

Adding Conditional Control to Text-to-Image Diffusion Models

2022314182 / 박수연

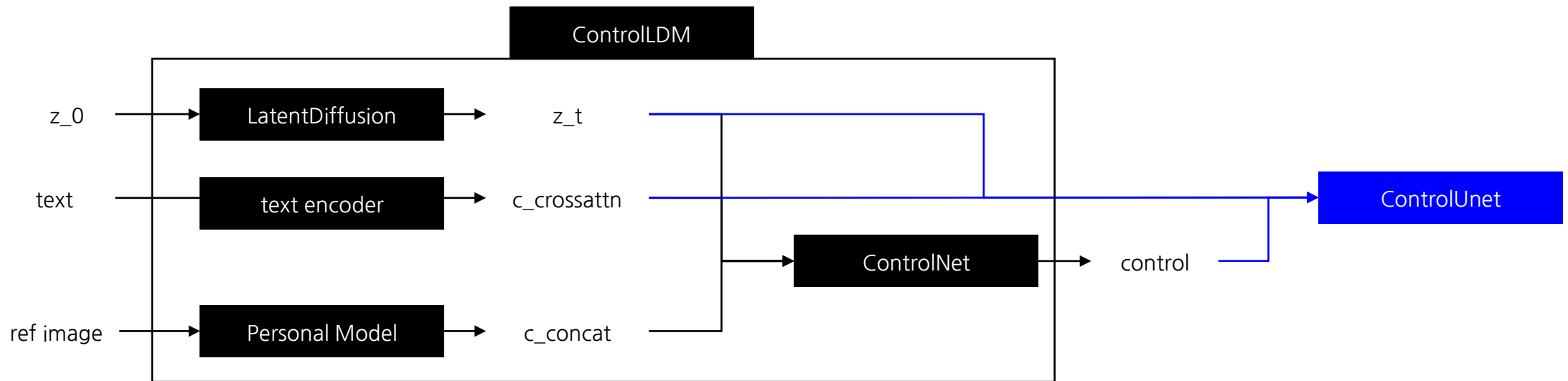
1. 모델의 전체 구조

전체 모델의 구동은 아래와 같다.

pure한 이미지의 latent 는 Latent Diffusion Scheduler에 의해서 noise latent 가 된다.

control 은 각각 concat 요소와 crossattn 요소가 ControlNet 에 의해서 control 로 도출이 된다.

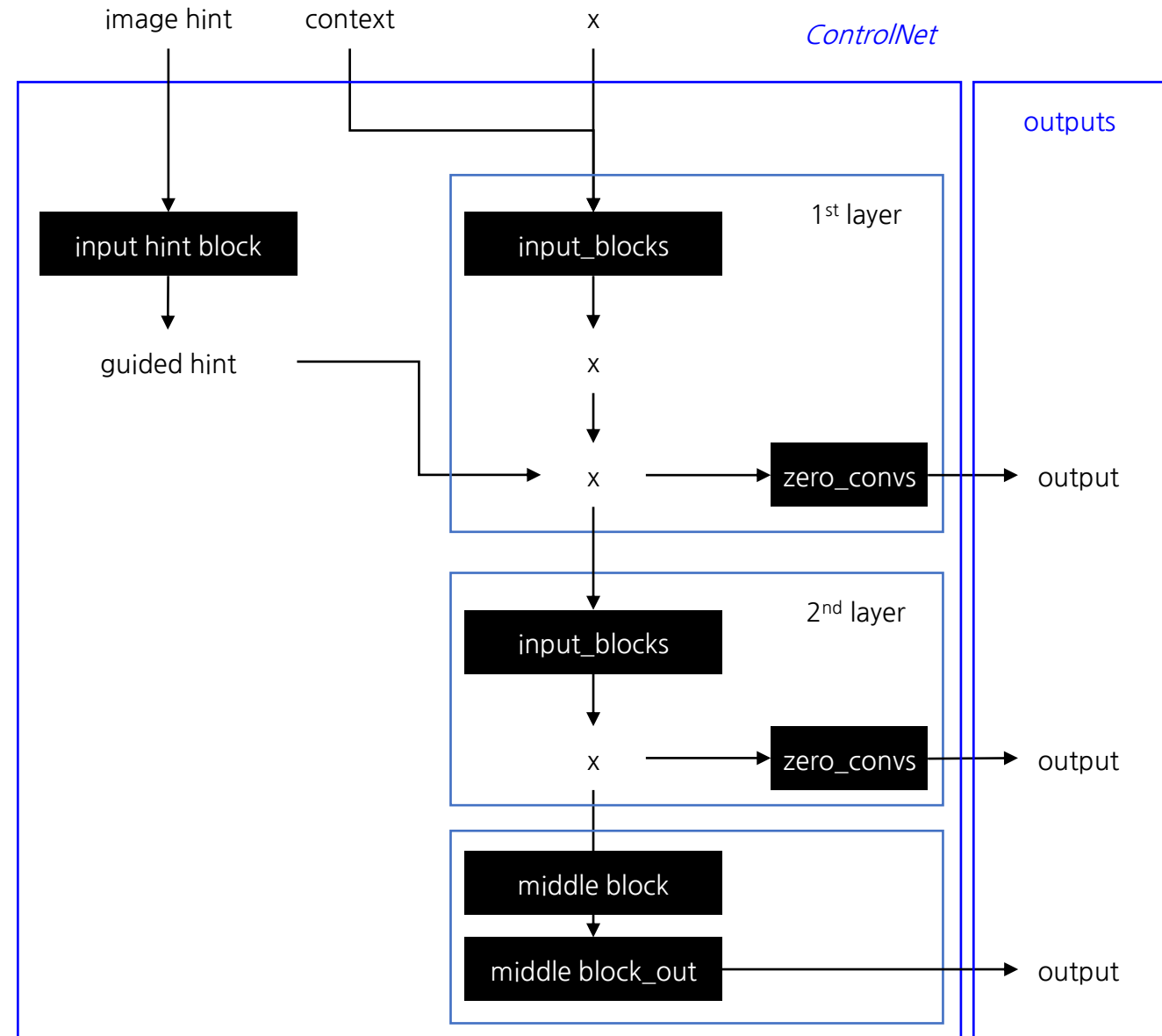
이는 Unet 에서 noise를 predict 하는 방식으로 모델은 학습하게 된다.



2번 모듈, ControlNet

Controlnet 은 zero convolution이 함께 있다. 여기서 중요한 포인트는 두가지 이다.

1. Controlnet 은 input 와 middle block 으로만 이루어지는데, 각 층에서의 결과가 통합되지 않고, list 에 쌓이게 된다.
2. guided hint 는 1st layer 에서만 활용이 되고, 그 뒤에서부터는 쓰이지 않는다.



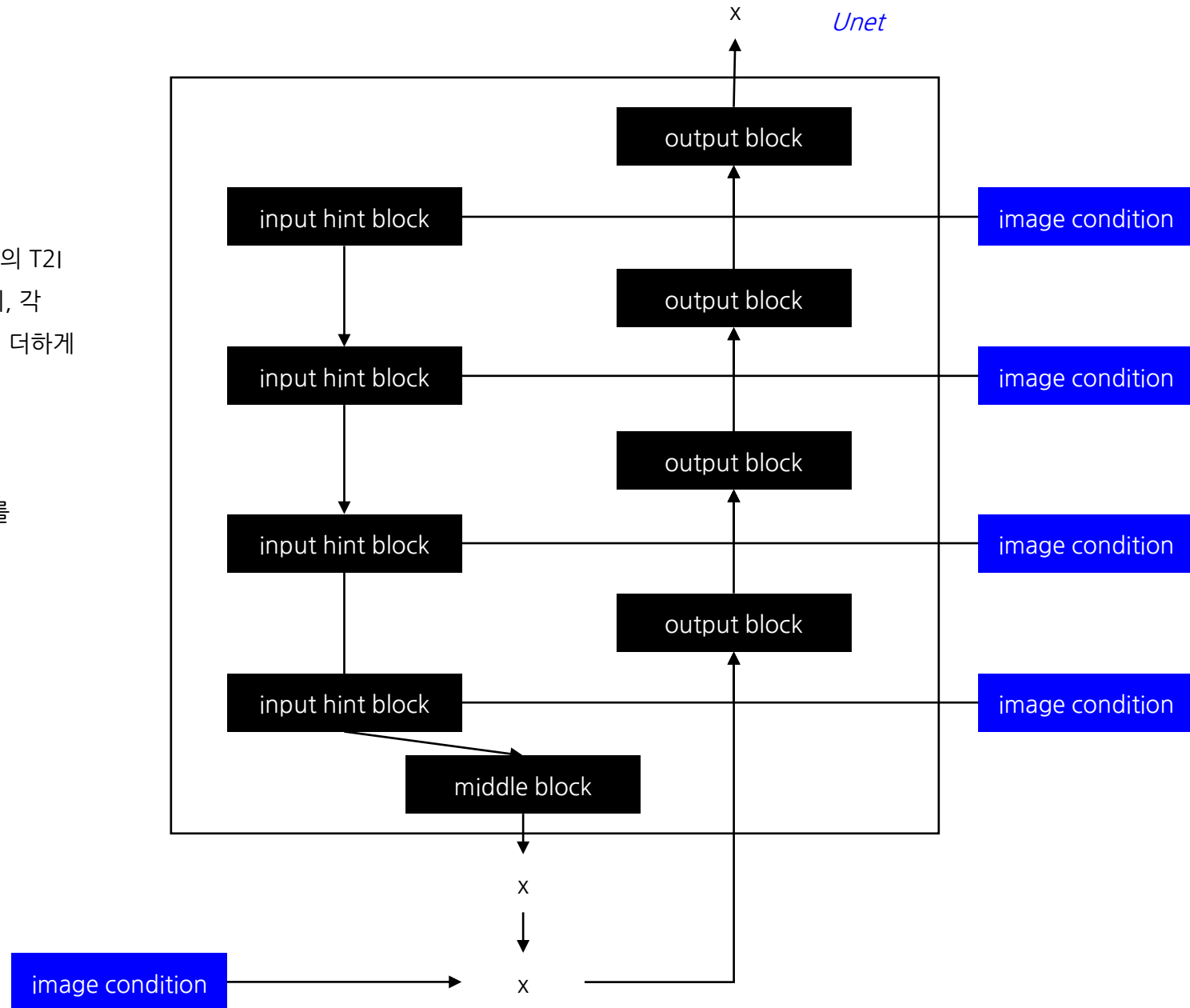
Unet Module

Controlnet 의 Unet 은 input blocks / middle blocks / output blocks 로 이루어진다. 먼저 input block 에서는 이미지 조건과 무관하게 text to image 의 T2I 과정이 일어난다. 이후 middle block부터 output block 에서는 controlnet 의, 각 층에서의 결과를 list 에서 pop 을 통해서 하나씩 빼내면서 unet 에서의 결과에 더하게 된다.

Unet 의 특성 상, input block 에서 down sampling 을 하면서 얻어진 결과는 output block 에서 upsampling 하는 역과정과 동일하며 이는 input 의 결과를 output 에 matching 시킬 수 있음을 의미한다.

이에 따라서, output 을 형성해 나갈 때는

- (1) Controlnet 의 결과를 더하면서
 - (2) input block 에서의 결과를 더하는
- 두가지 작업이 추가가 된다.



THANK YOU