FacadeNet: Conditional Facade Synthesis via Selective Editing

Yiangos Georgiou[1] Marios Loizou[1] Tom Kelly[2] Melinos Averkiou[1]
[1]Univesity of Cyprus/CYENS CoE, Cyprus {ygeorg01, mloizo11, maverk01}@ucy.ac.cy
[2]KAUST, Saudi Arabia thomas.kelly@kaust.edu.sa

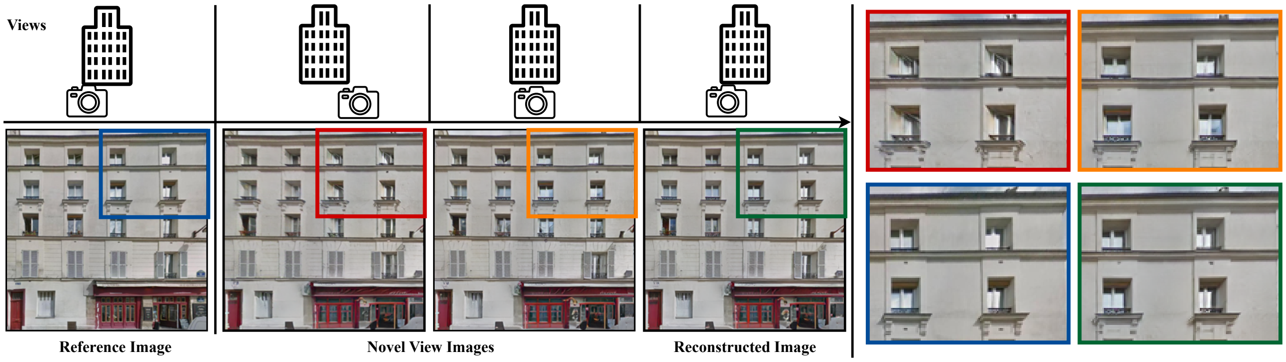

Figure 1. *Left:* Utilizing a reference facade image (bottom row, left column) and relative camera position information (top row), our method generates novel facades from varied viewpoints, all while preserving the reference image's structure and style (centered columns). Additionally, our approach faithfully reconstructs the reference facade (right column). *Right:* Zoomed facade regions highlight our method's capacity to modify critical facade elements, like windows, across diverse viewpoints (red and orange regions). Furthermore, our approach accurately reconstructs (green region) the reference facade (blue region).

Abstract

We introduce FacadeNet, a deep learning approach for synthesizing building facade images from diverse viewpoints. Our method employs a conditional GAN, taking a single view of a facade along with the desired viewpoint information and generates an image of the facade from the distinct viewpoint. To precisely modify view-dependent elements like windows and doors while preserving the structure of view-independent components such as walls, we introduce a selective editing module. This module leverages image embeddings extracted from a pretrained vision transformer. Our experiments demonstrated state-of-the-art performance on building facade generation, surpassing alternative methods.

1. Introduction

In urban planning, architectural design, and historical preservation, there is an increasing demand for rich visual representations of building facades, as it facilitates more comprehensive visual analyses, interactive 3D visualizations such as virtual tours, and digital archiving of structures [10, 37]. A conventional approach to capturing building facades involves taking photographs from different viewpoints. However, this approach is often constrained by practical limitations, such as the unavailability of multiple vantage points, especially in densely built urban environments. Moreover, obtaining a large dataset of images from varying views is time-consuming and expensive. In light of these challenges, synthesizing novel views of building facades from a single image emerges as a compelling alternative.

We tackle the problem of synthesizing images of a building's facade from novel viewpoints, given a single image of the facade taken from an arbitrary viewpoint. Synthesizing novel views of a facade given a single image entails estimating the appearance of the facade as seen from viewpoints different from the original image. This problem has been a subject of interest in computer graphics and computational photography due to its wide range of applications and inherent challenges associated with geometric and photometric consistency [41, 48]

Traditional methods have relied on 3D reconstruction techniques [15], which involve extensive manual intervention and are not easily scalable. Learning-based approaches, especially conditional Generative Adversarial Networks (cGANs), have recently been explored for view synthesis [12, 23]. However, state-of-the-art methods such as StyleGAN2 [24] and Swapping Autoencoder [34] that rely on style-content separation, often fail to decouple view information from structural properties of the facade.

In contrast, our approach, *FacadeNet*, addresses these challenges through a novel *selective editing module*, which enables finer control over the generation process. It guides the generation by computing a *selective editing mask*, allowing the alteration of view-dependent elements (e.g. windows) while keeping view-independent elements (e.g. walls) intact. FacadeNet takes as input a single image of a building facade together with the desired view information in the form of a view tensor, and uses a conditional GAN equipped with our novel selective editing module to synthesize an image of the facade from a different viewpoint.

Computing a selective editing mask could be simplified, if a semantic segmentation of the input facade image is available, as in SPADE [33] and SEAN [49]. Unfortunately such semantic segmentation masks are not available for the vast majority of facade images captured in the wild. Using a pretrained network to generate such semantic masks is possible, but these will always be imprecise, especially at boundaries. Inspired by recent advances in the explainability of large pretrained vision transformer models [3], we hypothesize that a *selective editing mask* could be computed by combining, in a learnable network module, features from deep layers of such models. Our novel *selective editing module* (Section 3) takes as input image features obtained by a pretrained DINO model [5], and learns an optimal weighting of them in order to compute a selective editing mask. This mask then drives the synthesis of a facade image where view dependent parts such as windows and doors are edited according to the provided view information, while view independent parts such as the walls remain fixed.

Through a series of experiments (Section 4), we demonstrate that our method outperforms quantitatively and qualitatively competing works, as well as strong baselines such as having access to ground truth semantic segmentation masks, on the large LSAA facade image dataset [50]. Further ablations motivate the design choices for our selective editing module as well as some technical implementation details.

In summary, the main contributions of this paper are:

1. Introducing a novel selective editing module within a conditional GAN, that enables real-time synthesis of novel facade views from a single arbitrary image.
2. Demonstrating through rigorous evaluation that FacadeNet outperforms state-of-the-art alternatives in single-image facade view synthesis.
3. Present a comprehensive ablation study, unveiling the importance of different components in the FacadeNet architecture.
4. Showcasing two applications of our novel approach in (i) eliminating rectification artifacts in facade images extracted from panoramic street-view images, and (ii) real-time texturing of simple 3D building models with dynamic camera-dependent facade views.

2. Related Work

Our work lies within the broader realm of view synthesis, with a particular focus on synthesizing novel views of building facades via conditional GANs. In this section, we briefly review key areas of related work, including facade image analysis, traditional view synthesis, learning-based view synthesis, and conditional GANs for image generation.

Facade Image Analysis. Building facades have been a subject of interest due to their role in urban planning, architectural design, and 3D modeling. Debevec et al. [10] used facade images for architectural scene modeling, while Remondino and El-Hakim [37] emphasized image-based 3D modeling for archival and historical preservation, focusing on facades. Datasets like eTrims [28], CMP Facade dataset [46], and Graz50 [36], have been developed to facilitate research on facade analysis and modeling, aiding tasks such as facade segmentation, object detection, and 3D reconstruction [45]. However, these datasets predate recent deep learning advancements and are relatively small, thus are unsuitable for our purposes. The recent LSAA dataset of Zhu et al [50] provides a large set of rectified facade images with various metadata including annotated view information and is thus used to train and evaluate our network.

Traditional View Synthesis View synthesis involves generating new scene images from viewpoints different from the available ones. Classic methods include view morphing by Seitz and Dyer [41], a technique that generates intermediate views of a scene by blending and warping two or more images, and Debevec et al. [10], who focused on creating photorealistic models of architectural scenes from images. Traditional approaches focused on geometric and photometric consistency to synthesize novel views. Hartley and Zisserman [15] provide an extensive exploration of the geometry involved in multiple view synthesis. McMillan and Bishop [30] proposed the image-based rendering technique, which involved blending of different views. While these methods were ground-breaking, they often require extensive manual effort and are not easily scalable. In contrast, our method is fully-automated, scalable, works in real-time, and does not rely on any hand-engineered features or correspondences to generate novel views of a facade.

Learning-based View Synthesis The advent of deep learning led to learning-based methods gaining popularity for view synthesis tasks. Hedman and Kopf [16] used a deep neural network to synthesize motion blur and refocus images. Flynn et al. [12] introduced DeepStereo, which predicts new views from large natural imagery datasets. Zhou et al. [48] advanced this domain with a multiplane image representation for stereo magnification. Neural Radiance Fields [4, 29, 31] have recently taken the view synthesis and reconstruction research areas by storm. However, they necessitate training on multiple images per facade, requiring training from scratch for each new facade, and often lack real-time speed, although recent methods offer vast improvements [17]. Diffusion models [11, 19, 32, 43] are a valid recent alternative to GANs, but they are slower and harder to control. Diffusion-based generative models [20, 44] have gained substantial attention for surpassing GANs in FID scores, particularly in unconditionally generated tasks like ImageNet [21] and super-resolution [40]. However, image editing poses a more intricate challenge for diffusion models. Recent advancements in both conditional [39] and unconditional [7] diffusion models have tackled this, yielding high-quality results. Here, we undertake a comparative analysis between the performance of Palette [39] and FacadeNet. In contrast, our method offers real-time performance, generalizes to unseen facades, and only requires one facade image as input.

Conditional Generative Adversarial Networks GANs have been used for various image generation tasks, and conditional GANs, in particular, have been successful in image-to-image translation tasks. The Pix2Pix network by Isola et al. [23] is a well-known example of using conditional GANs for image translation. StyleGAN [25] introduced a novel approach that utilizes a learned constant feature map and a generated latent code z to control the output image features. StyleGAN2 [24] further enhances this concept with AdaIN [22] layer, which adjusts image channels to unit variance and zero mean, retaining channel statistics. It then incorporates style through scaling and shifting, guided by conditional information. Karras et al [26] introduced an alternative approach that retains scale-specific control, eliminating undesired artifacts while preserving result quality. Image2StyleGAN [1, 2] aimed to overcome a disadvantage of StyleGAN-based approaches to manipulate reference images. However, these approaches are slow due to a preprocessing step that iteratively predicts the latent code to reconstruct the reference image.

Another line of work to disentangle an image's latent space was introduced in the Swapping Autoencoder paper [34] where two discriminators are used to disentangle style and structure from a reference input image. IcGAN [35] on the other hand used two inverse encoders to allow changing of the conditional and latent vectors separately. StarGAN [8] is one of the most common multi-domain image-to-image translation method which tries to improve the scalability of GAN models by using a single generator to generate all available domains. To extend StarGAN [8] to a multi-modal approach, StarGANv2 [9] replaced the domain label with a domain specific style latent code that can represent diverse styles for each of the available domains. The SPADE network [33] introduced a novel conditional batch normalization specifically modified for images. An improved approach that enables multi-modal controllable style change for each different input was described in the SEAN network [49]. Our method is most closely related to StyleGAN2 [24] and Swapping Autoencoder [34], that rely on style-content decoupling. Encoding facade view information as *style* is not enough however, since conditional GANs often fail to decouple view information from structural properties of the facade. As a result, they tend to hallucinate elements (e.g. windows or balconies), or add noise and other artifacts. FacadeNet tackles the shortcomings of these methods through a novel *selective editing module*, which computes a *selective editing mask* based on pretrained DINO [5] features to guide the generation, allowing it to focus on altering only the view-dependent elements (e.g., doors, balconies, windows).

Currently, 3D-aware synthesis techniques [6, 14, 47] exhibit training efficiency and sampling capabilities comparable to 2D Generative Adversarial Networks (GANs). However, their effectiveness relies on meticulously curated datasets with aligned structures and scales, such as those for human or cat faces. A recent study [42], circumvented previous challenges, notably dependence on known camera poses. This workaround trains models to learn pose distributions from single-view data, rather than relying on multi-view observations. Similarly, our proposed approach eliminates the need for known camera poses and can be trained on sparse single-view image data.

3. Method

Our approach focuses on synthesizing building facades with varying viewing angles, based on a reference facade image $\mathbf{f}_{ref} \sim \mathbf{F} \subset \mathbb{R}^{HxWx3}$ and horizontal and vertical angle vectors $\theta_h \sim \Theta_H \in [-1, +1]^W$ and $\theta_v \sim \Theta_V \in [-1, +1]^H$. By conditioning the generation process on the new view direction controlled by the input angle vectors, our method, called *FacadeNet*, produces facades that exhibit realistic modifications in semantic components like windows and doors. To ensure authenticity, we incorporate a discriminator D that enforces plausible changes in these components. Preserving the overall facade structure is crucial. Thus, our network is designed to faithfully reconstruct areas such as walls that remain unchanged regardless of the viewing direction. To achieve this, we utilize the feature embeddings of a self-supervised vision transformer [5] to construct a semantic-aware mask, which guides the reconstruction process. Moreover, we employ a reconstruction loss to enhance the structural accuracy of the generated facades. In the following sections, we first provide an overview of our proposed architecture in Section 3.1. We then delve into how we enforce the generation of structurally aware and novel building facades in Section 3.2 and Section 3.3, respectively. Finally, in Section 3.4, we present the implementation details of our approach.

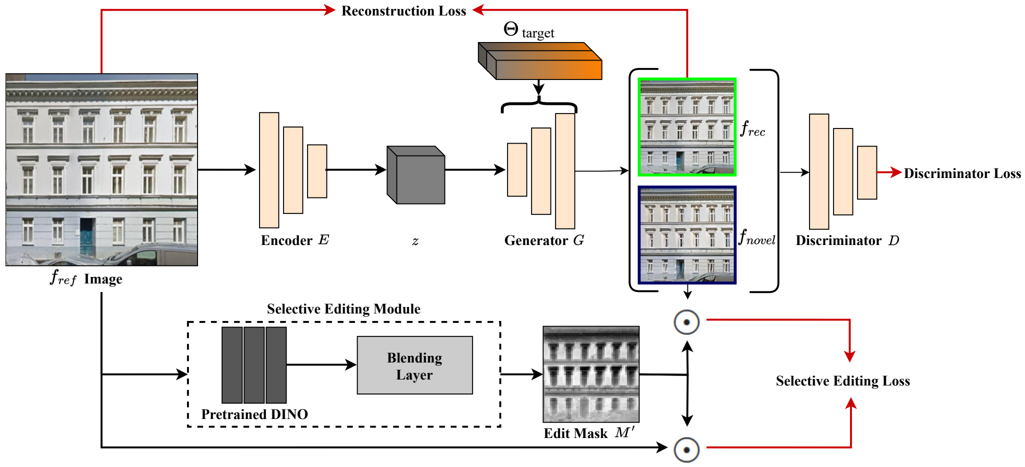

Figure 2. Our training procedure creates latent tensors $z = E(f_{ref})$ which capture the reference facade's style and structure. These z tensors form the basis for diverse view generation from a single reference image f_{ref} with conditional information θ. The generator utilizes the latent code z and the target view direction θ to produce new samples $G(z, \theta_{target})$. Our approach accurately reconstructs f_{ref} by aligning θ_{target} with the same viewpoint of f_{ref}, while varying the θ_{target} it generates novel views f_{novel}. To ensure high-quality and consistent facade reconstruction from various viewpoints, our model employs multiple loss functions. These encompass L_1 reconstruction and conditional discriminator losses, enhancing fidelity and structural awareness. Furthermore, we introduce a selective editing module that employs prior information (DINO ViT features) to extract a selective editing mask. This mask designates editable components during novel view synthesis, contributing to the precision of the process.

3.1. FacadeNet architecture

To generate building facades with specified viewing directions based on a reference image $\mathbf{f}_{ref}$ and guided by the target viewing vector $\theta_{target} = [\theta_h, \theta_v]$, we employ a *task-specific conditional* GAN architecture [23]. As illustrated in Figure 2, FacadeNet comprises an autoencoder network. First, the encoder E, inspired by [34], takes the reference building facade as input and produces a latent tensor $\mathbf{z}$ with spatial dimensions, encoding both the structural and texture information present in the image $\mathbf{f}_{ref}$:

$$\mathbf{z} = E(\mathbf{f}_{ref}) \tag{1}$$

Next, the conditional generator G, following [26], utilizes the latent tensor $\mathbf{z}$ to synthesize a facade that aligns with the desired viewing direction, using the target viewing vector θ_{target}:

$$\mathbf{f}_{novel} = G\big(E(\mathbf{f}_{ref}), [\theta_h, \theta_v]\big) \tag{2}$$

The generated facade $\mathbf{f}_{novel}$ is crafted by combining the encoded reference image embedding $\mathbf{z}$ and the target viewing information. To ensure realistic results, the discriminator D assesses the novel facade. The target viewing vector is also incorporated into the discriminator to enforce that the generated facade aligns with the desired viewing direction, by adopting the conditional discriminator idea of [23].

3.2. Structural awareness

In our specific task, we aim to modify the high-frequency areas within the facade image, as they have a greater impact on influencing the viewing direction of the building being represented. At the same time, we want to preserve the low-frequency structural components typically found in building facades, such as flat surfaces like walls. These components remain visually consistent regardless of the observer's viewing position.

Accurate reconstruction. In line with the principles of the classic autoencoder [18], we aim to acquire a mapping between the latent code $\mathbf{z} \sim \mathbf{Z}$ and the image $\mathbf{f}_{ref} \sim \mathbf{F} \subset \mathbb{R}^{HxWx3}$. To this end, we employ an image reconstruction loss that compares the input facade $\mathbf{f}_{ref}$ with the reconstructed facade $\mathbf{f}_{rec}$ conditioned on the *ground truth* target viewing vector θ^{gt}_{target}, which corresponds to the direction of the camera that originally captured the input image:

$$\mathcal{L}_{rec}(E, G; \theta^{gt}_t) = \mathbb{E}_{\mathbf{f}_{ref} \sim \mathbf{F}}\left[|\mathbf{f}_{ref} - G\big(E(\mathbf{f}), [\theta^{gt}_h, \theta^{gt}_v]\big)|\right] \tag{3}$$

During the optimization process of the autoencoder, this loss facilitates the acquisition of accurate and informative latent representations for building facades.

Selective Editing Module. Our primary goal is to selectively modify areas within the input facade that have a significant impact on the building's viewing direction while disregarding regions that maintain visual consistency relative to the capturing camera's orientation. To accomplish this, we introduce the *Selective Editing Module* that constructs the *selective editing mask*, effectively isolating these areas within the facade image. Drawing inspiration from the findings of [3], we initially pass the input facade through a pre-trained self-supervised vision transformer [5] and extract the key representations of the last attention layer. To retain high-frequency information from the underlying data distribution, we employ principal component analysis

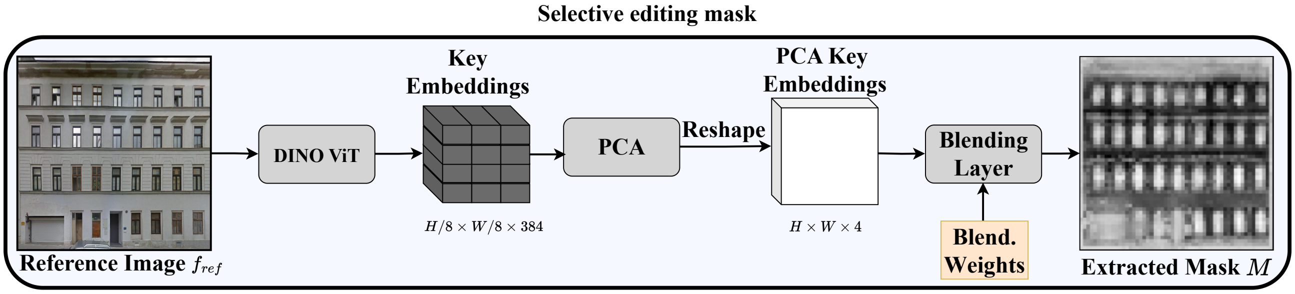

Figure 3. This figure depicts the step-by-step procedure for calculating selective editing masks for a given image f_{ref}. We commence by employing a pre-trained DINO ViT model to extract localized key embeddings that encapsulate crucial visual features within the image. Once DINO key embeddings are obtained, we apply PCA (Principal Component Analysis) to retain high-frequency information from these embeddings. Lastly, our adaptive blending layer utilizes the PCA embeddings, enabling optimal feature combination extraction. This process yields a single-channel selective mask denoted as $\boldsymbol{M}$.

(PCA) on these embeddings By selecting the top four principal components $\mathbf{V}_{pc} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4]$, we capture the most salient features. To create the selective editing mask, we blend these selected principal components using a linear combination approach. More specifically, we employ a set of learnable blending weights $\mathbf{W} \in \mathbb{R}^4$ and apply the sigmoid function $\sigma(\cdot)$ to obtain the final mask:

$$\mathbf{M} = \mathbf{V}_{pc} \cdot \sigma(\mathbf{W}) \tag{4}$$

This formulation enables the network to learn the optimal linear combination of the principal components, effectively highlighting the most prominent elements within the facade, such as windows and doors (see Figure 3). The extracted selective editing mask is semantic-aware, since it captures semantic areas in the facade image that are view-dependent.

In addition to the reconstruction loss (Eq 3), which penalizes modifications across the entire reconstructed image, we introduce the *Selective Editing Loss* that capitalizes on the selective editing mask $\mathbf{M}$. This loss compares the masked input facade $\mathbf{f}_{ref}$ with the masked novel facade $\mathbf{f}_{novel}$ whose viewing direction is altered according to the conditional target viewing direction θ_{target}^{novel}, rather than the ground truth direction of the reference facade. While the extracted mask $\mathbf{M}$ highlights editable areas within the facade, the primary objective of this loss is to retain the appearance of view-independent components. To achieve this, we utilize the complement of the mask, denoted as $\mathbf{M}' = 1 - \mathbf{M}$. By employing $\mathbf{M}'$, we effectively ignore high-frequency areas, enabling the network to preserve the viewing direction of the view-independent structural elements during the synthesis of novel facades with varying viewing directions:

$$l_{edit}(E, G; \theta_t^n) = ||\mathbf{f}_{ref} \odot \mathbf{M}' - G\big(E(\mathbf{f}_{ref}, \theta_t^n)\big) \odot \mathbf{M}'|| \tag{5}$$

Furthermore, as part of our training process, we synthesize n novel facades for each input image and apply the selective editing loss. This step is crucial in ensuring that the appearance of view-independent components remains consistent across various novel facades with different viewing directions:

$$\mathcal{L}_{edit}(E, G; [\theta_t^{n_1}, \theta_t^{n_2}, \cdots, \theta_t^{n_k}]) = \frac{1}{k}\sum_{i=1}^{k} l_{edit}^{(i)} \tag{6}$$

3.3. Novel reconstruction

When it comes to image editing, a crucial aspect is to modify the latent representation of an input image in a way that ensures the novel reconstruction appears both authentic and aligns with the provided conditional information. Simultaneously, this representation should be able to faithfully and easily reconstruct the input image. To address these requirements, we introduce two key loss components: the *View-dependent loss* and the *View-consistent loss*. These losses play a vital role in guiding the network to generate realistic novel facades while preserving the fidelity of the reference facade and its original viewing direction, enforced by the discriminator D.

View-dependent loss. This loss facilitates the network's ability to synthesize novel facades that appear authentic and visually consistent with the specified viewing direction. By incorporating the conditional information, the network learns to modify the relevant components of the facade, ensuring the alterations align with the intended changes in the viewing perspective. Following the synthesis process during of the selective editing loss, we utilize a conditional adversarial loss that guides the generation of multiple novel facades based on the conditional target viewing directions $\theta_{target}^{novel_i}$:

$$\mathcal{L}_{GANd}^{dep}(E, G, D; \theta_t^{n_i}) = \mathbb{E}_{\mathbf{f}_{ref} \sim \mathbf{F}, \theta \sim \Theta}[-log(D(\mathbf{f}_n^i, \theta))] \tag{7}$$

View-consistent loss. In parallel, the View-consistent loss serves to maintain the faithfulness of the reconstruction process for the reference facade. By minimizing the impact of viewing direction changes on the entire image, we ensure the reconstruction remains as close as possible to the original facade, thereby preserving its original viewing direction. The non-saturating adversarial loss [13] for the generator G and the encoder E is computed as:

$$\mathcal{L}_{GAN}^{cons}(E, G, D; \theta_t^{gt}) = \mathbb{E}_{\mathbf{f}_{ref} \sim \mathbf{F}}[-log(D(\mathbf{f}_{rec}, \theta_t^{gt}))] \tag{8}$$

3.4. Implementation Details

In this section we provide the implementation details regarding our training process. For the construction process of the target viewing vectors please see our supplementary material.

To optimize our network, we combine the structural awareness and novel reconstruction losses using linear weights::

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{edit} + \lambda_3\mathcal{L}_{GAN}^{dep} + \lambda_4\mathcal{L}_{GAN}^{cons} \quad (9)$$

We empirically set the weights as $\lambda_1 = \lambda_2 = 3$ and $\lambda_3 = \lambda_4 = 0.5$, effectively balancing the importance of each loss component. For training, we employ the Adam optimizer with a learning rate of 0.001. Our network is trained on four NVIDIA V100 GPUs using a batch size of 32 images at a resolution of 256×256. The training process consists of 15 million iterations, taking approximately 4 days to complete. Remarkably, our trained model achieves an impressive average generation speed of 62 frames per second (fps) for a batch size of 32 images. We also refer readers to our project page with source code for more details. [1]

4. Evaluation

This section presents a comprehensive evaluation of FacadeNet's performance, combining qualitative and quantitative analysis. We begin by discussing the dataset used for our experiments in Section 4.1, providing important context for the subsequent evaluations. In Section 4.2, we conduct a comparative analysis of our approach against state-of-the-art GAN-based and diffusion-based models. To assess performance, we employ various metrics, illustrating the superiority of FacadeNet. To conclude our evaluation, Section 4.3 presents two novel applications specifically tailored to urban environments. These applications serve as powerful demonstrations of the versatility and potential impact of FacadeNet. Through these innovative use cases, we highlight the practical value and wider implications of our research. Please see our supplementary material for an in-depth ablation study of our design choices, where we emphasize the effectiveness of our selective editing module. Moreover, we provide examples of facade interpolation under varying viewing directions.

4.1. Dataset

We make use of a rectified facade dataset sourced from [50] to obtain genuine rectified facade images. This dataset comprises approximately $23,000$ images. The dataset provides various attributes from four distinct categories: *Metadata* (including geographic properties like longitude, latitude, city, and building information), *homography* (encompassing view angle and homography error), *semantic attributes* (such as windows, balcony, and door area), and *semantic embedding* (capturing similarities between samples). These attributes offer valuable insights into various scenarios.

In our particular case, we leverage the homography attributes, particularly the view angle and the size of the cropped facade, to generate pairs of facade images and view direction targets (see supplementary). These targets represent the viewpoint for each given facade image, thereby enabling accurate processing of images regarding the angle information.

4.2. Comparisons with other methods

In this section, we perform a comparison between our approach and the state-of-the-art approaches related to our work. We use styleGAN2-ADA [24], Palette [39], 3DGP [42] and Swapping-autoencoder [34] to extract metrics that measure the quality and consistency of facade reconstruction and novel view synthesis. To assess the performance of facade reconstruction, we utilize several metrics including $PSNR$, $SSIM$, and FID_{rec}. These metrics allow us to evaluate both the image quality and the structural similarity between the reference and generated reconstructed images. Additionally, for evaluating the quality of novel view synthesis, we rely on the FID_{novel} score. In terms of consistency and perceptual similarity, we employ $LPIPS_{vgg}$ and $LPIPS_{alex}$ metrics [27], which aim to measure the smoothness of viewpoints interpolation. A low $LPIPS$ score indicates that the image patches are perceptually similar which implies fewer alternations in structure and style between viewpoints.

To measure StyleGAN2 [24] and 3DGP [42] reconstruction quality, we project the reference image to the latent space in an iterative manner. We set the maximum number of optimization iterations to 1000 and 4000 respectively. We further modified the Swapping-Encoder [34] model to disentangle structure and novel direction, thereby enabling novel view manipulation and synthesis, and trained Palette [39] given as conditional input the angle-view maps and the reference image to follow our task's approach. Further, we introduce two versions of our network: $FacadeNet_{base}$ serves as the baseline, which relies solely on L1 and adversarial losses to optimize its performance. On the other hand, $FacadeNet_{full}$ represents the enhanced version of $FacadeNet$, incorporating all the design choices discussed in section 3, including the utilization of the selective editing module.

To assess the capabilities of StyleGAN2 [24] and 3DGP reconstruction [42], we employ an iterative approach to project the reference image into the latent space. We set the maximum optimization iterations to 1000 for StyleGAN2 and 4000 for $3DGP$. Furthermore, we have customized the

[1] *ygeorg01.github.io/FacadeNet includes our code and trained models.*

Method	LPIPS-alex↓	LPIPS-vgg↓	FID$_{rec}$ ↓	FID$_{novel}$ ↓	PSNR↑	SSIM↑
StyleGAN2-ADA [24]	–	–	22.52	–	20.591	0.467
Palette [38]	0.259	0.401	23.367	22.829	17.881	0.332
3DGP [42]	0.186	0.347	35.918	33.005	14.673	0.201
Swapping-AE [34]	0.198	0.386	10.55	15.64	22.24	0.668
$FacadeNet_{base}$	0.174	0.296	10.59	9.91	**24.13**	0.69
$FacadeNet_{full}$	**0.119**	**0.240**	**9.601**	**8.327**	23.866	**0.714**

Table 1. This table presents a comprehensive comparison between our model baseline $FacadeNet_{base}$, StyleGAN2 [26], Palette [38], 3DGP [42], swapping-autoencoder [34] and $FacadeNet$. The results clearly demonstrate the superiority of our task-specific model across various evaluation criteria, including reconstruction quality, novel view synthesis quality, and consistency. To assess the reconstruction image quality, we employ FID_{rec}, $PSNR$, and $SSIM$ metrics. Regarding novel view image quality we rely on FID_{novel}, while we measure the inter-view consistency with $LPIPS - \{alex, vgg\}$ metrics. Our final model $FacadeNet$ outperforms previous approaches by a significant margin.

Swapping-Encoder model [34] to disentangle structure and novel direction, thereby enabling novel view manipulation and synthesis. Additionally, we have trained Palette [39] using angle-view maps and the reference image as conditional inputs, aligning with our task's approach. Moreover, we introduce two versions of our network $FacadeNet_{base}$ serves as the baseline, relying solely on L1 and adversarial losses to optimize its performance. Conversely, $FacadeNet_{full}$ represents the enhanced iteration of FacadeNet, incorporating all the design choices discussed in Section 3, including the utilization of the selective editing module.

Table 1 showcases a comprehensive comparison of various models. Among them, $FacadeNet_{base}$ and $FacadeNet_{full}$ achieves higher reconstruction quality, outperforming alternative approaches with $PSNR$ scores of 24.13 and 23.86, and $SSIM$ values of 0.69 and 0.714, respectively. Notably, the performance of $Swapping - AE$ [34] and $Palette$ [39] surpasses that of $3DGP$ [42].

In terms of FID scores, our $FacadeNet$ approach generates more realistic samples than other models. Particularly in novel view facade synthesis, the FID score difference increases noticeably. Our design choices effectively elevate facade synthesis, resulting in robust artifact-free novel views. $FacadeNet_{full}$ outperforms swapping-AE [34] by 1.04 FID score points for FID_{rec}, and this gap widens to 7.32 for novel view synthesis FID_{novel}.

A central advancement of $FacadeNet$ centers around the coherence among diverse viewpoints. To quantify this aspect, we employ the $LPIPS$ metric. Notably, compared to other methods, $3DGP$ [42] excels in consistency, by achieving the highest scores. Building upon this, $FacadeNet_{full}$ further surpasses $3DGP$ [42] by improving $LPIPS_{alex}$ and $LPIPS_{vgg}$ scores by 0.05 and 0.88 respectively.

4.3. FacadeNet Applications

Problematic Rectified Facade Improvement Our main assumption to improve problematic facade images, is that 0-view angle difference facades are closer to ortho-rectified images that contain minimum distortion or other artifacts. We define as problematic the group of facades that their mean view-angle value is higher than 60°. Those facades have extreme orientations either to the left or right and contain large areas of missing information on doors and windows or other assets.

Moreover, our methodology manages to generate hidden information that is not visible in the reference facades(top images) due to the originally captured view angle. Our generator G manages to generate new information for the unseen parts of the input facade images with identical style and structure which is essential for the consistency between input and output facade images. Examples are provided in the supplementary.

Real-Time Textures for Urban Scenes Following the facade view interpolation experiments we developed an application where we represent a 3D city environment by using interactive textures created by our generative model. More specifically we create a large scene that contains simple cubes $b \sim B$ that represent the buildings and we use reference facades $f \sim F$ as textures. As we navigate around the city the view direction targets between the camera and the points are computed. The computed view angle maps are used as conditional view direction targets $\theta_{target} = \{\theta_h, \theta_v\}$ to alter the orientation of each texture map accordingly. The equations to compute the view direction target in 3D environment are as follows:

$$d = \quad p - c \tag{10}$$

$$\theta_h^f(d, n) = \quad (d \odot [1, 0, 1]) \cdot (n \odot [1, 0, 1]) \tag{11}$$

$$\theta_v^f(d, n) = \quad (d \odot [0, 1, 1]) \cdot (n \odot [0, 1, 1]) \tag{12}$$

$d \in \mathbb{R}^3$ stands for the viewing direction vector that starts from the camera positions c and points on a facade point p, the ray direction from the camera to a specific point is computed with the following equation $d = c - p$. The target maps are computed as the dot product of the viewing direction vector d and the facade's surface normal vector n. θ_h^f and θ_v^f denote the target vectors for the horizontal and

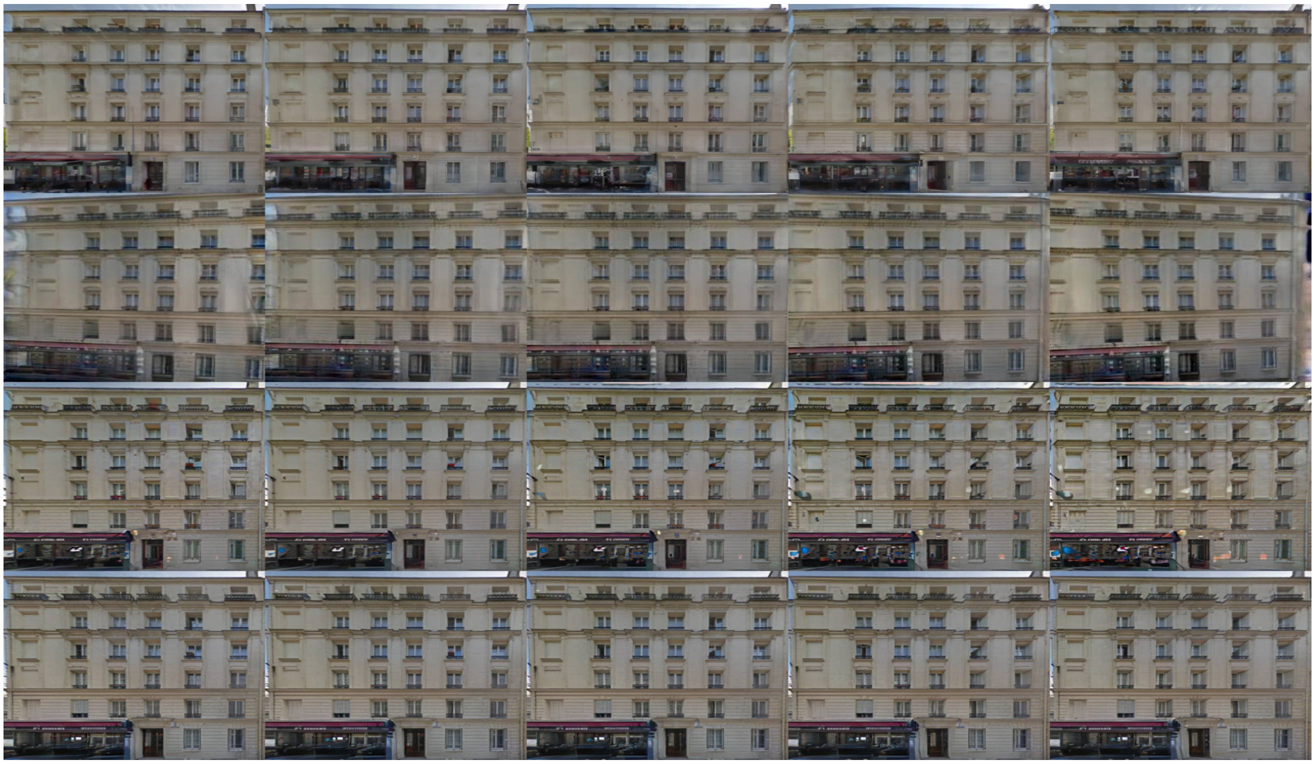

Figure 4. This figure presents qualitative comparisons between *Palette* [39] (1^{st} row), *3DGP* [42] (2^{nd} row), *swapping-AE* [34] (3^{rd} row) and $FacadeNet_{full}$ (4^{th} row). *Palette* and *3DGP* are unable to generate fine details as the generation is combined with novel view synthesis. 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.

vertical axis respectively. θ_h^f and θ_v^f target maps differ on the angle that is computed on each occasion, for θ_h^f targets we consider x and z axis to compute the angle difference on the horizontal axis, while, for vertical maps θ_v^f we consider y and z axis to output the difference on the vertical axis. To cancel out a specific axis, the Hadamard product $\odot$ is used to isolate either the angle difference on the $x - axis$ for horizontal target vector θ_h^f or the difference on $y - axis$ for vertical target vector θ_v^f.

This enables a real-time manipulation of textures for an urban 3D scene, more precisely, for each building b we randomly assign a reference image f which serves as texture for this building(cube). In each rendering iteration, our application updates the textures for each building b according to camera location c. Given, facade point p their normal n, and the camera position c we create the horizontal and vertical view targets θ_h and θ_v. Then, our model generates the new texture accordingly $t = G(E(f), \theta_{target}^f)$. For more examples check the video in the supplementary materials.

5. Conclusion and Future Work

In this paper, we presented FacadeNet, a novel conditional GAN that synthesizes building facade images from different viewpoints given a single input image. By introducting the selective editing module, FacadeNet effectively focuses on view-dependent facade features, leading to high-quality synthesized images with fewer artifacts compared to existing methods. Our experimental evaluations demonstrate state-of-the-art performance on standard metrics and appealing qualitative results.

Future Work FacadeNet offers several avenues for future work. First, exploring methods to handle highly complex facades with intricate geometric patterns could enhance the model's and applicability to a wider range of architectural styles. Moreover, integrating additional context, such as surrounding buildings and natural elements, could improve the visual coherence of the synthesized images in urban environments. FacadeNet could also be extended to support the synthesis of interiors, which would be beneficial for virtual reality applications. Finally, incorporating temporal information for dynamic scene elements (e.g., varying lighting conditions) could make FacadeNet applicable to time-varying view synthesis.

Acknowledgements This project has received funding from the EU H2020 Research and Innovation Programme and the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy (GA 739578).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 3

[3] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 2, 4

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 4

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 3

[7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 3

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 3

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 3

[10] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 1996. 1, 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[12] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 2016. 1, 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 5

[14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[15] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 1, 2

[16] Peter Hedman and Johannes Kopf. Deep blurring: Learning to refocus and synthesizing motion blur using a deep neural network. *arXiv preprint arXiv:1806.05666*, 2018. 2

[17] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *ICCV*, 2021. 3

[18] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313, 2006. 4

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1), 2022. 3

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 3, 4

[24] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 1, 3, 6, 7

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3, 4, 7

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 6

[28] F. Korc, H. Riemenschneider, and L. Van Gool. etrims image database for interpreting images of man-made scenes. In *Technical report*, 2009. 2

[29] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 2

[30] Leonard McMillan and Gary Bishop. Plenoptic modeling: An image-based rendering system. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995. 2

[31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 3

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2, 3

[34] 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. 1, 3, 4, 6, 7, 8

[35] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible Conditional GANs for image editing. In *NIPS Workshop on Adversarial Training*, 2016. 3

[36] G. Reitmayr and T. Drummond. Building facade interpretation from image sequences. In *ECCV*, 2008. 2

[37] Fabio Remondino and Sabry El-Hakim. Image-based 3d modelling: A review. *The Photogrammetric Record*, 21(115), 2006. 1, 2

[38] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 7

[39] 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. 3, 6, 7, 8

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 3

[41] Steven M. Seitz and Charles R. Dyer. View morphing. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 1996. 1, 2

[42] 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. 3, 6, 7, 8

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[44] Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models. In *NeurIPS*, 2020. 3

[45] Olivier Teboul, Loic Simon, Panagiotis Koutsourakis, and Luc Van Gool. Segmentation of building facades using procedural shape priors. In *CVPR*, 2010. 2

[46] R. Tylecek and R. Sara. A new image representation for recognizing imaged objects. In *13th Computer Vision Winter Workshop*, 2008. 2

[47] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3

[48] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1, 2

[49] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020. 2, 3

[50] 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. 2, 6