

# Supplementary Material: Semantic Image Synthesis of Anime Characters Based on Conditional Generative Adversarial Networks

XuHui Zhu

zhuxuhui@cqu.edu.cn

Feng Jiang

jiangfeng@stu.cqu.edu.cn

Jing Wen

wj@cqu.edu.cn

Yi Wang

Yiwang@cqu.edu.cn

Qiang Gao

gaoqiang@cqu.edu.cn

College of Computer Science

Chongqing University

Chongqing, China

---

## 1 Additional Ablation Study

**Character identity tensor dimension ablation.** Table 1 presents the performance when selecting 3, 11, 12, and 64 channel dimensions for the character identity tensor. It can be observed that the performance is better at 11 and 12 dimensions. Notably, semantic label maps, when encoded as one-hot, has exactly 11 dimensions. Hence, we think that the character identity tensor not only guides the generator in generating specific characters but also supplements the feature information needed to generate the character to some extent. When the dimension is small, such as 3 dimensions, the feature information content within the identity tensor is insufficient. Conversely, when the identity tensor dimension is large, such as 64 dimensions, it tends to interfere with the generator’s recognition of semantic label map. Therefore, we conclude that selecting the identity tensor dimension comparable to that of the semantic label map is a better choice, as it appropriately supplements feature information without interfering with the generator’s label map recognition.

**Conditional noise types ablation.** We experimented with both conditional noise and unconditional noise (adding noise without distinguishing character identity, similar to StyleGAN[2, 3]). As shown in Table 2, the performance is better when using conditional noise, indicating that conditional noise can better fit the feature distribution of the generated images.

**Edge label type ablation.** We investigated two different edge labeling strategies: annotating only the character’s hair, and annotating the entire character excluding the background. As shown in Table 3, annotating only the character’s hair outperforms on all metrics. The

Dataset	Quintuplets			Zero Two		
Dim	FID↓	KID×100↓	LPIPS↓	FID↓	KID×100↓	LPIPS↓
3	74.01	1.37	0.361	82.76	1.21	0.363
11	<b>68.38</b>	<b>0.97</b>	<b>0.358</b>	82.56	1.27	0.363
12	69.06	0.99	0.362	<b>82.11</b>	1.23	<b>0.360</b>
64	71.11	1.00	0.363	82.21	<b>1.20</b>	0.365

Table 1: Quantitative results of using character identity tensors with different channel dimensions as generator inputs.

Dataset	Quintuplets			Zero Two		
Noise type	FID↓	KID×100↓	LPIPS↓	FID↓	KID×100↓	LPIPS↓
Condition	<b>67.85</b>	<b>0.75</b>	<b>0.359</b>	<b>82.26</b>	1.40	<b>0.362</b>
Non-condition	69.02	1.02	0.361	82.38	<b>1.18</b>	0.368

Table 2: Quantitative results when adding different types of noise to the generator.

Dataset	Quintuplets			Zero Two		
Edge label type	FID↓	KID×100↓	LPIPS↓	FID↓	KID×100↓	LPIPS↓
Only hair	<b>67.91</b>	<b>0.82</b>	<b>0.359</b>	<b>80.15</b>	<b>1.09</b>	<b>0.363</b>
Whole character	68.41	1.05	0.360	83.95	1.33	0.373

Table 3: Quantitative results of using two different annotation methods as edge label maps.

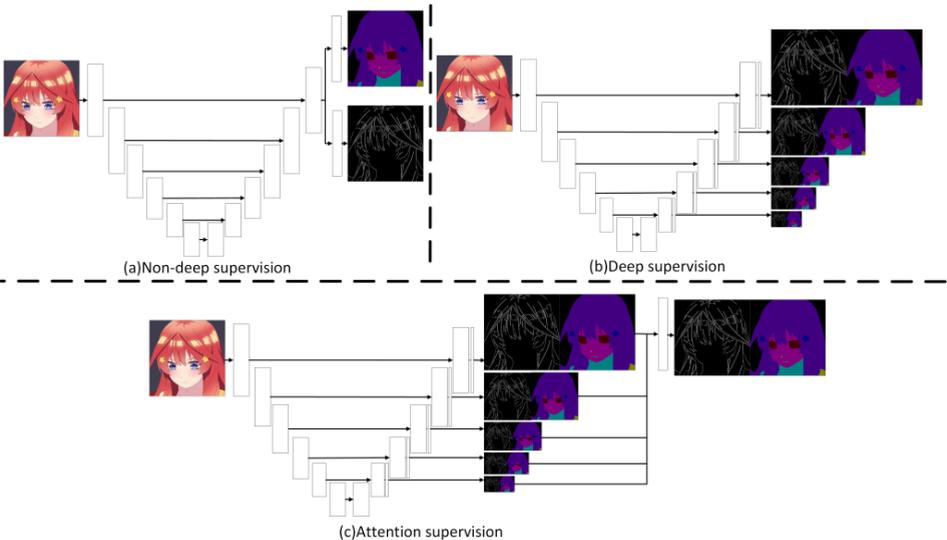


Figure 1: Three different discriminator structures.

Dataset	Quintuplets			Zero Two		
Discriminator structure	FID↓	KID×100↓	LPIPS↓	FID↓	KID×100↓	LPIPS↓
None deep	<b>67.91</b>	<b>0.82</b>	0.359	<b>80.15</b>	<b>1.09</b>	<b>0.359</b>
Deep	73.31	1.62	<b>0.358</b>	85.02	1.45	0.360
Attention	71.37	1.12	0.359	87.45	1.55	0.362

Table 4: Qualitative results of training generator using three different discriminator structures.

Dataset	Quintuplets			Zero Two		
$\alpha$	FID↓	KID×100↓	LPIPS↓	FID↓	KID×100↓	LPIPS↓
0.1	70.03	0.98	0.360	84.83	1.43	0.364
0.3	<b>67.91</b>	0.82	<b>0.359</b>	<b>80.15</b>	1.09	<b>0.359</b>
0.5	69.45	<b>0.81</b>	0.361	80.18	<b>0.95</b>	0.365
1.0	70.50	1.11	0.357	80.92	1.11	0.364

Table 5: Qualitative results of the impact of different values of  $\alpha$  used for edge detection on the quality of generated images.

reason for our analysis is that using hair as the edge detection target helps the discriminator pay more attention to the hair texture of the image. Conversely, for regions such as the character’s face, eyes, and eyebrows, the semantic boundaries in the semantic label map already provide sufficient boundary information. Therefore, additional edge detection in these regions is redundant.

**Discriminator structure ablation.** As shown in Figure 1, we used three different discriminator structures to train our generator: (a) None deep supervision, corresponding to the Unet[14] architecture, where the discriminator only uses the final layer for prediction; (b) Deep supervision, similar to HED[15], where predictions are made at each layer of the decoder; and (c) Attention supervision, similar to the HED-Unet[16], where an attention mechanism is used to merge predictions from each layer. Table 4 shows that (a)None deep supervision performs better. We attribute this to the fact that, compared to real-humans images, the texture composition of anime characters is simpler. This simplicity facilitates segmentation and edge detection. In contrast, using more complex discriminator structures tends to cause overfitting and instability during training.

**Binary cross entropy loss coefficient ablation.** To investigate the impact of the binary cross-entropy loss coefficient  $\alpha$  used for edge detection on the quality of generated images, we conducted experiments with values of 0.1, 0.3, 0.5, and 1.0. Table 5 demonstrates that when  $\alpha=0.3$ , the performance is optimal in terms of FID and LPIPS metrics, while  $\alpha=0.5$  yields the best results for KID. We analyze that when  $\alpha$  is small, such as 0.1, the optimization effect on texture is not significant, whereas when  $\alpha$  is large, such as 1.0, it tends to interfere with the discriminator’s recognition of semantics, especially in the early stages of training. Therefore, 0.3 is more appropriate.

## 2 Additional Results

In Figures 2 and 3, we present additional synthesis results of our method on the Quintuplets and Zero Two datasets, and compare them with results from other methods. Overall, our method generates images that are closer to ground truth, with higher visual quality in terms of character clothing, background, and hair.

Figure 4 shows additional results of our method generating different characters by changing the character identity tensor while using the same semantic label map, and generating same characters by using the same character identity tensor while using different semantic label map.

## References

- [1] Konrad Heidler, Lichao Mou, Celia Baumhoer, Andreas Dietz, and Xiao Xiang Zhu. Hed-unet: Combined segmentation and edge detection for monitoring the antarctic coastline. *IEEE transactions on geoscience and remote sensing*, 60:1–14, 2021.
- [2] 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.
- [3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.
- [4] 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.
- [5] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.



Figure 2: Comparison between our method and other methods on the Quintuplets and Zero Two dataset.



Figure 3: Comparison between our method and other methods on the Quintuplets and Zero Two dataset, especially hair textures.



Figure 4: The same semantic label map is used to generate different character images and the different character images are generated in the same semantic label map by character identity tensors. The upper left corner of the semantic label map is Ground Truth.