3D building model edit with generative AI

Student: Yingxin Feng

Supervisors: Nail Ibrahimli Dr. Ken Arroyo Ohori

Co-reader: Dr. Liangliang Nan

TU Delft

Content

- Introduction
- Related work
- Methodology and result
- Conclusion

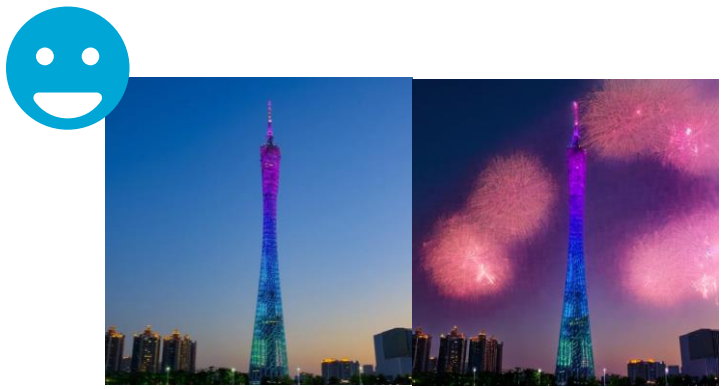
TU Delft

Introduction: motivation

- Promising future of generative AI

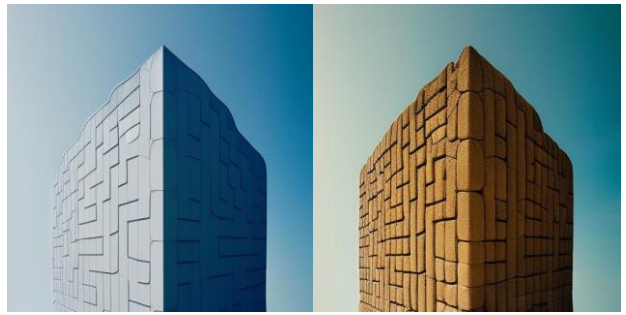
- Success in image generation and edit for general cases
  - limited identity-preserving level or geometric deformation scope for buildings

Examples (InstructPix2Pix)



“Add fireworks”

“Make it a castle”



“Make it ancient”

Introduction: motivation

- Partial success in 3D model edit (based on 2D pre-trained models)
  - Limited to certain object types and viewing angles

Posterior Distillation Sampling (Implicit based)

"Roses"

Text2Tex (Explicit based)

"Wooden barrel"

"Metal CD player"

X-Mesh (Explicit based)

"Colorful candy vase"

"Blue Whale"

Image Reference: [Koo et al., 2023] [Chen et al., 2023a] [Ma et al., 2023]

Introduction: motivation

Lack of attention in 3D building models domain

- Limitations from 2D models: challenging in dealing with complex and specific buildings and prompts, inconsistency and bias in views

2D Challenging cases (Stable Diffusion)

“A zoomed out DSLR photo of a two-Storey red townhouse with small windows and grey roof, five connected”

“A four-storey office building with perforated brickwork and plant decorated façade”

2D View inconsistency (InstructPix2Pix)

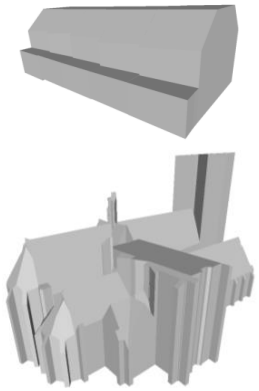
Original

Edited

“Make it a church”

Introduction: research objective

- Explore the potential of generative AI based 3D edit in building models
  - What existing pipelines are promising in the building model edit field?
  - How do these chosen pipelines perform in different building cases?
  - How to develop a new pipeline or modify existing ones to make the edit results better comply with user guidance and have higher fidelity?
  - What are the user scenarios and limits of the existing and modified edit pipelines?



Uncoloured building meshes

Text or image prompt

Promising existing edit pipelines and modified ones

Target mesh with texture and small-scale geometric changes

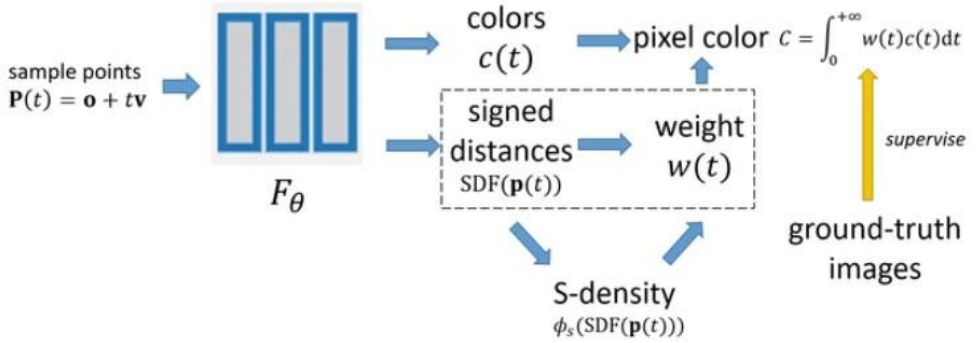
Target mesh with texture

Related work: 3D representation

- Implicit:

  Neural Implicit Surfaces (NeuS)

  - Able to extract a high-quality surface and render images regardless of resolution limit
  - Geometry and texture information influence mutually: more difficult to control



- Explicit: Mesh
  - Relatively compact, can represent both the geometry and texture explicitly
  - Wide application and easy for further processing
  - Limited by resolution

```
####
v 1.000000 -1.000000 -1.000000
v 1.000000 -1.000000 1.000000
v -1.000000 -1.000000 1.000000
v -1.000000 -1.000000 -1.000000
v 1.000000 1.000000 -0.999999
v 0.999999 1.000000 1.000001
v -1.000000 1.000000 1.000000
v -1.000000 1.000000 -1.000000
# 8 vertices, 0 vertices normals

f 2 3 4
f 8 7 6
f 5 6 2
f 6 7 3
f 3 7 8
f 1 4 8
f 1 2 4
f 5 8 6
f 1 5 2
f 2 6 3
f 4 3 8
f 5 1 8
# 12 faces, 0 coords texture
```

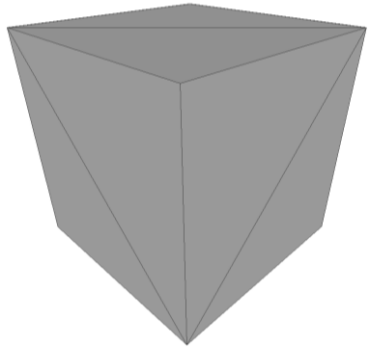
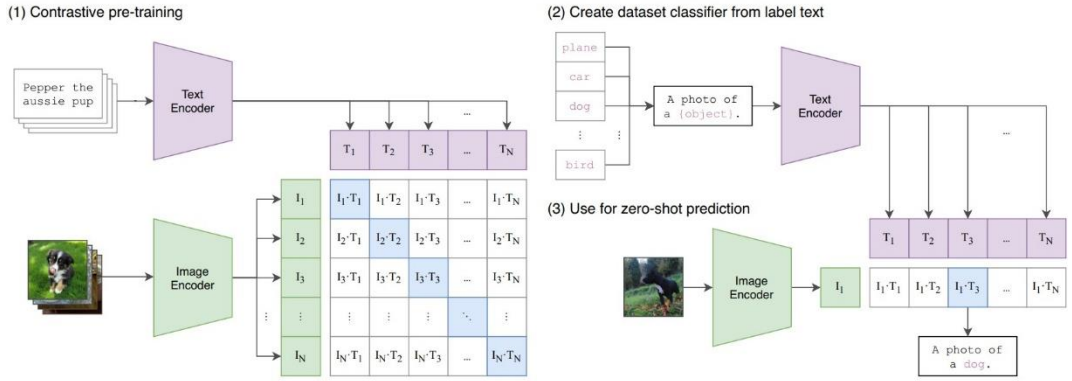


Image reference: [Wang et al., 2021]

Related work: Generative AI in image

- CLIP
  - Connect text and image
  - Encoders serve as bases for other pipelines



- Diffusion model
  - Add Gaussian noise at each time step in the training stage and reverse the process in the inference stage
  - Stable Diffusion: text-to-image, depth-to-image generation
  - Others: InstructPix2Pix (text-guided image edit), SDEdit (denoising), Zero-1-to-3 (view-aware image generation)

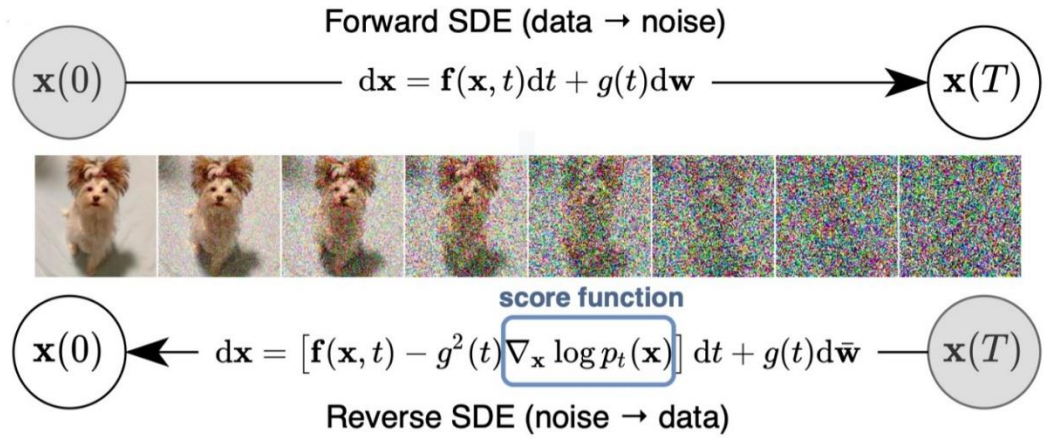


Image Reference: [Radford et al., 2021] [Po et al., 2023]

Related work: Generative AI in 3D

- Utilize 2D text-guided pre-trained model to guide 3D content

Implicit

- Edit the base multi-view images dataset
  - Instruct-nerf2nerf: iteratively dataset update
- Backpropagate the image loss to the 3D implicit neural model
  - Dreamfusion: Score Distillation Sampling (SDS) loss
  - Posterior Distillation Sampling: PDS loss
  - Zero123: view aware image generation

Explicit (mesh)

- Geometry and texture joint edit
  - Text2Mesh
  - CLIP-Mesh
  - X-Mesh
- Texture generation
  - TANGO
  - Latent-paint
  - Fantasia3D
  - TEXTure and Text2Tex
  - Paint-it

Methodology and result: overview

General idea

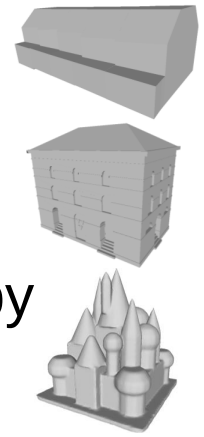
- First experiment with implicit 3D representation (NeuS) based edit: **fail**
- Focus on explicit 3D representation (Mesh) based edit instead
- Guidance engineering in 2D space
- Chosen representative existing pipelines: Latent-Paint, Text2Tex, X-Mesh
- Modifications (6): based on Text2Tex and X-Mesh
- Evaluations: qualitative and quantitative (user study)

Data (implicit)

- DTU MVS
- NeRF-Synthetic

Data (explicit)

- 3D BAG (LoD 2.2)
- 3D Warehouse
- Sample provided by the X-Mesh paper



Tool

- Programming language: Python
- Mesh processing and visualization: Meshlab

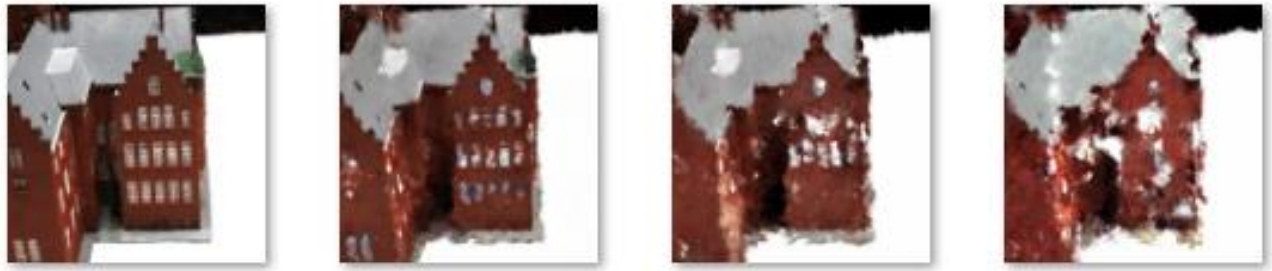
Implicit 3D representation (NeuS) based edit

Original image

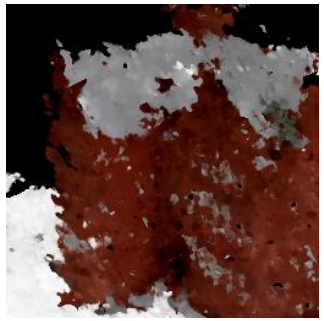
Text prompt: Make it a church

- Iteratively update image dataset

Iteration increases



Basic: update one image at a time

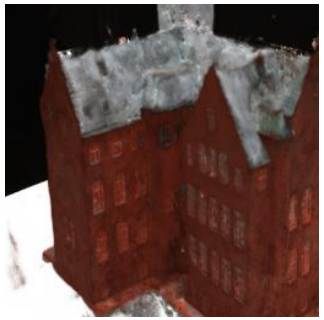


Update the whole dataset

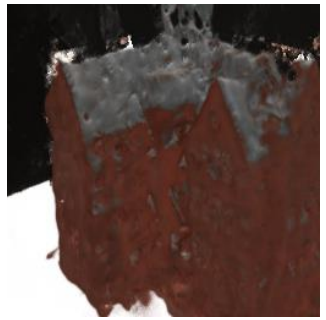
Manual view selection

Combine SDEdit

- Incorporate 2D image loss to 3D model

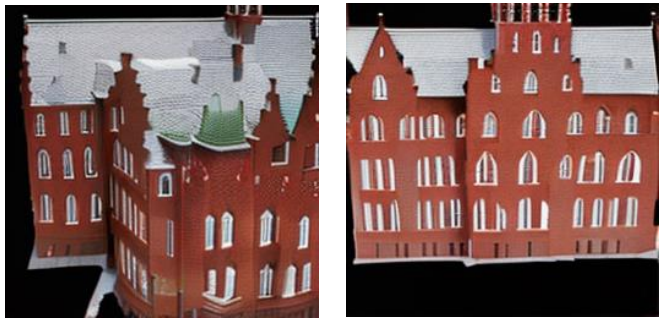


SDS loss



PDS loss

Away from the input view



Stable Zero123

Representative existing pipelines

Name	Latent-Paint	Text2Tex	X-Mesh
Texture creation	yes	yes	yes
Geometry edit	no	no	yes
Pre-trained model	text-to-image Stable Diffusion	depth-to-image Stable Diffusion	CLIP

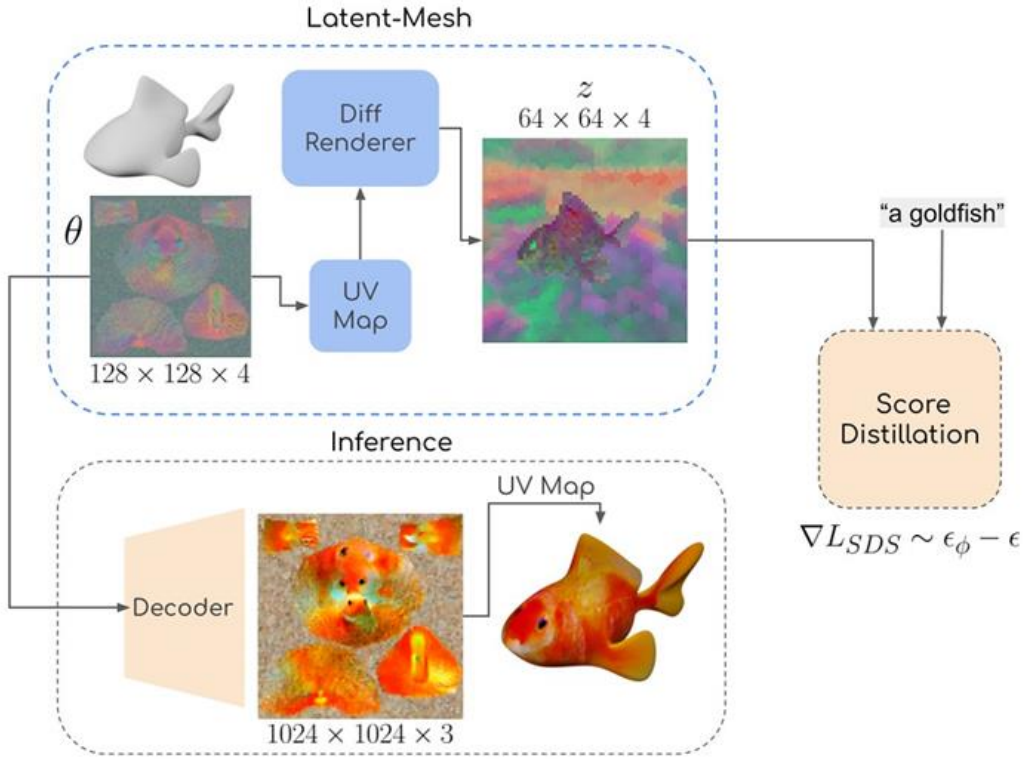


Image reference: [Metzer et al., 2023]

Representative existing pipelines

Name	Latent-Paint	Text2Tex	X-Mesh
Texture creation	yes	yes	yes
Geometry edit	no	no	yes
Pre-trained model	text-to-image Stable Diffusion	depth-to-image Stable Diffusion	CLIP

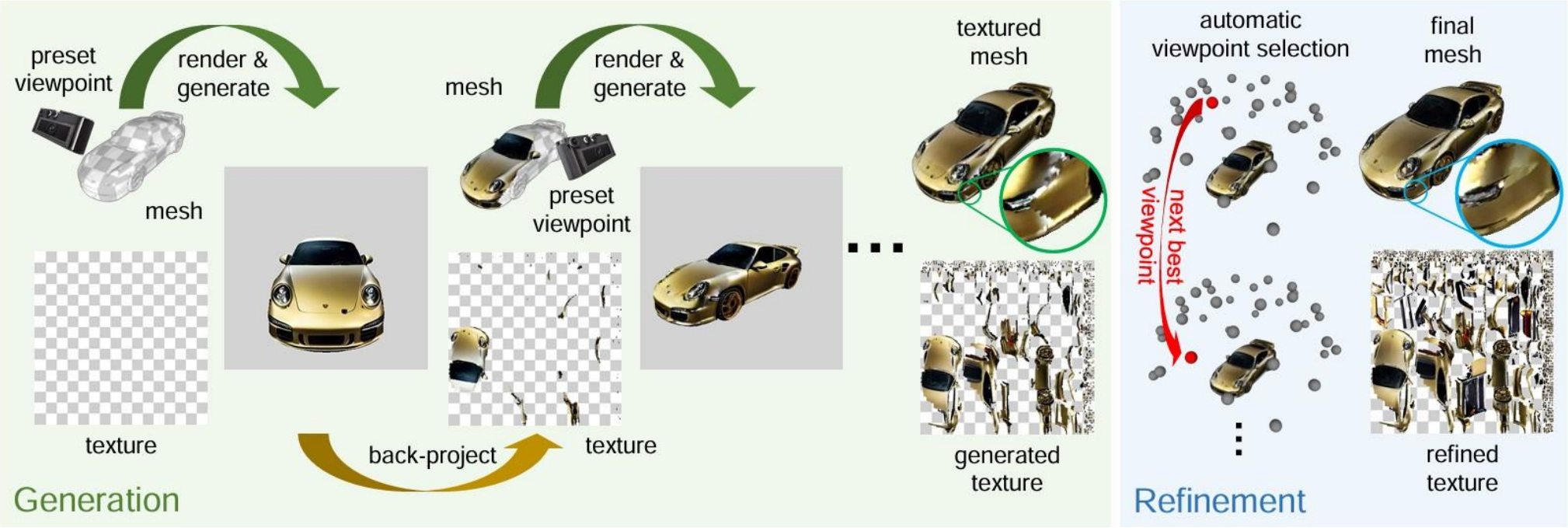
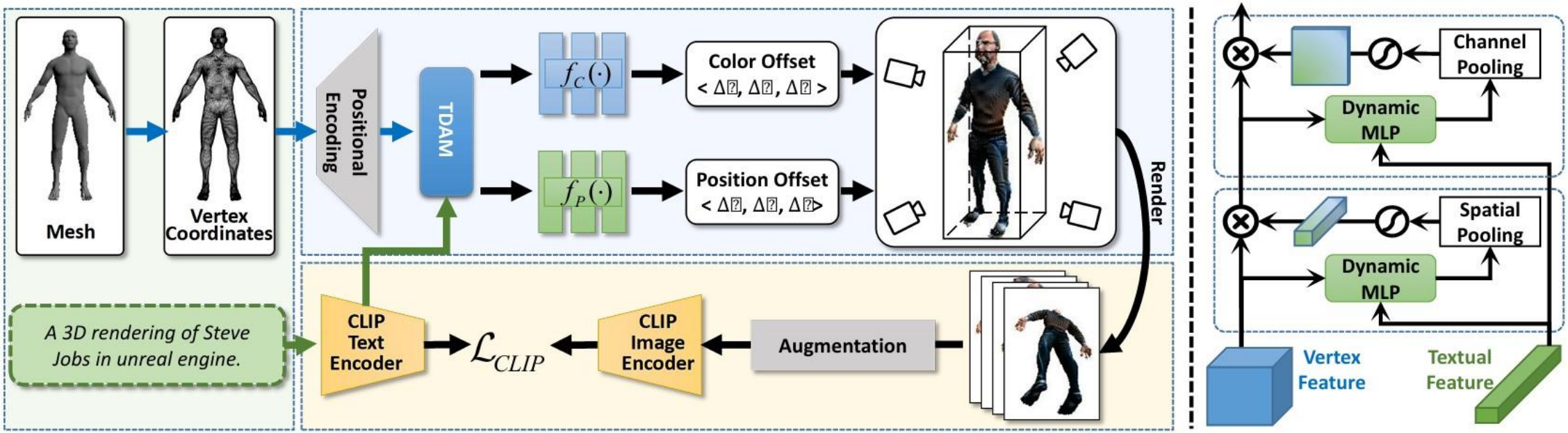


Image Reference: [Chen et al., 2023a]

Representative existing pipelines

Name	Latent-Paint	Text2Tex	X-Mesh
Texture creation	yes	yes	yes
Geometry edit	no	no	yes
Pre-trained model	text-to-image Stable Diffusion	depth-to-image Stable Diffusion	CLIP



(a) The proposed X-Mesh

(b) Text-Guided Dynamic Attention

Image Reference: [Ma et al., 2023]

Guidance engineering in 2D space

Text for CLIP and Stable Diffusion

- clear and specific
- keywords for realistic style
- keywords for camera location

"An exterior four-storey red apartment with grey roof"

"Apartment"

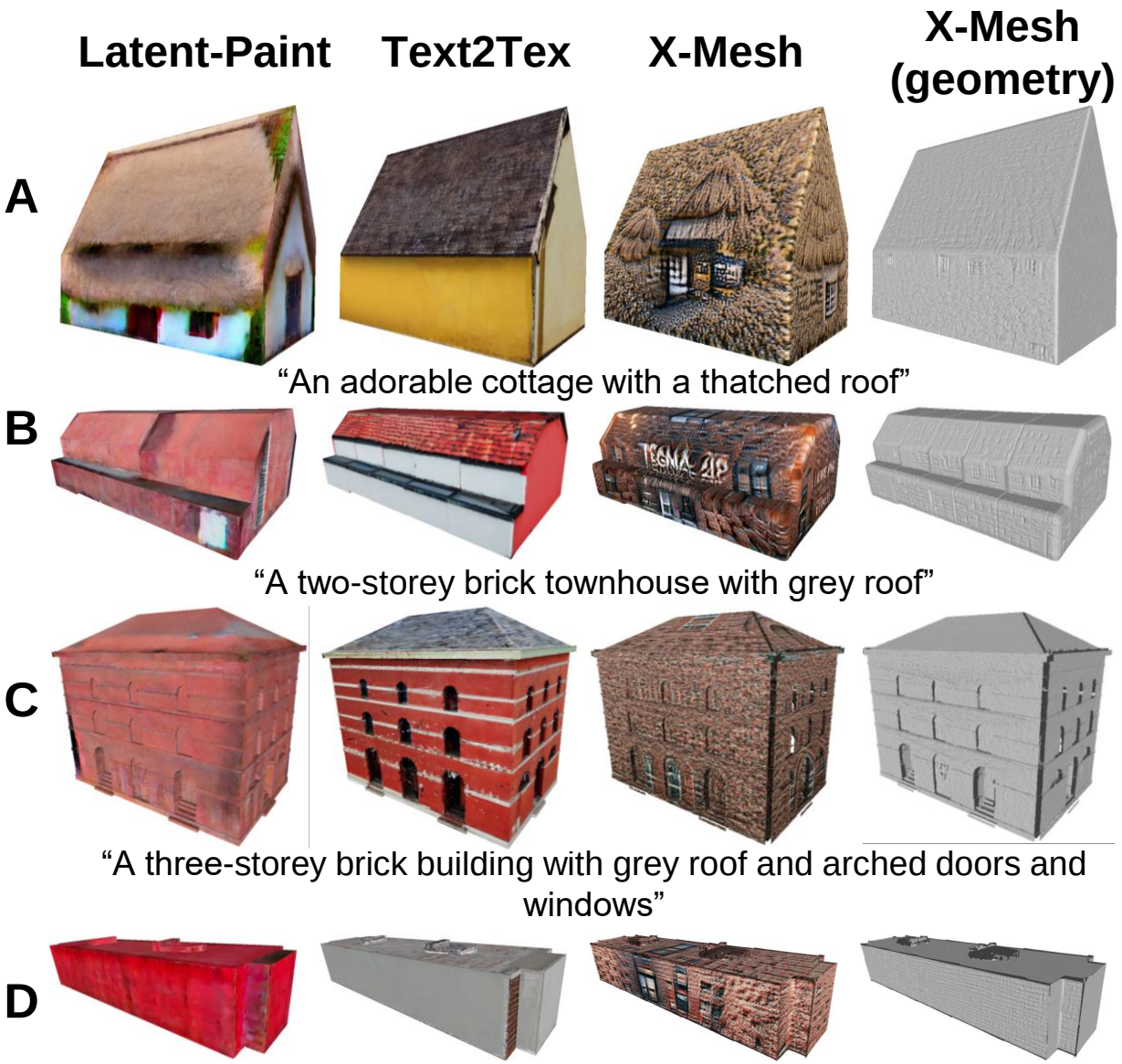
Image (for Image control X-Mesh)

- close-up and unobstructed buildings
- CLIP Interrogator: can not return perfectly matching text

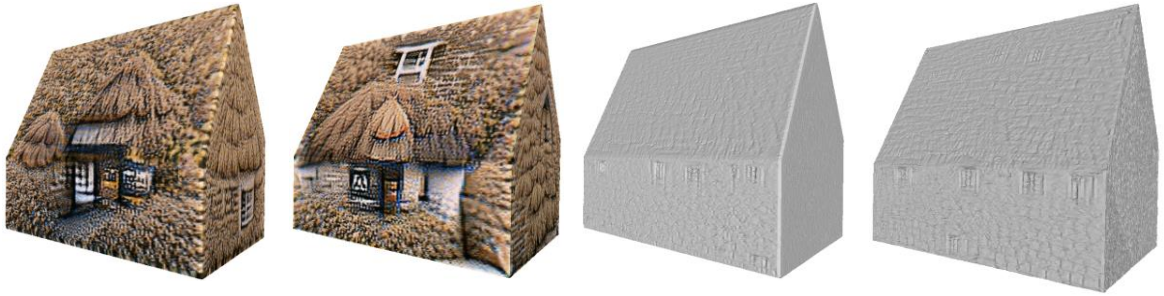
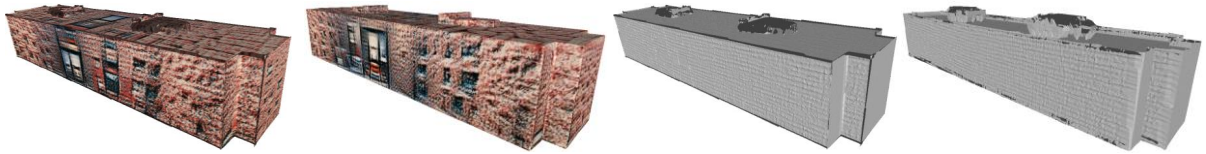
A proper image prompt

Image generated by Stable Diffution using Text prompt returned by CLIP interrogator

Representative existing pipelines



Representative existing pipelines

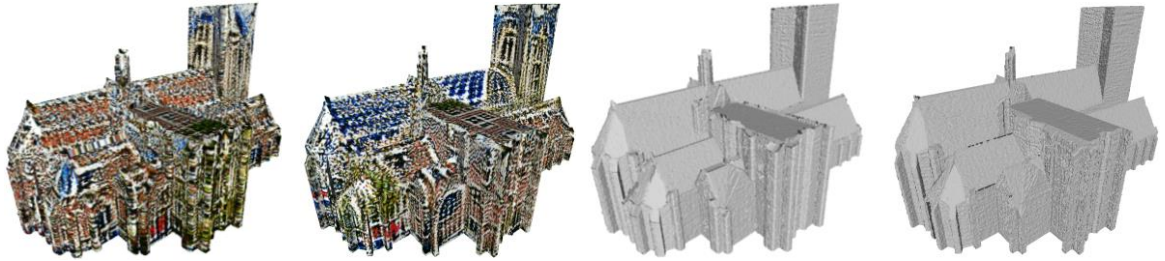
	Original X-Mesh	X-Mesh with only photometric loss	Original X-Mesh (geometry)	X-Mesh with only geometry loss
A				
B				
C				
D				

“An adorable cottage with a thatched roof”

“A two-storey brick townhouse with grey roof”

“A three-storey brick building with grey roof and arched doors and windows”

“An exterior brick apartment”

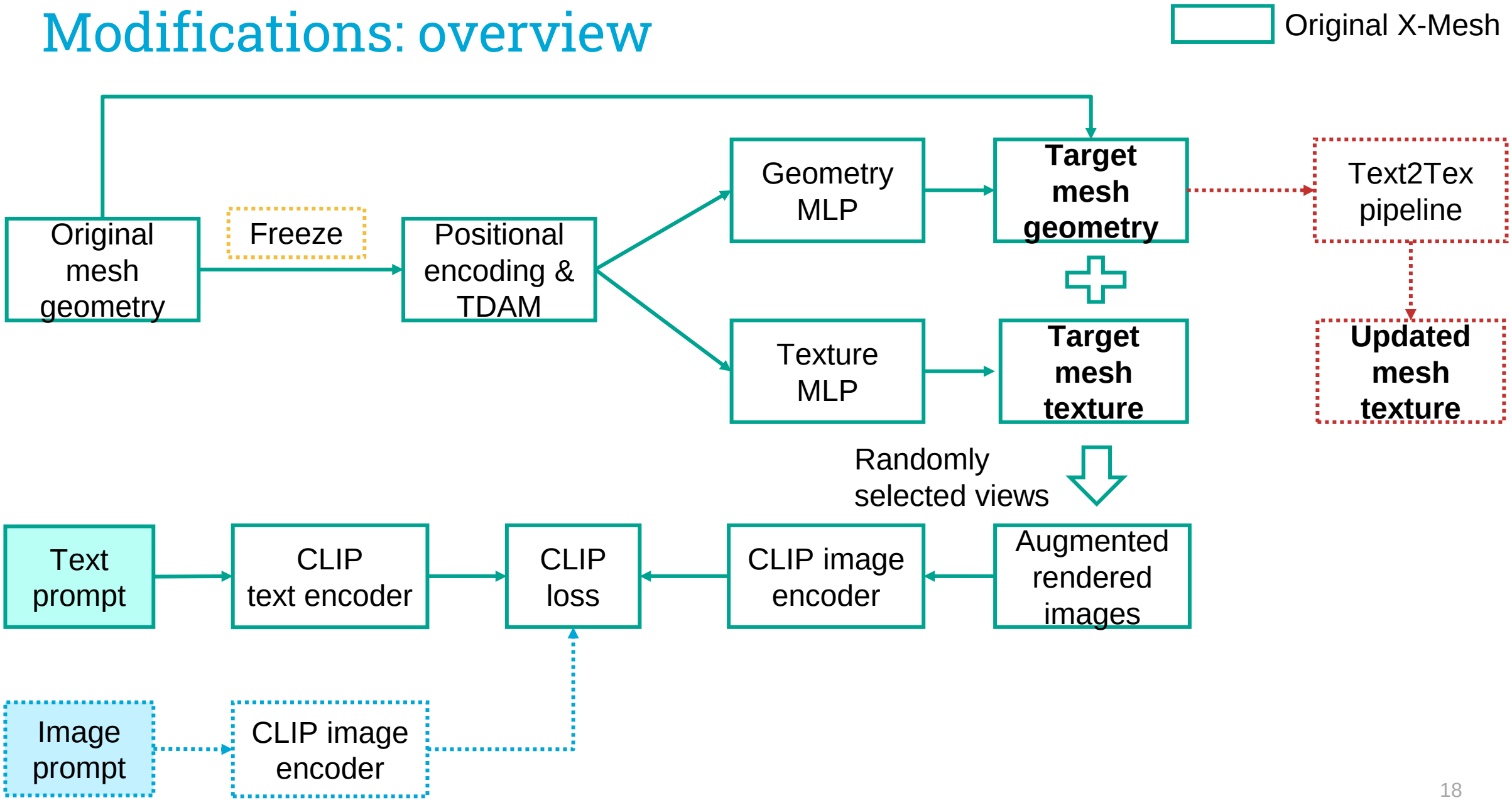
	Original X-Mesh	X-Mesh with only photometric loss	Original X-Mesh (geometry)	X-Mesh with only geometry loss
E				
F				
G				

“An exterior modern high glass window office”

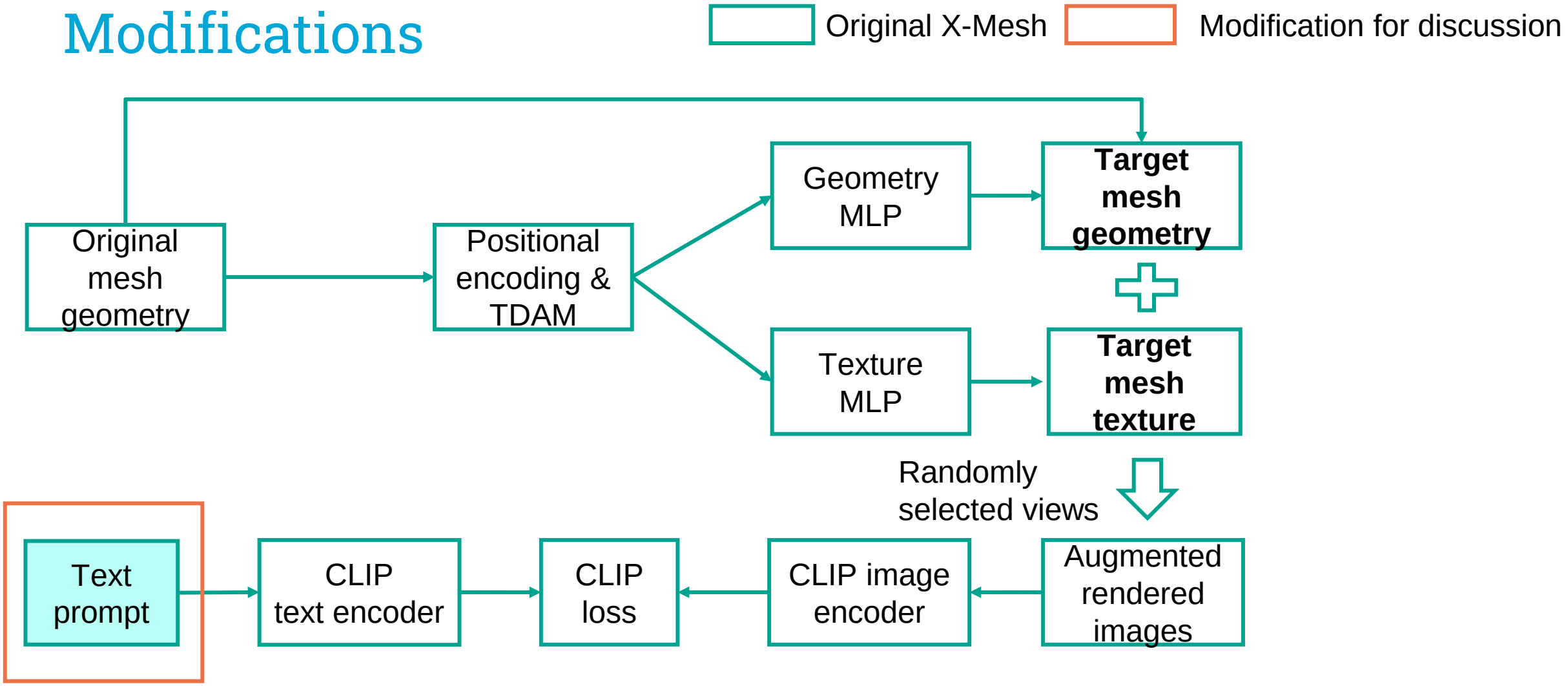
“An old church delft”

“A brick castle”

Modifications: overview



Modifications

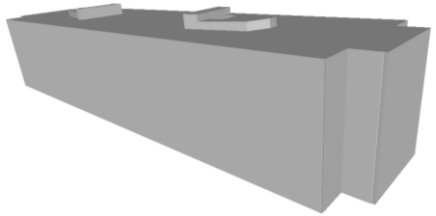


Modifications

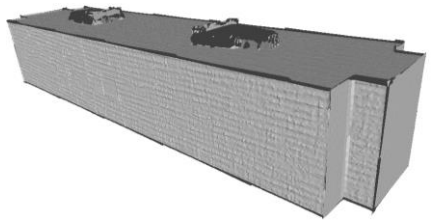
Text prompt

Additional procedure: add view specification prompt

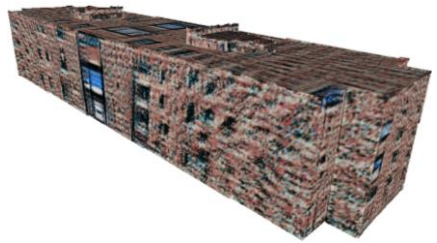
“An exterior brick apartment”

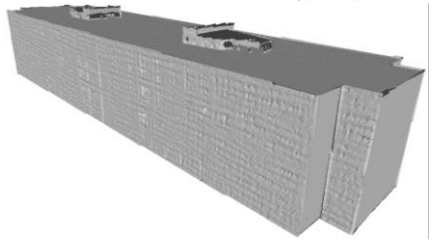


Original mesh :
An exterior brick apartment

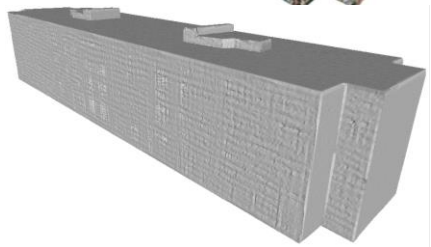


Without view specification





With view specification

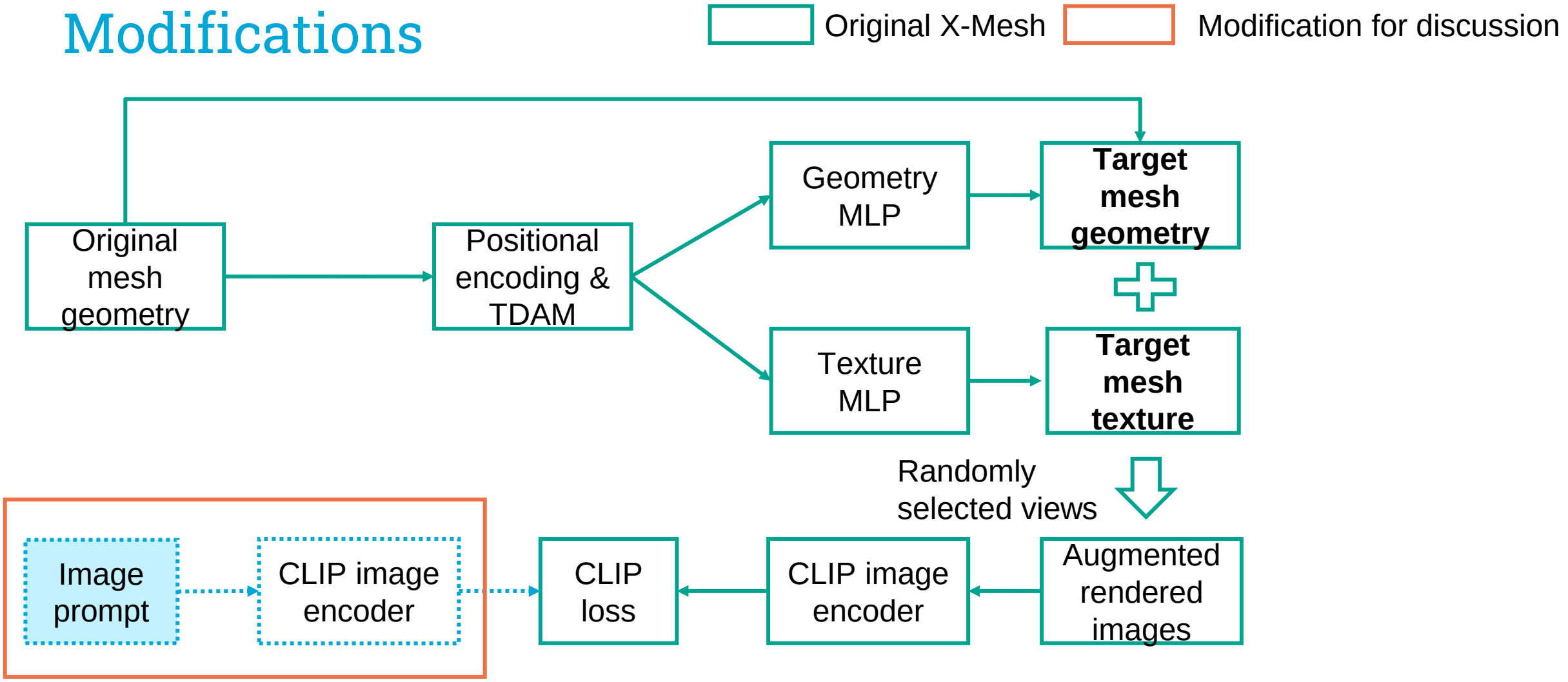


With view specification (Building-related descriptions)

X-Mesh

X-Mesh (geometry)

Modifications



Modifications: Image control X-Mesh

Image control module

Alternative module: use image prompt

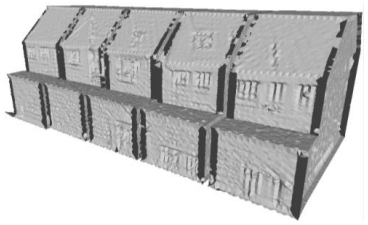
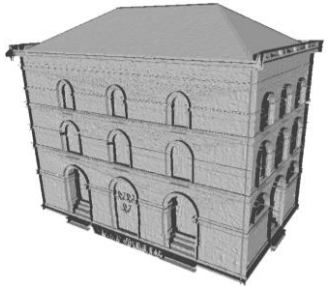
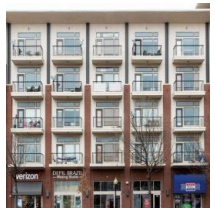
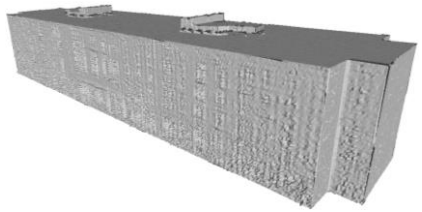
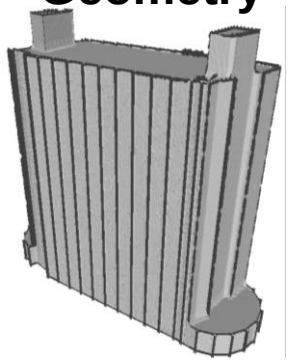
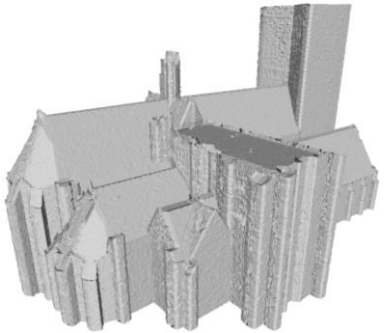
	Image prompt	Output	Geometry
A			
B			
C			
D			

	Image prompt	Output	Geometry
E			
F			
G			

Modifications: Image control X-Mesh

Image control module

Additional procedure: weight specification on input image view

Original mesh

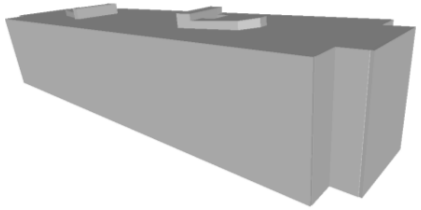


Image prompt

Without view weight specification

With view weight specification

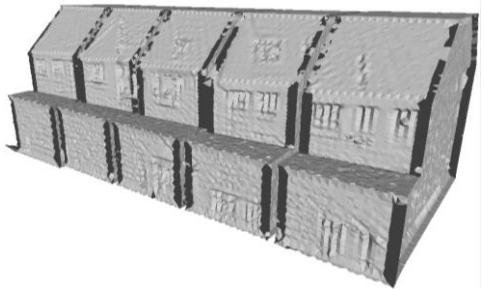
Modifications: Image control X-Mesh

Image control module

Additional procedure: edit façade and roof separately

Complete image prompt

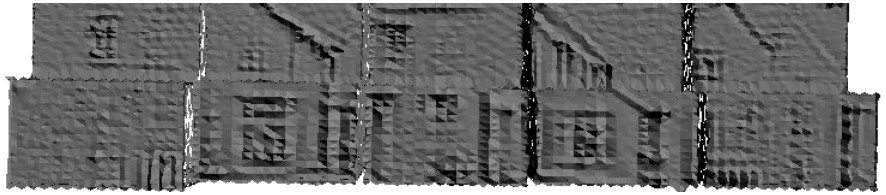
Complete result



Complete result (geometry)

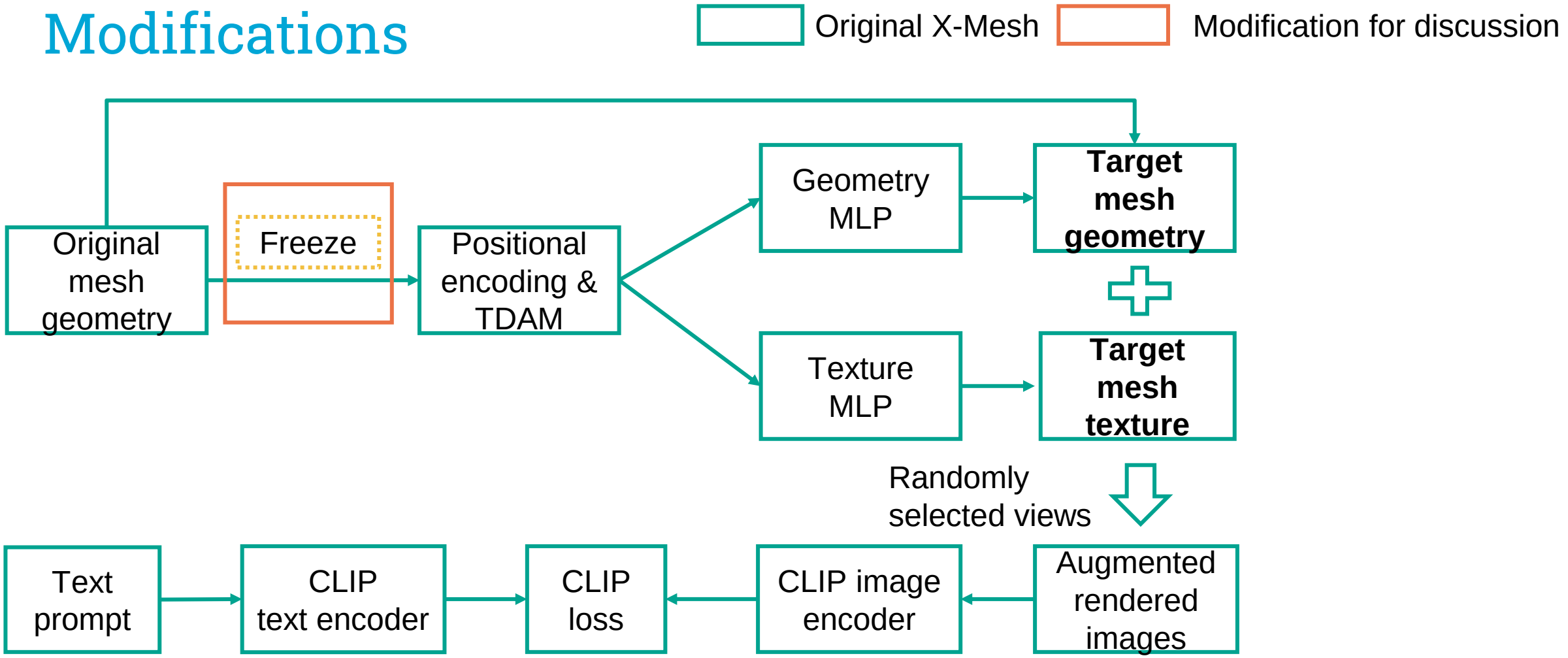
Facade image prompt

Façade result



Façade result (geometry)

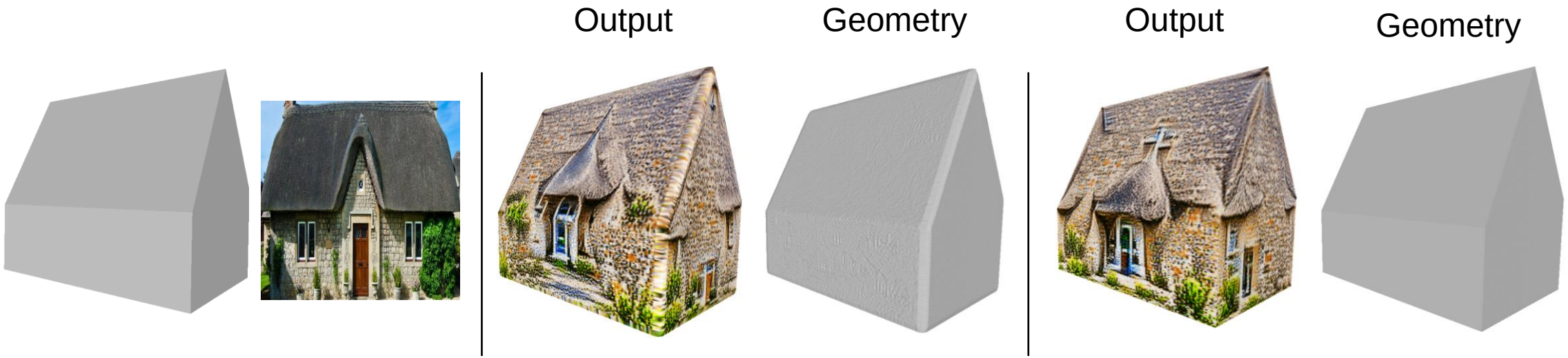
Modifications



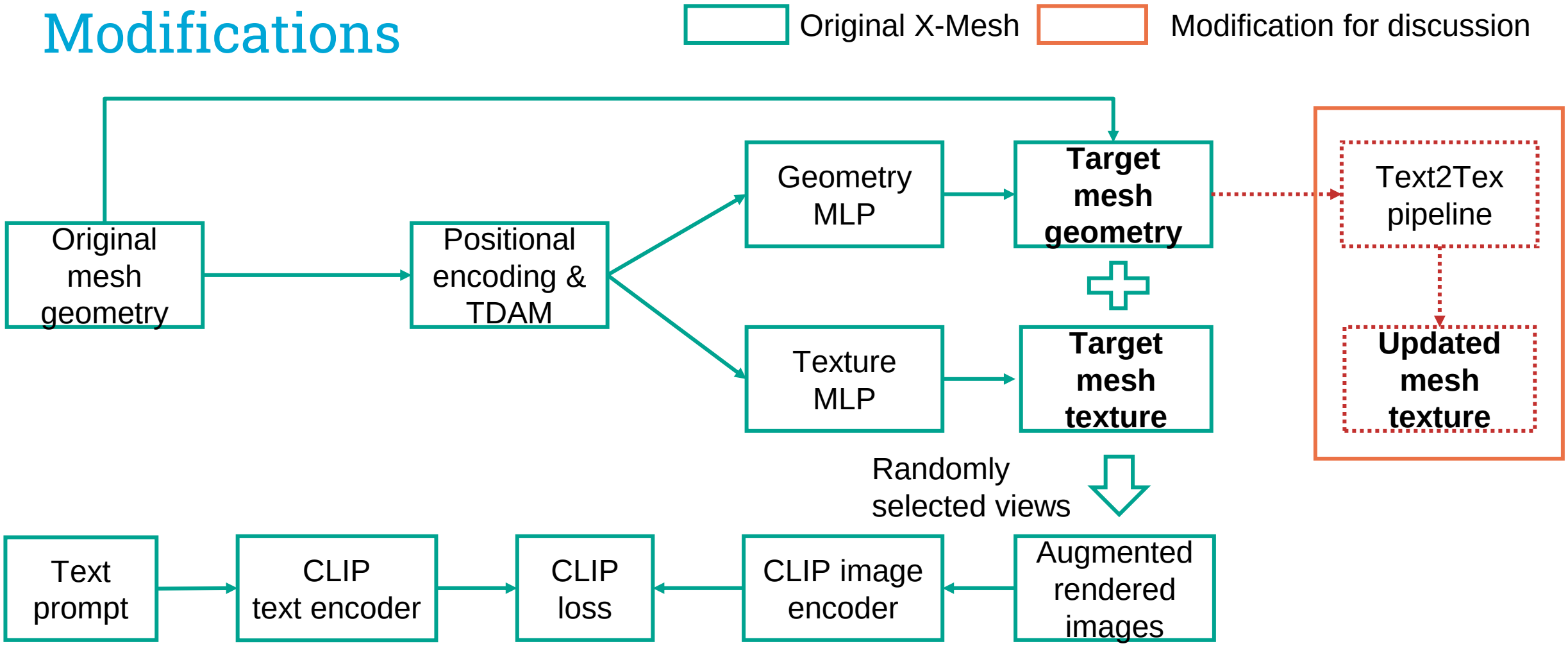
Modifications

Freeze

Additional procedure: freeze randomly sampled vertices geometry



Modifications



Modifications: Combine X-Mesh and Text2Tex

Text2Tex module

Additional module: use Text2Tex to update texture

A

“An adorable cottage with a thatched roof”

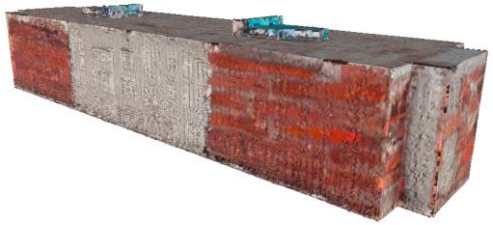
B

“A two-storey brick townhouse with grey roof”

C

“A three-storey brick building with grey roof and arched doors and windows”

D



“An exterior brick apartment”

E

“An exterior modern high glass window office”

F

“An old church delft”

G

“A brick castle”

Quantitative results

- Score on how realistic the image is (1 lowest, 5 highest)
- 50 respondents

Overall average user score

- Latent-Paint: 1.98
- Text2Tex: 2.73
- X-Mesh: 2.31
- Image control X-Mesh: **2.79**
- Combination of X-Mesh and Text2Tex: **2.86**

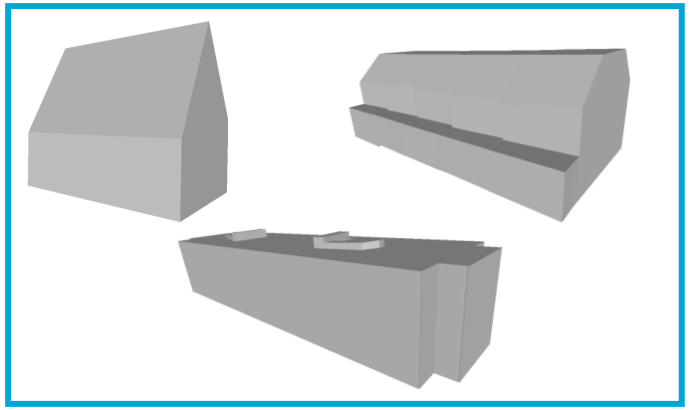
Quantitative results

Ranking of separate average user score

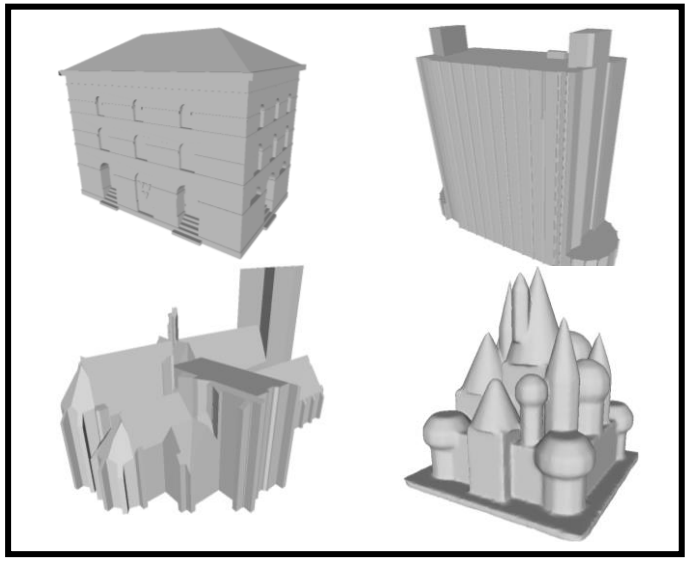
(1 highest ranking, 5 lowest ranking)

	Latent -Paint	Text2 Tex	X-Mesh	Image control X-Mesh	Combination of X-Mesh and Text2Tex
Model A	4	5	3	1	2
Model B	5	3	4	2	1
Model D	5	4	2	1	3
Model C	5	2	4	3	1
Model E	4	1	5	2	2
Model F	3	1	4	5	2
Model G	4	1	5	2	3

Simple model

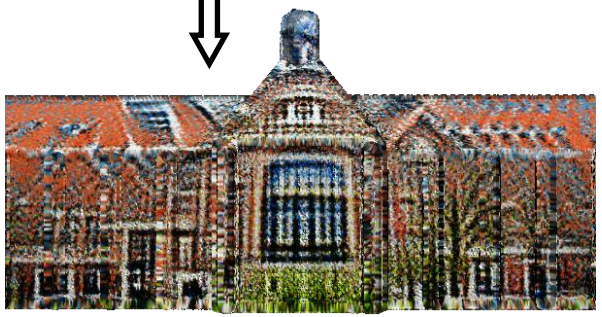


Complex model

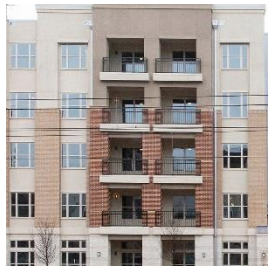


Application examples: Image control X-Mesh

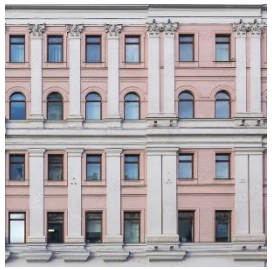
- Single building edit
- Scene edit by combing



Faculty building

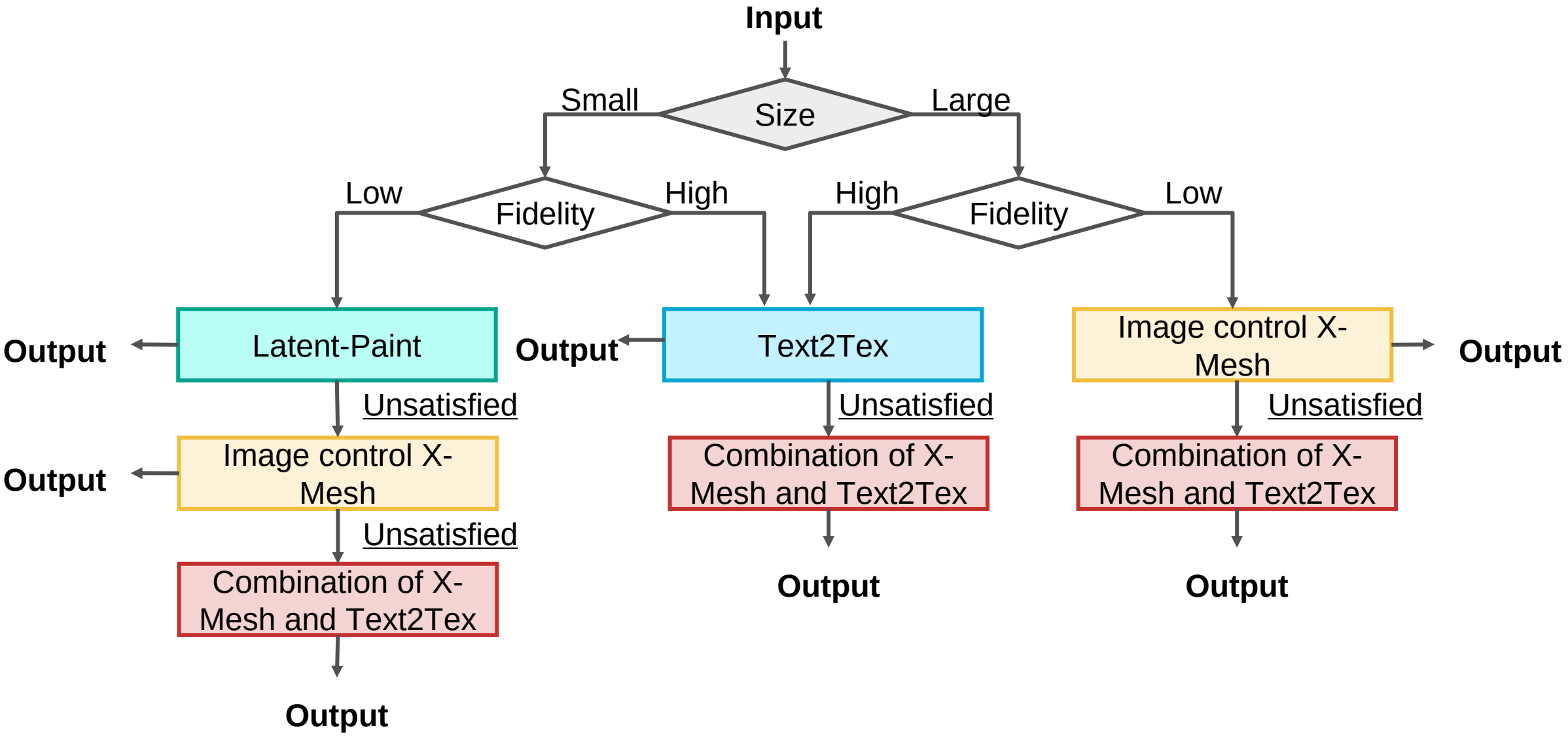


Modern style residential area



Classical style residential area

Conclusion


Input
Size
Small
Large
Fidelity
Low
High
Fidelity
High
Low
Latent-Paint
Output
Unsatisfied
Image control X-Mesh
Output
Unsatisfied
Combination of X-Mesh and Text2Tex
Output
Text2Tex
Output
Unsatisfied
Combination of X-Mesh and Text2Tex
Output
Image control X-Mesh
Output
Unsatisfied
Combination of X-Mesh and Text2Tex
Output

Conclusion

✓ Contributions

- ✓ Evaluate performances of representative 3D edit pipelines in building domain
- ✓ Make modifications on them to better fit the demand of users and generate high-quality results

☐ Limitations

- ☐ Suitable for simple cases: limited geometric edit scope and detailed level of textures
- ☐ Inherit the limits of the 2D pre-trained models: limited generalization ability and view consistency problems
- ☐ High computational demand: limited performances especially for large and complex buildings

GitHub repository

- ✓ Easy-to-use codes for two successful modifications:
  - Image Control X-Mesh
  - Combination of X-Mesh and Text2Tex
- ✓ Example data and results

https://github.com/fengyingxin/MSc-Thesis

Reference

Armandpour, M., Zheng, H., Sadeghian, A., Sadeghian, A., and Zhou, M. (2023). Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*.

Biljecki, F., Ledoux, H., and Stoter, J. (2016). An improved lod specification for 3d building models. *Computers, Environment and Urban Systems*, 59:25–37.

Brooks, T., Holynski, A., and Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. *arXiv e-prints*, pages arXiv–2211.

Chao, C.-K. T. and Gingold, Y. (2023). Text-guided image-and-shape editing and generation: A short survey. *arXiv preprint arXiv:2304.09244*.

Chen, D. Z., Siddiqui, Y., Lee, H.-Y., Tulyakov, S., and Nießner, M. (2023a). Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18558–18568.

Chen, M., Xie, J., Laina, I., and Vedaldi, A. (2023b). Shap-editor: Instruction-guided latent 3d editing in seconds. *arXiv preprint arXiv:2312.09246*.

Chen, R., Chen, Y., Jiao, N., and Jia, K. (2023c). Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*.

Chen, Y., Chen, R., Lei, J., Zhang, Y., and Jia, K. (2022). Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems*, 35:30923–30936.

Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G., et al. (2008). Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pages 129–136. Salerno, Italy.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Dong, J. and Wang, Y.-X. (2023). Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fang, S., Wang, Y., Yang, Y., Tsai, Y.-H., Ding, W., Zhou, S., and Yang, M.-H. (2023). Editing 3d scenes via text prompts without retraining. *arXiv e-prints*, pages arXiv–2309.

Guo, Y.-C., Liu, Y.-T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.-H., Zou, Z.-X., Wang, C., Cao, Y.-P., and Zhang, S.-H. (2023). threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio.

Haque, A., Tancik, M., Efros, A. A., Holynski, A., and Kanazawa, A. (2023). Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hong, S., Ahn, D., and Kim, S. (2023). Debiasing scores and prompts of 2d diffusion for view-consistent text-to-3d generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jpcy (2022). Xatlas. https://github.com/jpcy/xatlas.

Kamata, H., Sakuma, Y., Hayakawa, A., Ishii, M., and Narihira, T. (2023). Instruct 3d-to-3d: Text instruction guided 3d-to-3d conversion. *arXiv preprint arXiv:2303.15780*.

Koo, J., Park, C., and Sung, M. (2023). Posterior distillation sampling. *arXiv preprint arXiv:2311.13831*.

Li, C., Zhang, C., Waghwase, A., Lee, L.-H., Rameau, F., Yang, Y., Bae, S.-H., and Hong, C. S. (2023a). Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131*.

Li, W., Chen, R., Chen, X., and Tan, P. (2023b). Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., and Vondrick, C. (2023). Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309.

Ma, Y., Zhang, X., Sun, X., Ji, J., Wang, H., Jiang, G., Zhuang, W., and Ji, R. (2023). X-mesh: Towards fast and accurate text-driven 3d stylization via dynamic textual guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2749–2760.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2021). Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Metzer, G., Richardson, E., Patashnik, O., Giryes, R., and Cohen-Or, D. (2023). Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673.

Michel, O., Bar-On, R., Liu, R., Benaim, S., and Hanocka, R. (2022). Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., and Wang, W. (2021). Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, Q., Zhang, B., Birsak, M., and Wonka, P. (2023a). Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. (2023b). Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., and Lipman, Y. (2020). Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502.

Youwang, K., Oh, T.-H., and Pons-Moll, G. (2023). Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. *arXiv preprint arXiv:2312.11360*.

Mohammad Khalid, N., Xie, T., Belilovsky, E., and Popa, T. (2022). Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8.

Müller, T., Evans, A., Schied, C., and Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Pharmapsychotic (2023). clip-interrogator. https://github.com/pharmapsychotic/clip-interrogator.

Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J. T., Bermano, A. H., Chan, E. R., Dekel, T., Holynski, A., Kanazawa, A., et al. (2023). State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204*.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., and Cohen-Or, D. (2023). Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510. IEEE.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Shen, T., Gao, J., Yin, K., Liu, M.-Y., and Fidler, S. (2021). Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101.

Shi, Y., Wang, P., Ye, J., Long, M., Li, K., and Yang, X. (2023). Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Song, L., Cao, L., Gu, J., Jiang, Y., Yuan, J., and Tang, H. (2023). Efficient-nerf2nerf: Streamlining text-driven 3d editing with multiview correspondence-enhanced diffusion models. *arXiv preprint arXiv:2312.08563*.

Stabilityai (2023). Stable zero123. https://huggingface.co/stabilityai/stable-zero123.

3D building model edit with generative AI

Student: Yingxin Feng

Supervisors: Nail Ibrahimli Dr. Ken Arroyo Ohori

Co-reader: Dr. Liangliang Nan

TU Delft