ControlNet-XS: Rethinking the Control of Text-to-Image Diffusion Models as Feedback-Control Systems

Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother

Computer Vision and Learning Lab, IWR, Heidelberg University, Germany
{name.surname}@iwr.uni-heidelberg.de

Abstract. The field of image synthesis has made tremendous strides forward in the last years. Besides defining the desired output image with text-prompts, an intuitive approach is to additionally use spatial guidance in form of an image, such as a depth map. In state-of-the-art approaches, this guidance is realized by a separate controlling model that controls a pre-trained image generation network, such as a latent diffusion model [64]. Understanding this process from a control system perspective shows that it forms a feedback-control system, where the control module receives a feedback signal from the generation process and sends a corrective signal back. When analysing existing systems, we observe that the feedback signals are timely sparse and have a small number of bits. As a consequence, there can be long delays between newly generated features and the respective corrective signals for these features. It is known that this delay is the most unwanted aspect of any control system. In this work, we take an existing controlling network (ControlNet [88]) and change the communication between the controlling network and the generation process to be of high-frequency and with large-bandwidth. By doing so, we are able to considerably improve the quality of the generated images, as well as the fidelity of the control. Also, the controlling network needs noticeably fewer parameters and hence is about twice as fast during inference and training time. Another benefit of small-sized models is that they help to democratise our field and are likely easier to understand. We call our proposed network ControlNet-XS. When comparing with the state-of-the-art approaches, we outperform them for pixel-level guidance, such as depth, canny-edges, and semantic segmentation, and are on a par for loose keypoint-guidance of human poses. All code and pre-trained models will be made publicly available.

Keywords: Text-to-Image Generation · Controlling Image Generation Models · Feedback-Control Systems

1 Introduction

Using Generative Artificial Intelligence to synthesize new images is a topic that has received large attention in social media, press, research and industry. It started off in 2014 with the introduction of Generative Adversarial Networks

Fig. 1: Image synthesis using our approach with text-prompts, as well as, a guidance image in form of a depth map, canny-edges image, semantic map, and human pose. The two results on the left-hand side were generated by the production-quality model of Stable Diffusion XL [53], and the remaining by Stable Diffusion Version 1.5.

(GAN) [19] that were able to synthesize small-sized images of a given class [56], *e.g.* celebrity faces. Today, we have commercial and non-commercial products, such as Midjourney [1] and Stable Diffusion XL [53], which are able to generate large-sized images (up to 1024×1024) with almost arbitrary content, *e.g.* ranging from professional photographs to Manga Art. This can be considered as a truly disruptive technology, for which the development is still far from being completed. One ongoing development is to create control tools, with which users can steer the image generation process towards their desired output. A common control mechanism is text-prompts. Another choice is to use a *guidance image*, which defines the desired output image in an abstract form such as a sketch or a depth map. This control mechanism is known as image-to-image translation, *e.g.* [30]. The control mechanisms can also be combined by adding one or many guiding images to a text-to-image model, *e.g.* [88]. Our work falls into this class of methods. There are in general two different choices for implementing the image guidance in a text-to-image model.

On one hand, there is the approach of *fine-tuning* a generative model with a new control mechanism at hand, *e.g.* [79], or training it from scratch *e.g.* [64]. Such methods use guidance images as additional input. However, such an end-to-end learning approach is challenging since oftentimes there is a large imbalance between the original training data for the generative process, *e.g.* $\sim$ 3B images for training Stable Diffusion [64], in contrast to only 1M images with known control, as in [88]. Such an imbalance can lead to effects like "catastrophic forgetting" [46], which means that known properties of the generative model disappear after fine-tuning. Additionally, fine-tuning often requires access to a large computing cluster.

On the other hand, there are approaches that lock the parameters of the generative network and then train a *separate controlling network*. In general, the idea of two networks communicating with each other has been shown to be beneficial for various computer vision tasks [42, 43, 76, 80]. In our context, this particular design choice [47] is currently most popular and also utilised in this work.

When analysing this approach from a control system perspective, we see that nearly all existing approaches are feedback-control systems, also known as closed-loop control systems. The controlling network is the *controller* and the generation processes is a dynamically changing system. The only exception is the T2I-Adapter approach [47] since it forms an open-loop control system where

the controlling network does not receive any feedback from the generating network. From a control system perspective, it is of paramount importance that the controller receives feedback from the generation process as often as possible (*i.e.* high-frequency) and also as much useful information as possible (*i.e.* large-bandwidth). Furthermore, the controller should send as often as possible a control signal back to the generation process. In a typical hardware control system, such as a manufacturing plant, there are physical limitations to achieve this. However, with software, we do not have such limitations. Unfortunately, to the best of our knowledge, all existing guided image generation methods which work as feedback-control systems [27, 54, 88, 90] fall short in terms of designing a bidirectional, high-frequency and large-bandwidth communication between the controller and the generation process. The main contribution of our work is to construct a control system with these properties. For this, we use ControlNet [88] as our initial controlling network. While improving the communication mechanism, we observe that we can scale down the number of parameters of ControlNet by a factor of 6.5 (or even more), and at the same time improve the quality of the generated images as well as the fidelity of the control. As a result, we are about two times faster than ControlNet with respect to inference and training time. We call our network *ControlNet-XS*. Another benefit of small-sized models is that they help to democratise our field and are likely easier to understand. It is important to note that our main contribution, a bidirectional, high-frequency and large-bandwidth communication, can also be integrated into all other existing guided image generation approaches that are designed as a feedback-control system [27, 54, 88, 90]. Furthermore, also approaches in related fields of generative Artificial Intelligence which utilise a feedback-control system could benefit from our contribution, such as video translation [15, 92] or controlled 3D object generation [29, 87].

Let us consider an example which illustrates the importance of having a high-frequency communication between the controller and the generation process. Assume that the generation process is a vehicle, which is equipped with a controller that is a satellite navigation system. The controller receives the current position of the vehicle, and gives back control signals such as "turn left at next junction". Assume that the target or input to the system is to drive to a specific address. A crucial requirement of this control system is, obviously, that the navigator knows always the exact position of the vehicle. However, if the navigator were to know the position of the vehicle with a delay of ten seconds, then the vehicle might have already passed the junction where it should have turned left. Hence, the controller has to be smart and predict where the vehicle may be in ten seconds time in order to give commands that are sensible for the vehicle at that point of time in the future. This is exactly what happens in all existing approaches for guided image generation that are designed as a feedback-control system. Due to delayed feedback of the generation process to the controlling network, it has to guess what the generation network is doing until it sends its controlling signal. Hence, for this task, the controlling network needs generative power.

In summary, our contributions are as follows: (1) Demonstrating the importance of controlling a text-to-image generation model with a bidirectional, high-frequency and large-bandwidth communication. (2) Training a straight-forward, small-sized controlling network that outperforms the state-of-the-art for pixel-level image guidance (depth, canny-edges, semantic map) and is on a par for loose keypoint guidance of human poses. (3) Controlling the production-quality Stable Diffusion XL model [53] with 2.6B parameters with a control network that has only 20M parameters.

2 Related Work

2.1 Image Generation and Translation

Generative Adversarial Networks (GANs) [19] are probably the most established generative models for unconditional [33–35], class-conditional [71] and text-conditional [31, 62, 63, 70, 77, 82, 86, 94] image generation as well as domain adaptation [48, 81, 93]. While achieving state-of-the-art results for particular semantic classes [32, 67, 78], generalizing GANs to synthesize images of arbitrary content remains an active area of research. With GigaGAN [31] demonstrating state-of-the-art performance, generic text-to-image synthesis is still dominated by autoregressive networks like DALL-E [59], Parti [85], CogView [13], Make-A-Scene [16] and Diffusion Models in particular. Since their introduction, **Image Diffusion Models** [73] rapidly became one of the best performing model-families [12, 23, 24, 36, 50, 74]. Diffusion models learn to transform a point from a simple, high-dimensional distribution, such as a Gaussian, to a complex distribution, like the space of all images. This transformation is done by iteratively applying a network that gradually removes the Gaussian noise in the image. This process allows to theoretically model arbitrary complex data distributions [73].

Conditioning Image Synthesis Models. To generate a desired output image, one popular choice is to use text-prompts as guidance. This is done by conditioning a generic image synthesis model on a textual-embedding, provided by pre-trained text-encoders like BERT [11], T5 [57] or CLIP [55]. Such conditioning has led to impressive results for complex text-to-image generation task by models like Stable Diffusion [64], DALL-E [3, 58], Imagen [69] and many others [49, 53, 83]. However, text-prompts alone provide very little control over the exact details within the generated image. To address this problem, the concept of *guidance images* became popular. Guidance images describe the desired output scene in an abstract form. This ranges from very loose guidance, such as bounding boxes and keypoints for human poses, on to slightly more precise controls like semantic maps or sketches, and up to pixel-accurate guidance, such as depth maps, normal maps, or edge maps.

As mentioned in the introduction, one line of work is to utilize guidance images as conditioning before training the model [14, 28, 38, 68, 83]. Another variant is to adapt a pre-trained model by fine-tuning it with new guidance images [26,79]. However, the drawbacks of both approaches are that they require substantial computational resources to train, and also that conditional modalities

cannot be changed without excessive re-training. The more popular approach is to train a separate controlling network which is combined with a pre-trained generation model, as discussed in the next section.

Image Editing and Subject-Driven Image Generation. There have been many works leveraging the rich internal representation of pre-trained text-to-image models to customise the output. This usually involves the editing of an existing image [2,9,10,18,37] or the insertion of a specific subject instance to the generated output [17, 51, 66, 84]. The approach of DiffEdit [10] takes an existing image and utilises local masks to manipulate the content. InstructPix2Pix [4] trains an image generation model to edit images according to human instructions. This is done by learned image-editing instructions. Dreambooth [66], on the other hand, fine-tunes a pre-trained text-to-image generation model to a set of photographs representing a specific subject, binding it to a text-token which can be used to generate the desired subject in a new environment. In general, such approaches can be considered as complementary to works that use pixel-accurate image guidance, such as ours, since these approaches do not control the generated, or edited, content on a pixel-level.

2.2 Controlling Pre-Trained Networks

With the increasing size of generative models, it has become popular to leave the generative base model unaltered in order to keep its generative capabilities. A straight-forward approach is to employ weight-adaptation methods like "Low-Rank-Adaptation" (LoRA) [25] to add a learnable offset to the pre-trained weights, which are approximated by the multiplication of two low-rank matrices. After training, the weights can be merged without the need to add new parameters to the network. Beyond this simple approach, there are in general three concepts for controlling image generation models by the addition of new network components, although these concepts cannot always be clearly distinguished.

Adapters. One control mechanism is to use so-called adapters, which insert new trainable modules, *e.g.* neural blocks, to the pre-trained generative network. However, in contrast to LoRA the inference time increases. Adapters are popular in natural language processing [45, 52, 75] and have been transferred to image generation models [61, 65], vision transformers (ViT) [39], and are also used for dense predictions in the form of ViT-Adapters [8]. In the context of our work, there is one adapter-based approach, T2I-Adapter [47], which uses a guidance image to control a pre-trained text-to-image Stable Diffusion model. It is an open-loop control system where features, derived from the guiding image, are added to the generation model. There is no feedback signal from the generation process back to the controlling network. We validate experimentally that it is on average inferior to feedback-control systems.

Image Control with Attention Maps. Another popular control mechanism is to manipulate the attention maps of the diffusion model [7,20,40,44,91]. One example of such an approach is GLIGEN [40], to which we compare experimentally. GLIGEN introduces new learnable gated attention layers to incorporate the guidance. While this gives impressive results for loose guidance, such

as bounding boxes, we validate experimentally that for pixel-accurate guidance, like depth maps, it performs rather poorly compared to methods discussed next.

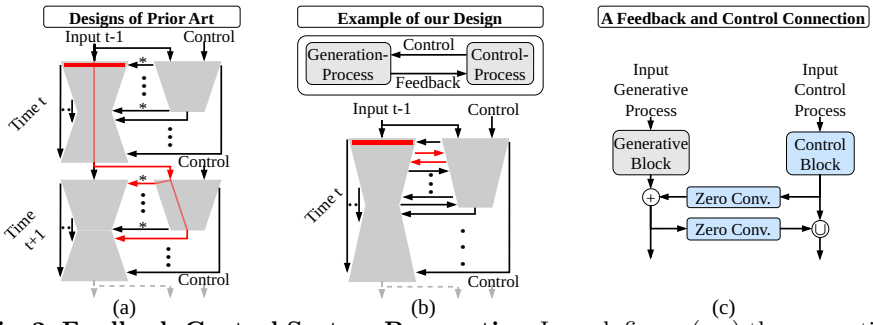
Image Control with a Separate Controlling Network. The final approach for controlling a pre-trained image generation model is to train a separate controlling network, which generates features that are combined with the generation model. This is also our approach. The first work in the context of diffusion models was ControlNet [88] which trains a separate controlling network for each kind of guidance image. Building on the design of ControlNet, three recent works appeared that allow to use multiple guidance images jointly within one controlling model. These are UniControl [54], Uni-ControlNet [90] and Cocktail [27]. The main focus of these works is to design neural networks that merge multiple control signals. For instance, UniControl uses a Mixture-of-Experts (MOE) Adapter in combination with a task-aware network. By doing so, each of these works add additional "concepts" on top of the initial ControlNet model: i) an additional text-embedding in the control signal [54], ii) a global control adapter [90], iii) additional manipulation of attention maps of the encoder of the generation model [27]. In contrast to these works, we focus on single image guidance. We also use ControlNet as initial model, but then only reduce its size without adding any other "concept". Our key contribution is a new communication mechanism between the controlling and generating networks, which is different to all four works [27, 54, 88, 90]. By doing so, we are able to improve over the state-of-the-art. It is important to note that our improved communication mechanism can also be integrated into Uni-ControlNet, UniControl and Cocktail.

3 Method

We start with a brief introduction to the Stable Diffusion [64] architecture, which serves as our generative model (Section 3.1). The pre-trained generative model is controlled by a controlling network. In Section 3.2 we analyse the design of existing controlling network architectures from a feedback-control system perspective. In this work, we build upon ControlNet [88] as a controlling network which we describe in Section 3.3. Lastly, in Section 3.4, we introduce our ControlNet-XS network and its training procedure in Section 3.5.

3.1 Stable Diffusion

Stable Diffusion [64] is a U-Net based diffusion model for text-to-image generation. As a conditional diffusion model, it receives a text embedding from a separate text encoder, as well as a learned time embedding. The output image is reconstructed from noise by iteratively running the U-Net over, for example, 50 time-steps. The U-Net generator is composed of a sequence of neural blocks involving cross-attention mechanisms for text conditioning. The image signal is processed by the encoder in four layers with diminishing resolution and three neural blocks per layer. Through the mirrored structure of the decoder and one middle block in between, the U-Net has a total of 25 neural blocks. The output of each neural block can be influenced individually by a controlling network.



(a) (b) (c)

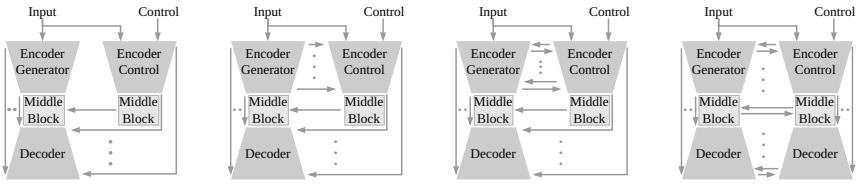
Fig. 2: Feedback-Control System Perspective. In each figure (a-c) the generation process is on the left-hand side, and the control process on the right-hand side. The focus of this illustration is on the communication (directed arrows) between the generation and controlling process. (a) Feedback-control system for approaches [27,54,88,90], where links denoted by * are only present in [27]. (b) An example of our communication design. (c) Zoom into the connections between a generative encoder block and a ControlNet-XS block. Please find the explanation for this figure in Section 3.2

3.2 Feedback-Control System Perspective

In Figure 2a-b we analyse two different design choices for controlling a generation process. Feedback-control systems for approaches [27,54,88,90] are shown in (a), where links denoted by * are only present in [27].[1] Note that we only illustrate the generation and controlling networks and their respective communication links. The approaches [27, 54, 90] have additional networks and communication links (see Section 2) which are omitted here, since it is not relevant to our analysis. The main drawback of design (a) is that there can be generated features, illustrated by the red rectangle in the encoder, at time t of the generation process, that evolve in the generative U-Net and receive a control signal only at time $t+1$ (two red arrows pointing towards U-Net). However, by that time, they have travelled through 35 generative blocks (or 25 blocks if * links are present). The two red paths show possible flows through the network which start at the generated features (red rectangle) and pass through the controlling network until they form the control signals for the generative process at time $t+1$ (red arrow with * or without *). There is no earlier control signal that is aware of the generated features (red rectangle). In our design (b), this drawback is eliminated and such features (red rectangle) receive a control signal after one generative block.

Besides implementing a high-frequency communication, the bandwidth of the connection may also play a crucial role. When measuring the bits that travel through the networks, we notice that rather few bits go from one time-step to the next time-step of the generation process (precisely 524K bits for a latent image of size 64×64×4). This is the only feedback to the controlling network in design (a). In contrast, the total number of bits entering the controlling network in (b) is 212M bits, which is over 404 times more.[2] The conclusion of this analysis

[1] The * links adapt the attention maps of the generative encoder.
[2] The size of the features is measured where they leave the generative model.



(a) ControlNet [88] **(b)** Our Type A **(c)** Our Type B **(d)** Our Type C

Fig. 3: Architectural choices. Different design-sketches for controlling a U-Net based generation process with a controlling network. The generation process is in each example on the left-hand side and the control process on the right-hand side. (a) The architecture of ControlNet [88]. (b-c) Three new architectures (Type A-C) proposed in this work. We verify experimentally that model Type B performs better than Type A, and is on a par with Type C. We choose Type B as our final architecture, and call it ControlNet-XS, since it has fewer parameters than Type C.

is that in (a) the controlling model may face an even more challenging task, compared to design (b), for computing the appropriate control signal since it receives far less input from the generative process.

3.3 ControlNet

The ControlNet [88] architecture is sketched in Fig. 3a. It starts with a pre-trained generative model, here the U-Net of Stable Diffusion [64]. The control model copies the encoder of the U-Net and hence has a representation that is capable of generating images by itself. The control encoder receives the control signal, *e.g.* in form of a depth map, as well as the intermediate, noisy generated image. It outputs control signals that are fed into the different decoder blocks of the generative process. The connections from the control model to the generation model are initialized by so-called zero-convolutions, which have the effect that the generative capabilities of the controlled U-Net are not diminished at the beginning of training. During training, the encoder can learn to provide useful control signals to the generative process. The training objective is the same as for Stable Diffusion, *i.e.* image denoising (see Section 3.5). While these may seem like reasonable design choices at first glance, they are sub-optimal from the perspective of a controlling system, as illustrated in Figure 2a and discussed in Section 3.2. In brief, the control model has two jobs at once: i) It has to process the feedback signal in order to make it useful for the generation process; ii) It has to anticipate what the generation process is going to do until the control signal is received by the generation model. We remedy the second job, and hence the control model can focus on the first one.

3.4 ControlNet-XS

In Section 3.2 we have motivated our communication mechanism between controlling and generating network. The key idea is that the two encoders have an interaction with high-frequency. Based on this, we design three variants Fig. 3b-Fig. 3d. They vary in terms of connectivity between the two encoders and the two decoders, respectively. From a feedback-control system perspective, designs (c)

and (d) are good, since they have a high-frequency communication between the two encoder networks. In design (b) there are also many so-called control-loops, however, for each loop the generative network does still uncontrolled processing within the loop. We validate experimentally that model Type B is superior to A and on a par with Type C. Hence, we choose Type B as our final architecture since it has fewer parameters than Type C. We call Type B architecture *ControlNet-XS*. A detailed illustration of the architecture of Type B is in the supplement. The key building block for connecting the generation network and ControlNet-XS is shown in Figure 2c. The calculated features are processed by zero-convolutions and added to the calculated features of the counterpart. Features coming from the generative block can be either added or concatenated to the features of the control block. Because we train the controlling network from scratch we utilise concatenation. Our new design allows to drastically reduce the size of the controlling network, by consistently changing the number of channels in each control layer. We validate experimentally that even a model with as little as 1.7M parameters performs on a par with ControlNet [88] with 361M parameters. Note that in ControlNet, a version of ControlNet with fewer parameters, called ControlNet-light, was evaluated but found to perform inferior.

3.5 Training

As in related works, all weights of the generative model are frozen during training and we only learn the weights of the controlling network. Due to our improved design (Type A-C), we do not need the generative power of the controlling network, and hence all parameters are initialized randomly. We observe that the zero convolutions (see Figure 2c) help to stabilize training. We train a separate controlling network for each kind of guidance. As training data, we use one million images from the Laion-Aesthetics dataset [72]. For getting guidance images, we follow ControlNet [88] and extract canny-edges, use ground truth segmentation maps, predict the depth maps with MiDaS [60], or predict human keypoints with OpenPose [6]. The standard diffusion model objective remains unchanged:

$$\mathcal{L} = \mathbb{E}_{z_0,t,c_t,c_f,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_c)\|_2^2\right], \tag{1}$$

with the target image z_0, the noisy image z_t, the timestep t, the text conditioning c_t and the control conditioning c_c.

4 Experiments

We start by defining the evaluation metrics (Section 4.1), and then analyse variations of our new model in terms of architecture and size (Section 4.2). Section 4.3 is an in-depth, quantitative comparison to state-of-the-art approaches. The next Section 4.4 discusses a semantic bias induced by large control models. Finally, to demonstrate the versatility of our approach, we apply it to a larger generative model, namely Stable Diffusion XL (Section 4.5). If not stated differently, we use Stable Diffusion Version 1.5 as the generative model.

4.1 Evaluation Metrics

We evaluate performance by estimating the fidelity of the control and the quality of the generated images, to ensure that the generated quality does not reduce with respect to uncontrolled image generation. For quality evaluation, we use the CLIP-Score [21] which approximates the similarity between a given text-prompt and an image, and the CLIP-Aesthetics score [72] which approximates the aesthetic appearance of an image as perceived by humans. The fidelity of the control is evaluated implicitly with the Learned Perceptual Image Patch Similarity (LPIPS) [89] and explicitly by a distance measure between two images, which is MSE for depth control, denoted by MSE-depth, mIoU for semantic map control, and the Hausdorff distance (HDD) for canny-edges and human poses (*i.e.* keypoints). Here, the first image is the reference control image, *e.g.* the depth map, and the second image is the extracted control, *e.g.* the depth map from the generated image. The extraction algorithm is the same as the one used to generate the training data, *i.e.* MiDaS [60] for depth extraction. Note that for improved readability, the MSE-depth values are scaled by 10^3 and the Hausdorff distance is scaled by 10^{-1}. We also compute the Fréchet Inception Distance (FID) [22]. We evaluate the semantic map control with the COCO-Stuff [5] validation set of 5000 images and all other controls with the COCO [41] validation dataset of 5000 images.

Note that for pixel-accurate guidance (*i.e.* depth and edges) the FID score measures both quality and fidelity of control since the control signal comes from a target image of the respective COCO validation set and hence the features of the generated image are expected to be similar. However, more loose controls like semantic maps and human poses in particular do not contain precise positional information about the features and hence the FID score only measures quality. The same applies to the similarity metrics LPIPS which is less relevant for guidance with semantic maps and not applicable to human pose guidance.

4.2 Ablation Study: Architecture

We conduct an ablation study in Tab. 1 for the four architectures shown in Fig. 3. The size of our Type A-C are chosen to be about 20% of ControlNet. For two metrics, quality (FID) and control (MSE-depth), all of our architectures are clearly superior to ControlNet. For other metrics, our design is marginally superior or on a par, respectively. Furthermore, Type B performs better than Type A for all measures. This is not surprising, since Type A has no control-signals in the encoder part (see Fig. 3b). The performance of Type B and C are on a par. However, Type C has the drawback of effectively doubling the model size. We explain this lack of quantitative improvement of Type C in a sensitivity analysis in the supplement material. We choose type B as our final architecture for ControlNet-XS and use it in all remaining experiments.

In the next experiment we evaluate whether changing the parameter size of ControlNet-XS influences the performance, see Table 2. We examined ControlNet-XS (*i.e.* Type B) with 491M, 55M, 11.7M and 1.7M parameters, respectively. We roughly see the same trend for all metrics that when varying the sizes of

Table 1: Ablation study for four different architectures illustrated in Fig. 3. We see that all of our designs (Type A to Type C) outperform ControlNet [88] (CN), both in terms of quality (FID) and control (MSE-depth). We select Type B as our ControlNet-XS architecture for all remaining experiments, since it performs best, on average, and has fewer parameters than Type C.

		Both	Control		Quality	
	Method	FID ↓	MSE-d ↓	LPIPS ↓	CLIP-Sc ↑	CLIP-Ae ↑
Depth	CN (361M)	19.01	29.1	0.532	28.96	6.08
	Our Type A (53M)	17.11	20.9	0.492	29.00	6.02
	Our Type B (55M)	16.36	**19.6**	**0.468**	**29.21**	6.09
	Our Type C (117M)	**16.24**	20.2	0.476	29.14	**6.10**

Table 2: Ablation study of ControlNet-XS (Type B) in terms of size. On average, we see that the performance increases slightly from 491M to 55M and decreases afterwards for smaller model sizes, up to 1.7M. Please see discussion in Section 4.2.

		Both	Control		Quality	
	Method	FID ↓	MSE-d ↓	LPIPS ↓	CLIP-Sc ↑	CLIP-Ae ↑
	Stable Diffusion	22.69	(69.7)	(0.618)	28.40	6.16
Depth	CN-XS (491M)	16.91	21.4	0.487	29.09	6.07
	CN-XS (55M)	**16.36**	**19.6**	**0.468**	**29.21**	6.09
	CN-XS (11.7M)	17.90	28.6	0.525	28.83	6.10
	CN-XS (1.7M)	18.45	29.9	0.526	28.73	**6.12**

ControlNet-XS, the performance increases slightly from 491M to 55M and decreases afterwards for smaller model sizes, up to 1.7M. Hence, we choose the 55M model as our best model and show qualitative results in Fig. 1. In terms of control, it means that smaller models have reduced fidelity of the control. We show qualitative results of this effect in Fig. 4. The general decrease in performance for smaller model sizes can be explained as follows. Control models with fewer parameters have less power and hence perform more similarly to the uncontrolled generative model, *i.e.* Stable Diffusion, which performs worse in general. Note that the CLIP-Aesthetic score is highest for Stable Diffusion. Hence, our 1.7M model performs best for this score.

(a) Control **(b)** Original Image **(c)** CN-XS (55M) **(d)** CN-XS (11.7M) **(e)** CN-XS (1.7M)

Fig. 4: The fidelity of the control reduces with smaller model sizes of ControlNet-XS. In the 55M parameter model the complex structure of the street junction is identical to the one in the original image, as well as the skyscrapers in the upper-left corner. Smaller models with 11.7M and 1.7M parameters, respectively, are still guided by the control but less rigorously.

4.3 Quantitative Comparison

Tab. 3 compares our ControlNet-XS with six state-of-the-art control methods. We only use the officially published weights to not bias the comparison. We examined four different guidance types: pixel-accurate depth and canny-edge guidance, more loose semantic map guidance, and very loose human pose guidance. Among all competitors, Uni-ControlNet is on average the best performing

Table 3: Quantitative comparison of seven different approaches. For each guidance type the best method is marked **bold** and the two worst are marked in red. The results for ControlNet [88] with a semantic map guidance are shown in grey brackets since the authors state that the test images were part of the training set.

		Quality + Control	Control		Quality		
	Method	FID ↓	MSE-d ↓	LPIPS ↓	CLIP-Sc ↑	CLIP-Ae ↑	SD
	Stable Diffusion	22.69	-	(0.618)	28.40	6.16	v1.5
Depth	ControlNet (361M)	19.01	29.1	0.532	28.96	6.08	v1.5
	GLIGEN (231M)	19.15	21.2	0.490	29.03	5.81	v1.4
	T2I-Adapter[32] (77M)	20.29	31.4	0.526	28.80	5.98	v1.5
	UniControl (374M)	26.80	19.9	0.487	28.04	5.97	v1.5
	Cocktail (378M)	/	/	/	/	/	/
	Uni-ControlNet (459M)	18.38	26.1	0.524	29.00	5.90	v1.5
	ControlNet-XS (55M)	**16.36**	**19.6**	**0.468**	**29.21**	**6.09**	v1.5
			HDD ↓				
Canny Edges	ControlNet (361M)	21.18	18.52	0.544	29.01	**6.17**	v1.5
	GLIGEN (231M)	27.24	**15.09**	0.446	29.20	5.76	v1.4
	T2I-Adapter (77M)	18.34	16.66	0.459	29.14	5.66	v1.5
	UniControl (374M)	33.12	16.02	**0.416**	28.13	5.79	v1.5
	Cocktail (378M)	/	/	/	/	/	/
	Uni-ControlNet (459M)	17.37	15.94	0.460	29.28	5.87	v1.5
	ControlNet-XS (55M)	**15.13**	15.22	0.417	**29.61**	5.98	v1.5
			mIoU ↑				
Semantic Map	ControlNet (361M)*	(35.35)	(0.32)	(0.590)	(30.22)	(5.85)	v1.5
	GLIGEN (231M)	29.83	0.25	0.608	29.82	5.84	v1.4
	T2I-Adapter (77M)	23.76	0.22	0.613	30.31	5.69	v1.4
	UniControl (374M)	39.26	**0.31**	0.552	27.85	6.02	v1.5
	Cocktail[19] (378M)	26.07	0.19	0.604	31.25	**6.11**	v2.1
	Uni-ControlNet (459M)	22.26	0.25	0.588	31.16	5.88	v1.5
	ControlNet-XS (55M)	**17.70**	**0.31**	**0.519**	**31.26**	5.95	v1.5
			HDD ↓				
Hum-Poses	ControlNet (361M)	23.82	8.58	-	28.14	6.18	v1.5
	GLIGEN (231M)	/	/	-	/	/	/
	T2I-Adapter (77M)	23.16	8.81	-	28.21	6.05	v1.5
	UniControl (374M)	54.63	**8.53**	-	24.72	5.87	v1.5
	Cocktail[19] (378M)	26.44	9.87	-	**28.28**	6.15	v2.1
	Uni-ControlNet (459M)	**22.80**	8.98	-	28.06	5.84	v1.5
	ControlNet-XS (55M)	23.58	8.67	-	28.15	**6.21**	v1.5

method, since it rarely has negative outliers and scores mostly among the top methods. For pixel-accurate depth guidance our ControlNet-XS outperforms all other approaches for all metrics. This includes our baseline model ControlNet. For pixel-accurate canny-edge guidance, ControlNet-XS is either the best or second best performing. It is important to note that there is no clear runner-up method. For instance GLIGEN, which is second best for the HDD score, performs very poorly with respect to quality, *e.g.* FID score. In general, GLIGEN and UniControl exhibit a trade-off between control (HDD/MSE-d, LPIPS) and quality (FID score) since they are not able to consistently exert proper control without diminishing image quality. For guidance with a semantic map, which is a more loose control, ControlNet-XS clearly outperforms all competitors, apart from the CLIP-Aesthetics score, where it ranks third. Especially noticeable is the tremendous gain in FID. Runner-ups are Uni-ControlNet and UniControl, although UniControl performs very poorly with respect to FID score. For the most loose control, *i.e.* keypoint guidance for human poses, there is in general not much control that has to be enforced by the controlling model. All compared

models perform similarly in terms of HDD score and keep the FID score in the vicinity of the score of the unguided base model. The two exceptions appear to be UniControl and Cocktail, which show a major drop in performance for the FID score. In general, the quantitative results for human poses have to be taken with a large grain of salt, given considerably less training data.

Table 4: Comparison of inference and training times of our ControlNet-XS and ControlNet [88], trained to control depth. Inference times are averaged over seven runs and we evaluate for 50 DDIM steps with a batch size of 10. The training time is given in NVIDIA A100 GPU hours.

Method	Inference ↓	Training ↓
ControlNet (361M)	1min 11sec	∼ 500h (A100)
ControlNet-XS (55M)	38sec	∼ 200h (A100)

In summary, we see that our improved communication mechanism plays an important role when it comes to pixel-accurate image guidance (depth and canny-edges) as well as more loose guidance (semantic map). For very loose guidance, such as human pose, the communication mechanism may play a less important role. However, even then our approach performs well across all metrics without any negative outlier. In general, only Uni-ControlNet and our approach show the behaviour of no negative outliers, while all other approaches seem to sometimes trade-off image quality for more accurate control.

Tab. 4 compares inference and training times of ControlNet-XS and ControlNet. For both, we increase the speed by about a factor of 2.

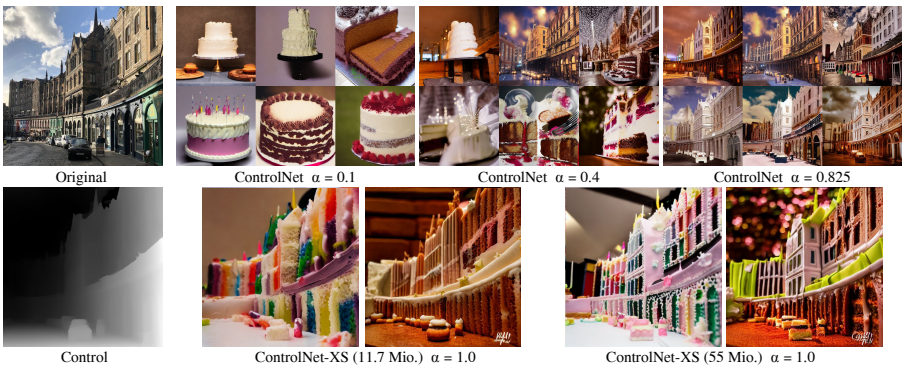


Fig. 5: Semantic bias for depth control. Given the control depth map of a street scene and an unrelated text-prompt: "high quality photo of a delicious cake, 4k image". The large-sized (361M) ControlNet [88] has a semantic bias and is unable to produce a cake scene with the given depth, independent of control strength α. Our small-sized models with 11.7M and 55M respectively mitigate this bias.

4.4 Semantic Bias of Large Control Models

We have seen already that the large-sized ControlNet [88] needs generative power to produce good results. However, this can induce a semantic bias as shown in

Fig. 5. The images are generated with a control depth map of a street scene and an unrelated text-prompt: "high quality photo of a delicious cake, 4k image". Note that these are not contradicting control inputs but the inputs rather challenge the generative process to produce a creative solution with a cake in form of a street scene. We see that ControlNet-XS with 11.7M parameters is able to produce impressive results, followed by results of the 55M model. In contrast, ControlNet [88] is not able to produce satisfying results, even when adjusting the control strength α.[3] Note that $\alpha = 0.825$ is the default for ControlNet. With this default value, ControlNet shows proper house facade textures, while ControlNet-XS shows typical cake textures such as "sponge", "marzipan" or "icing". We conjecture that the reason is a semantic bias induced by large control models. A large control model can use its power to add semantic meaning to input depth maps. This semantic bias cannot be removed by adapting α. Here $\alpha = 0.4$ was the "sweet spot" where ControlNet suddenly transitions from producing images of a cake to images of a street scene. In the supplement, we show that a large-sized ControlNet-XS also has this semantic bias.

4.5 Evaluation with Stable Diffusion XL

We evaluate our ControlNet-XS model with Stable Diffusion XL [53] as generative model. Stable Diffusion XL has about 2.6B parameters and hence is over three times larger than its predecessor Stable Diffusion. We are able to train a ControlNet-XS for depth control which has only 20M parameters, *i.e.* less than 1% of parameters of the generative model. Our model provides good control, *i.e.* MSE-d score is 22.6 in contrast to 123.2 of the uncontrolled Stable Diffusion XL. Furthermore, we achieve high quality results with a low FID score of 18.75. ControlNet-XS is also considerably superior to the T2I-Adapter [47] with an FID score of 61.03 and MSE-d score of 49. A qualitative result is shown in Fig. 1. We refer to the supplement for more results and a discussion.

5 Conclusion and Limitation

We have analyzed existing approaches for controlling pre-trained text-to-image Diffusion Models with respect to their communication mechanism with the generative model. We proposed a new bidirectional, high-frequency and large-bandwidth communication. This led to the development of ControlNet-XS, a small-sized controlling network that outperforms the state-of-the-art for pixel-level image guidance. One major limitation in this field is a missing unifying benchmark with consistent evaluation protocols and ideally a metric that truly represents human judgment. There are many exciting directions for future work. One next step is to integrate our approach into the multi-image guidance approaches, which are based on a feedback-control system, but also approaches in related fields such as video translation [15, 92] and controlled 3D object generation [29, 87] should benefit from our method.

[3] The output signals of the controlling network are added with a global weighting α to the output signals of the generation network at the respective neural blocks. This weighting can be adjusted at test time.

Acknowledgements

We thank Nicolas Bender for his help in conducting experiments. The project has been supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) funded by the German Academic Exchange Service (DAAD). The project has also been supported by the Trilateral DFG Research Program (Germany-France-Japan). The project was also support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

1. Midjourney (2023), https://www.midjourney.com/
2. Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., Dekel, T.: Text2live: Text-driven layered image and video editing. In: European Conference on Computer Vision. pp. 707–723 (2022)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving Image Generation with Better Captions (2023)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
7. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024)
8. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022)
9. Choi, J., Choi, Y., Kim, Y., Kim, J., Yoon, S.: Custom-edit: Text-guided image editing with customized diffusion models. arXiv preprint arXiv:2305.15779 (2023)
10. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., Tang, J.: CogView: Mastering Text-to-Image Generation via Transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan,

J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 19822–19835. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/a4d92e2cd541fca87e4620aba658316d-Paper.pdf
14. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
15. Feng, R., Weng, W., Wang, Y., Yuan, Y., Bao, J., Luo, C., Chen, Z., Guo, B.: Ccedit: Creative and controllable video editing via diffusion models. arXiv preprint arXiv:2309.16496 (2023)
16. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In: Avidan Shai and Brostow, G., Moustapha, C., Maria, F.G., Tal, H. (eds.) Computer Vision – ECCV 2022. pp. 89–106. Springer Nature Switzerland, Cham (2022)
17. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
18. Goel, V., Peruzzo, E., Jiang, Y., Xu, D., Sebe, N., Darrell, T., Wang, Z., Shi, H.: PAIR-Diffusion: Object-Level Image Editing with Structure-and-Appearance Paired Diffusion Models (2023), http://arxiv.org/abs/2303.17546
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
20. He, Y., Salakhutdinov, R., Kolter, J.Z.: Localized Text-to-Image Generation for Free via Cross Attention Control. arXiv preprint arXiv:2306.14636 (2023)
21. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
22. 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. Advances in neural information processing systems **30** (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
25. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (2021)
26. Hu, H., Chan, K.C.K., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W.: Instruct-Imagen: Image generation with multi-modal instruction. arXiv preprint arXiv:2401.01952 (2024)
27. Hu, M., Zheng, J., Liu, D., Zheng, C., Wang, C., Tao, D., Cham, T.J.: Cocktail: Mixing Multi-Modality Control for Text-Conditional Image Generation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
28. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
29. Huang, T., Zeng, Y., Zhang, Z., Xu, W., Xu, H., Xu, S., Lau, R.W.H., Zuo, W.: Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. arXiv preprint arXiv:2312.06439 (2023)
30. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

31. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
32. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
33. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34** (2021)
34. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
35. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
36. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Advances in neural information processing systems **34**, 21696–21707 (2021)
37. Li, T., Ku, M., Wei, C., Chen, W.: DreamEdit: Subject-driven Image Editing. arXiv preprint arXiv:2306.12624 (2023)
38. Li, W., Xu, X., Liu, J., Xiao, X.: UNIMO-G: Unified Image Generation through Multimodal Conditional Diffusion. arXiv preprint arXiv:2401.13388 (2024)
39. Li, Y., Mao, H., Girshick, R., He, K.: Exploring Plain Vision Transformer Backbones for Object Detection. In: Computer Vision – ECCV 2022, pp. 280–296. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-20077-9_17, http://dx.doi.org/10.1007/978-3-031-20077-9_17
40. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
41. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
42. Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L.: Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4823–4833 (2021)
43. Lu, C., Xia, M., Qian, M., Chen, B.: Dual-branch network for cloud and cloud shadow segmentation. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2022)
44. Lukovnikov, D., Fischer, A.: Layout-to-Image Generation with Localized Descriptions using ControlNet with Cross-Attention Control. arXiv preprint arXiv:2402.13404 (2024)
45. Mao1, Y., Mathias, L., Hou, R., Almahairi, A., Ma, H., Han, J., Yih, W.t., Khabsa, M.: UNIPELT: A Unified Framework for Parameter-Efficient Language Model Tuning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers. vol. 1, pp. 6253–6264. ACL (2022)
46. McCloskey, M., Cohen, N.J.: Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. vol. 24, pp. 109–165. Academic Press (1989). https://doi.org/https://doi.org/10.1016/S0079-7421(08)60536-8, https://www.sciencedirect.com/science/article/pii/S0079742108605368

47. Mou, C., Wang, X., Xie, L., Wu, Y., 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)
48. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4500–4509 (2018)
49. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
50. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
51. Patel, M., Jung, S., Baral, C., Yang, Y.: λ-ECLIPSE: Multi-Concept Personalized Text-to-Image Diffusion Models by Leveraging CLIP Latent Space. arXiv preprint arXiv:2402.05195 (2024)
52. Pfeiffer, J., Kamath, A., Ruckl, A., Cho, K., Gurevych1, I.: AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. pp. 487–503. ACL (2021)
53. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis (2023)
54. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S.: UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. arXiv preprint arXiv:2305.11147 (2023)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Others: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
56. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)
57. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
58. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents (2022)
59. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-Shot Text-to-Image Generation. CoRR **abs/2102.1** (2021), `https://arxiv.org/abs/2102.12092`
60. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
61. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Efficient Parametrization of Multi-Domain Deep Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
62. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International conference on machine learning. pp. 1060–1069. PMLR (2016)
63. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Advances in neural information processing systems **29** (2016)

64. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022), https://github.com/CompVis/latent-diffusion
65. Rosenfeld, A., Tsotsos, J.K.: Incremental Learning Through Deep Adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(3), 651–663 (2020). https://doi.org/10.1109/TPAMI.2018.2884462
66. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
67. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
68. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-Image Diffusion Models. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. ACM (2022). https://doi.org/10.1145/3528233.3530757, http://dx.doi.org/10.1145/3528233.3530757
69. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf
70. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515 (2023)
71. Sauer, A., Schwarz, K., Geiger, A.: StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022). https://doi.org/10.1145/3528233.3530738
72. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
73. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
74. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
75. Stickland, A.C., Murray, I.: BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 5986–5995. PMLR (2019), https://proceedings.mlr.press/v97/stickland19a.html
76. Tang, H., Bai, S., Zhang, L., Torr, P.H., Sebe, N.: Xinggan for person image generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 717–734. Springer (2020)

77. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: Df-gan: A simple and effective baseline for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16515–16525 (2022)
78. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
79. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pre-training is All You Need for Image-to-Image Translation (2022)
80. Wang, W., Guo, R., Tian, Y., Yang, W.: Cfsnet: Toward a controllable feature space for image restoration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4140–4149 (2019)
81. Xia, Y., Monica, J., Chao, W.L., Hariharan, B., Weinberger, K.Q., Campbell, M.: Image-to-Image Translation for Autonomous Driving from Coarsely-Aligned Image Pairs. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7756–7762. IEEE (2023)
82. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1316–1324 (2018)
83. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7754–7765 (2023)
84. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
85. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Others: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
86. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 5907–5915 (2017)
87. Zhang, H., Chen, B., Yang, H., Qu, L., Wang, X., Chen, L., Long, C., Zhu, F., Du, K., Zheng, M.: Avatarverse: High-quality & stable 3d avatar creation from text and pose. arXiv preprint arXiv:2308.03610 (2023)
88. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (2023)
89. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
90. 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. Advances in Neural Information Processing Systems **36** (2024)
91. Zhao, Y., Peng, L., Yang, Y., Luo, Z., Li, H., Chen, Y., Zhao, W., Liu, W., Wu, B.: Local Conditional Controlling for Text-to-Image Diffusion Models. arXiv preprint arXiv:2312.08768 (2023)
92. Zhao, Y., Xie, E., Hong, L., Li, Z., Lee, G.H.: Make-A-Protagonist: Generic Video Editing with An Ensemble of Experts. arXiv preprint arXiv:2305.08850 (2023)

93. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
94. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5802–5810 (2019)