OmniControlNet: Dual-stage Integration for Conditional Image Generation

Yilin Wang[*,1] Haiyang Xu[*,2,4] Xiang Zhang[4] Zeyuan Chen[4]
Zhizhou Sha[1] Zirui Wang[3] Zhuowen Tu[4]
[1] Tsinghua University [2] University of Science and Technology of China
[3] Princeton University [4] University of California, San Diego

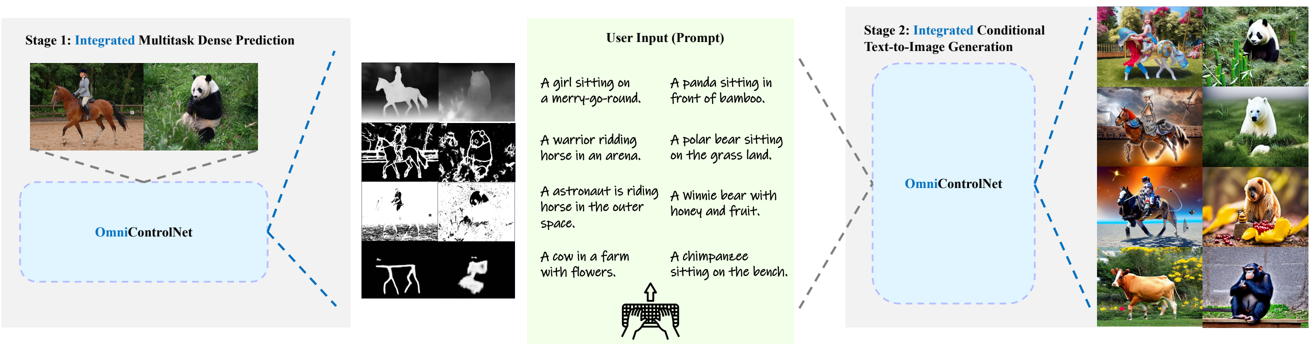


Figure 1. **Given an input image, our single, integrated OmniControlNet** extracts its control features and generates high-quality images. From the first to the last row in the middle, the feature visualization represents Depth, HED, Scribble, and Animal Pose respectively.

Abstract

We provide a two-way integration for the widely adopted ControlNet by integrating external condition generation algorithms into a single dense prediction method and incorporating its individually trained image generation processes into a single model. Despite its tremendous success, the ControlNet of a two-stage pipeline bears limitations in being not self-contained (e.g. calls the external condition generation algorithms) with a large model redundancy (separately trained models for different types of conditioning inputs). Our proposed OmniControlNet consolidates 1) the condition generation (e.g., HED edges, depth maps, user scribble, and animal pose) by a single multi-tasking dense prediction algorithm under the task embedding guidance and 2) the image generation process for different conditioning types under the textual embedding guidance. OmniControlNet achieves significantly reduced model complexity and redundancy while capable of producing images of comparable quality for conditioned text-to-image generation.

1. Introduction

The exploding development of diffusion [35, 93, 94] based text-to-image generators [66, 71, 81, 83, 85] has led to a recent wave of generative model progressing beyond traditional models such as VAE [44] and GAN [27, 98].

The ControlNet [116] further promotes the popularity of text-to-image generation by introducing additional user controls as the conditioning input available in a myriad of forms including edges [9, 107], line segments [28], human pose [11], normal map [100], depth map [73], segmentation map [118], and user scribble. With the additional image-level input beyond the text prompts, ControlNet can greatly expand the scope of application domains for text-to-image generation to real-world workflows in various areas, including design, architecture, gaming, art, manufacturing, animation, and human-computer interaction.

ControlNet [116] is a two-stage pipeline comprising 1) a condition generation stage and 2) a text-to-image generation stage conditioned on the output from the first stage. Despite the great success ControlNet has achieved, it still suffers from the issue of large model redundancy in two means: 1) in stage 1, a specific external algorithm is executed to create each type of image-level condition, and 2) in stage 2, a separate diffusion model is trained for each type of conditional input. Fig. 3 gives an schematic illustration for the ControlNet method [116].

In this paper, we aim to alleviate the algorithm and model redundancy problem in ControlNet [116] by proposing OmniControlNet, which provides a dual-stage integration. That is, in stage 1, instead of calling the external algorithms,

* equal contribution. Work done during the internship of Yilin Wang, Haiyang Xu, Zhizhou Sha, and Zirui Wang at UC San Diego.

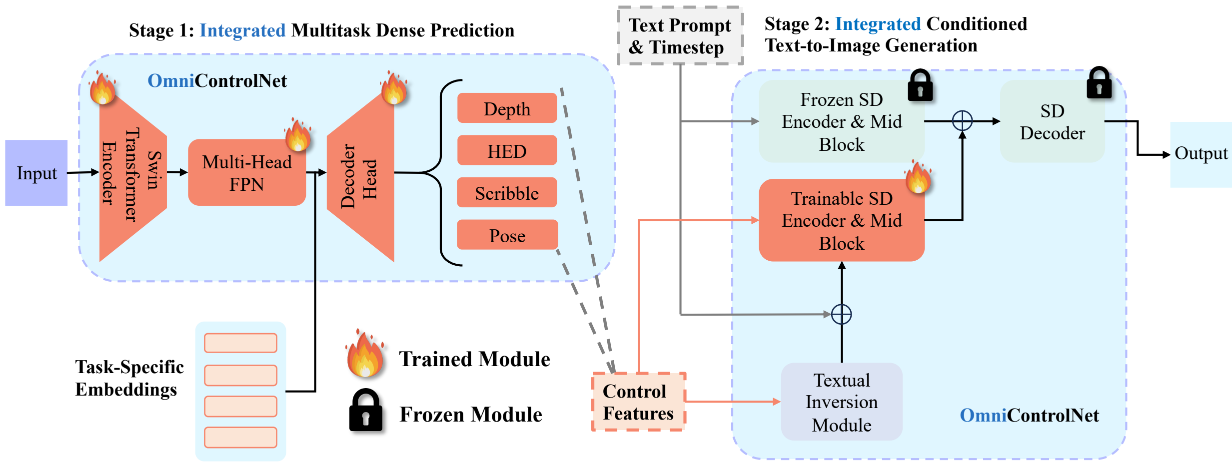


Figure 2. **Our OmniControlNet model.** From condition generation to image synthesis, while the ControlNet model has to deal with all the features separately, our model can handle the tasks within an integrated pipeline.

we develop an integrated dense image prediction method to perform edge detection, depth map generation, animal pose estimation, and scribble generation in a single multi-tasking framework under the guidance of task prompts; in stage 2, instead of training separate image generation models for different conditioning input types, we train a single model for four kinds of image-level conditional control under the textual inversion guidance. We observe a large model, parameter, and memory redundancy reduction, compared with the existing approaches, while being able to generate comparable image quality. The contribution of our work can be summarized as follows.

The contribution of our paper is summarized as follows:

- We develop a new module to integrate four dense image prediction tasks, including edge detection, depth estimation, scribble segmentation, and animal pose estimation, under the task embedding guidance.
- We develop a new module to perform conditioned text-to-image generation that integrates four different types of conditional input under the textual inversion guidance.
- Combining the above two modules yields OmniControlNet, which greatly reduces algorithm complexity for conditional text-to-image generation. OmniControlNet points to a promising direction for condition text-to-image generation under an integrated pipeline.

2. Related Works

2.1. Text-to-Image Generation

The task of text-to-image generation [18, 53, 71, 114] is to generate an image matching the provided text prompts using deep learning models. Before the wide use of diffusion models, the task was primarily achieved by GAN [27] based models [78, 110, 115]. The work *Generative Adversarial Text to Image Synthesis* [78] applied an encoder to encode the texts and concatenated the encoded features to the image features before inserting them into the GAN model, which was among the first works to tackle the task. After the introduction of diffusion models [35, 94], lots of diffusion-based models appeared [4, 8, 13, 23, 29, 32, 33, 38, 57, 61, 90, 99, 102, 103], which mainly used cross attention to combine the image and text features in the UNet [82] backbone. DALLE-2 [72] and Stable Diffusion [81] are among the outstanding literature in the field. Many works, including T2I-Adapter [62], ControlNet [116] our OmniControlNet model, are based on the Stable Diffusion model.

2.2. Image-to-Image Generative Model

Image-to-image generation involves transferring an image from one domain to another. For example, in ControlNet [116], additional features provided as images are fed into the model to generate the required images. Before the widespread use of diffusion models, GAN-based models [27] such as [1, 15, 25, 41, 42, 64, 65, 80, 104, 119, 120] and Transformer-based models [21, 71, 101] were commonly adopted. CycleGAN [119] was one of the foremost models for image-to-image transfer, utilizing a GAN-based approach for style transfer with cycle consistency. With the introduction of diffusion models [35, 94], many [12, 84, 95] have demonstrated the significant potential of diffusion models in this task. Recently, several works [37, 51, 62, 68, 116, 117] have combined text and image conditions within diffusion models, enabling the generation of high-quality images. ControlNet is a notable example, taking text prompts and additional features as constraints to guide image generation.

2.3. Condition Generation

ControlNet has demonstrated its performance in conditional image generation across various conditions, including Depth Map [73], Canny Edge [9, 107], OpenPose [11], Normal Map [100], User Scribble, and Segmentation [64], *etc*. In this section, we delve into four representative tasks: Depth Map, HED Edge, User Scribble, and Animal Pose, along with the expert models associated with each.

Generating depth maps to represent relative distances is a fundamental challenge in computer vision and 3D scene understanding tasks. Numerous methods have been proposed, ranging from traditional stereo matching algorithms [47, 86] to deep learning-based approaches [24, 48, 56, 109]. We use MiDAS [7, 76] as our expert model, which exhibits exceptional performance and generalization capabilities.

Image edge detection plays a major role in tasks such as object segmentation and visual salience. Early methods [3, 45, 46, 59, 60] relied on manual design for edge detection. However, with the advent of deep learning, learning-based methods [36, 67, 88, 96, 105] have demonstrated great potential in handling edge detection tasks. A classic benchmark in this field is Holistically-Nested Edge Detection (HED) [107], and we take it as our expert model.

User scribbles serve as user-defined guidance for image generation tasks, enabling users to convey their intentions and preferences to the generative model. In ControlNet, this involves a simple mapping of pixels with values greater than 127 to 255, and the rest to 0 in an image.

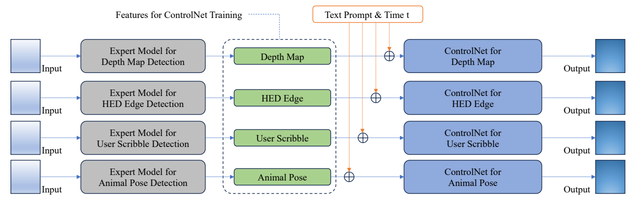
Generating pose maps, which encode spatial information about the arrangement of objects or characters in images, is crucial for tasks like image-to-image translation, particularly in human or object pose manipulation. Human pose estimation models [10, 63, 97, 106, 111] are designed to describe human skeletons. Notable benchmarks in this domain include PoseNet [43] and OpenPose [11]. In terms of animal pose estimation, which often presents more diversity and challenges than human pose estimation, datasets like AP-10K [113] and APT-36K [112] are considered mainstream references.

3. Background

3.1. ControlNet

The ControlNet model [116] presents an efficient framework for fine-tuning the Stable Diffusion model [81]. It introduces an additional control feature (*e.g.* depth map or edge detection) to the generative process, ensuring that the generated images adhere to both the textual prompt and the control condition. In our approach, the weights of the Stable Diffusion model (SD-v1.5) are fixed, while a trainable duplicate of the weights from the 12-layer U-Net encoder and middle block is created. The additional features are integrated into this trainable duplicate via a zero-convolution layer (a 1×1 convolution layer with all-zero initial weights).

We denote the encoder in the frozen part as $\mathcal{E}$, the encoder of the trainable copy as $\mathcal{E}'$, the middle block and the



* ControlNet for <*feature*>is similar to our OmniControlNet's stage 2 in Fig. 2, except that there's no input from the textual embedding module.

Figure 3. **Original ControlNet [116] model.** For different features, we have to use different expert models for condition generation, and we have to train ControlNet on each of the features.

decoder of the frozen part as $\mathcal{M}$ and $\mathcal{D}$, respectively. Let the CLIP-encoded additional feature be c_f, the input of the model as z, time as t, and the CLIP-encoded text prompt as c_t. With $\mathcal{Z}_1, \mathcal{Z}_2$ representing two trainable zero convolution layers, the output of the trainable copy should be $\mathcal{E}'(\mathcal{Z}_1(c_f)+z,t,c_t)$. Consequently, the output of the model, ϵ_{pred}, which also estimates the noise in the denoising process, should be

$$\epsilon_{pred} = \mathcal{D}(\mathcal{M}(\mathcal{E}(z,t,c_t)+\mathcal{Z}_2(\mathcal{E}'(\mathcal{Z}_1(c_f)+z,t,c_t)))) \quad (1)$$

During training, suppose the noise of a diffusion step be ϵ, then the training loss should be

$$\mathcal{L}_{\text{diff}} = \|\epsilon - \epsilon_{\text{pred}}\|_2^2 \quad (2)$$

4. Our Method

4.1. Stage 1: Multi-task Dense Image Prediction

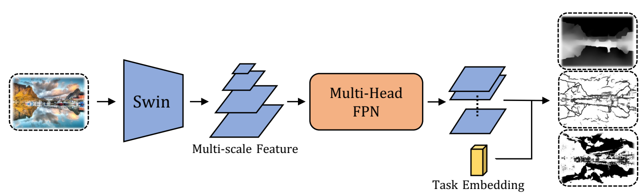


Figure 4. **An overview of our multi-task dense image prediction pipeline.** First, we leverage a Swin Transformer to extract multi-scale features and propose a multi-head FPN to get full-resolution feature maps. Finally, we utilize task-specific embeddings to decode dense predictions from the feature maps.

As depicted in Fig. 4, our multi-task dense image prediction model is architecturally divided into three components: a backbone structure, a Multi-Head Feature Pyramid Network (FPN) [55], and a Decoder Head.

Initially, we employ a pre-trained Swin Transformer [58] to extract multi-scale image features. Considering the resolution of the input image as $1\times$, the extracted features at each stage correspond to resolutions of $\frac{1}{4}\times$, $\frac{1}{8}\times$, $\frac{1}{16}\times$, and $\frac{1}{32}\times$, with a uniform feature channel count of 256.

Subsequently, a Multi-Head FPN is employed to harness rich semantic information from these multi-scale features. To foster feature diversity across various task types, the FPN is structured in a parallel configuration with $\mathbf{m}$ distinct heads, each representing a variant of the original FPN architecture. Specifically, each FPN head undergoes an additional transposed convolution layer to upscale the resolution to $1\times$ while simultaneously reducing the channel dimension to $\mathbf{C}$. The concatenated outputs of all $\mathbf{m}$ heads yield a comprehensive, full-resolution multi-task output feature with channel dimension $\mathbf{mC}$.

In the final stage, task-specific embedding is leveraged to decode the target condition from the aforementioned output. The flexibility in the type of task embedding is noteworthy; both one-hot and clip text embeddings derived from the task name are effective. We employ a Multilayer Perceptron (MLP) to project the task embedding into a latent space with an embedding dimension of $\mathbf{mC}$, subsequently unsqueezing the channel dimension to 1. A cross-product operation is then executed between the output of the Multi-Head FPN and the encoded task embedding, culminating in the decoder output, followed by a Sigmoid.

4.2. Stage 2: Conditioned T2I Generation

Fig. 5 provides an overview of our conditioned text-to-image generation (stage 2) pipeline.

For different tasks, such as depth map or hed edge as an additional feature, we initially apply the textual inversion [26], using 16 random images for each feature to learn the corresponding new "words" (represented by forms such as <depth> or <hed>). Subsequently, we add these new "words" into the CLIP [34] embedding space so that when they are used in text prompts, the CLIP encoder can recognize their specific meanings.

After acquiring these new embeddings, we adapt the prompts for each (prompt, feature, image) triplet. For instance, if the feature for a given triplet is the depth map of the image and the original prompt is "a motorcycle in front of a tree", the revised prompt would be "Use <depth> as a feature, a motorcycle in front of a tree". The modified triplets are fed into the trainable copy, while the corresponding original triplets are fed into the frozen part. Following this, the model is trained with a methodology similar to ControlNet, where the triplets are fed into the model undifferentiated, without separating them by features.

Tab. 1 provides the comparison of the model size as well as the data scale when compared to other integrated models, including UniControl [68], and Uni-ControlNet [117], and our model demonstrates several advantages.

When compared to UniControl, our model, following the structure of ControlNet, requires no additional parameters. In contrast, UniControl incorporates an additional mixture-of-experts (MoE) module, resulting in a substantially larger model (20M more parameters than other models, including ControlNet, Uni-ControlNet, and our model). During training, an increase of 1 in batch size leads to a ∼3 Gigabytes increase in GPU memory usage.

	Extra Parameters	**Extra Data**
Uni-ControlNet	0	$\times n$
UniControl	20M	None
Ours	0	None

*n refers to the number of datasets we combine. In our work, $n = 2$.

Table 1. Comparison of parameters and data scale between OmniControlNet and competing works. *Extra Parameters* refers to the number of extra parameters compared to the original ControlNet, while *Extra Data* refers to the increased amount of data during training. Uni-ControlNet needs to fill the blanks of the mixed datasets with black images, which will double the scale of the data.

In contrast to Uni-ControlNet, our model does not need to perform channel-wise concatenation of multiple additional features. In our configuration, different features originate from varying sets of images. Whereas for Uni-ControlNet, when an image provides a feature such as a depth map but lacks another (*e.g.* animal pose), the corresponding channels for the animal pose are filled with zeros, yielding a larger data scale.

4.3. Textual Inversion Module

Textual Inversion [26] is an approach for extracting and defining new concepts from a few example images, which is the inversion process of text-to-image generation. This method creates new "words" or tokens in the embedding space of the text encoder within the text-to-image generation pipeline, such as Stable Diffusion [81]. Once established, these unique tokens can be integrated into textual prompts, allowing for precise control over the characteristics of the images produced.

We leverage Stable Diffusion as our base model. For the set of images provided, the prompt is set to s = "an image of $<w>$", while the embedded feature v of the "word" $<w>$is our target. For the frozen SD model, suppose c is the encoded feature of s, then we can express $c = c(v)$, as c is determined by v. Therefore, the optimization goal should be

$$v^* = \arg\min_v \mathbb{E}_{z\sim\varepsilon(x),\epsilon\sim\mathcal{N}(0,1),c(v),t}||\epsilon - \epsilon_\theta(z_t, t, c(v))||_2^2.$$

where θ is the weight of the UNet in the SD model and is frozen, and therefore we can directly simulate v in this approach.

4.4. The Whole Integrated Model

Initially, we train the multi-task dense image prediction (stage 1) model, which can generate various features with a

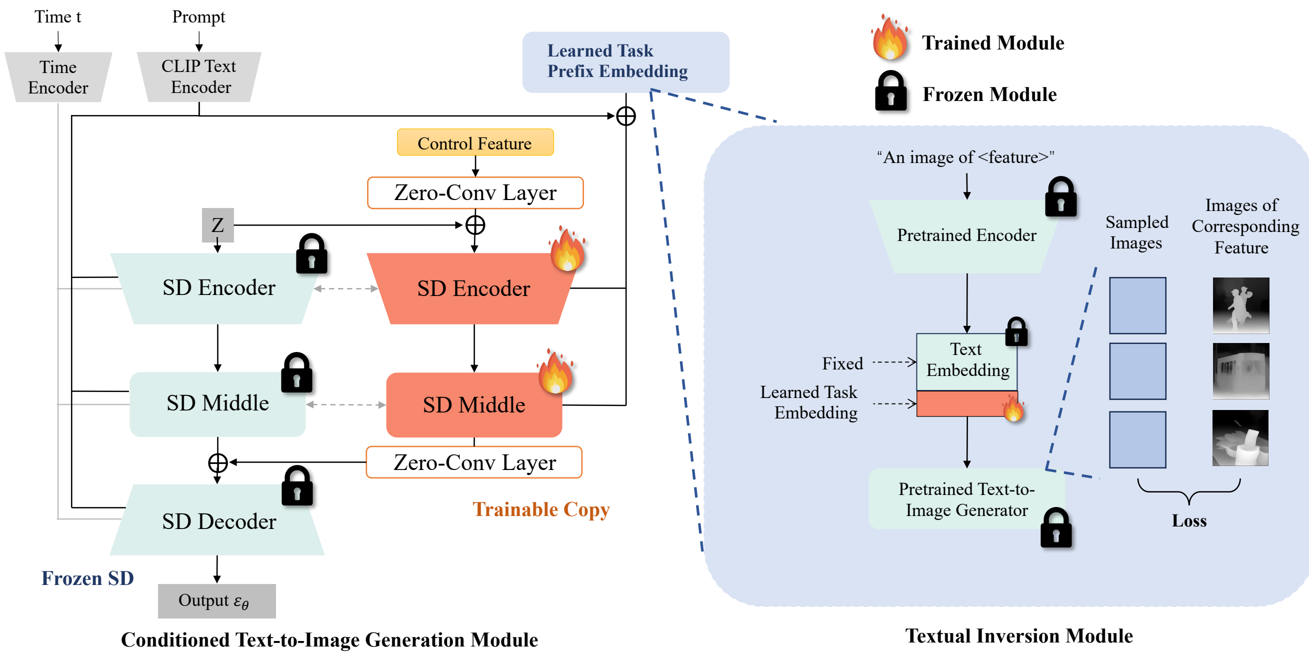


Figure 5. **An overview of our conditioned text-to-image generation pipeline.** Beginning with the original ControlNet structure [116], we utilize the textual inversion to learn task embeddings. Subsequently, we append the prefix *use <feature> as feature* to the prompt and feed the result into the trainable copy. The left side of the figure provides an overview of the conditioned text-to-image generation model, while the right side illustrates the process of learning the CLIP embedding for the new "word" with textual inversion [26].

single model. Subsequently, the samples generated by the Stage 1 model serve as the training data for the conditioned text-to-image generation (stage 2) model. During inference, images are input into the Stage 1 model, whose output is then forwarded to the Stage 2 model for further processing. By utilizing this stage 1 model, we can directly sample different features from a single model without needing multiple expert models. Then, we can use these sampled features to generate images that share similar features with the original one but with specified semantic meanings. Fig. 2 shows the structure of the whole pipeline.

5. Experiments

For our OmniControlNet and the competing works, we perform training and inference on 4 tasks, including Depth, HED, Scribble, and Animal Pose.

5.1. Implementation Details

5.1.1 Datasets

Training. The dataset for both multi-task dense image prediction (stage 1) and conditioned text-to-image generation (stage 2) training consists of 2 different parts. Features depth map, HED edge, and user scribble are from the first part, while the feature animal pose is from the second part. In the first part, we first use YOLOv5 [77] model to detect all the humans in the images from the Laion-5B [87] dataset and choose the first 50,000 images that consist at most 1 human. We directly sample user scribbles from the images, employ an HED boundary detection model [108] to generate HED edges, and use the Midas depth detector [75] to produce depth maps. The captions of the images are taken from the origin Laion-5B dataset. In the second part, we utilize the AP-10K dataset [113] and use the MMPose [16] model to generate the animal poses of the animals. The captions are generated by the BLIP2 [50] model. In order to make the 2 parts contain approximately the same number of images, we duplicate each image in the second part 5 times.

Sampling and Testing. For the features depth map, HED edge, and user scribble, we utilize the validation split of the COCO2017 [54] dataset and obtain the corresponding feature in the same way as the training set. We use the first caption for each image in the dataset. For the animal pose, we utilize the APT-36K dataset [112] and choose the first image from each frame as the dataset. We sample the animal poses the same way as the training set and use the BLIP2 [50] model to perform the image captioning.

5.1.2 Training Details

For our multi-task dense image prediction (stage 1) model, we assign distinct loss functions and associated weights for four different conditions. The depth map generation utilizes L1 loss, while binary cross-entropy loss is

Method	FID Score				CLIP_t Similarity			
	Depth ↓	HED ↓	Scribble ↓	Animal Pose ↓	Depth ↑	HED ↑	Scribble ↑	Animal Pose ↑
Disunified Model								
T2I-Adapter [62]	20.85	18.31	19.79	45.56	0.3099	0.3072	0.3094	0.3327
ControlNet [116]	24.24	24.33	21.97	57.14	0.3076	0.2760	0.3091	0.3160
Unified Stage 2								
Uni-ControlNet [117]	33.71	28.56	30.24	**47.71**	0.3011	**0.3072**	0.3028	**0.3321**
UniControl [68]	25.34	**21.03**	25.82	54.10	0.3020	0.3006	**0.3043**	0.3105
Ours	**23.20**	27.26	**25.79**	53.28	**0.3055**	0.2988	0.3002	0.3292
Unified Stage 1 + 2								
Ours	34.86	36.57	36.63	51.10	0.3024	0.2971	0.2971	0.3269

Table 2. **Quantitative results** of our model, including single stage 2 (conditioned text-to-image generation) model and integrated stage 1 (multi-task dense image prediction) + integrated stage 2 (conditioned text-to-image generation) models. Although methods that utilize different models (T2I-Adapter and ControlNet) tend to perform better, our framework demonstrates competitive results among the integrated models. The numbers in bold indicate the best performance among the integrated methods. The **bold** numbers represent the best score among integrated methods.

employed for the other three scenarios. The assigned loss weights for depth, HED edge, user scribble, and pose are 0.5, 1, 5, and 5, respectively. We resize all the images to 512×512 and take a batch size 16. The model employs an SGD Optimizer with an initial learning rate of 1e-6, which subsequently decreases to 9e-7 following a polynomial decay pattern after 120k iterations. The entire training process takes about 20 hours on 8 NVIDIA RTX 3090 GPUs.

For the textual inversion module, each of the new "word" of a corresponding feature is trained on 8 NVIDIA RTX 3090 GPUs for about 1 hour.

For our conditioned text-to-image generation (stage 2) model, the number of DDIM diffusion steps is set to 50. We adopt the AdamW optimizer and set the learning rate to 1e-5. We train the model on 8 NVIDIA RTX 3090 GPUs with batch size 2 for 50,000 iterations (4 epochs), which takes about 40 hours.

5.1.3 Evaluation Metrics

For our multi-task dense image prediction (stage 1) model, various metrics are adopted to evaluate different aspects of the model's performance. For depth estimation, the Root Mean Square Error (RMSE) is utilized. For edge detection, three distinct metrics are adopted: the fixed contour threshold (ODS), per-image best threshold (OIS), and average precision (AP). The ODS is a metric that evaluates edge detection performance by considering a fixed threshold value across all images, thereby providing a universal performance measure. On the other hand, OIS varies the threshold for each image to find the optimal threshold for that particular image, offering a more adaptive measure of performance. Lastly, AP is a commonly used metric in edge detection tasks. It computes the average precision value for recall values over the interval [0, 1].

For our conditioned text-to-image generation (stage 2) model and the integrated model, we adopt FID score [69] and CLIP_t [34] similarity score as our metrics. For the FID score, we utilize a widely used inception model to measure the similarity between synthesized and real images. For the CLIP_t similarity score, for each pair of generated image and corresponding caption, we use ViT-B/32 [20] CLIP to encode them, and calculate the inner product of them as the CLIP_t similarity score. We report the average of the inner products of all the image-caption pairs.

5.2. Experiment Results

Fig. 1 and Fig. 6 display the visual results for both the multi-task dense image prediction (stage 1), the conditioned text-to-image generation (stage 2), and the combined model. According to the figure, it is evident that the models from both stages and the combined one can generate high-quality results.

Stage 1: Integrated Dense Prediction. To demonstrate the ability of our stage 1 model, we show the result on the depth benchmark NYUDv2 [17] and the HED benchmark BSDS500 [3].

For depth estimation, we compare our result with $\text{DPT}_{\text{hybrid}}$'s contemporary work, including DeepLabv3+ [30], RelativeDepth [49], ACAN [14], ShapeNet [70] and $\text{DPT}_{\text{hybrid}}$ [74]. As shown in Tab. 3, our result outperforms all the models except for $\text{DPT}_{\text{hybrid}}$.

For edge detection, we compare with classic methods including [2, 5, 6, 9, 19, 22, 31, 39, 40, 52, 79, 88, 91, 92, 107]. As illustrated in Tab. 4, our model surpasses all the models except for HED [107].

Stage 2: Integrated Conditioned Text-to-Image Gener-

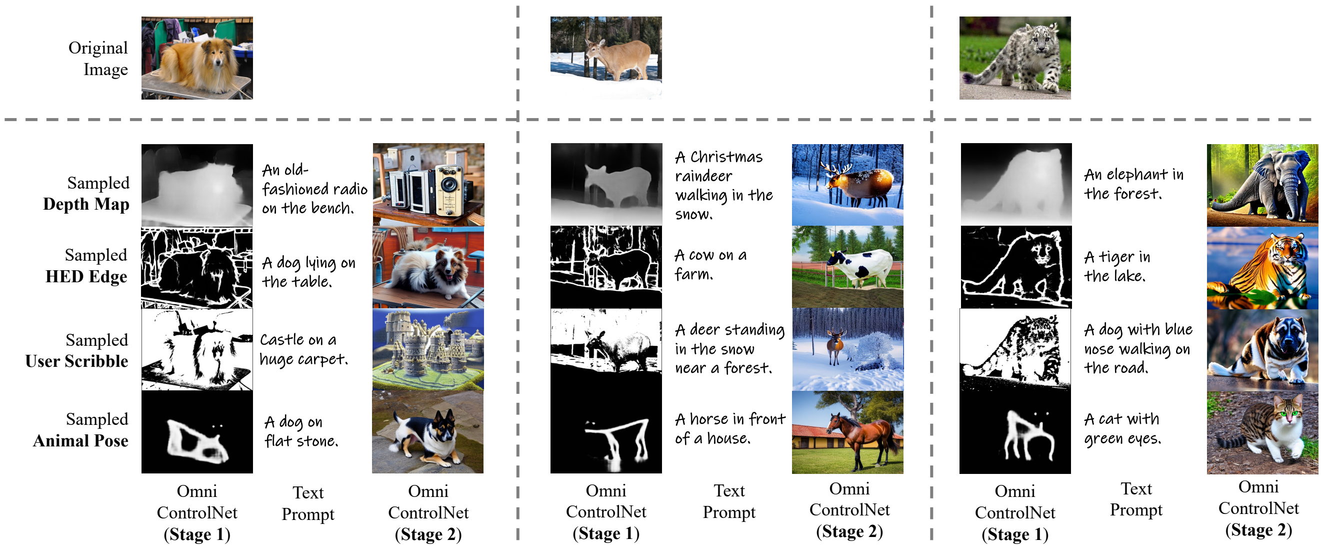


Figure 6. Features and images generated by our OmniControlNet model.

Method	RMSE ↓
DeepLabv3+ [30]	0.575
RelativeDepth [49]	0.538
ACAN [14]	0.496
ShapeNet [70]	0.496
DPT_{hybrid} [74]	0.357
Ours	0.472

Table 3. Depth performance of our multi-task dense image prediction (stage 1) model. Our model utilizes the output of DPT_{hybrid} as the training data; therefore, it is acceptable for surpassing all other methods except for DPT_{hybrid}.

ation. We compare the quantitative results on the metrics FID score and $CLIP_t$ similarity score with other methods, including ControlNet [116], T2I-Adapter [62], Uni-ControlNet [117] and UniControl [68]. The latter two methods build an integrated pipeline that can use a single model to generate images with different additional features, while for the first two methods, a new model must be trained for each different additional feature.

Tab. 2 presents the numerical results for the FID score and the $CLIP_t$ similarity score across various additional features and methods. Although methods that utilize different expert models for different features perform better, our method ranks among the best-performing methods within the category of integrated models.

Integrated Model Results. In the integrated model, similar to the stage 2 model, we once again compare the quantitative results using metrics such as FID score and $CLIP_t$ similarity score with methods including T2I-Adapter [62], ControlNet [116], UniControl [68], and Uni-ControlNet [117]. The quantitative results are presented in Tab. 2. It can be observed that although the overall performance of the integrated model is slightly inferior to methods directly utilizing features from multiple expert models, it still manages to generate images of promising quality.

Method	ODS ↑	OIS ↑	AP ↑
Canny [9]	0.600	0.640	0.580
Felz-Hutt [22]	0.610	0.640	0.560
gPb-owt-ucm [2]	0.726	0.757	0.696
SCG [79]	0.739	0.758	0.773
Sketch Tokens [52]	0.727	0.746	0.780
PMI [40]	0.741	0.769	0.799
SE [19]	0.746	0.767	0.803
OEF [31]	0.746	0.770	0.820
MES [92]	0.756	0.776	0.756
DeepEdge [5]	0.753	0.772	0.807
CSCNN [39]	0.756	0.775	0.798
MSC [91]	0.756	0.776	0.787
DeepContour [89]	0.757	0.776	0.800
HFL [6]	0.767	0.788	0.795
HED [107]	0.788	0.808	0.840
Ours	0.761	0.782	0.811

Table 4. HED performance of our multi-task dense image prediction (stage 1) model. For the three metrics, ODS, OIS, and AP, the larger the number, the better the performance. We can see that our method achieves competitive performance.

6. Ablation Studies

To demonstrate the effectiveness of our model, OmniControlNet, and to reveal the impacts of certain structural designs, we conducted several ablation studies: **1)** Injecting learned task prefix embedding into different parts of the conditioned text-to-image generation module; **2)** Learning weights of the zero-convolution layers with an MLP while the model is trained with the learned task prefix embedding; and **3)** Comparing different encoding methods and the number of heads in the multi-head Feature Pyramid Network. For **1)** and **2)**, we report the results based on our unified stage 2 setting. For **3)**, we report the results based on our unified (stage 1 + stage 2) setting.

6.1. Prefix Injection

In our original framework, only the text prompts fed into the trainable copy of the SD model contain prefixes such as "Use <depth> as feature." In this ablation study, we added the prefix to both parts of the model. The results are shown in Tab. 5. We observe that adding the prefix only to the trainable part yields better results.

	FID Scores			
Method	**Depth ↓**	**HED ↓**	**Scribble ↓**	**Animal Pose ↓**
Prefixes in both parts	80.17	91.73	58.08	172.29
OmniControlNet (Ours)	**23.20**	**27.26**	**25.79**	**53.28**

	CLIP$_t$ Similarity Score			
Method	**Depth ↑**	**HED ↑**	**Scribble ↑**	**Animal Pose ↑**
Prefixes in both parts	0.2321	0.2404	0.2676	0.1843
OmniControlNet (Ours)	**0.3055**	**0.2988**	**0.3002**	**0.3292**

Table 5. Quantitative comparison of different prefix injection strategies. *Prefixes in both parts* refers to adding a prefix to text prompts that are fed into both parts (frozen and trainable copy) of the model.

6.2. Learning Zero-Conv with MLP

	FID Scores			
Method	**Depth ↓**	**HED ↓**	**Scribble ↓**	**Animal Pose ↓**
Learn weight by MLP	32.06	32.17	32.04	72.21
OmniControlNet (Ours)	**23.20**	**27.26**	**25.79**	**53.28**

	CLIP$_t$ Similarity Scores			
Method	**Depth ↑**	**HED ↑**	**Scribble ↑**	**Animal Pose ↑**
Learn weight by MLP	**0.3102**	**0.3085**	**0.3101**	0.3266
OmniControlNet (Ours)	0.3055	0.2988	0.3002	**0.3292**

Table 6. Quantitative results of generating zero-conv weights via textual inversion embeddings. *Learn weight by MLP* refers to the model using an MLP to learn the weight of the first zero-convolution.

In our original framework, the zero-conv layers are initialed with zeros and updated during each training step by backpropagation, where multiple tasks share the same zero-conv weights. In the ablation study, we use an MLP to generate the weights of the first zero-conv layer from the textual inversion embedding of each task. The results are presented in Tab. 6. We observe that directly training the first convolution layer instead of using the MLP yields a better FID score, yet generating the weights dynamically via MLP produces an overall higher CLIP$_t$ score.

6.3. Different Task Encoding and Number of Heads

In our foundational framework, a multi-head Feature Pyramid Network (FPN) is employed to process multi-scale features, while one-hot encoded task embeddings are utilized for extracting target conditions. Our ablation study investigates the indispensability of the multi-head FPN and the efficacy of one-hot encoding. We implement two variations: one model with a single FPN head and another leveraging complex text embeddings generated by the CLIP [65] text encoder. The comparative results are detailed in Tab. 7. Results show that integrating one-hot encoding with multiple FPN heads yields superior performance, demonstrating the effectiveness of our design.

Method	**HED Edge**			**Depth Map**
	ODS ↑	**OIS ↑**	**AP ↑**	**RMSE ↓**
Text Embedding	0.600	0.640	0.580	0.558
Single Head	0.610	0.640	0.560	0.520
Ours	**0.761**	**0.782**	**0.811**	**0.472**

Table 7. Quantitative comparisons of different design choices of OmniControlNet. *Text Embedding* refers to the model with CLIP [65] text encoded task embeddings. *Single Head* delineates using a single-head Feature Pyramid Network (FPN).

7. Conclusion and Limitations

In this paper, we propose OmniControlNet, a streamlined approach that combines multiple external condition image generation processes into a cohesive one. This integration addresses the limitations of ControlNet's two-stage pipeline, which relies on external algorithms and has separate models for each input type. With OmniControlNet, we have a multitasking algorithm for generating conditions like edges, depth maps, and poses and an integrated image generation process guided by textual embedding. This results in a simpler, less redundant model capable of generating high-quality text-conditioned images.

Limitations. **1)** When adding an additional task condition, it's required to train a new embedding for the task. **2)** With the integrated stage 1 model, the training complexity will increase, and image generation quality will decrease compared to using separate expert models as the stage 1 model.

Acknowledgement. This work is supported by NSF Award IIS-2127544. We are grateful for the constructive feedbacks from Yifan Xu and Zheng Ding.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 2

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 6, 7

[3] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. 3, 6

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[5] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4380–4389, 2015. 6, 7

[6] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 504–512, 2015. 6, 7

[7] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 3

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[9] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 1, 3, 6, 7

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 3

[12] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2

[13] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[14] Yuru Chen, Haitao Zhao, Zhengwei Hu, and Jingchao Peng. Attention-based context aggregation network for monocular depth estimation. *International Journal of Machine Learning and Cybernetics*, 12:1583–1596, 2021. 6, 7

[15] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[16] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 5

[17] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 6

[18] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[19] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1558–1570, 2014. 6, 7

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[22] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 6, 7

[23] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 2

[24] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[25] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2

[26] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 4, 5

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 2014. 1, 2

[28] Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. Towards light-weight and real-time line segment detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1

[29] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[30] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2019. 6, 7

[31] Sam Hallman and Charless C Fowlkes. Oriented edge forests for boundary detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1732–1740, 2015. 6, 7

[32] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2

[33] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4, 6

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[36] Linxi Huan, Nan Xue, Xianwei Zheng, Wei He, Jianya Gong, and Gui-Song Xia. Unmixing convolutional features for crisp edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6602–6609, 2021. 3

[37] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[38] Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*, 2023. 2

[39] Jyh-Jing Hwang and Tyng-Luh Liu. Pixel-wise deep learning for contour detection. *arXiv preprint arXiv:1504.01989*, 2015. 6, 7

[40] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Crisp boundary detection using pointwise mutual information. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 799–814. Springer, 2014. 6, 7

[41] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[42] Oren Katzir, Vicky Perepelook, Dani Lischinski, and Daniel Cohen-Or. Multi-level latent space structuring for generative control. *arXiv preprint arXiv:2202.05910*, 2022. 2

[43] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 3

[44] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2013. 1

[45] Josef Kittler. On the accuracy of the sobel edge detector. *Image Vis. Comput.*, 1:37–42, 1983. 3

[46] S. Konishi, A.L. Yuille, J.M. Coughlan, and Song Chun Zhu. Statistical edge detection: learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):57–74, 2003. 3

[47] Lubor Ladický, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014. 3

[48] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 3

[49] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019. 6, 7

[50] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 5

[51] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[52] Joseph J Lim, C Lawrence Zitnick, and Piotr Dollár. Sketch tokens: A learned mid-level representation for contour and object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3158–3165, 2013. 6, 7

[53] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021. 2

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[55] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[56] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015. 3

[57] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 2

[58] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[59] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 3

[60] D.R. Martin, C.C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004. 3

[61] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[62] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 6, 7

[63] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 3

[64] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2, 3

[65] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2, 8

[66] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2022. 1

[67] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. Edter: Edge detection with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1402–1412, 2022. 3

[68] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 2, 4, 6, 7

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[70] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6, 7

[71] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1, 2

[72] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[73] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020. 1, 3

[74] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 6, 7

[75] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 5

[76] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3

[77] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5

[78] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016. 2

[79] Xiaofeng Ren and Liefeng Bo. Discriminatively trained sparse code gradients for contour detection. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1*, pages 584–592, 2012. 6, 7

[80] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or.

Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2

[81] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 4

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[83] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, 2023. 1

[84] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[85] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[86] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009. 3

[87] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294. Curran Associates, Inc., 2022. 5

[88] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3982–3991, 2015. 3, 6

[89] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3982–3991, 2015. 7

[90] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[91] Amos Sironi, Vincent Lepetit, and Pascal Fua. Multiscale centerline detection by learning a scale-space distance transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2697–2704, 2014. 6, 7

[92] Amos Sironi, Vincent Lepetit, and Pascal Fua. Projection onto the manifold of elongated structures for accurate extraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 316–324, 2015. 6, 7

[93] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2015. 1

[94] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2

[95] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 2

[96] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021. 3

[97] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[98] Zhuowen Tu. Learning generative models via discriminative approaches. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1

[99] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[100] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. Diode: A dense indoor and outdoor depth dataset, 2019. 1, 3

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[102] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[103] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[104] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image

synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[105] Yupei Wang, Xin Zhao, and Kaiqi Huang. Deep crisp boundaries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3892–3900, 2017. 3

[106] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 3

[107] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In *Proceedings of IEEE International Conference on Computer Vision*, 2015. 1, 3, 6, 7

[108] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 5

[109] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 2018. 3

[110] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[111] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *proceedings of the IEEE international conference on computer vision*, pages 1281–1290, 2017. 3

[112] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. APT-36k: A large-scale benchmark for animal pose estimation and tracking. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3, 5

[113] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3, 5

[114] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[115] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[116] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 3, 5, 6, 7

[117] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 2023. 2, 4, 6, 7

[118] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 1

[119] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[120] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 2