FacadeNet: Conditional Facade Synthesis via Selective Editing – Supplementary Material –

Appendix A: Target viewing vectors

Here we discuss the construction process of the target viewing vectors θ^{gt}_{target}, which are crucial for optimizing FacadeNet. To achieve this, we leverage the pre-processed rectified facades of panoramic images found in the the Large Scale Architectural Asset dataset [5]. Building facades were already extracted from each panoramic street-view image, during the development of the LSAA dataset. Moreover, for each panorama two rectified planes, Π^{left}_{rect} and Π^{right}_{rect} are constructed, that cover the entire panoramic image. Each rectified plane has a predefined width W_Π and height H_Π and a horizontal and vertical field of view that span in the range of $[-75°, +75°]$. Following a standard image rectification process, each extracted facade is mapped to one of the two rectified planes, according to its location in the panorama, by maintaining its viewing angles $\theta^{\mathbf{f}}_h$ and $\theta^{\mathbf{f}}_v$. These angles denote the viewing directions of the center pixel of the rectified facade image along the horizontal and vertical axes, w.r.t. to a camera that is positioned in the center of the rectified plane, .

Utilizing this information, for a rectified plane we construct two horizontal and vertical viewing vectors according to their corresponding field of views, normalized in the range $[-1, +1]$, and the spatial dimensionality of each axis, denoted as $\theta^{\Pi}_h \in [-1, +1]^{W_\Pi}$ and $\theta^{\Pi}_v \in [-1, +1]^{H_\Pi}$ respectively. Each element of these viewing vectors, captures the horizontal and vertical viewing directions of each pixel in the rectified plane. Then, we construct the horizontal target viewing vector $\theta^{\mathbf{f}}_h$ and vertical target viewing vector $\theta^{\mathbf{f}}_v$ of the rectified facade, whose dimensionality is equal to the width $W_{\mathbf{f}}$ and $H_{\mathbf{f}}$ of the facade. Finally, by treating the rectified plane's viewing vectors as lookup tables, we find the position of each facade in these, based on the values of $\theta^{\mathbf{f}}_h$ and $\theta^{\mathbf{f}}_v$, and extract the viewing angles for the facade's target viewing vectors (see Figure 1). For the synthesis of novel facades, we modify the target viewing vectors of the reference facade by adding a constant negative or positive offset, in order to influence its viewing direction from left to right and top to bottom.

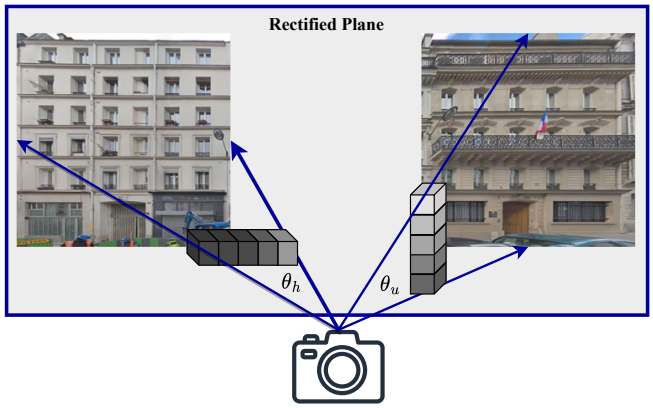

Figure 1. Facades can occupy various positions on the rectified plane, each position denoting specific horizontal θ_h and vertical θ_v view direction targets for the facade image. These view direction targets encompass the location of the facade within the rectified plane, effectively determining the perspective from which it is viewed. By referring to the accompanying figure, it becomes evident that the values of the horizontal and vertical vectors (h and v) undergo gradual changes as the points shift across the rectified plane. The spatial location of the rectified plane captures identical view directions for distinct facades

Appendix B: FacadeNet analysis

Ablation. We conducted a comprehensive ablation study, presented in Table 1, to examine the influence of our design choices on both image synthesis quality and content consistency with respect to reference images. Our model is based on an Encoder-Decoder(Generator) architecture where the reference facade image f is encoded into a latent tensor z that captures the structure and the style of the input image. Subsequently, leveraging the latent matrix z, our model generates a coherent representation of the reference facade from various viewpoints by conditioning the generative process using view target vectors.

$FacadeNet_{base}$ serves as our baseline and represents the existing design choice without the selective editing module. Utilizing the selective editing mask (**SEM**) $FacadeNet_A$ achieves higher image quality in novel view synthesis, reducing the metric from 10.59 to 9.74. In the case where the **features** used for the selective editing module is set to **semantics**, we refer to *fixed masks* that utilize

Method	SEM	features	# Views	LPIPS-alex↓	LPIPS-vgg↓	FID$_{rec}$ ↓	FID$_{novel}$ ↓	PSNR↑	SSIM↑
$FacadeNet_{base}$	–	–	1	0.174	0.296	10.59	9.91	24.13	0.69
$FacadeNet_A$	✓	semantics	1	0.147	0.265	9.74	9.28	24.13	0.693
$FacadeNet_B$	✓	semantics	2	0.143	0.262	9.66	8.89	24.25	0.698
$FacadeNet_C$	✓	semantics	4	0.135	0.247	10.32	8.63	24.62	0.717
$FacadeNet_D$	✓	semantics	6	0.136	0.255	9.96	9.64	24.55	**0.718**
$FacadeNet_E$	✓	DINO	1	0.143	0.261	9.626	8.971	23.80	0.708
$FacadeNet_F$	✓	DINO	2	0.128	0.250	9.645	8.769	**24.77**	0.712
$FacadeNet_G$	✓	DINO	4	**0.119**	**0.240**	**9.601**	**8.327**	23.866	0.714
$FacadeNet_H$	✓	DINO	6	0.145	0.265	12.39	11.453	23.91	0.708

Table 1. Our ablation study aims to assess the performance of various design choices we employed in our work. Specifically, we investigate the impact of **selective editing mask (SEM)**, the choice of **features** used as priors for computing the selective mask and the number of novel views per iteration (**#Views**). From our findings, we conclude that SEM yields to the most significant improvement in terms of both novel view consistency and image quality. Notably, we observe a substantial enhancement when employing DINO features and learnable weights to combine them, as opposed to manually selected semantic groups as editing masks. Moreover, using more than 1 view for each iteration improves the results regarding novel inter-view consistency even further.

semantic maps and a manually selected group of classes to create a binary mask assigning 1 to selected classes and 0 otherwise. Notably, it significantly outperforms the baseline in terms of consistency, with the $LPIPS - alex$ and $LPIPS - vgg$ metrics improving from 0.174 to 0.147 and from 0.296 to 0.265, respectively.

$FacadeNet_E$ demonstrates that learnable masks are better suited for the novel view synthesis task. $DINOViT$ [1] proves to be a valuable resource, offering meaningful and useful localized features that can effectively be utilized in methodologies similar to ours without the need for supervision. The binary format of fixed masks seems to have a disadvantage in contrast to the continuous representation of masks that are extracted from $DINOViT$ features [1]. Moreover, $DINOViT$ features provide the freedom to the network to choose the group of features that are required to alter for novel view synthesis in contrast to fixed $semantics$ features. In terms of quality $FacadeNet_E$ enhances the FID_{novel} value for novel view synthesis from 9.28 of $FacadeNet_A$ to $8,971$. Additionally, it achieves better scores, reducing the $LPIPS - alex$ metric from 0.147 to 0.143 and the $LPIPS - vgg$ metric from 0.265 to 0.261.

Furthermore, it is evident that incorporating multiple views during training brings significant benefits. Intuitively, this approach provides a multi-view consistency, preventing the model from being misled and generating incompatible results between different views. The multi-view version of $FacadeNet$ outperforms their single-view counterparts, regardless, of the other attributes being used in the model. Multi-view training primarily enhances the quality and consistency of novel facade synthesis, while also yielding slight improvements in reconstruction tasks. As illustrated $FacedeNet_{B,C,D}$ clearly outperform $FacedeNet_A$ regarding $LPIPS - alex$, $LPIPS - vgg$ and FID_{novel} while the same observation stands for $FacedeNet_{F,G,H}$ in contrast to $FacedeNetE$. Among our models, $FacadeNet_H$ emerges as the best-performing version. It combines the selective editing module, learnable edit masks and multi-view training, resulting in superior performance compared to other versions of $FacadeNet$. $FacadeNet_H$ is referenced as $FacadeNet_{full}$ in the main paper.

Selective editing improvements. To visually assess the effectiveness of our selective editing module, we conducted a comparison between our $FacadeNet_{full}$ and our base model, $FacadeNet_{base}$, in order to validate the improvements in consistency. In Figure 2, we present an interpolation of view angles for two facades, emphasizing the enhancements we have achieved.

While both models generate high-quality and believable center images, it is evident that the model trained with the selective editing module exhibits significantly superior consistency. The improvements in maintaining coherence and smooth transitions between the generated views are remarkable when compared to our base model.

In Figure 2, we present the results of $FacadeNet_{full}$ and $FacadeNet_{base}$ in pairs of rows. The top row in each sample corresponds to the outputs generated by $FacadeNet_{full}$, while the rows below depict the results from $FacadeNet_{base}$. Upon careful observation, it becomes apparent that $FacadeNet_{base}$ introduces various artifacts between different view angles. In contrast, $FacadeNet_{full}$ demonstrates a higher level of robustness and maintains the integrity of facades' detailed areas (see highlighted area in the green and red boxes in Figure 7).

In the second sample (rows 3 & 4), $FacadeNet_{full}$ exhibits the ability to discern insignificant features, such as the car that is present in the image, and preserves them consistently across different views. However, $FacadeNet_{base}$

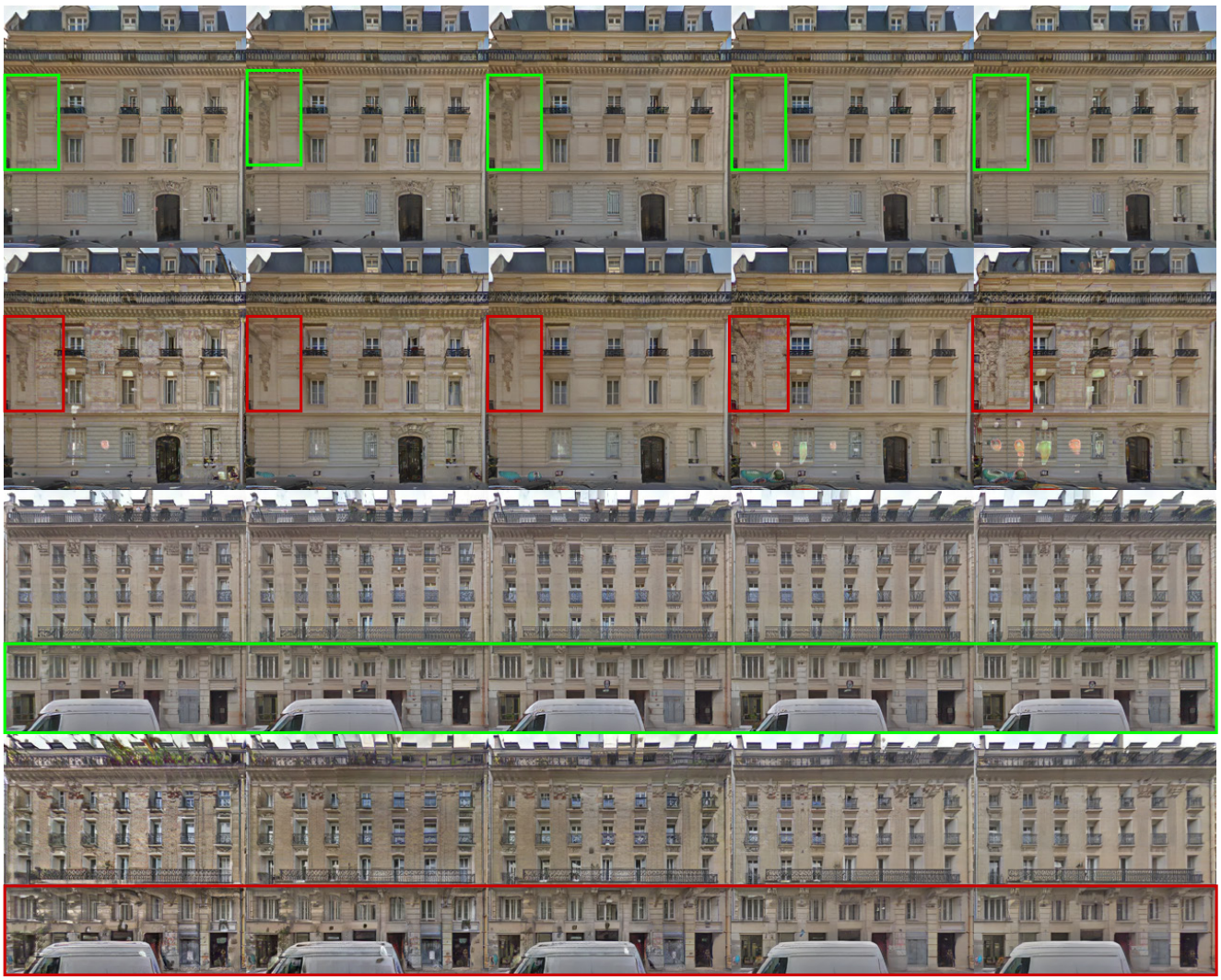

Figure 2. In this figure we display visual comparisons between $FacadeNet_{full}$ and $FacadeNet_{base}$. We observe the emergence of various artifacts when generating examples using $FacadeNet_{base}$ (bottom row of each sample) from different view angles. In contrast, $FacadeNet'_{full}s$ results (top row of each sample) demonstrate a higher level of robustness, effectively preserving the structural details. This distinction is highlighted by the annotated area, where green and red boxes indicate the differences in inter-view consistency.

fails to maintain such details, resulting in distorted and peculiar outcomes (see last row red box in figure 2).

Appendix C: Facade view interpolation

Here, we demonstrate the impact of our horizontal vector θ_h and vertical vector θ_v on specific input facade images, showcasing their influence. By utilizing our encoder, we obtain a latent matrix z based on a reference facade image f_{ref}. Subsequently, our generator employs the same latent matrix z along with different combinations of horizontal θ_h and vertical θ_v targets to generate samples.

Figure 4 illustrates the camera movement on the horizontal axis, moving from right to left (represented by the camera's motion in the top row). Similarly, on the vertical axis, the camera moves from bottom to top (represented by the camera's motion in the left column). This visual representation exemplifies the disentangled controllability of the view angle vectors for both axes. As evident from the results, the generated content precisely aligns with the target vectors.

Figures 7 and 8 present additional results that serve to exemplify the effectiveness of our approach. These visual examples showcase our model's ability to successfully handle a diverse range of structures and architectural styles. Notably, our method maintains the overall style coherence while preserving the distinctive style of individual windows throughout the interpolation process. Moreover, we observe

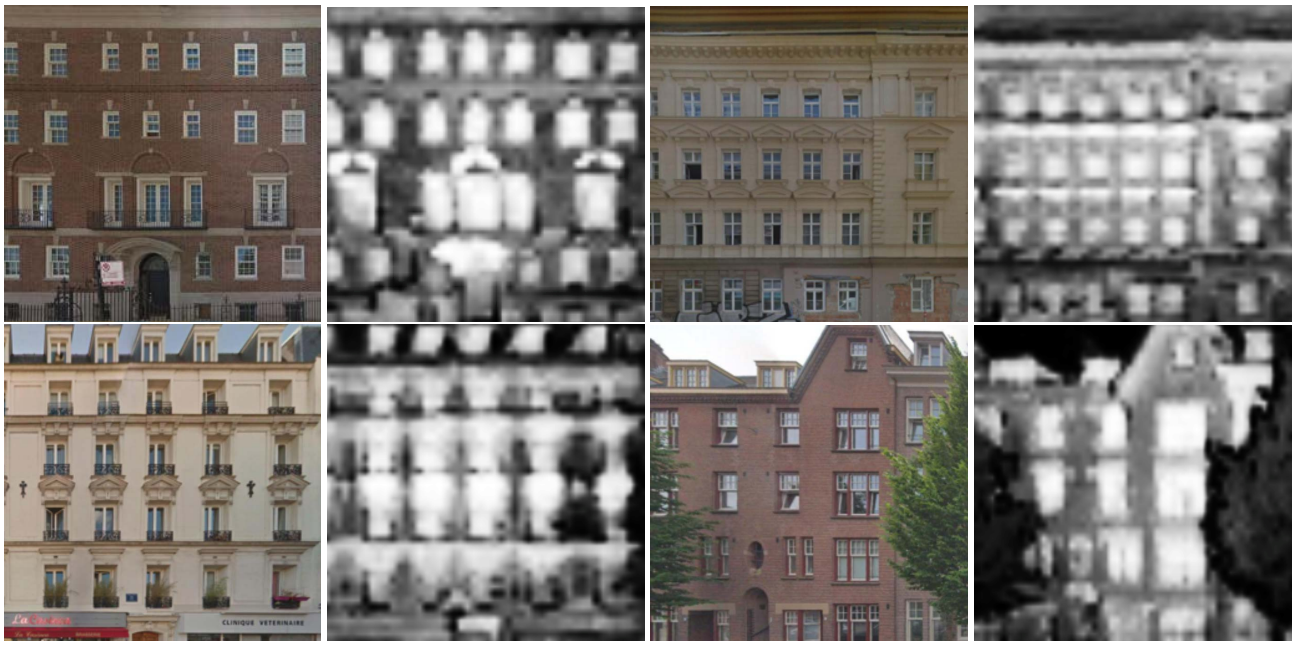

Figure 3. Example of edit masks that were extracted from mask extraction module.

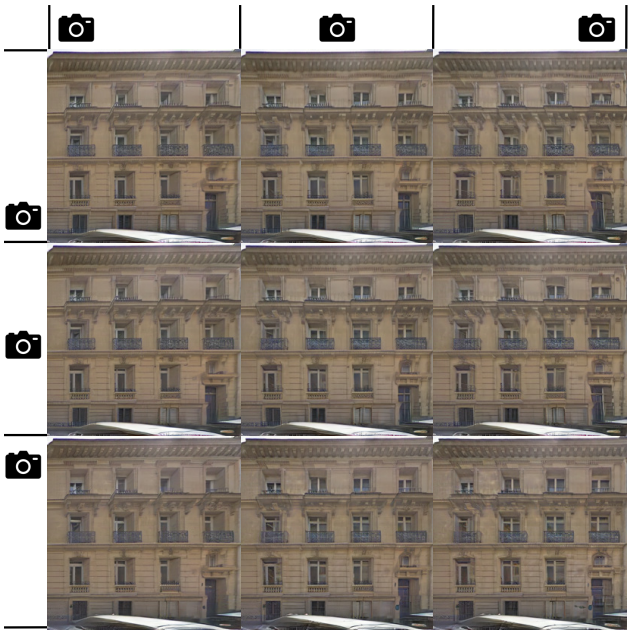

Figure 4. We present an interpolated representation of camera movement in both horizontal and vertical directions. The spatial code z remains fixed, while the output dynamically adjusts according to the camera's motion, which modifies the conditional information of the view direction target, θ_{target}. In the accompanying figure, we demonstrate the camera's movement from left to right (top row) along the horizontal axis and from top to bottom (left column) along the vertical axis. Furthermore, we showcase the disentangled controllability achieved on both axes.

the robustness of our model in challenging scenarios where the image contains noise or when facade details are partially occluded by trees. These results highlight the adaptability and reliability of our approach in real-world situations.

Appendix D: FacadeNet applications results

Problematic rectified facade improvement. Figure 6 showcases pairs of facade images for visual comparison. The top row displays the problematic rectified facade images f_{ref}, while the bottom row exhibits the 0-view improved generated facade images $f_{novel} = G(E(z_f, \theta_0))$. We observe that our generative approach successfully reconstructs an identical appearance to the reference facade images f_{ref} while maintaining structure and style. Additionally, the approach effectively translates the components of the facades to align with a 0-angle viewpoint. This transformation results in fewer problematic areas in the generated facade images.

Real-time textures for urban scenes. In Figure 5 we render examples of our approach from 2 different view angles. This example contains 4 buildings whose textures are changing simultaneously but differently based on their location in the 3D world and the position of the camera. We illustrate that our application can create multiple plausible results in real-time

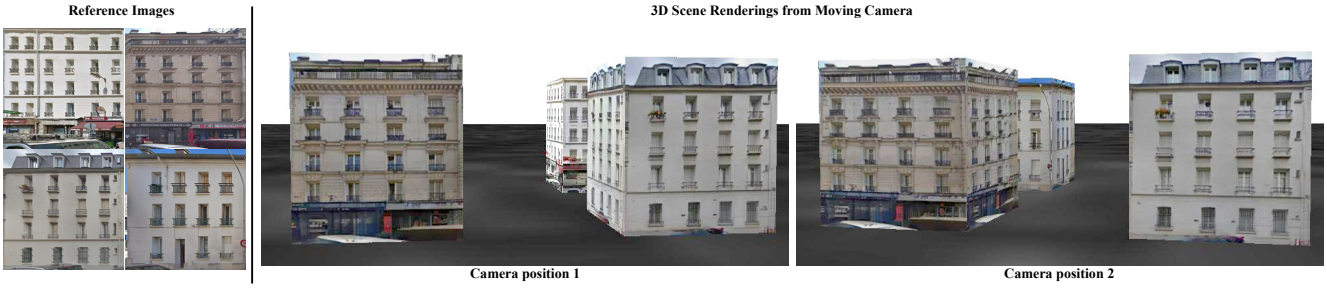

Figure 5. Urban Scenes Renders from the in Real-Time interactive textures application

Figure 6. We display examples of problematic facade image improvement. We display pairs of the reference images (left) and the θ_0(center image) reconstruction images(right). We observe that our model can rotate the facade to a better view orientation in contrast to the reference while at the same time, it achieves a high similarity of style and structure.

Appendix E: Additional visualizations

Edit mask examples. Figure 3 provides a visual representation of the edit mask obtained through our selective editing module. This module leverages information from DINO ViT to extract plausible edit masks by employing learnable weights that blend the input features into a 1-channel edit mask. The purpose of this edit mask is to guide the network in manipulating the reference image in a targeted manner, thereby enhancing the consistency of novel views across different view angle targets.

A notable observation in Figure 3 is the consistent pattern exhibited by the selective editing module. It assigns high values to semantic areas such as windows, doors, and balconies, indicating their significance in the editing process. Moderate values are assigned to various facade details, particularly those found on the ground floor. Areas with plain walls or minimal details receive low importance values, while the sky and trees receive the lowest importance values in the context of the novel view synthesis task.

The underlying rationale behind the use of masks is to group areas in the reference image that are crucial for our task and focus on modifying them while leaving the remaining areas intact. This approach allows us to selectively and effectively alter specific regions of the image to achieve our desired outcomes.

Qualitative results. Figures 9, 10, 11, 12 presents qualitative comparisons between *Palette* [3] (1^{st} row), *3DGP* [4] (2^{nd} row), *swapping-AE* [2] (3^{rd} row) and $FacadeNet_{full}$ (4^{th} row). *Palette* and *3DGP* are unable to generate fine details as the generation is combined with novel view synthe-

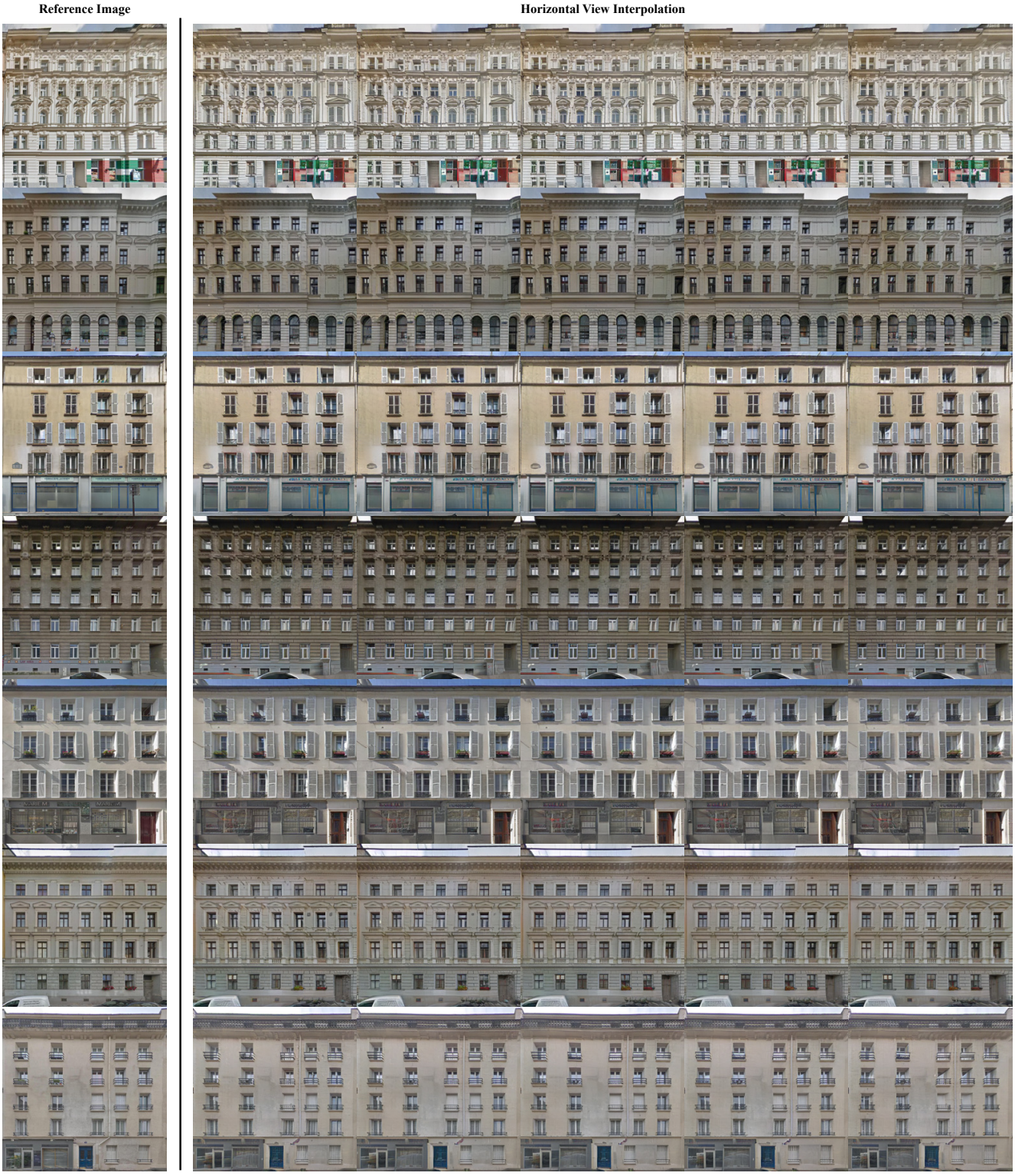

Figure 7. Examples of 5-step image interpolation on the horizontal axis. Given the reference images (left column), we can reconstruct the novel view from different angles as it is illustrated in the images of columns 2-6.

Figure 8. Examples of 5-step image interpolation on the horizontal axis. Given the reference images (left column), we can reconstruct the novel view from different angles as it is illustrated in the images of columns 2-6.

sis. Notably, artifacts become apparent in the output generated by the $swapping - AE$ model across varying viewing angles. In contrast, $FacadeNet_{full}$'s results demonstrate a higher level of robustness, effectively preserving the structural details. More results are displayed in the supplementary.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 2

[2] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *NeurIPS*, 2020. 5, 9, 10

[3] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021. 5, 9, 10

[4] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3D Generation on ImageNet. *arXiv preprint arXiv:2303.01416*, 2023. 5, 9, 10

[5] Peihao Zhu, Wamiq Reyaz Para, Anna Frühstück, John Femiani, and Peter Wonka. Large-scale architectural asset extraction from panoramic imagery. *IEEE Transactions on Visualization and Computer Graphics*, 28(2), 2022. 1

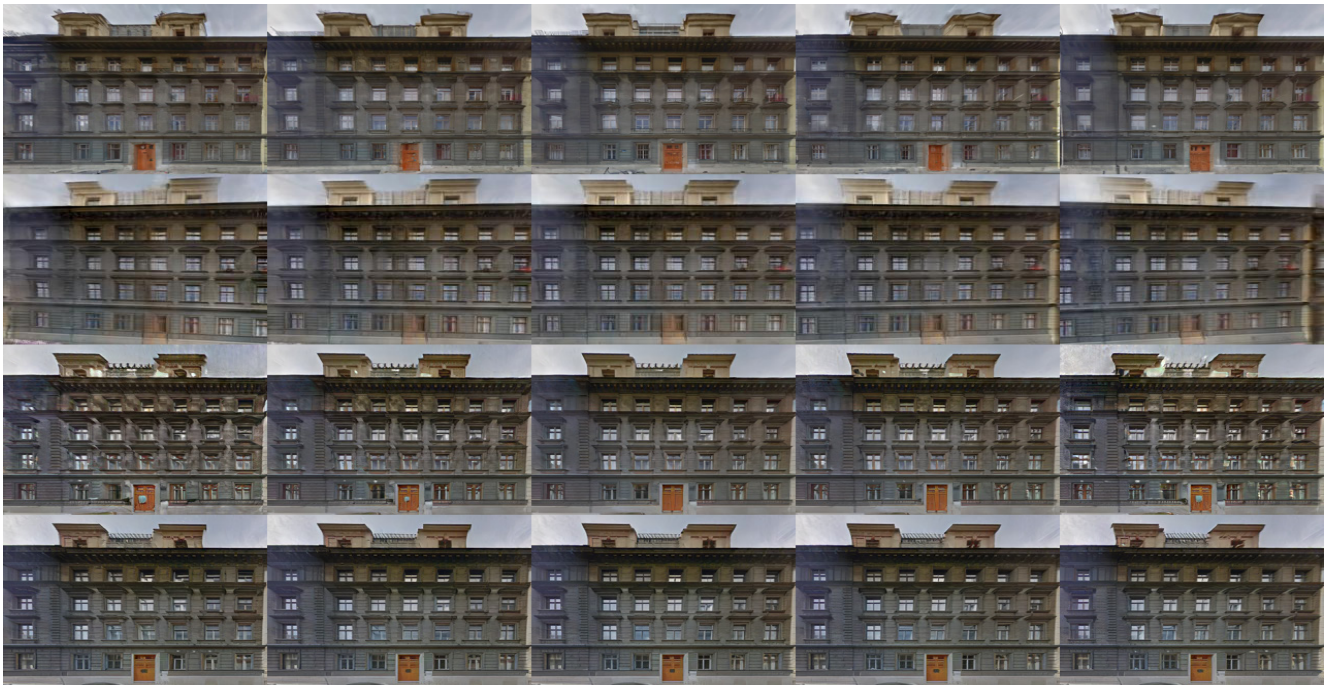

Figure 9. Novel view interpolation comparison between textitPalette [3] (1^{st} row), *3DGP* [4] (2^{nd} row), *swapping-AE* [2] (3^{rd} row) and $FacadeNet_{full}$ (4^{th} row)

Figure 10. Novel view interpolation comparison between textitPalette [3] (1^{st} row), *3DGP* [4] (2^{nd} row), *swapping-AE* [2] (3^{rd} row) and $FacadeNet_{full}$ (4^{th} row)

Figure 11. Novel view interpolation comparison between textitPalette [3] (1^{st} row), *3DGP* [4] (2^{nd} row), *swapping-AE* [2] (3^{rd} row) and $FacadeNet_{full}$ (4^{th} row)

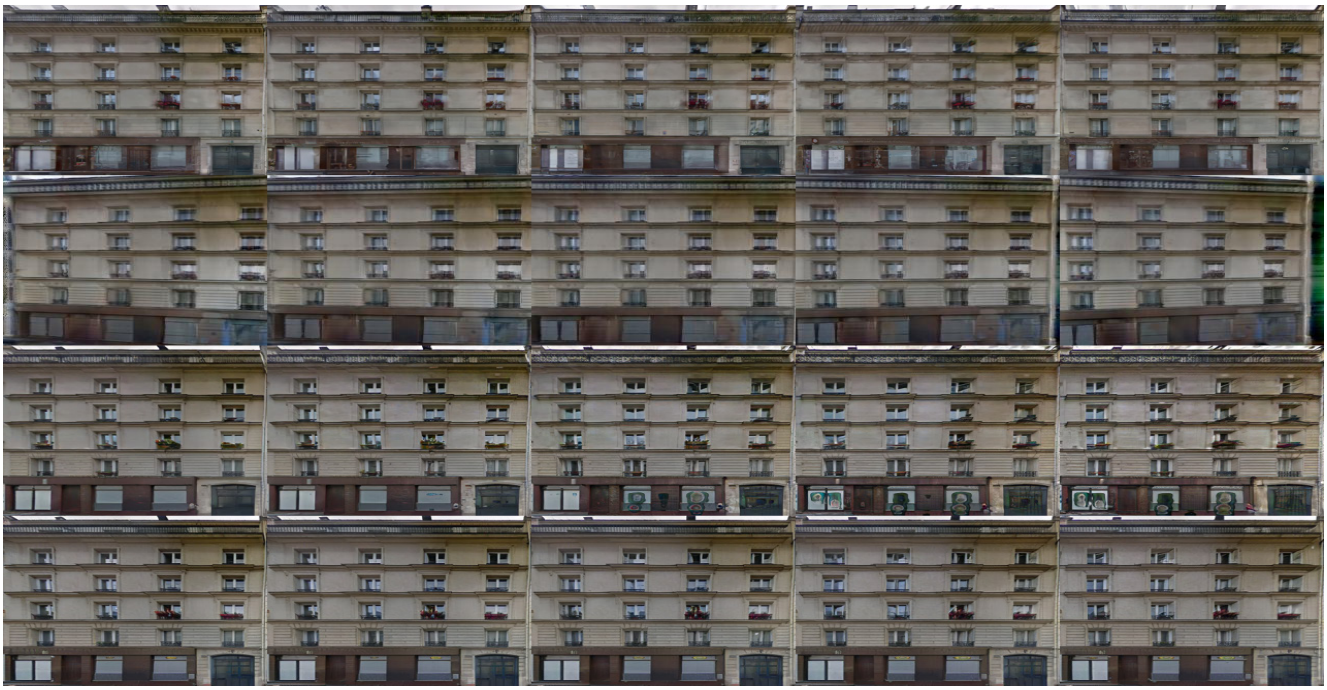

Figure 12. Novel view interpolation comparison between textitPalette [3] (1^{st} row), *3DGP* [4] (2^{nd} row), *swapping-AE* [2] (3^{rd} row) and $FacadeNet_{full}$ (4^{th} row)