Label-free Neural Semantic Image Synthesis

Jiayi Wang[1], Kevin Alexander Laube[1], Yumeng Li[1,2], Jan Hendrik Metzen[1], Shin-I Cheng[1], Julio Borges[1], and Anna Khoreva[1]

[1] Bosch Center for Artificial Intelligence
{jiayi.wang2, kevinalexander.laube, yumeng.li, janhendrik.metzen, shin-i.cheng, julio.borges, anna.khoreva}@de.bosch.com
[2] University of Mannheim

Abstract. Recent work has shown great progress in integrating spatial conditioning to control large, pre-trained text-to-image diffusion models. Despite these advances, existing methods describe the spatial image content using hand-crafted conditioning inputs, which are either semantically ambiguous (e.g., edges) or require expensive manual annotations (e.g., semantic segmentation). To address these limitations, we propose a new label-free way of conditioning diffusion models to enable fine-grained spatial control. We introduce the concept of *neural semantic image synthesis*, which uses neural layouts extracted from pre-trained foundation models as conditioning. These layouts provide rich descriptions of the desired image, containing both semantics and detailed geometry of the scene. We experimentally show that images synthesized via neural semantic image synthesis achieve similar or superior pixel-level alignment of semantic classes compared to those created using expensive semantic label maps. At the same time, they capture better semantics, instance separation, and object orientation than other label-free conditioning options, such as edges or depth. Moreover, we show that images generated by neural layout conditioning can effectively augment real data for training various perception tasks.

Keywords: Diffusion Models · Neural Representation · Controllable Synthesis · Text-to-Image

1 Introduction

Controllable image synthesis enables users to specify the desired image content, while relying on a generative model to fill in details that align with the distribution of natural images. One way to express the content is through natural language, as popularized by the recent advances of large-scale text-to-image (T2I) diffusion models [1,3,27,29]. However, precisely and comprehensively specifying a complex scene (such as in Fig. 1) including layout, pose, shape, and orientation of objects can be a tedious task. To obtain a satisfactory image, a user often has to spend an extensive amount of time modifying prompts.

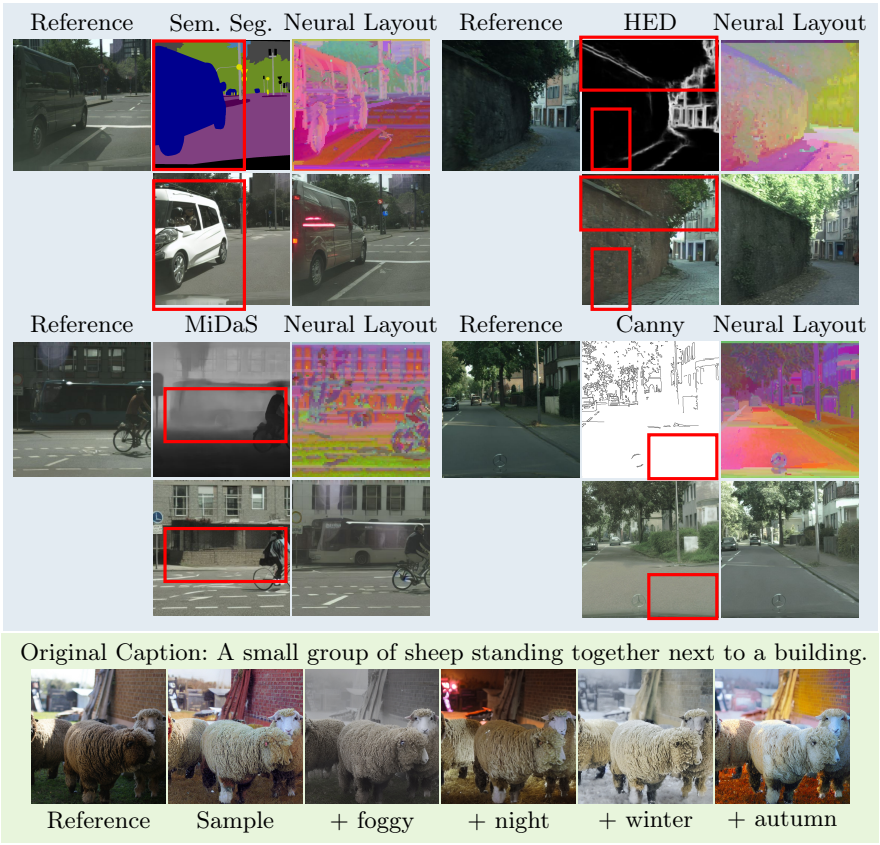


Fig. 1: Our proposed neural layout conditioning enables *neural semantic image synthesis*. Neural layout simultaneously specify semantic and spatial concepts (e.g., scene geometry, object semantics and orientation) without requiring expensive pixel-wise label annotations. In contrast, commonly used image conditions are geometrically (Sem.Seg.) or semantically (MiDaS, Canny, HED) ambiguous, as highlighted with the red boxes.

This becomes prohibitive when images need to be generated on a large scale without human-in-the-loop; for instance, when using them for training models in downstream perception tasks. For applications such as 2D/3D object detection, pose estimation, or semantic segmentation, it is crucial to ensure both *diversity* of generated images and their *faithfulness* to spatial details provided in the conditioning, e.g., correct position, size, pose and orientation of objects.

Recent work [19, 22, 25, 47, 49] proposed to introduce adapters to integrate spatial control to directly specify content. These methods have shown that segmentation, edge, depth, and normal maps as well as skeletal poses can be used to condition the composition of diffusion models. Given this variety, it is natural to ask what descriptor is best suited as spatial and semantic descriptor of scenes. We argue that two properties are key to the general applicability of a descriptor:

richness of semantic and spatial content and the *ease to obtain descriptor-image pairs* for fine-tuning pre-trained T2I diffusion models.

Semantic segmentation maps are a popular descriptor choice [20, 30, 37, 43], being interpretable high-level abstractions. However, creating them for real images requires either costly and tedious pixel-wise manual annotation or a suitable pre-trained network to generate noisy pseudo-labels. Even more so, since segmentation maps are ambiguous with respect to object pose, orientation, or geometry, errors such as cars driving the wrong way down a street (see Fig. 1) are quite common. Similarly, one can employ pretrained edge detector [41] or depth estimator [28] models to cheaply obtain image *edge* and *depth* conditions from unlabeled images [47]. While these conditions are clearer with respect to object geometry, they remain ambiguous with respect to semantic content. For example, both "cat" and "blanket" are plausible interpretations for the object boundaries seen in Fig. 4. Similarly for depth maps, the same geometric shape seen in Fig. 1 can be interpreted as either "building" or "bus". In short, existing conditioning descriptors can not satisfy both desired properties at once.

In this work, we propose a new way of conditioning T2I diffusion models to enable fine-grained spatial control without requiring expensive human annotations. We introduce the concept of *neural semantic image synthesis*, which derives its conditioning from dense neural features extracted from foundation models (FMs). Recent work [17, 23, 50, 52] has shown that these features preserve the semantic content and geometry of the images, and are thus rich spatial descriptors of the desired scene. However, these features encode nuisance appearance variations which must be removed to ensure diverse synthesis. We introduce a *semantic separation* step using PCA decomposition to extract only the desired information. The resulting features are refered to as *"neural layout"* (see Fig. 2).

Using neural layouts for conditioning reduces the need to create detailed textual descriptions or costly pixel-level annotations for training images. This makes it easier to scale up the training data needed to learn conditional control of T2I diffusion models, and thus further boost the quality of image synthesis. At the same time, neural layout conditioning ensures both faithfulness to semantic and spatial details and diversity of appearance. Providing both is vital for synthetic data augmentation for training different downstream perception tasks.

To showcase the benefits of neural layout conditioning, we propose the LUMEN model which stands for **L**abel-free ne**U**ral se**M**antic imag**E** sy**N**thesis. LUMEN builds upon ControlNet [47] and uses neural layouts extracted from an image's Stable Diffusion features [29] for conditioning (see Fig. 2). We show that images generated by LUMEN achieve superior alignment in semantic layout to the reference image even when compared to those created using expensive ground truth semantic label maps (see Tab. 2). In comparison to other conditioning inputs obtained from pretrained networks (e.g. predicted edges and depth), images generated with neural layouts capture better the semantics and geometry of the scene (see Fig. 4). Furthermore, we experimentally verify that this improvement in image alignment and quality results in performance gains when used to train downstream perception tasks (see Tab. 4 and Tab. 5).

2 Related Work

Controllable Image Synthesis. In the realm of generative adversarial networks (GANs), previous methods [15, 24, 33, 34, 38, 39, 53] have attempted to utilize various types of conditioning information to specify the spatial and semantic content of the synthesized images. With the increasing popularity of large-scale text-to-image (T2I) diffusion models, e.g., DALL-E 2 [27], eDiff-I [3] and Stable Diffusion (SD) [29], more recent works seek to incorporate additional control mechanism into the T2I models, to better steer the generation process. One popular direction is to leverage extra image specifications such as label maps, edges, or depth [19, 22, 43, 47, 49] to provide a more detailed description of the desired spatial composition, and thus gain finer control over the synthesis process.. FreestyleNet [43] finetunes the entire SD model to rectify the cross attention maps based on the semantic labels. T2I-Adapter [22] and ControlNet [47] both introduce an adapter network on top of a frozen SD backbone to accommodate the conditioning information. Uni-ControlNet [49] extends ControlNet by accepting multiple conditioning inputs via two adapters, further enhancing its controllability.

Different from these prior works and complementary to them, we focus on improving the representation of the image descriptor used by these mechanisms. We introduce the novel concept of "neural semantic image synthesis", which makes use of the rich spatial and semantic knowledge within large-scale pretrained foundation models (FMs) as conditioning input for T2I diffusion models. In contrast to existing representations, our proposed neural layouts can capture high-level semantic and geometry information, without human labeling efforts.

Dense Neural Features. Benefiting from advanced pretraining techniques such as self-supervised learning [6, 23] and joint vision-language pretraining [26, 29], large-scale pretrained foundation models (FMs) like CLIP [26], DINO [6], DINOv2 [23], and Stable Diffusion (SD) [29] itself can serve as robust feature extractors for diverse tasks. For example, several works [2, 12, 46] have discovered that these intermediate features can provide well-localized semantic correspondence. Others have used it for various perception tasks, e.g., object detection, semantic segmentation and depth estimation [17, 23, 50, 52]. A few recent works attempt to leverage these features for image synthesis in the context of image editing. PnP-Diffusion [35] represents the original image using SD features, which is more amenable to text prompting through the use of cross-attention. Other works [7, 44] use neural features to encode the exact identity of a subject so that it can be replicated in novel scenes.

Image Editing. Unlike image editing task [16, 42] that relies on prompt edits to change the image, the proposed "neural semantic image synthesis" aims to generate *various diverse images*, all of which satisfy the high-level semantics encoded by both the text and image conditions. Hence, instead of capturing the identical appearance, the proposed method focuses on keeping the overall semantics and spatial composition of the scene intact, meanwhile maintaining the freedom to synthesize diverse results, e.g., varying appearances.

3 Method

In this section, we introduce the concept of *neural semantic image synthesis*. Instead of using ad-hoc conditioning to describe the desired output, neural semantic image synthesis makes use of *neural layouts* derived from the dense features of pretrained foundation models. This semantically and spatially rich representation extracted from a reference image can better specify the desired semantic content and scene geometry without requiring annotations for finetuning.

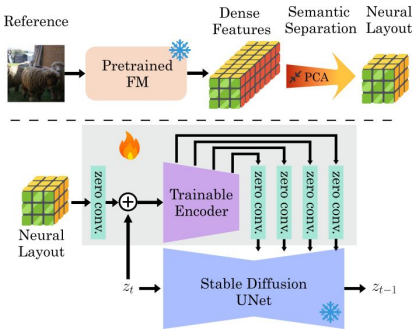


Fig. 2: LUMEN uses foundation models (FMs) features to extract neural layouts as conditioning input.

Our model LUMEN (Fig. 2) takes advantage of the observation that information is stored in easily separable form inside the dense features of foundation models [2, 6, 23, 35]. It has been shown that even a simple linear projection is sufficient to extract diverse contents such as semantic segmentation, depth estimates, and part correspondences [12, 23, 50, 52]. Analogously, we observe that semantic and spatial scene composition useful for specifying a desired image can be separated from the exact object appearance using a simple linear projector. Through this, LUMEN can align the generation with the spatial and semantic content of a reference image while providing meaningful variations in appearance through the sampling process.

We first introduce how neural layouts are extracted from a reference image in Sec. 3.1. We then explain how this can be used for conditional image synthesis with diffusion models in Sec. 3.2. Finally, we will provide the best practices for extracting semantically and spatially meaningful features from different foundation models in Sec. 3.3.

3.1 Neural Layout

Dense Feature Extraction. Modern foundation models make heavy use of self-attention (SA) and cross-attention (CA) modules [36]. Following the convention by Amir et al. [2], we introduce the query, key, value, and token features available at each attention layer that is associated with different patches of an input image $\mathbf{x_{ref}}$. At layer l, a SA module takes the tokens of the previous layer t^{l-1}, and linearly projects them into queries q^l_{SA}, keys k^l_{SA}, values v^l_{SA} using matrices W^l_q, W^l_k, W^l_v:

$$q^l_{SA} = W^l_q \cdot t^{l-1}, k^l_{SA} = W^l_k \cdot t^{l-1}, v^l_{SA} = W^l_v \cdot t^{l-1}.$$

CA modules instead compute the features based on another conditioning input y:

$$q^l_{CA} = W^l_q \cdot t^{l-1}, k^l_{CA} = W^l_k \cdot y, v^l_{CA} = W^l_v \cdot y.$$

The resulting features are combined through a series of scaled dot-product attention, normalization, multi-layer perceptron, and residual connections to obtain the token feature $t^l = \text{TransformerBlock}(q^l, k^l, v^l, t^{l-1})$.

Once extracted, these features are reshaped into a dense features map f corresponding to the original image patches. It was noted by several works [2, 46] that dense features can have different properties depending on where they are extracted. In Sec. 3.3, we detail the conventions we followed for different foundation models we investigated.

Semantic Separation. Retaining the entire dense feature map f reveals too detailed information about the reference image $\mathbf{x}_{ref}$ and thus impairs diversity (see Fig. 3), as generated samples are highly similar. To prevent this, we need to separate semantic and spatial features from appearance details. Based on existing works [23], we hypothesize that the principal directions in the dense features should at least partially correspond to what humans intuitively understand as spatial and semantic image content. We thus employ Principal Component Analysis (PCA) to obtain a linear projector that can remove nuisance variations.

To obtain the neural layout $\mathbf{c}_i$, used as conditioning in LUMEN, we retain only the information in the top N PCA components, where N is a hyperparameter of our method. In practice, we compute the PCA projection on a random sample of $40,000$ feature vectors extracted from images in the training set.

3.2 Conditional Image Synthesis

Neural semantic image synthesis generates the images $\mathbf{x}$ based upon text prompts $\mathbf{c}_t$ and a neural layout $\mathbf{c}_i$. While this task is compatible with a variety of conditional synthesis frameworks, we build LUMEN upon ControlNet [47], due to its popularity, ease of comparison, and its compatibility with most existing image conditions. ControlNet makes use of a frozen Stable Diffusion model [29] for T2I synthesis and adds a trainable adapter to incorporate the conditioning.

To train ControlNet, we start with an input image $\mathbf{x}_0$ and encode it into its latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ using the encoder $\mathcal{E}$. We add noise to $\mathbf{z}_0$ and obtain $\mathbf{z}_t$, where t is a time step specifying the scale of the noise added. The corresponding neural layout $\mathbf{c}_i$ is extracted from $\mathbf{x}_0$ and used as conditioning input. We train a denoiser ϵ_θ, to predict the noise ϵ added to $\mathbf{z}_t$, with the following loss $\mathcal{L}$:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}_0,\mathbf{t},\mathbf{c}_t,\mathbf{c}_i,\epsilon\sim\mathcal{N}(0,1)}\Big[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{t}, \mathbf{c}_t, \mathbf{c}_i))\|_2^2\Big].$$

3.3 Foundation Model Backbones

We considered 4 different foundation models for extracting neural layouts $\mathbf{c}_i$. This covers features from state-of-the-art methods trained in a self-supervised fashion (DINO [6] and DINOv2 [23]), vision-language models trained contrastively (CLIP [26]), as well as text-to-image synthesis (Stable Diffusion [29]). Since the output resolution of these features varies, all features are upscaled using exact nearest neighbor interpolation to match the image resolution.

DINO. We follow established work [2] and use the "key" feature from SA layers when processing $\mathbf{x_{ref}}$ as these features contain information useful for challenging semantic correspondence problems. We use the last SA layer and exclude the key corresponding to the CLS token to compute the neural layout.
DINOv2. As a scaled-up version of DINO that is trained on a large amount of curated data, DINOv2 potentially contains more generalizable features. Here, we use the non-CLS tokens of the last transformer layer for neural layout [46].
CLIP. We use the CLIP vision encoder to encoding $\mathbf{x_{ref}}$ and keeping all but the CLS tokens from the last feature layer before the final pooling [52]. These features were shown to be well-suited for segmentation tasks.
Stable Diffusion. We first use BLIP [18] to obtain text prompts $\mathbf{c}_t$ for the respective image $\mathbf{x_{ref}}$, and encode $\mathbf{x_{ref}}$ in the latent space of Stable Diffusion 1.5 (SD): $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_{ref})$. Next, we perform a single denoising step $\epsilon_\theta(\mathbf{z}_0, \mathbf{0}, \mathbf{c}_t)$ and extract the intermediate activations from layer 2, 5, and 8 of SD's U-Net, upsample them to match the resolution of layer 8, and concatenate them along the channel dimension. These SD features [46] were shown to have a strong sense of spatial layout, which makes them a promising candidate for neural layouts.

4 Experiments

In this section, we first explain our experimental setup and evaluation metrics. We show the impact of our design choices for neural layout conditioning in Sec. 4.1 and demonstrate the advantages of neural layouts compared to existing conditioning inputs in Sec. 4.2. Finally, we validate the quality of images synthesized by LUMEN by showing its impact on downstream applications in Sec. 4.3.
Evaluation Metrics. To measure the image synthesis quality of our method, we use FID [13] for perceptual quality, average LPIPS [48] between generated samples for diversity, and TIFA [14] for text controllability. We additionally evaluate how well each conditioning captures the semantic composition and geometry of the scene. For alignment with semantic layouts, we use mIoU between ground truth segmentation label and those predicted by a pretrained segmenter; here we use Mask2Former [8] for COCO-Stuff [4] and ADE20k [51], DRN [45] for Cityscapes [9]. However, since mIoU does not contain 3D information, we introduce the scale-invariant depth error (SI depth) [10] as a metric for geometric consistency. SI depth is computed on depth estimated using MiDaS [28] from the reference image and the synthesized image. Further evaluation and implementation details are available in the supplementary material.

4.1 Neural Layout Design Space

We explore the design space of neural layouts on the diverse COCO-Stuff dataset [4] to determine how to best extract descriptive semantic and spatial information from a given reference image.
Best Features for Neural Layout. For the task of neural semantic image synthesis, we want neural layouts to preserve semantic and geometry content but

Features	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑
CLIP	41.7	25.7	15.6	**0.58**	**0.84**
DINO	49.5	22.2	12.8	0.41	0.77
DINOv2	51.1	22.0	12.8	0.42	0.78
SD	**51.4**	**21.5**	**12.2**	0.42	0.79

(a) Comparison of FMs for extracting neural layouts using 10 PCA components. SD features produces the best overall image quality.

Comp.	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑
1	45.0	25.5	14.5	0.53	0.87
5	50.0	22.8	12.9	0.44	0.81
10	51.4	21.5	12.2	0.42	0.79
20	52.9	21.1	11.8	0.36	0.66
100	54.6	18.6	8.8	0.25	0.47
Image	56.1	15.6	5.9	0.15	0.30

(b) Impact of the number of PCA components on image quality for SD features.

Table 1: We ablate the choice of FM feature (a) and the number of PCA components (b) on the COCO dataset. We observe that the number of SD components can be used to select the desired trade-off between faithful generation (mIoU, FID, SI Depth) and sample diversity (LPIPS)/text controllability (TIFA). At large number of components, the performance approaches using the reference 'Image' itself as conditioning.

to discard appearance details. Thus, we want a backbone that extracts features where this is readily separable. To evaluate this, we compare in Tab. 1a the image quality of DINO, DINOv2, CLIP, and SD features at $N = 10$ PCA components.

We observe that SD features provide the best perceptual image quality and also retain the semantic content best, while DINOv2 is a close second. Although CLIP conditioning can generate more varied images, this diversity is due to the weak semantic and spatial constraints imposed during synthesis. This is likely because CLIP's image-level training objective is less suitable to capture precise pixel-level information without further processing.

Number of PCA Components. Since discarding PCA components removes information, useful or not, the resulting synthesis task also becomes less constrained (See Fig. 3). This is reflected in Tab. 1b where we compare synthesis results across different numbers of PCA components. We observe a clear trade-off between metrics linked to semantic layout and geometry (mIoU and SI Depth), image diversity (LPIPS), and text editability (TIFA). As additional PCA components impose further constraints, it becomes increasingly likely that these constraints conflict with the text description. In these cases, the generator prioritizes the more constraining image conditioning over the more ambiguous text conditions. Note that when the number of components is chosen large enough, conditioning with neural layout simply reproduces the reference image and the performance approaches that obtained from directly using the reference image as conditioning input (see Tab. 1b).

These results demonstrate that the number of components utilized in the neural layout provides an intuitive mechanism to balance the trade-off between faithfulness, diversity, and text responsiveness. This level of control is absent in most other layout conditions.

LUMEN. Given these results, we design LUMEN to use neural layout extracted from SD features using $N = 10$ and $N = 20$ PCA components, as this result in a satisfactory range of diversity while remaining faithful to the reference image. However, this can be further fine-tuned to suit specific downstream tasks.



Fig. 3: Visual comparison of four random samples drawn using LUMEN trained with a different number of PCA components. Using fewer trades fidelity for diversity.

		COCO-Stuff					ADE20K				
Conditioning	Pix. Annot.	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑	mIoU ↑	SI Depth ↓	FID ↓	LPIPS ↑	TIFA ↑
Edges (Canny)	No	44.4	24.7	13.2	0.48	0.84	35.1	25.0	19.1	0.49	0.86
Edges (HED)	Indirect	49.3	21.4	12.1	0.39	0.74	41.8	**21.1**	17.4	0.37	0.73
Depth (MiDaS)	Indirect	45.3	24.0	14.3	0.53	**0.88**	34.0	22.1	21.2	0.52	**0.87**
Sem. Seg. (pred)	Indirect	42.9	27.7	16.0	0.64	**0.88**	33.0	24.6	23.4	0.62	0.85
Sem. Seg. (GT)	Yes	43.3	28.8	15.3	**0.65**	**0.88**	35.1	27.1	22.6	**0.63**	0.85
Neural Layout	No	**52.9**	**21.1**	**11.8**	0.36	0.66	**45.7**	**21.1**	**16.1**	0.33	0.67

Table 2: Comparison of using different conditioning options for ControlNet across datasets. Neural layout outperforms all conditioning options in terms of image quality (FID), as well as semantic and spatial alignment (mIoU, SI).

4.2 Comparison to Existing Conditioning

We compare LUMEN using neural layouts with $N = 20$ PCA components against different established image conditioning. For this, we consider the common ControlNet condition: Canny edges [5] (Canny), HED edges [41] (HED), MiDaS depth [28] (MiDaS), and semantic segmentation with ground truth (Sem. Seg. GT) and predicted labels (Sem. Seg. pred). For Sem. Seg. (pred), we use the same pretrained network as the one used for evaluating the mIoU metric.

It is important to first discuss the relative cost of obtaining the conditioning input for each of the types. Sem. Seg. (GT) requires expensive pixel-level annotations making it difficult to scale the generator training. HED, MiDaS, and Sem. Seg. (pred) can be cheaply obtained through pretrained estimators, which, however, require pixel-wise annotations in their training. In contrast, neural layout is either weakly supervised with much more affordable image captions or entirely self-supervised, depending on the backbone. This distinction could be important for applications with domain shifts where the feature extraction network itself need to be fine-tuned. The only other conditioning input that fulfills this criteria is Canny edges. These requirements for obtaining the conditioning is summarized in Tab. 2 (Pix. Annot.).

Effect of Conditioning Types. We evaluate the effects that different conditioning have on image synthesis using the challenging COCO-Stuff [4] and ADE20k [51] datasets. As Tab. 2 shows, using neural layout best preserves the semantic content, outperforming others in terms of both mIoU and SI Depth while achieving better image quality. Surprisingly, Sem. Seg. achieves only sim-

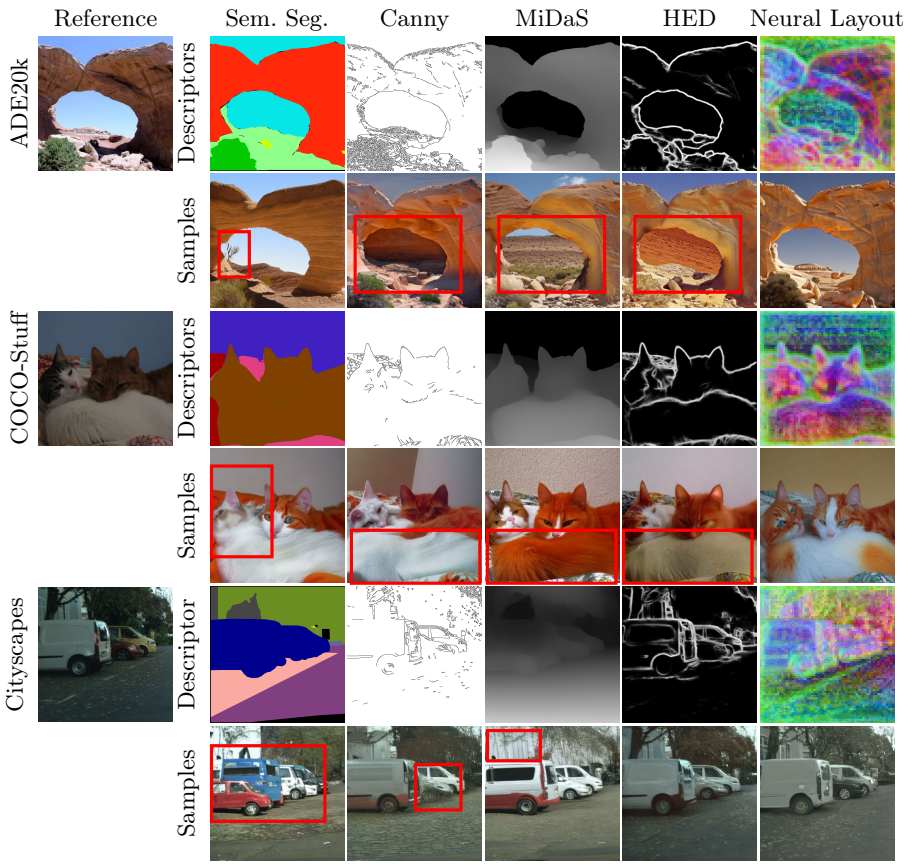


Fig. 4: Comparison of images generated with different conditioning types on ADE20k, COCO-Stuff and Cityscapes. Neural layouts provide rich description of the desired images, while other inputs contain limited information and are semantically ambiguous.

ilar or worse results in terms of mIoU compared to all other conditioning. We believe this is due to both datasets featuring a large number of difficult semantic classes with sometimes semantically ambiguous labels, and due to the long tail distribution on rare classes. We also see that HED performs second best overall in terms of FID, mIoU, and SI Depth, as it well encodes the object boundaries with additional image details being captured by soft edges. Still, the semantic class of the object within the boundary can be ambiguous, resulting in a lower mIoU compared to Neural Layout (also see Fig. 4).

Note that we observe again the trade-off between information content constraining the image and the diversity and editability of the results. Canny edge and MiDaS depth have low semantic content and sem. seg. does not constrain appearance or geometry, consequently, they often achieve the best LPIPS and TIFA at the cost of worse alignment or image quality. We also observe that despite



Fig. 5: Image manipulation through prompt editing on the COCO-Stuff validation set. We show an unedited sample and additional results from replacing the underlined words in the original caption (→) or appending additional words at its end (+).

the lower TIFA, LUMEN still responds well to a variety of out-of-distribution prompt edits (see Fig. 5). As we will see in Sec. 4.3, LUMEN provides enough image variation from sample diversity and text editability in practice for the downstream network to achieve the best performance.

Effect of Data Scaling. Our method can utilize additional unlabeled data to enhance synthesis quality. This is exemplified with Cityscapes [9], which contains only 2975 training images with pixel-level labels (L). However, there are ∼20k images without labels (U) from Cityscapes that can be leveraged for training.

The positive impact of this extra data is evident in Tab. 3. When trained with only a very limited number of labeled images, Sem. Seg. achieves the best alignment in terms of mIoU (57.9), slightly surpassing neural layout (57.6), but falls behind in terms of scene geometry (16.9 vs. 12.1). With the inclusion of more data (L + U), all methods except for Sem. Seg. improve in terms of mIoU. Notably, our neural layout simultaneously improves in mIoU, SI Depth, and FID, outperforming Sem. Seg. that relies only on manual labels. To allow Sem. Seg. to also leverage the extra data, we employ the same DRN segmenter [45] for pseudo labelling as for mIoU evaluation. However, we observed a significant drop in its mIoU performance, likely attributed to the noise in these pseudo labels.

Conditioning	Data	mIoU ↑	SI Depth ↓	FID ↓
Edges (Canny)	L	35.7	15.8	37.6
	L + U	38.8	16.0	43.2
Edges (HED)	L	51.3	11.7	31.0
	L + U	53.9	**11.4**	**29.1**
Depth (MiDaS)	L	44.2	16.4	37.8
	L + U	46.7	16.5	35.9
Sem. Seg.	L	57.9	16.9	44.8
	L + U*	53.3	16.2	*46.5*
Neural Layout	L	57.6	12.1	31.2
	L + U	**58.2**	11.8	30.8

Table 3: Impact of using *unlabelled* data (U) to augment labelled data (L) to for training the image generator.



Fig. 6: Example of cross-domain (COCO to NYUv2) synthesis. Images created using neural layout best align with the reference image. Other images are mismatched in terms of instance separation, room/object geometry, or object semantics.

	sem. seg		depth		surface normal				
Conditioning	mIoU ↑	pixAcc ↑	AbsErr ↓	RelErr ↓	mean ↓	median ↓	<11 ↑	<22 ↑	<33 ↑
Baseline	44.6	69.9	**1.17**	**0.56**	25.8	18.7	31.7	57.3	68.6
Edges (Canny)	45.2	70.2	1.26	0.58	26.0	18.8	31.2	57.0	68.2
Edges (HED)	45.7	70.7	1.27	0.59	25.3	18.0	32.6	58.7	69.7
Depth (MiDaS)	45.8	71.2	1.24	0.58	25.2	18.0	32.4	58.7	69.7
SemSeg (pred)	47.0	71.4	1.27	0.59	25.4	18.3	31.8	58.1	69.3
Neural Layout	**47.3**	**72.0**	1.25	0.58	**24.9**	**17.7**	**32.7**	**59.4**	**70.4**

Table 4: Synthetic images improve the multitask performance on the NYUv2, despite that the generator is trained on COCO-Stuff. Additional synthetic training data improves upon the baseline in all cases, but most so for neural layouts.

4.3 Downstream Applications

In this section, we show the applicability of data from LUMEN to different downstream applications. Since image diversity is crucial for model training, $N = 10$ PCA components are kept in the neural layout.

Cross Domain Multi-Task Training. Being label-free, we can leverage unsupervised fine-tuning on a large dataset in order to augment a labeled training set with diverse variants. We demonstrate this by using LUMEN finetuned on COCO-Stuff [4] to create synthetic variants of training data from NYUv2 [32]. NYUv2 [32] is a small dataset of indoor scenes commonly used for multi-task learning. LibMTL [21], with its standard setting, was used to train a multi-task network based on ResNet-50 [11] to predict the semantic segmentation, image depth, and surface normal. To obtain the baseline results, we used only the real training data to train the multi-task network. For all other methods, we use the stated conditioning extracted from reference images in the real training data to create a synthetic dataset of equal size. In the case of SemSeg conditioning, we treat the segmentation network as a cross domain feature extractor and sim-

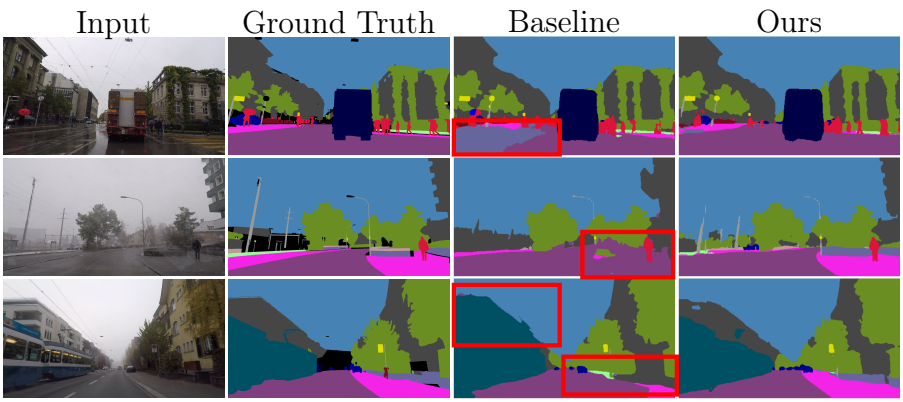


Fig. 7: Our additional synthesized data improves upon baseline of using only real Cityscapes data for domain generalization to the ADAC dataset.

ply use the predicted COCO-Stuff classes. The synthetic and real datasets are combined to train the multi-task network.

We report the standard metrics for NYUv2 multi-task learning in Tab. 4. Neural layout is able to outperform all other methods, including the real data baseline, for both semantic segmentation and surface normal estimation simultaneously as it preserves both semantics and geometry (see Fig. 6). In contrast, even though SemSeg conditioning performs well in the semantic segmentation, the lack of geometric information makes it worse than depth conditioning for estimating surface norms. Finally, despite the ill-posed nature of the monocular metric depth estimation due to depth-scale ambiguity and the out-of-domain image synthesis, our method performs on-par to explicit depth-based conditioning.

Domain Generalization. By choosing to use ControlNet as our backbone generator, we can use the text prompt to control the domain (mainly appearance) of the synthesized images while using neural layout to control the semantic and geometric content. We use this to perform domain generalization experiments from the daytime only Cityscapes to ACDC [31], containing adverse weather and lighting conditions (for more details see suppl. material). As shown in Tab. 5 and in Fig. 7, we verified that images generated by LUMEN can significantly improve the model's generalization ability upon the baseline, which is trained only on Cityscapes. We compare against other diffusion-based methods PnP-Diffusion [35], FreestyleNet [43], as well as ControlNet [47] with predicted Sem. Seg. conditioning. The prompt editing of PnP-Diffusion cannot generalize well to the image domain of

Method	CS	Rain	Fog	Snow	Night	Avg.
Baseline (CS)	67.9	50.2	60.5	48.9	28.6	47.0
PnP-Diffusion [35]	67.8	50.6	63.5	50.4	30.3	48.7
ControlNet [SemSeg]	66.8	50.4	64.7	51.7	34.6	50.3
FreestyleNet* [43]	**69.7**	52.7	**69.0**	54.3	32.9	52.2
LUMEN (Ours)	68.5	**53.4**	67.4	**55.6**	**35.1**	**52.9**
Oracle (CS + ACDC)	68.2	63.7	74.1	68.0	48.8	63.6

Table 5: Comparison of augmentation techniques for domain generalization via prompt edit. Quality is measured using mIoU of a SegFormer [40].



Fig. 8: LUMEN can create artistic stylized versions of an image using prompt editing.

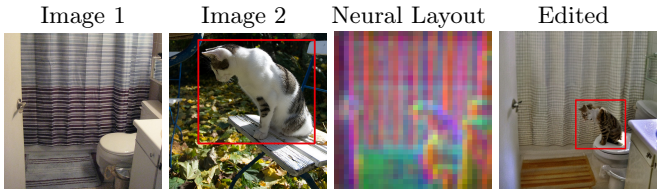


Fig. 9: Neural layouts can be combined to define desired layout in the final image.

Cityscapes, leading to little benefits for domain generalization. FreestyleNet require manual annotation to train the image generator, in contrast to our label-free LUMEN. Yet, our method still outperforms FreestyleNet overall.

Content Creation. Although we motivate our method by considering applications generation is automated (i.e. for creating training data), it is also suitable for interactive content creation. We show in Fig. 8 that LUMEN is not limited to synthesize realistic images, but is also versatile enough to generate images with a variety of different artistic styles. This suggests that the semantically separated neural layout is able to discard stylistic content of the reference. We further show in Fig. 9 that, with user interaction, object layouts and scene composition can be altered by combining different images. Although not as flexible as specifying composition from scratch, LUMEN still enables a new workflow where a creator can specify both the desired spatial and semantic layout of the output while relying on the diffusion model to fill in the appearance.

5 Conclusion

We introduced the concept of neural semantic image synthesis and established LUMEN as a strong label-free baseline that can simultaneous specify semantic and spatial concepts of the outputs. Neural semantic image synthesis has additional applications that we hope will be explored in future works. For example, targeted removal of irrelevant information could be possible by replacing PCA with a learned projector specialized to a downstream task. Similarly, one could use neural semantic image synthesis to create images in a target domain from images in a source domain if a projection direction can be learned to eliminate the domain differences. This allows the reuse of labels from one image domain to another in order to save costly annotation efforts.

References

1. Midjourney. https://www.midjourney.com/ (2023)
2. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 (2021)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018)
5. Canny, J.: A computational approach to edge detection. TPAMI (1986)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
7. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. NeurIPS (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: ICCV (2016)
12. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. arXiv preprint arXiv:2305.15581 (2023)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
14. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
16. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. ICLR (2024)
17. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: CVPR (2023)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
19. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
20. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Adversarial supervision makes layout-to-image diffusion models thrive. In: ICLR (2024)
21. Lin, B., Zhang, Y.: LibMTL: A Python library for multi-task learning. Journal of Machine Learning Research **24**(209), 1–7 (2023)
22. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)

23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
25. Qin, C., Yu, N., Xing, C., Zhang, S., Chen, Z., Ermon, S., Fu, Y., Xiong, C., Xu, R.: Gluegen: Plug and play multi-modal encoders for x-to-image generation. In: ICCV (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
27. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
28. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. TPAMI (2022)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
30. Saharia, C., Lee, C.A., Fleet, D.J., Chang, H., Ho, J., Norouzi, M., Salimans, T., Chan, W.: Palette: Image-to-image diffusion models. In: SIGGRAPH (2022)
31. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)
32. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760 (2012)
33. Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: Oasis: only adversarial supervision for semantic image synthesis. IJCV (2022)
34. Tan, Z., Chen, D., Chu, Q., Chai, M., Liao, J., He, M., Yuan, L., Hua, G., Yu, N.: Efficient semantic image synthesis via class-adaptive normalization. TPAMI (2021)
35. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR (2023)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
37. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952 (2022)
38. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
39. Wang, Y., Qi, L., Chen, Y.C., Zhang, X., Jia, J.: Image synthesis via semantic composition. In: ICCV (2021)
40. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)
41. "Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
42. Xu, S., Huang, Y., Pan, J., Ma, Z., Chai, J.: Inversion-free image editing with natural language. In: CVPR (2024)
43. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: CVPR (2023)

44. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
45. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: CVPR (2017)
46. Zhang, J., Herrmann, C., Hur, J., Cabrera, L.P., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In: NeurIPS (2023)
47. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
49. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. In: NeurIPS (2023)
50. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: ICCV (2023)
51. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
52. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV (2022)
53. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NeurIPS (2017)