More Robust and Faster ControlNet with Knowledge Distillation

Oliverio Theophilus Nathanael[1] , Julius Ferdinand[1] , Erio Yoshino[1] , Wawan Cenggoro[1]

[1]Universitas Bina Nusantara, Jakarta, Indonesia

Introduction

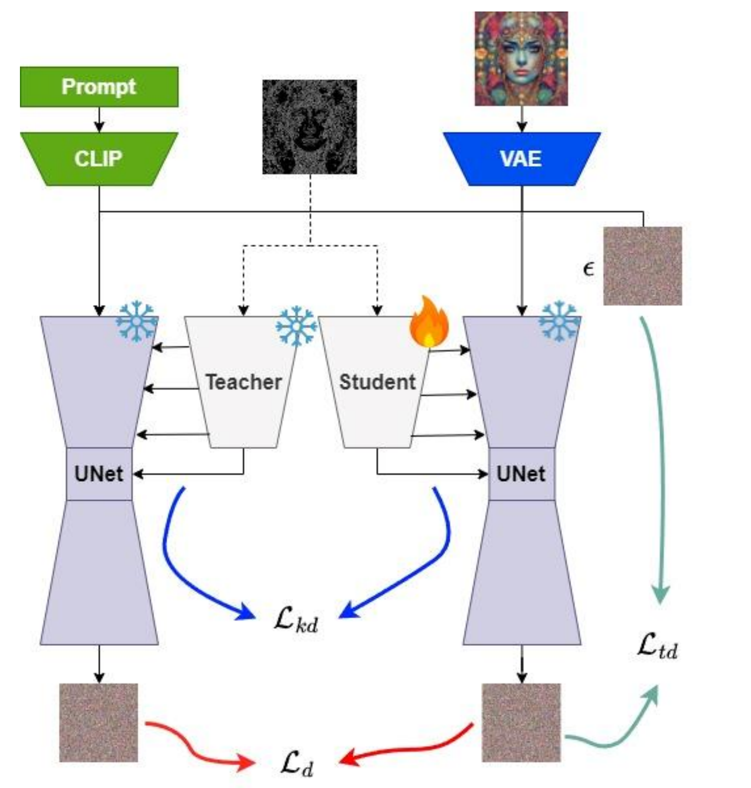
Motivations

- Model Size and Efficiency → nowadays controlnet [1] utilize the exact same architecture to its target backbone model, we believe that it is possible to develop a far lighter controlnet
- Training Behavior → Utilizing Knowledge Distillation [2] on controlnet is still rarely explored, we aimed to explore Knowledge Distilation training behavior as a starting point for future researches

Contributions

- We explore the strategies and effects of ControlNet Knowledge Distillation.
- We evaluate and analyze ControlNet Distillation qualitatively and quantitatively across different configurations and architecture.
- We propose a novel architecture to enable better knowledge transfer on lighter ControlNets.

Proposed Method



- Trained using Knowledge Distillation Scheme [2] where only the student controlnet is trained
- We use Stable Diffusion turbo model, a type of latent diffusion model as the base model

Diffusion Loss

$$\mathcal{L}_d = \mathbb{E}z_i, t, c, \epsilon \| \epsilon_\theta(z_i, t, c) - \epsilon \|_2^2$$

- Mean Squared Error of the predicted noise to the ground truth noise

Teacher Diffusion Loss

$$\mathcal{L}_{td} = \mathbb{E}_{z_i,t,c,\epsilon} \| \epsilon_\theta(z_i, t, c) - \epsilon_\Theta(z_i, t, c) \|_2^2$$

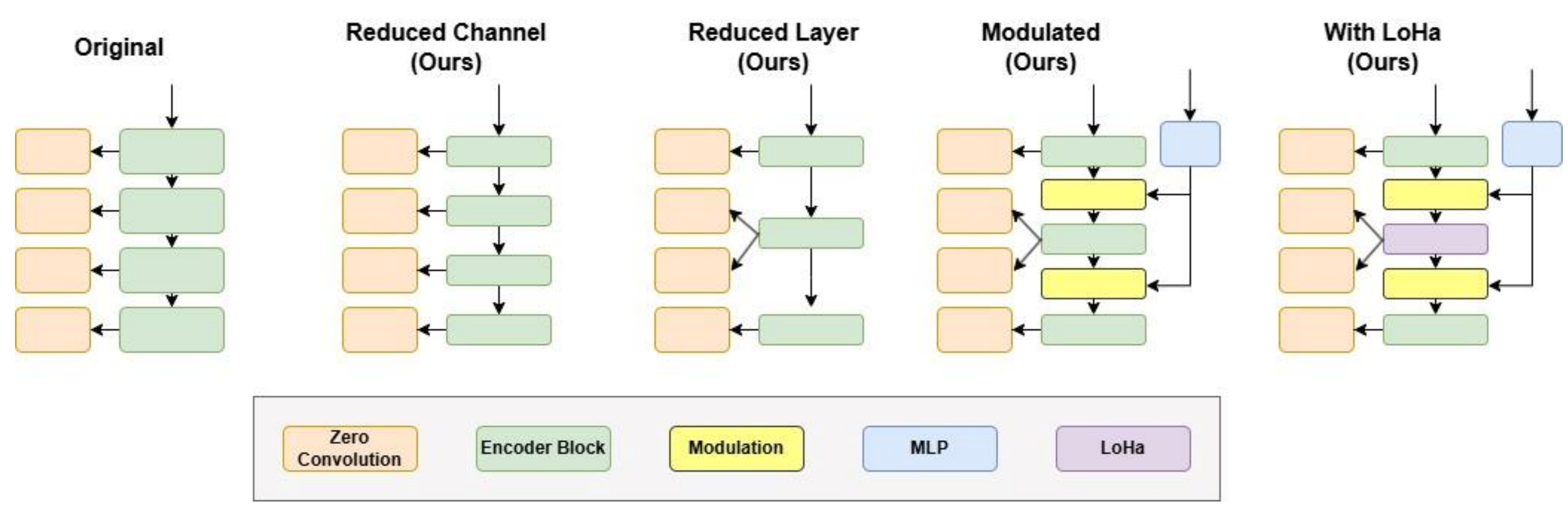
- Mean Squared Error of the predicted noise to the teacher's predicted noise

Layerwise Knowledge Distilation Loss

$$\mathbb{E}_{g,z_t,t,c,\epsilon,l} \| \frac{A_l^t - \mu_l^t}{\sigma_l^t} - \frac{A_l^s - \mu_l^t}{\sigma_l^t} \|_2^2$$

- Mean Squared Error for each controlnet output layer between the Teacher and Student, each output are normalized based w.r.t the teacher's mean and standard deviation

Proposed Architecture

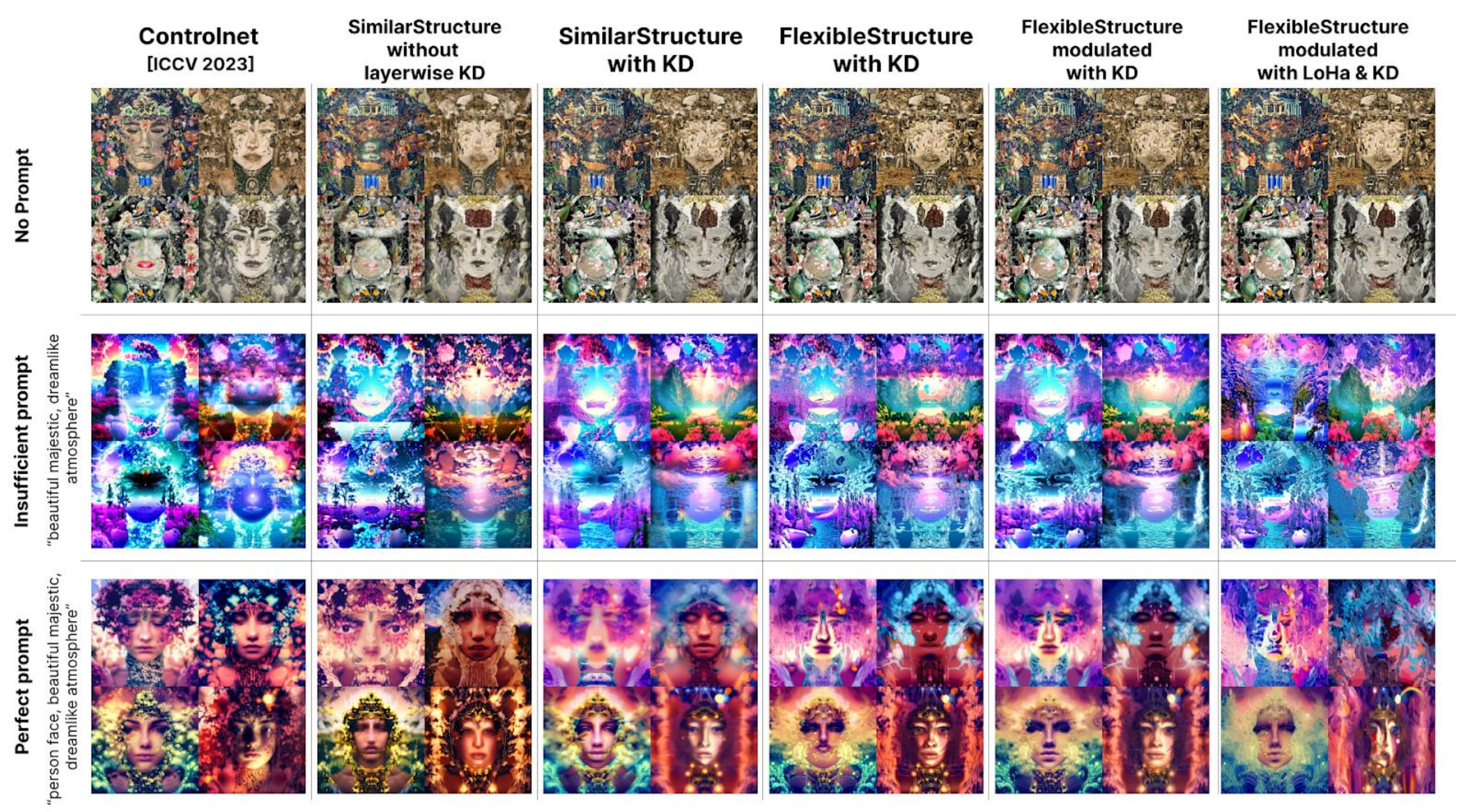


We introduce diverse strategies throughout the student controlnet shrinking process to address the challenges of each previous architecture (left to right).

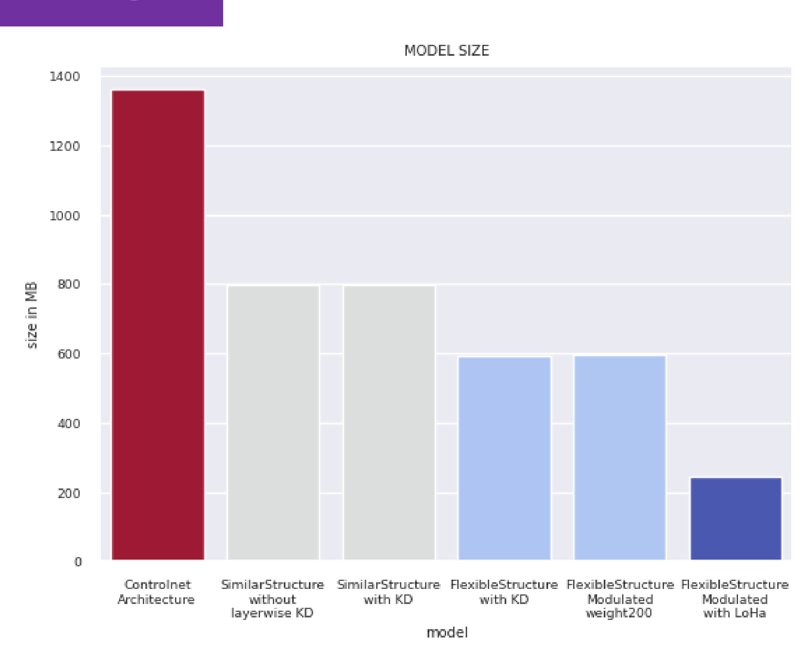
- **Original** → Present the exact same architecture to the target U-Net [1]
- **Reduced Channel** → Has the same amount of layer to the target U-Net with smaller number of filters/channel
- **Reduced Layer** → Has fewer layer, some output layers are fed from the same backbone layer to compensate the fewer layer
- **Modulated** → Added modulation layer after each layer, to address the weaker scaling capability on fewer layer model
- **LoHa** → some convolution layers are converted into LoRa Hadamard layer to further lighten the size

Results

Models	FID		SSIM		PSNR		LPIPS	
	Ground truth	Teacher	Ground truth	Teacher	Ground truth	Teacher	Ground truth	Teacher
Original Teacher	7.82781	4.49378	0.2206 ± 0.01041	0.19492 ± 0.00683	8.72913 ± 2.07138	9.37742 ± 2.10495	0.42781 ± 0.00593	0.38561 ± 0.00588
ControlNet [ICCV 2023]	6.65168	6.41135	0.23654 ± 0.01545	0.16631 ± 0.0071	8.48269 ± 1.57311	8.39484 ± 1.72188	0.46171 ± 0.00654	0.46616 ± 0.00582
Reduced channel	10.148	4.26166	0.205695 ± 0.011619	0.1566 ± 0.0061	8.33676 ± 1.5259948	8.3585 ± 1.6283	0.4844 ± 0.005856	0.4634 ± 0.00582
Reduced Layer	14.3793	3.5565	0.1852 ± 0.01045	0.1446 ± 0.00566	8.2995 ± 1.55535	8.3042 ± 1.42653	0.4921 ± 0.00514	0.4715 ± 0.0054
Modulated	11.6212	3.8268	0.19883 ± 0.01097	0.1497 ± 0.00557	8.4138 ± 1.60584	8.3697 ± 1.56499	0.4819 ± 0.00555	0.4667 ± 0.00597
LoRa Hadamard (LoHa)	18.193	4.0308	0.175 ± 0.00946	0.1341 ± 0.00483	8.4318 ± 1.54018	8.3667 ± 1.31751	0.5039 ± 0.00474	0.4832 ± 0.00541



Size



- We found that by applying our proposed architecture, it is possible to achieve 2-6 folds size reduction.

Conclusion

- **Knowledge Distilation Strategy**

We show an effective approach by incorporating a normalized layerwise knowledge distillation with the original supervision mechanisms

- **6x Lighter model**

Our proposed architecture along with our proposed distillation strategy successfully leverage a model 6x lighter than the teacher to be able to produce a competitive results.

- **Robust Student Model**

With only 48000 images we demonstrate that the trained student achived a competitive result both quantitatively and qualitatively

References

1. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models.
2. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.