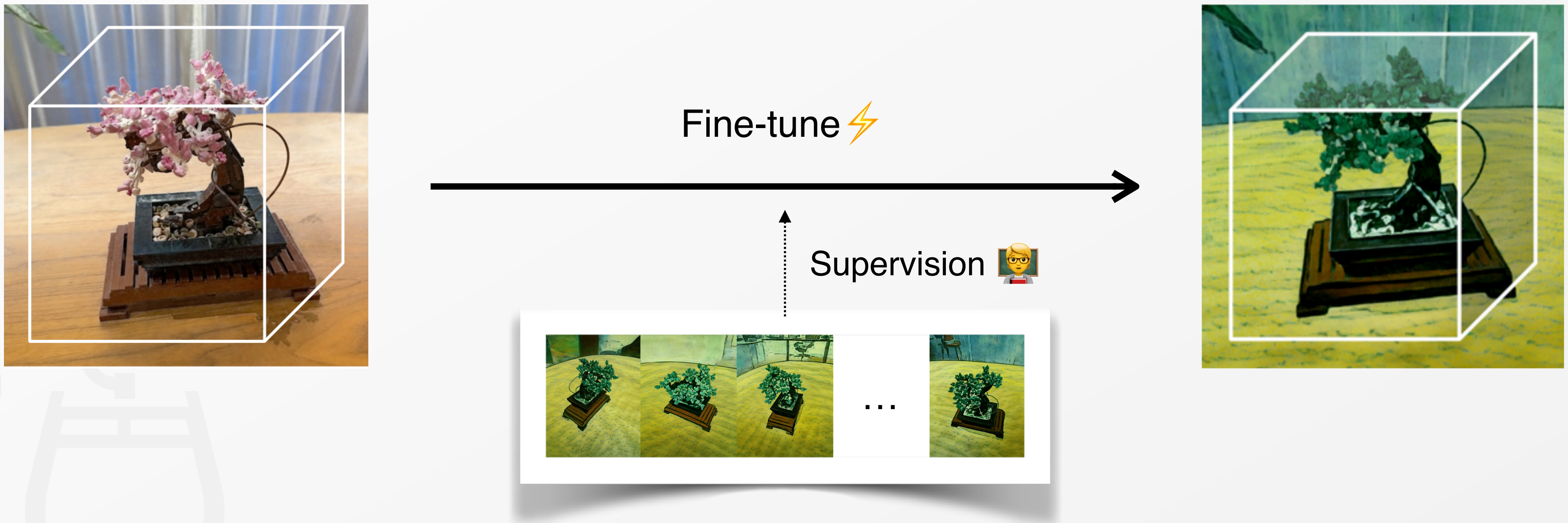


2025 TAIPEI

Previous Work

“Style-NeRF2NeRF” [Fujiwara+ SIGGRAPH Asia 2024]

Make stylized multi-views then finetine



Source 3D Refinement

- Represent images as discrete probability distributions of VGG features (L=12 layers)
- Style transfer by matching the distributions via Sliced Wasserstein distance loss

Source 3D Representation
(3DGS)

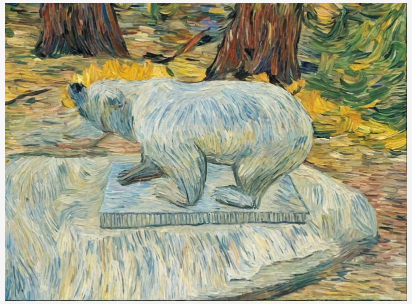
render

update

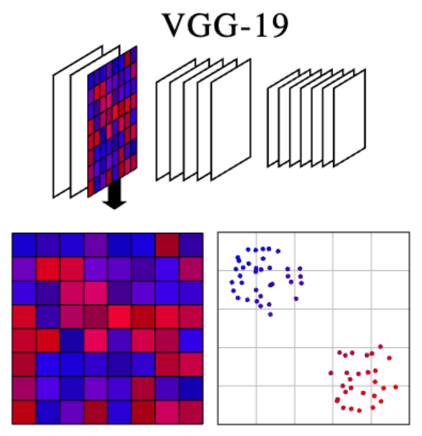
Rendered images (optimize target)

...

2D style reference views



...



$$p^l(x) = \frac{1}{M_l}\sum_{m=1}^{M_l}\delta_{F^l_m}(x)$$

$$\mathcal{L}_{SW} = \sum_{l=1}^{L}\mathbb{E}_V[\mathcal{L}_{SW1D}(p^l_V, \hat{p}^l_V)]$$

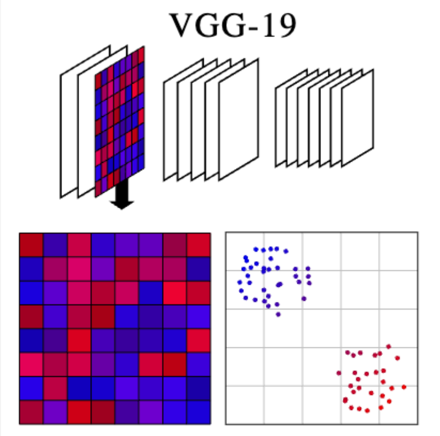


Image by "A Sliced Wasserstein Loss for Neural Texture Synthesis [Heitz+ CVPR21]

2025 TAIPEI

Summary of Contributions

In this paper, we address some limitations and make the following contributions:

1. Region-based control for 3D stylization

   Apply stylization selectively to distinct regions

2. Improved diffusion-based multi-view stylization pipeline

   Improve style-consistency of guidance multi-views

3. Efficient sliced Wasserstein distance loss

   Accelerate 3D fine-tuning

Contribution 1: Region-Control

Key Limitation: Previous work cannot apply 3D style transfer selectively (*e.g.* stylize only foreground)

original scene

w/o region control

w/ region control

"A polar bear in the woods"

Contribution 1: Region-Control

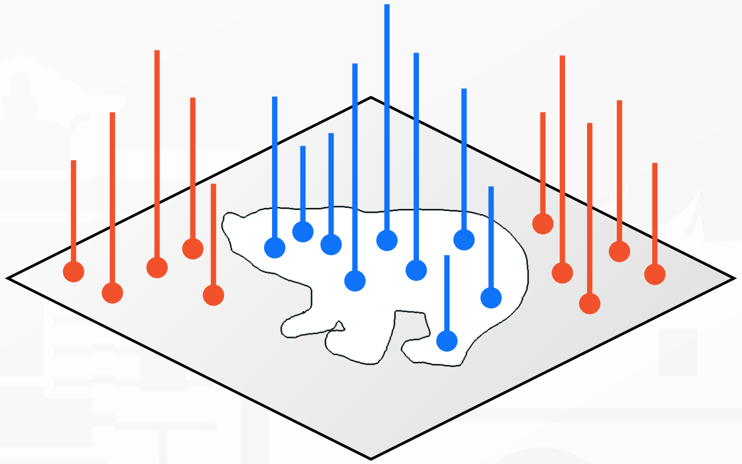
Key Limitation: Previous work cannot apply 3D style transfer selectively (*e.g.* stylize only foreground)

Solution: Split feature distribution with additional dimension based on segmentation masks

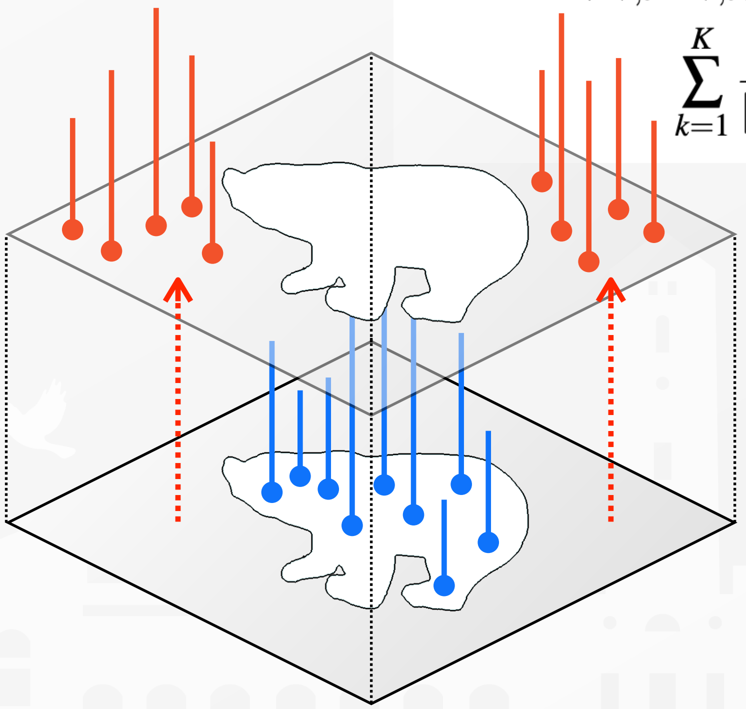
$$\mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l) = \frac{1}{|p_V^l|} \| \texttt{sort}(p_V^l) - \texttt{sort}(\hat{p}_V^l) \|^2$$

$$\mathcal{L}_{MR\text{-}SW1D}(p_{V,b}^l, \hat{p}_{V,b}^l) = \sum_{k=1}^{K} \frac{1}{|p_{V,k}^l|} \| \texttt{sort}(p_{V,k}^l) - \texttt{sort}(\hat{p}_{V,k}^l) \|^2$$

(K: num or regions)



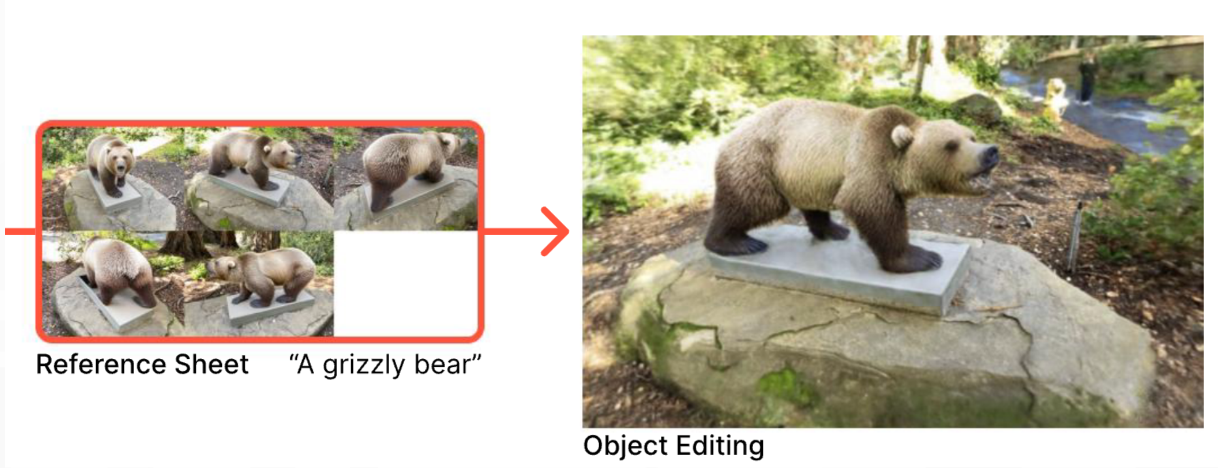
Before: Feature distribution in projected 1D



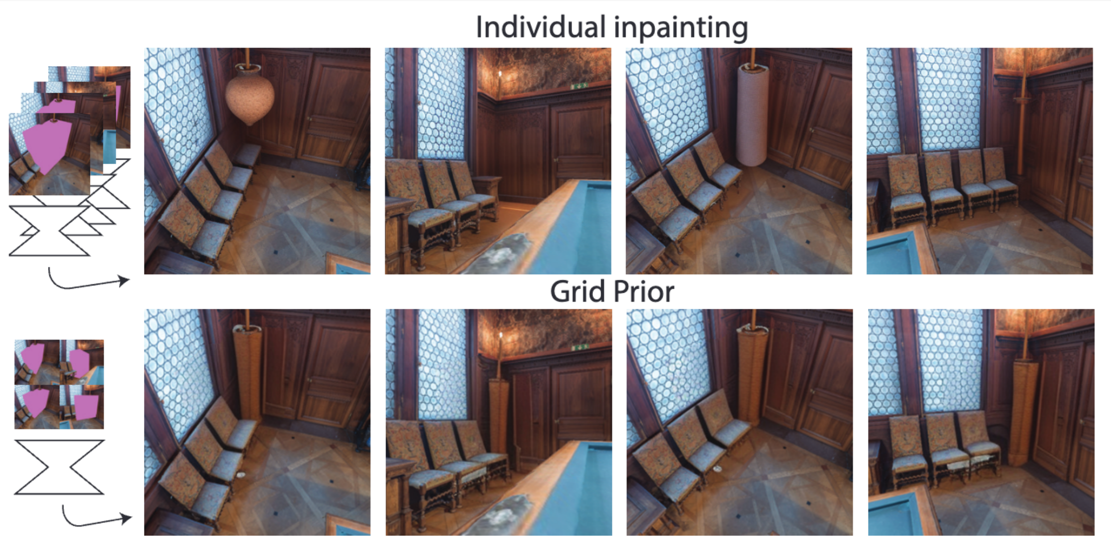
After: Add dimension and split feature distribution using mask

Contribution 2: Improved Multi-View Generation

- Better 2D stylized multi-views = Higher 3D stylization quality
- Recent 3D inpainting methods reveal that references tiled in a grid promotes 3D consistency
- Inspired by this idea, we upgrade the multi-view generation pipeline based on a tile reference of depth maps



SIGNeRF [Dihlmann CVPR2024]



NeRFiller [Weber+ CVPR2024]

1. Sample representative views (n=4) for reference tile
2. Pass reference+target depth tiles to diffusion pipeline with prompt
3. Generate target views with attention anchored on reference depth tile

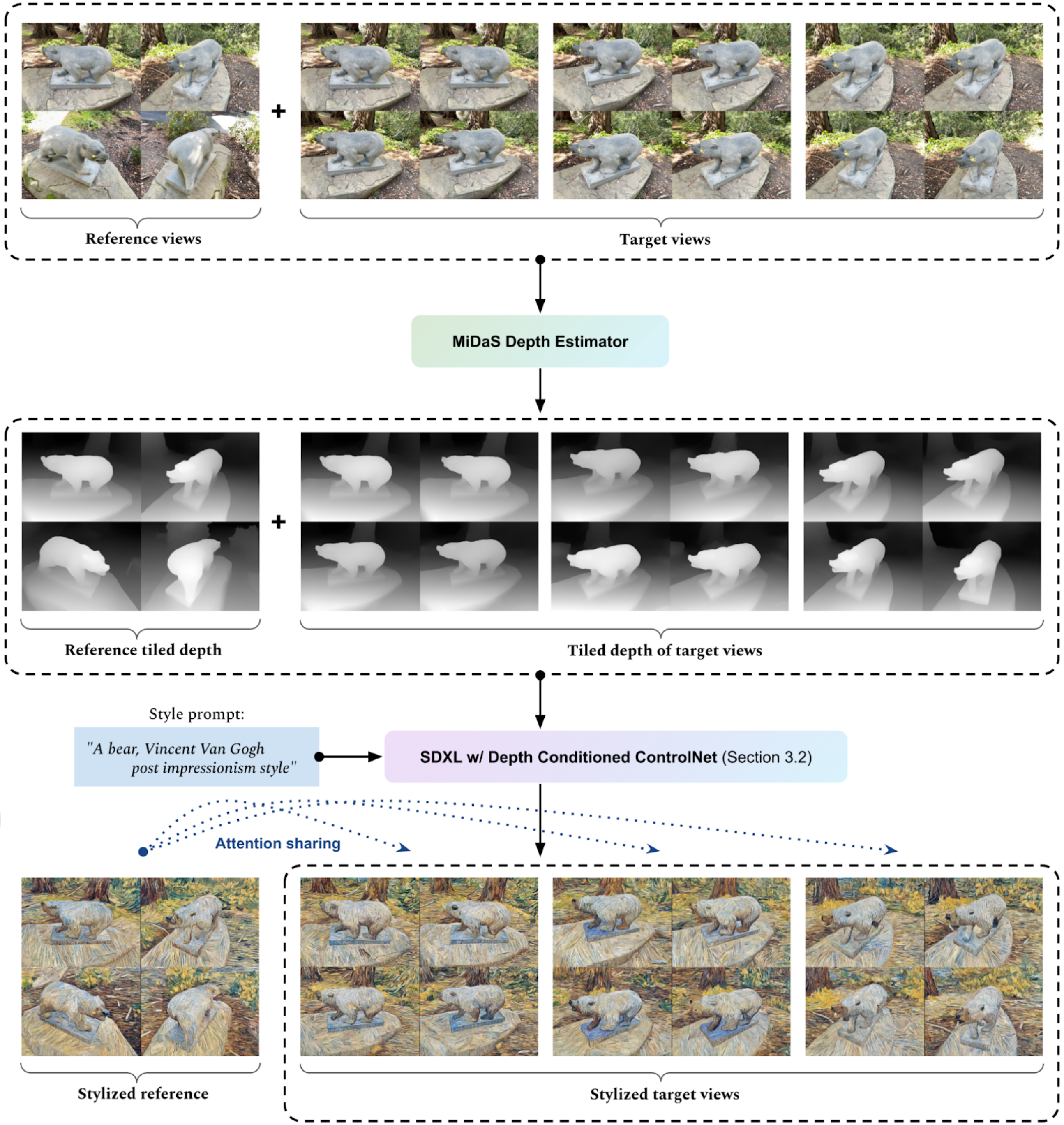
Concat the target keys and values in cross-attention (quite simple!!)

$$\texttt{Shared-Attn}(\hat{Q}_t, K_{rt}, V_{rt})$$

$$K_{rt} = [K_r, \hat{K}_t]^T, V_rt = [V_r, \hat{V}_t]^T$$

$$\hat{Q}_t = \texttt{AdaIN}(Q_t, Q_r), \hat{K}_t = \texttt{AdaIN}(K_t, K_r)$$

$$\texttt{AdaIN}(x, y) = \sigma(y)(\frac{x - \mu(x)}{\sigma(x)}) + \mu(y)$$



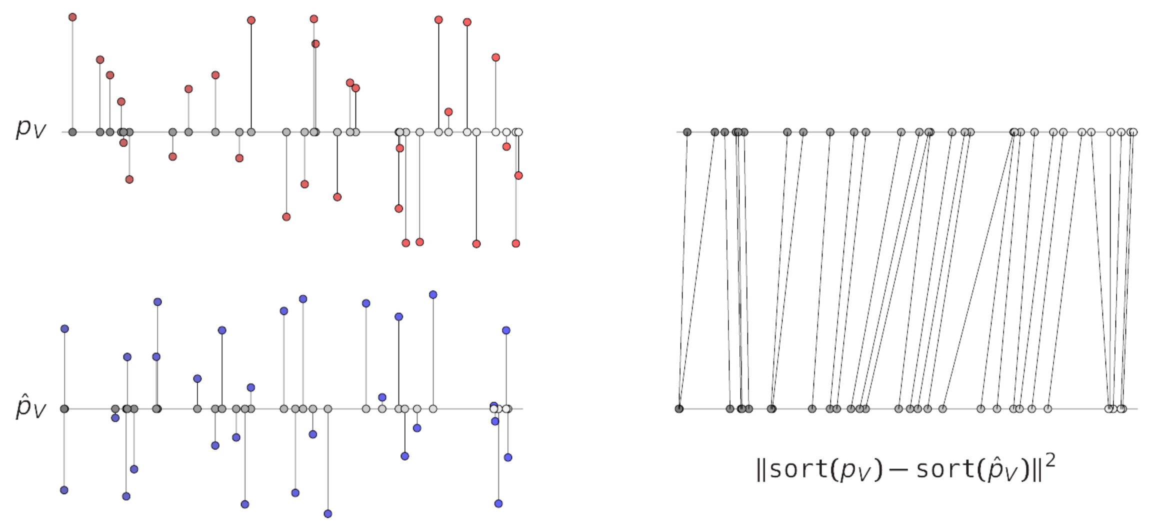
Contribution 3: Efficient 3D Finetuning

- Slicing directions are uniformly sampled in vanilla Sliced Wasserstein
- Higher 1D Wasserstein distance ≒ More informative direction
- Importance-weighting works well!!

$$\mathcal{L}_{style} = \sum_{l=1}^{L} \mathcal{L}_{SWD}(p^l, \hat{p}^l)$$

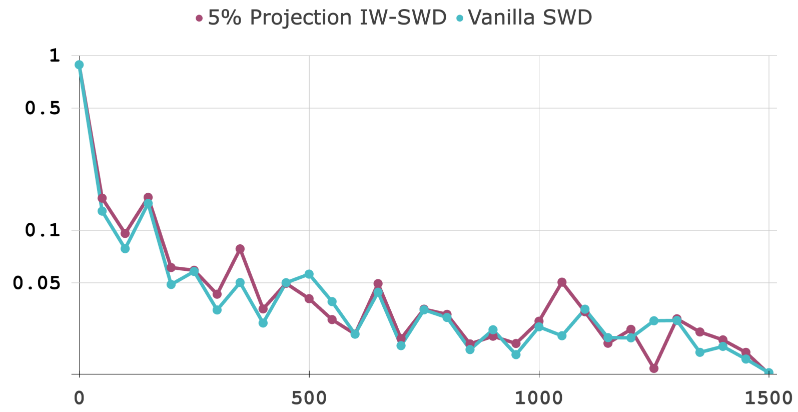
$$\mathcal{L}_{SWD} = \sum_{l=1}^{L} \mathbb{E}_V[\mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l)]$$

$$\mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l) = \frac{1}{|p_V^l|} \|\mathtt{sort}(p_V^l) - \mathtt{sort}(\hat{p}_V^l)\|^2$$



Contribution 3: Efficient 3D Finetuning

- Slicing directions are uniformly sampled in vanilla Sliced Wasserstein
- Higher 1D Wasserstein distance ≒ More informative direction
- Importance-weighting works well!!
    - Similar convergence with 5% projections



$$\mathcal{L}_{style} = \sum_{l=1}^{L} \mathcal{L}_{SWD}(p^l, \hat{p}^l)$$

$$\boxed{\mathcal{L}_{SWD} = \sum_{l=1}^{L} \mathbb{E}_V[\mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l)]}$$

Replace!

$$\boxed{\mathcal{L}_{IW\text{-}SWD} = \sum_{l=1}^{L} \sum_{V} w_V \mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l)}$$

$$\mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l) = \frac{1}{|p_V^l|} \|\mathrm{sort}(p_V^l) - \mathrm{sort}(\hat{p}_V^l)\|^2 \qquad w_V = \frac{\exp(\mathcal{L}_{SW1D}(p_V^l, \hat{p}_V^l))}{\sum_{V'} \exp(\mathcal{L}_{SW1D}(p_{V'}^l, \hat{p}_{V'}^l))}$$

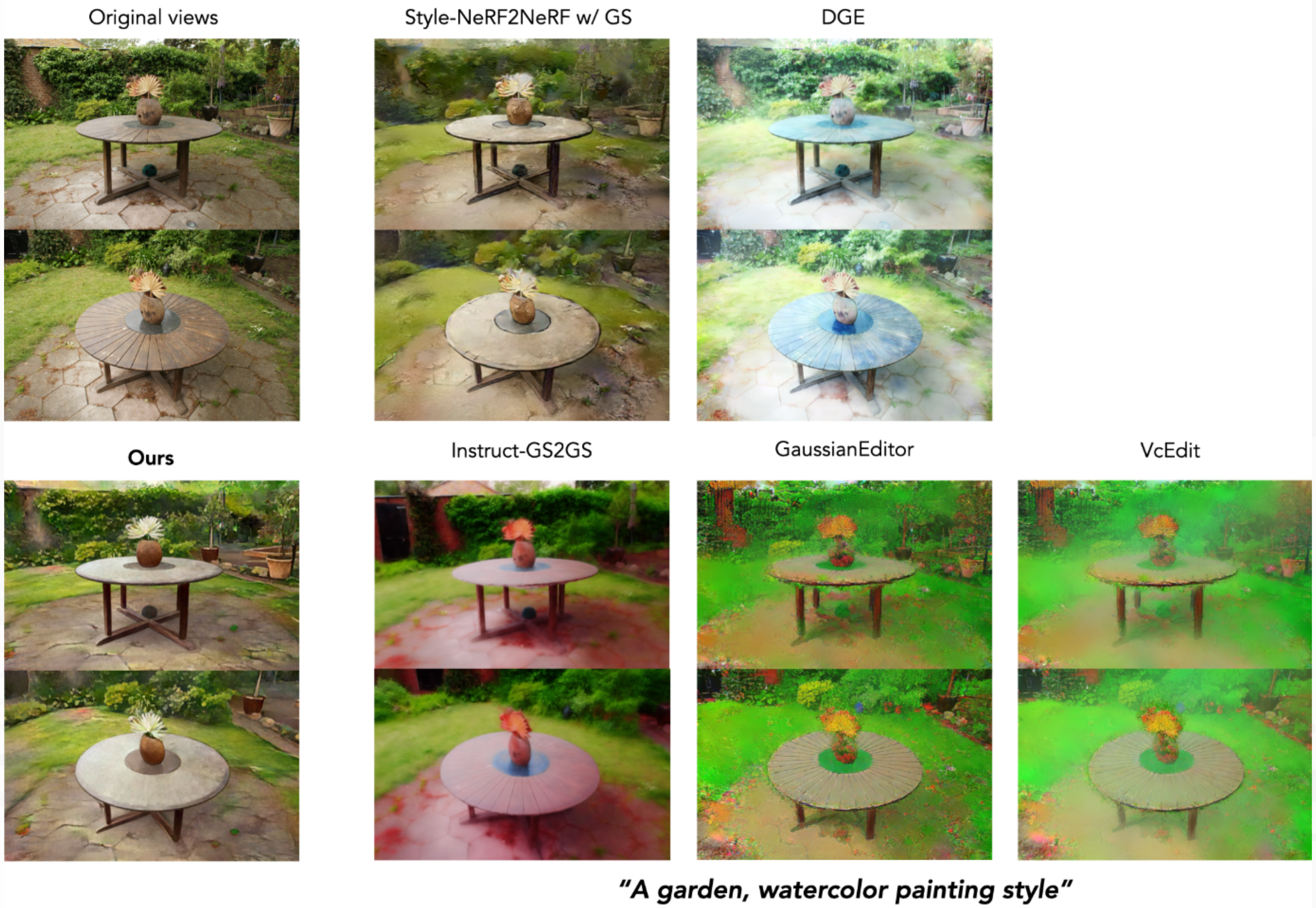
2025 TAIPEI

Results

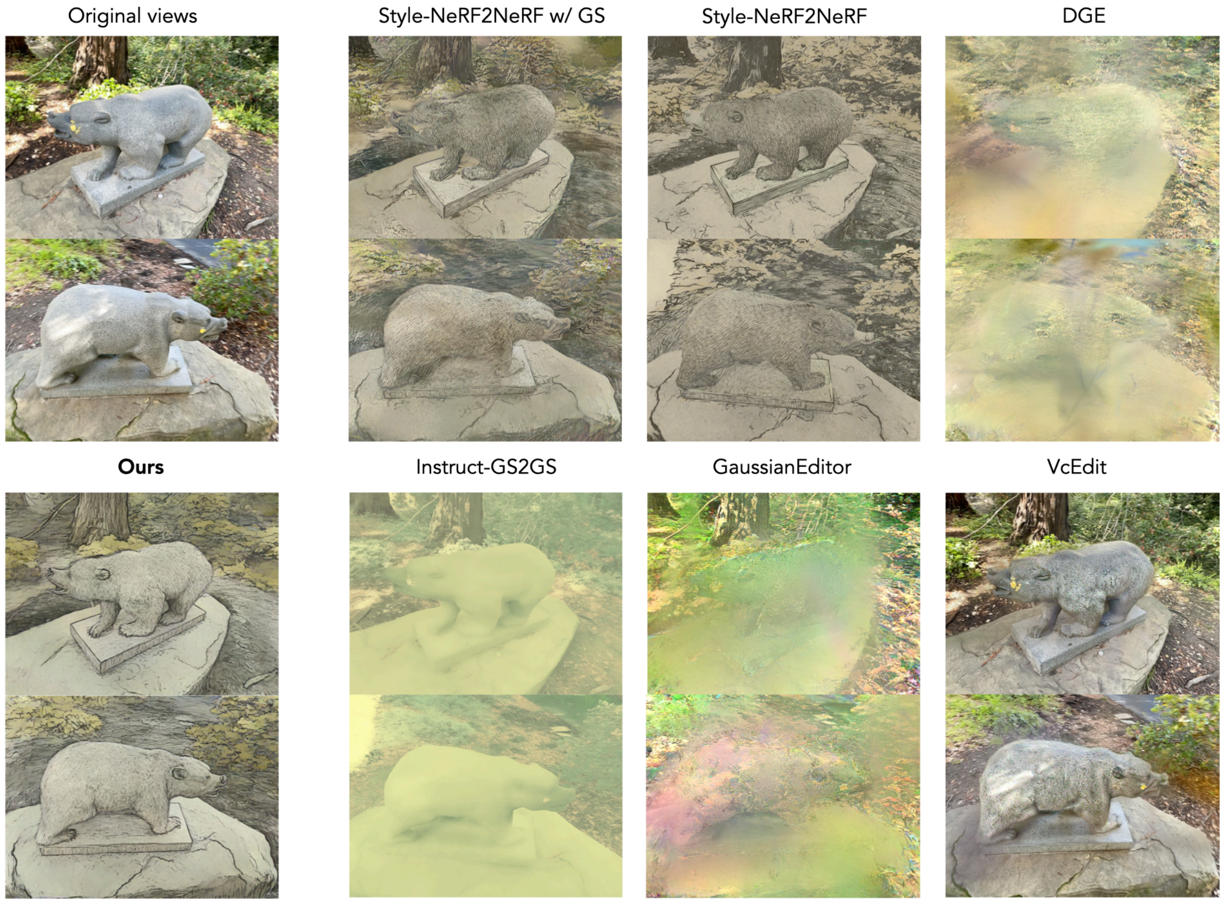
Improved 3D Scene Stylization via Generative Image Editing with Region-Based Control
Demo Video

H. Fujiwara [1], Y. Mukuta [1,2], T. Harada [1,2]
[1] The University of Tokyo, [2] RIKEN AIP

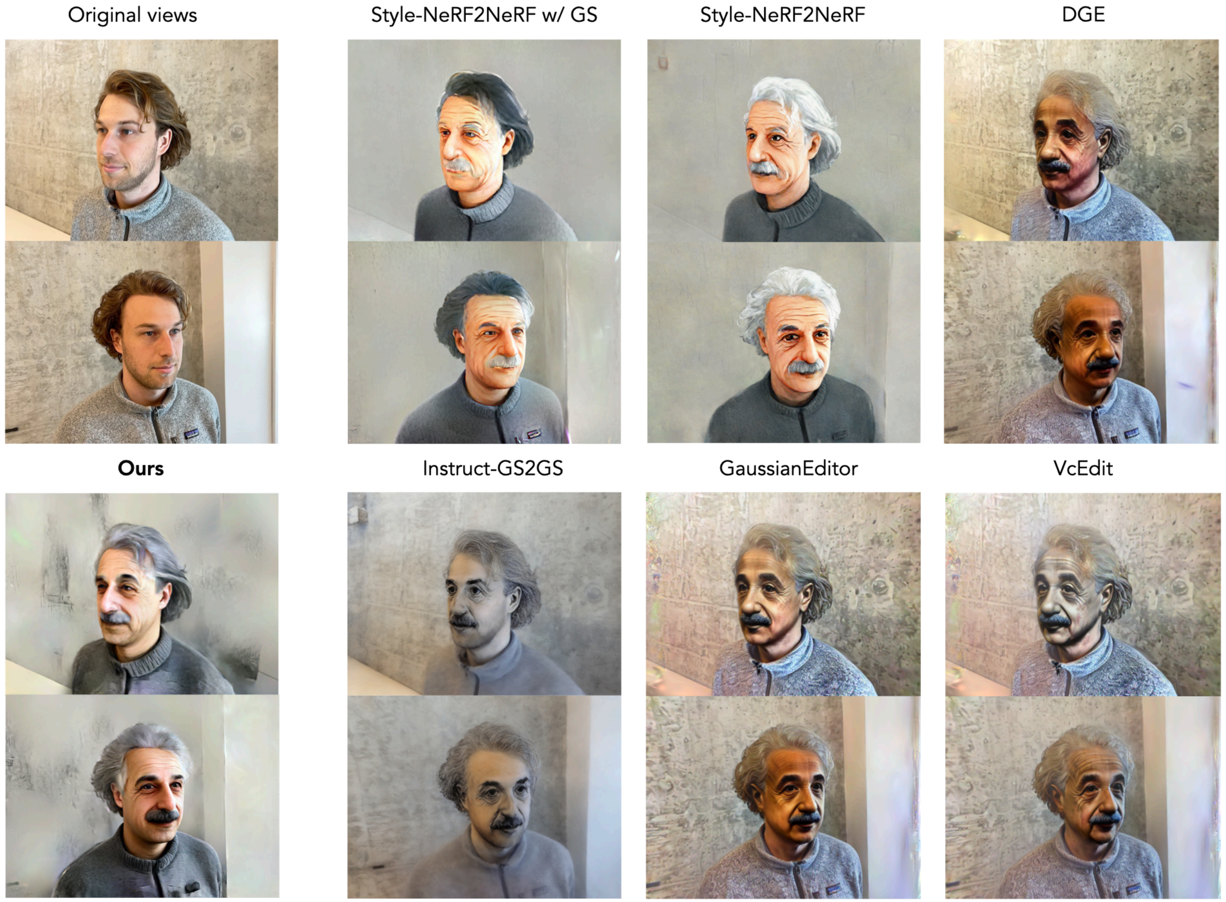
Method Comparison



Method Comparison



"A bear, Japanese ukiyo-e style"



"A person like Albert Einstein"

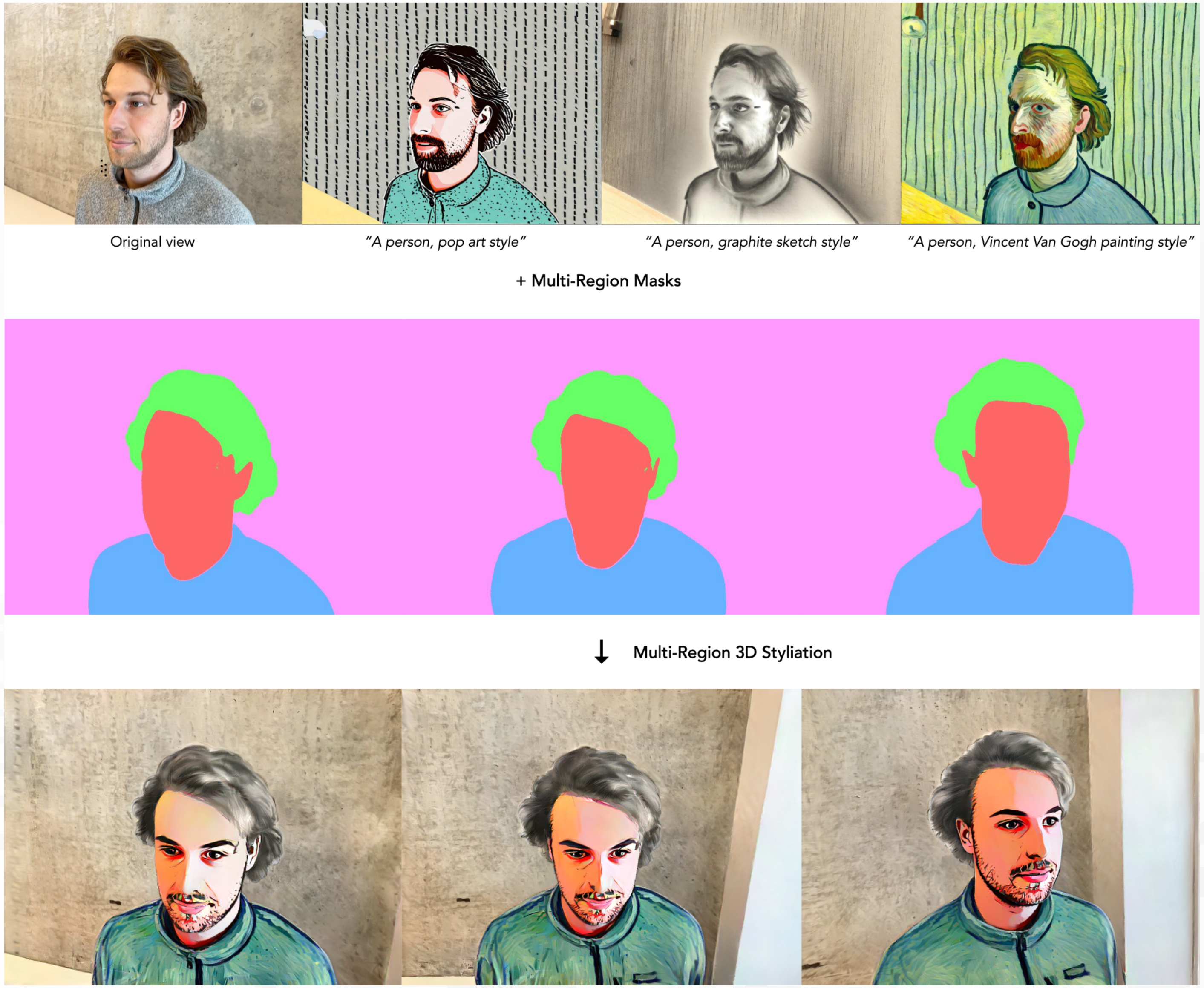
Ablation

Region information can achieve semantically coherent 3D stylization

Original view | **Ours** | w/o multi-region loss | w/o our multi-view gen. pipeline

“A blue bear in impressionism painting style”

Region-Based Stylization Example



2025 TAIPEI

Limitations and Future Work

- Due to depth-conditioning, our method cannot perform significant shape deformation
- We still rely on VGG19 as the style feature extractor
- Our method does not consider reflectance

Therefore in the future…

- Extend our method to alternative representations (*e.g.* hybrid of GS and mesh)
    - Support reflectance as well?
- Consider different architectures (DiT) and feature extractors (latents in diffusion?)

Thank You!