Supplementary material for "Tag2Pix: Line Art Colorization Using Text Tag With SECat and Changing Loss"

Hyunsu Kim*, Ho Young Jhoo*, Eunhyeok Park, and Sungjoo Yoo

Seoul National University

{gustnxodjs, mersshs, eunhyeok.park, sungjoo.yoo}@gmail.com

1. Ablation study with component and loss

We conducted three types of ablation studies to demonstrate the effects of each component and loss proposed in our Tag2Pix network. The experimental setting is the same as that in Section 5.3 of the paper. We calculated FIDs [3] while ablating each component and loss in our network. Tables 1, 2, and 3 show the contribution of each component and loss.

Network No.	**N1**	**N2**	**N3**	**N4**	**N5**
CIT feature	-	O	O	O	O
Guide decoder	-	-	O	O	O
SECat	-	-	-	O	O
Two-step	-	-	-	-	O
FID	57.81	52.78	48.74	42.29	**39.21**

Table 1. Incremental component ablation study. The rightmost column is our best network with all components.

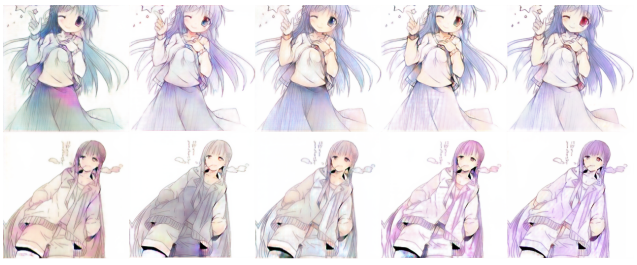

Figure 1. Images correspond to N1 to N5 from Table 1, left to right. Row 1 images were colorized with *blue_hair* and *red_eyes*; row 2 images were colorized with *purple_hair* and *pink_eyes*.

Table 1 shows that the FID decreases as more components are added. That is, colorization results gradually follow the color distribution of original images. As shown in Figure 1, from left to right, the quality of the image increases and the color of each segment becomes clearer and more similar to the given CVTs. Without CIT and the guide decoder (N1, N2), segmentation misses and color bleeding can occur. Without SECat and two-step training, as in the case of N3, the colors corresponding to the CVTs are not rendered well. N4, which is not trained by two-step training, is still insufficient, although some similar colors are painted in a particular position. After all components are used, the desired colors are appropriately painted in the desired areas.

Network No.	**N6**	**N7**	**N8**	**N4**	**N5**
CIT feature	-	O	O	O	O
Guide decoder	O	-	O	O	O
SECat	O	O	-	O	O
Two-step	O	O	O	-	O
FID	44.20	46.41	45.48	42.29	**39.21**

Table 2. Component ablation study removing one component at a time.

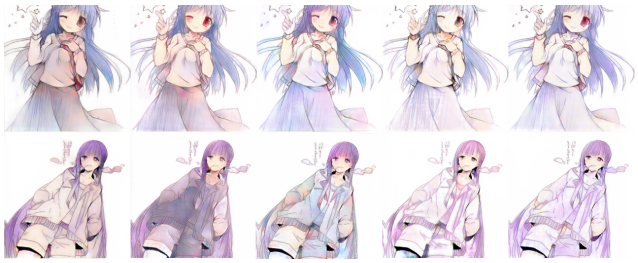

Figure 2. Ablation study colorization removing one component at a time. Images correspond to N6 to N5 from Table 2, left to right. Row 1 images were colorized with *blue_hair* and *red_eyes*; row 2 images were colorized with *purple_hair* and *pink_eyes*.

To show each component's effectiveness in more detail, we conducted additional experiments removing one component at a time to present quantitative and qualitative results. Table 2 and Figure 2 show that each component individually contributed to improving quality. The CIT feature ex-

*equal contribution

tractor provides color invariant information to the generator as a hint. Thus, the generator can localize well (N5 vs. N6 in Table 2 and the fifth vs. first column in Figure 2). The network needs all of the CIT features, SECat, and two-step training to colorize small feature like eyes (N5, N7). Without SECat, the colors are not uniform (N8). So far, we have seen shortages in coloring when not using any of the components.

Network No.	**N4**	**N5**	**N5**	**N5**
Adversarial loss	O	-	O	O
Classification loss	-	O	O	O
Reconstruction loss	O	O	-	O
FID	54.19	54.33	-(*)	**39.21**

Table 3. Loss ablation study. The rightmost column is the loss combination we used. (*) Failed to generate human-like image.

Figure 3. Colorization results of loss ablation study. Without classification loss (left), without adversarial loss (middle) and all loss contained (right). Row 1 images were colorized with *blue_hair* and *red_eyes*; row 2 images were colorized with *purple_hair* and *pink_eyes*.

In Table 3 and Figure 3, we used the same metric as in the component ablation studies. Note that, without classification loss, it is impossible to use a two-step method, so the ablation study with classification loss used the N4 network. Without classification loss or adversarial loss, the tone of color becomes dull overall and the small features, such as eyes, do not have the desired color. Eyes are instead generally painted in similar colors overall and are unnatural and lacking shade. Among loss functions, reconstruction loss (L1 loss) is the most important because it makes the network learn the overall distribution of real images so that the network can colorize somewhat naturally without classification loss or adversarial loss. However, our network colorizes small features such as eyes with the color specified by CVTs only when both adversarial loss and classification loss are in place.

2. User study details

20 users were recruited offline as evaluators without prior knowledge of our work. They evaluated networks in two sections, between sketch-based networks and text-based networks. In each section, they were given an hour to evaluate 30 sets of test images. Four categories of evaluation metrics were rated with a five-point Likert scale as follows:

- **Color Segmentation**
  The extent to which the colors do not cross to other areas, and individual parts are painted with consistent colors.
- **Color Naturalness**
  How naturally the color conforms to the sketch. The colors should match the mood of the painting.
- **Color Hints Accuracy**
  How well the hints are reflected. Output results should have red hair if a hint for red hair is given.
- **Overall Quality**
  The overall quality of the colorized result.

2.1. Sketch-based networks

We chose PaintsChainer [10], which is famous for colorizing an image with color strokes, and Style2Paints [7, 11], which is state-of-the-art in line art colorization. In PaintsChainer, we used the service provided through the official website. In Style2Paints, we created comparative images using the publicly available code (only V3 is publicly available). We prepared 140 real-world line arts and made 140 test sets. Each user evaluated 30 sets randomly selected out of the 140 test sets. Each test set consisted of PaintsChainer [10], Style2Paints [7], and Tag2Pix (ours).

Network	**PaintsChainer**	**Style2Paints**	**Tag2Pix**
Acc. Mean	3.28	3.73	**3.93**
Nat. Mean	2.44	3.47	**3.91**
Quality Mean	2.43	3.47	**3.86**
Seg. Mean	2.51	3.51	**3.94**
Acc. Std.	1.29	1.06	**0.98**
Nat. Std.	1.23	1.24	**1.05**
Quality Std.	1.19	1.15	**0.98**
Seg. Std.	1.28	1.24	**1.03**

Table 4. Mean and standard deviation of score in sketch-based networks.

Table 4 shows that Tag2Pix has the highest average scores and the lowest standard deviation in all evaluation metrics, indicating that our colorization network generates the most stable and highest quality images from line arts. Figure 4 also shows a box chart of the comparison results.

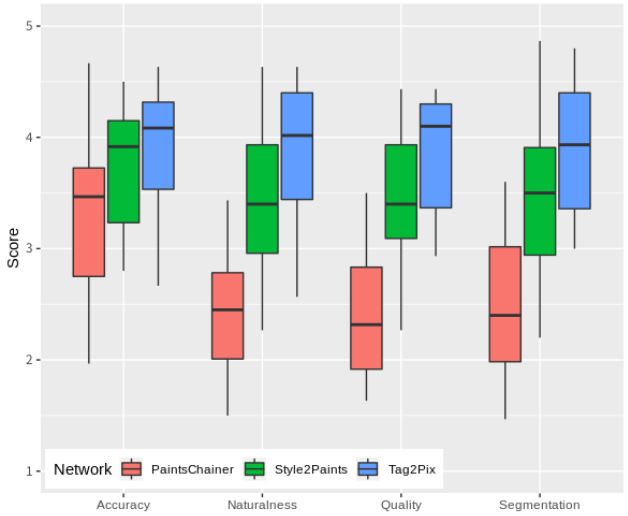

Figure 4. The box chart of sketch-based networks.

As PaintsChainer generally had much lower scores than Tag2Pix, we focused on comparing Style2Paints and Tag2Pix. Figure 8 shows evaluation images for which Tag2pix received a much higher rating than Style2Paints. Style2Paints images have very high contrast such that they look exaggerated. Additionally, strange colors spread on the arms and legs due to poor segmentation. Tag2pix is generally good at segmentation and provided natural colorization. However, Figure 9 shows evaluation images for which Style2Paints received a much higher rating than Tag2pix. As Style2Paints images have better distinct color and shade, some images that had proper segmentation got better scores. If the refinement stage of Style2Paints is added to our method, the image quality could be improved.

2.2. Text-based networks

There are comparable networks which use text as a hint for colorization. Chen *et al*. [1] colorizes a grayscale image using a text sentence hint. SISGAN [2] changes the color of certain parts of a color image as described in the sentence. However, Chen *et al*. [1] is a conceptual design, so fair comparison is difficult because it requires a separate implementation per dataset as done in the paper. In addition, SISGAN [2] completely failed to colorize, despite a small 74×74 RGB input because it was not suitable for our dataset and task. It failed to preserve the outline of the sketch and produced a strange result, thus we exclude them as a comparison target.

Instead, Manjunatha *et al*. [8] was selected as a comparison target, and the evaluation was conducted using the publicly available code [9]. Because Manjunatha *et al*. [8] colorizes the image using a sentence, we converted CVTs to sentences with fixed structure for both training and testing.

Network	Manjunatha *et al*.	Tag2Pix
Acc. Mean	3.27	**3.99**
Nat. Mean	3.46	**4.00**
Quality Mean	3.27	**3.86**
Seg. Mean	3.16	**4.13**
Acc. Std.	1.09	**0.95**
Nat. Std.	1.17	**0.99**
Quality Std.	1.09	**0.93**
Seg. Std.	1.17	**0.93**

Table 5. Mean and standard deviation of score in text-based networks.

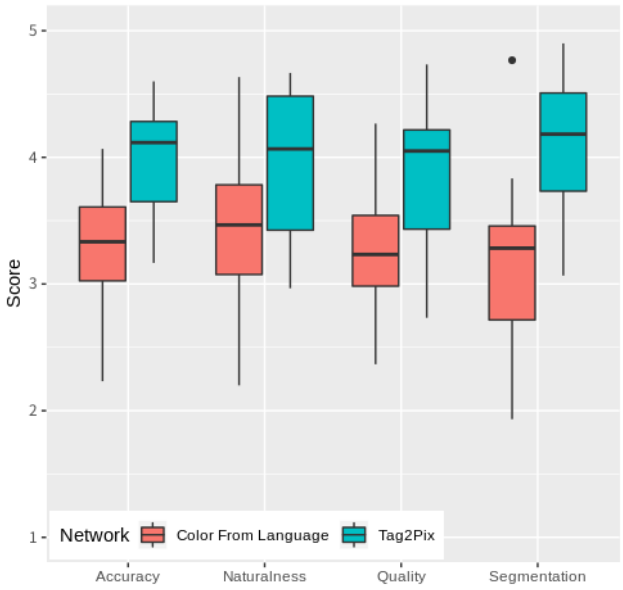

Figure 5. The box chart of text-based networks.

For example, *red_hair* and *blue_skirt* tags were converted to *a girl with red hair wearing blue skirt*.

As shown in Table 5 and Figure 5, even though our network's input image (line art) has much less information than Manjunatha *et al*. [8]'s input image (grayscale), ours has much better average scores and lower standard deviation in all evaluation metrics.

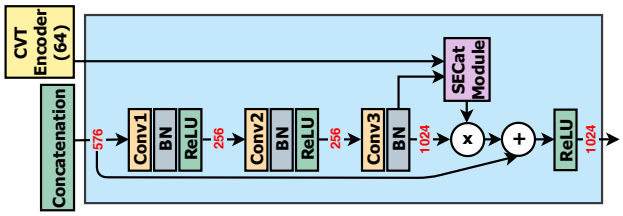

Figure 6. Proposed SECat-ResNeXt block structure.

3. SECat

Figure 6 illustrates the SECat-ResNeXt block structure. It shows how the decoder module, combining the Concatenation block, SECat-ResNeXt block, and PixelShuffle, decodes input features to output features with different dimensions using three convolution layers, *e.g.*, $576 \rightarrow 1024$ dimensions in the generator's first decoder module (Figure 3 in the paper).

As shown in Figure 5 in the paper, SECat module input is the output from the CVT encoder, and utilized to re-weight the residual block output feature maps. Figure 6 shows that CVT encoder output is only used by the SECat module, rather than by the residual block convolution layers. The SECat exploits weight re-balancing following SENet [4] and styleGAN [6], incorporating hints from the CVT encoder to re-calibrate the ResNeXt block output feature maps.

Compared to previous concepts [5, 12] for incorporating global information through concatenation, SECat can improve coloring quality by utilizing CVT information when re-calibrating output feature maps. Thus, SECat both incorporates CVT information into the network (see Figure 3 in the paper) and utilizes it to emphasize the features (shown in the Figure 6 above and Figure 5 in the paper).

In particular, SECat helps to significantly improve generation quality by enhancing fine detail colorization, including small objects such as eyes, compared with the concatenation method in Figure 11(c), as mentioned in Section 5.3 of the paper.

However, the FIDs in Table 3 of the paper struggle to indicate this improvement because the metric mainly focuses on overall colorization distribution. To show the SECat module's effectiveness on fine details, we conducted an additional user study to compare generated image output quality between embedding methods. We also released the SECat code[1] so that readers can reproduce the results.

3.1. Additional user study in CVT embedding schemes.

The experimental environment is similar to that in Section 2, employing 27 people without prior knowledge, and comparing three networks, each of which has a different hint embedding method. The evaluation time, method and number of evaluation sets are the same as in Section 2, but the evaluation data was the $6,545$ images used in Section 5.3 of the paper. We compared the three networks with the lowest FID in Table 3 of the paper, those being *SE-ResNeXt*, *Concat front*, and *SECat* (ours).

As shown in Table 6 and Figure 7, *SECat* is superior in all criteria. The mean is high, and the standard deviation is low. Comparing the results with *SE-ResNeXt* and other networks, the method of consistently inserting information about hints into all decoders is overwhelmingly effective. There is little FID difference between *Concat front* and *SECat* in Table 3 of the paper, but there is a somewhat meaningful difference in the user study. We have shown through user study that SECat is superior to other embedding methods.

Network	SE-ResNeXt	Concat front	SECat (ours)
Acc. Mean	2.75	3.51	**3.60**
Nat. Mean	2.85	3.40	**3.60**
Quality Mean	2.78	3.30	**3.54**
Seg. Mean	2.85	3.29	**3.51**
Acc. Std.	1.06	1.04	**1.00**
Nat. Std.	1.18	1.06	**1.00**
Quality Std.	1.17	1.08	**1.01**
Seg. Std.	1.19	1.13	**1.08**

Table 6. Mean and standard deviation of score of different CVT embedding schemes.

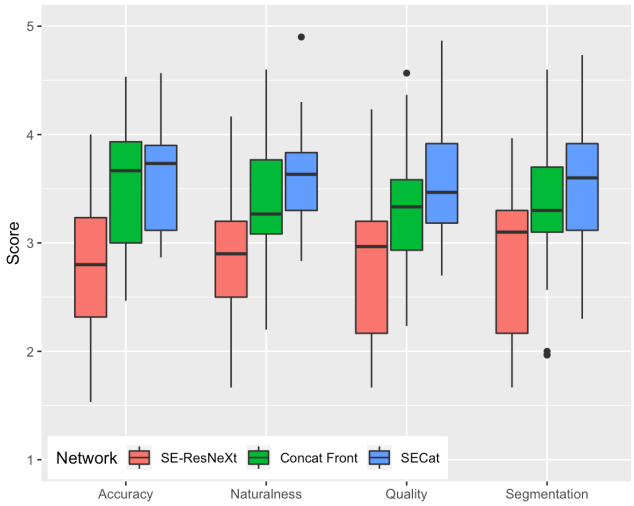

Figure 7. The box chart of different CVT embedding schemes.

4. More results and future work

Our proposed network has the potential for colorizing any pose in consecutive sketches with a single tag set, which makes Tag2Pix suitable for animation colorization. In Figures 10 and 11, we can see the feasibility of this idea. In the style-transfer colorization, which colorizes a line art with a reference image, Style2Paints cannot colorize properly unless the human pose of the line art is similar to the reference image. The color of the head and skin is especially misplaced even though we used the same character as a reference image in Figure 10. Through our method, we can colorize each part of the character properly regardless of the pose. This can be used as a tool in character design or

[1] https://github.com/blandocs/Tag2Pix

in the pre-production stage of the 2D animation production process.

We generated additional examples that can be seen in Figures 12 and 13. We changed the color of shirt, skirt, legwear, skin, and blush, as well as eyes and hair that were previously compared in the paper.

Acknowledgement

This work was supported by National Research Foundation of Korea (NRF-2016M3A7B4909604) and Samsung Electronics.

References

[1] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. *International Conference on Computer Vision (ICCV)*, 2017.

[3] 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. *Neural Information Processing Systems (NIPS)*, 2017.

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (SIGGRAPH)*, 2016.

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] lllyasviel. style2paints. `https://github.com/lllyasviel/style2paints`, 2018. Accessed: 2019-03-22.

[8] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. Learning to color from language. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[9] superhans. colorfromlanguage. `https://github.com/superhans/colorfromlanguage`, 2018. Accessed: 2019-03-22.

[10] Taizan Yonetsuji. Paintschainer. `https://paintschainer.preferred.tech/index_en.html`, 2017. Accessed: 2019-03-22.

[11] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage sketch colorization. *ACM Transactions on Graphics (TOG)*, 2018.

[12] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 2017.

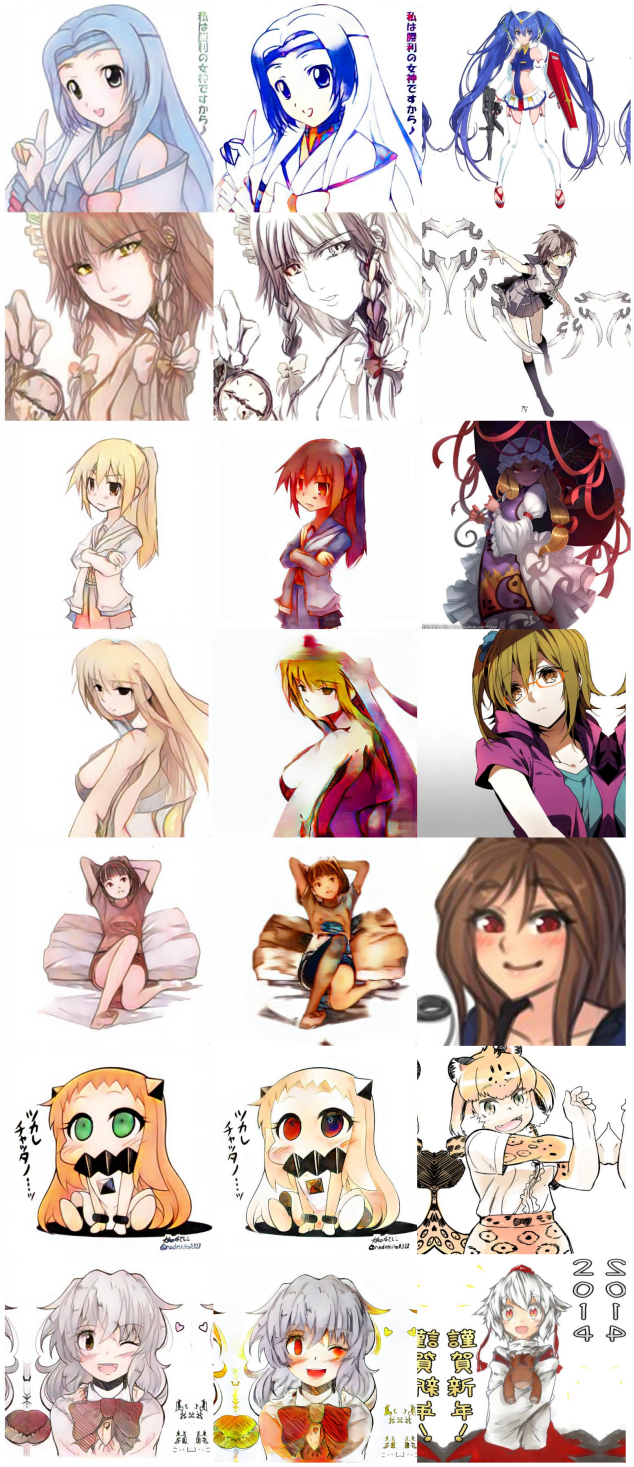

Figure 8. The images that Tag2pix (Ours) received a much higher rating than Style2Paints. Tag2pix colorization result (left), Style2Paints colorization result (middle) and Style2Paints reference images (right).

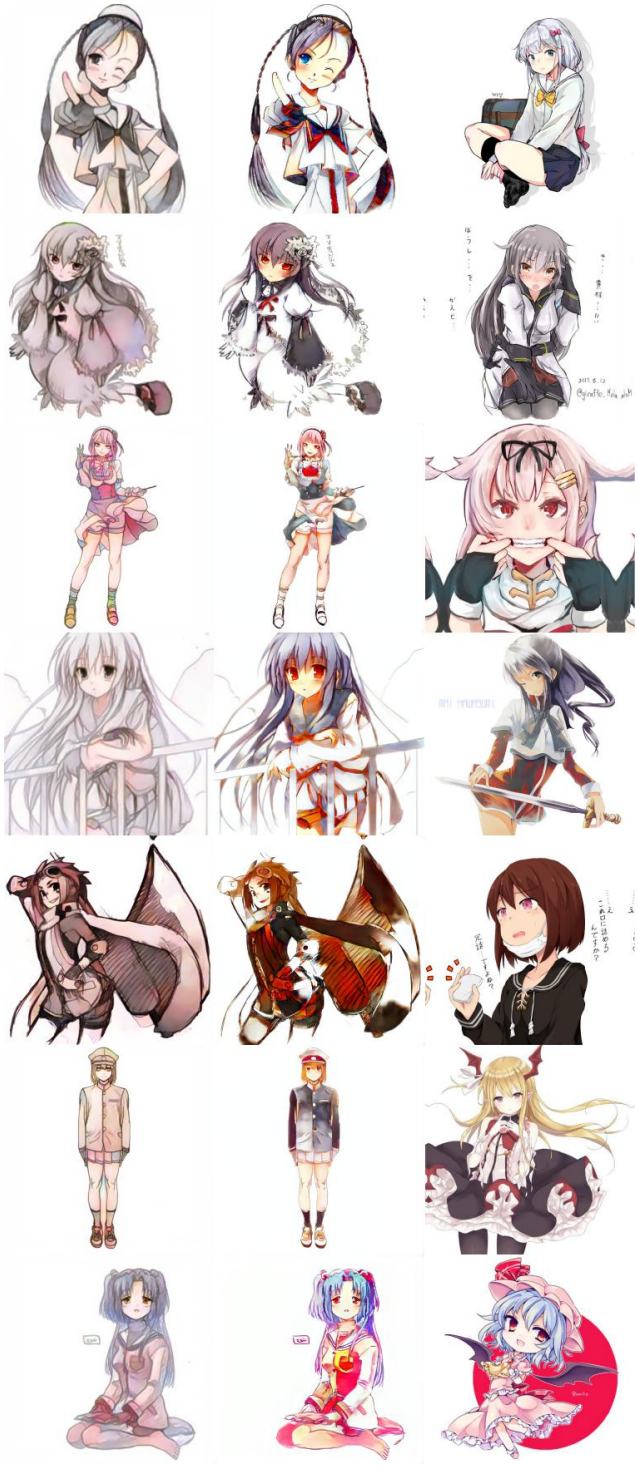

Figure 9. The images that Style2Paints received a much higher rating than Tag2pix (Ours). Tag2pix colorization result (left), Style2Paints colorization result (middle) and Style2Paints reference images (right).

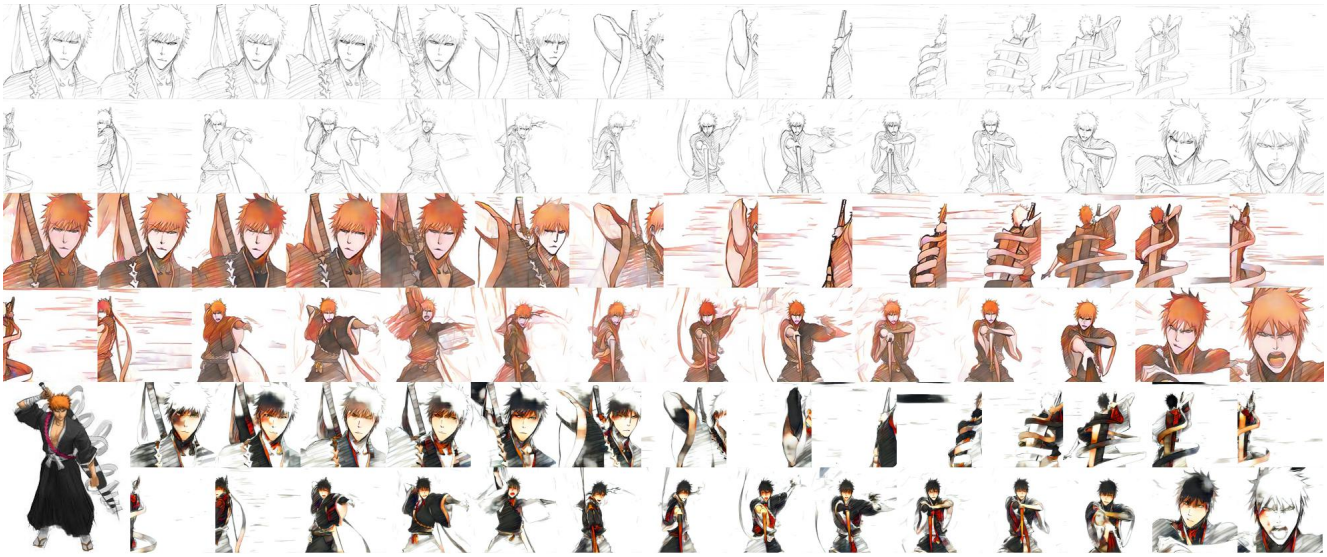

Figure 10. Animation colorization sample. The first and second rows are line arts in animation. The third and fourth rows are Tag2Pix (Ours) colorization results. We only give three tags to each frame, *white_background, orange_hair* and *black_dress*. The fifth and sixth rows are Style2Paints [7] colorization results. We used the leftmost image for reference.

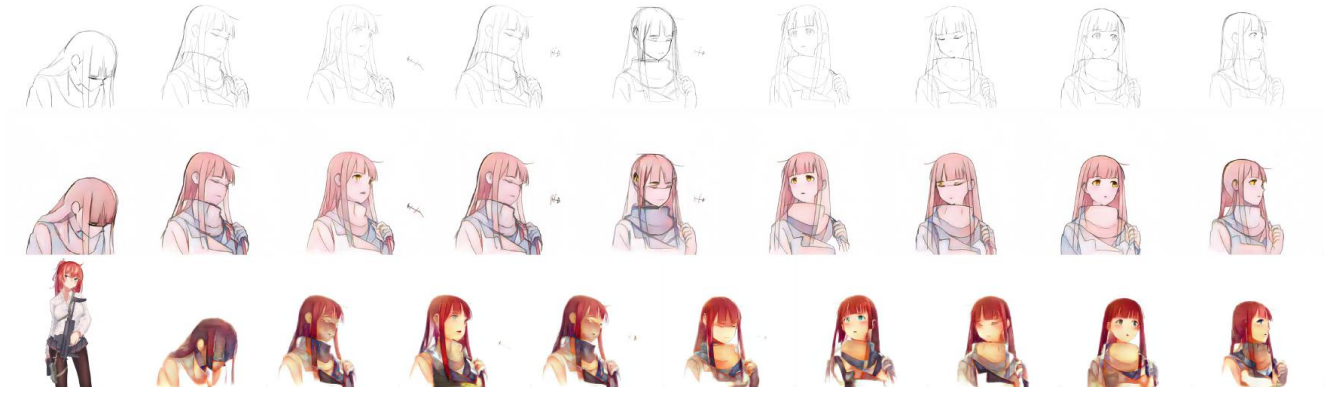

Figure 11. Animation colorization sample. The first row is line arts in animation. The second row is Tag2Pix (Ours) colorization results. We only give four tags to each frame, *white_background, pink_hair, yellow_eyes*, and *white_shirt*. The third row is Style2Paints [7] colorization results. We used the leftmost image for reference.

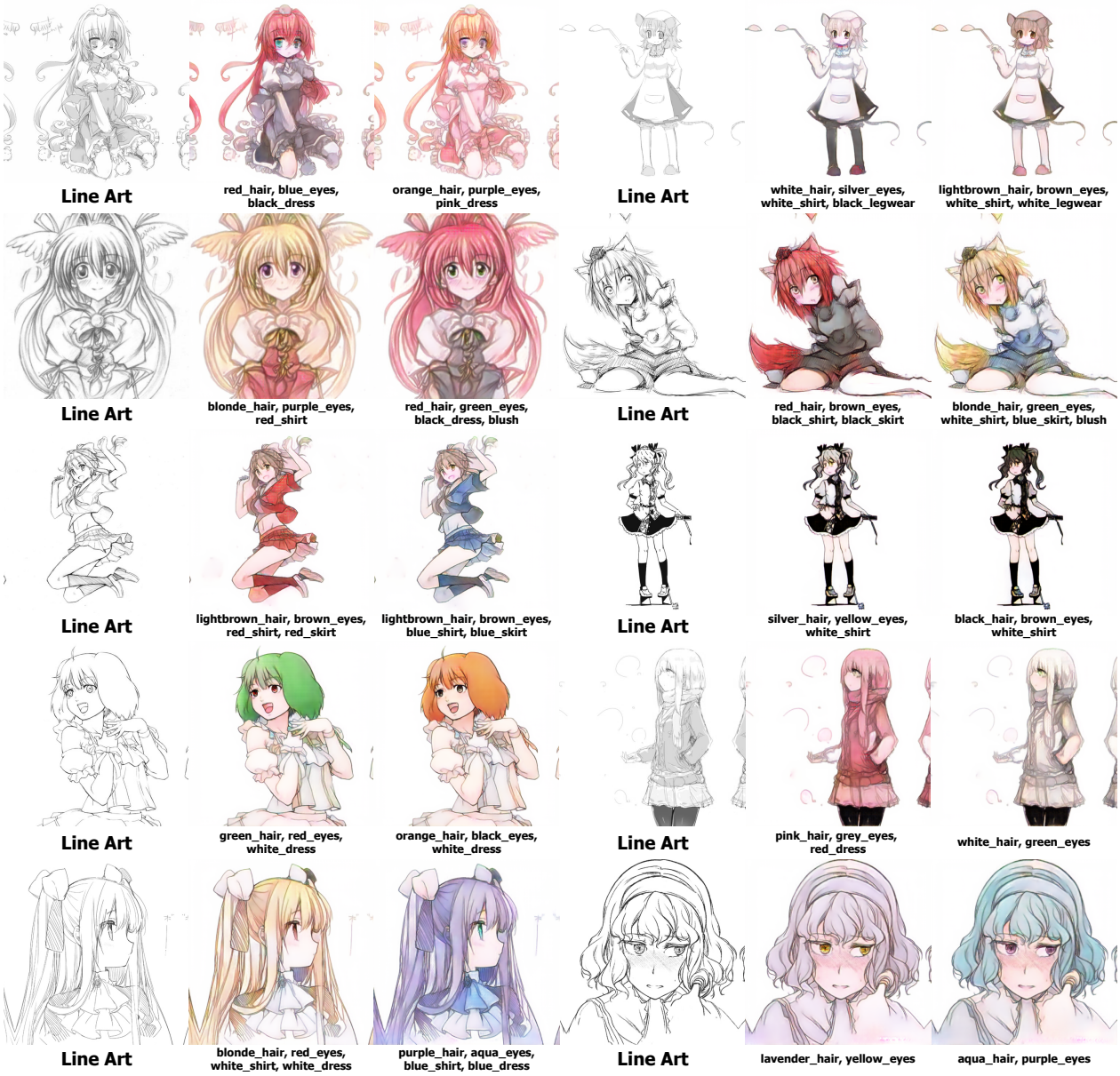

Figure 12. Results of our network. All images are colorized with the tag *white_background*.

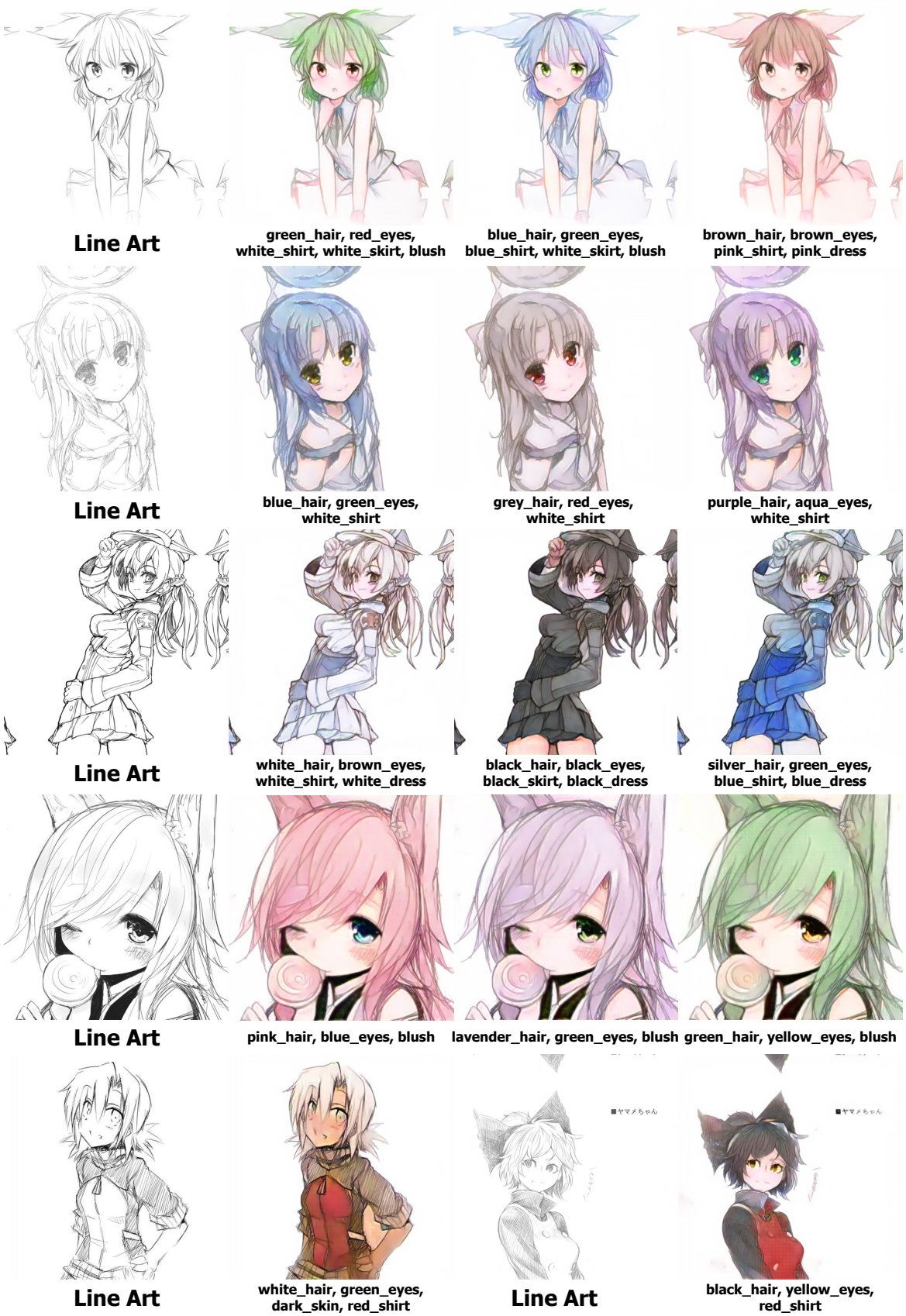

Figure 13. Results of our network. All images are colorized with the tag *white_background*.