A Comparative Study of Metric-Human Feedback Correlation in Image Colorization (March 2023)

Mario Amann, Felix Hamburger and Tobias Schmähling

***Abstract*—The paper compares different metrics for evaluating the quality of image colorization based on user feedback. Using a pre-trained classification network, we analyze the correlation between human feedback and distance metrics such as pixel distance, distances between convolutional layer outputs, and the distance of the output vector from the classification network itself. Furthermore, we apply the widely used classification accuracy, fréchet inception distance, structural similarity index measure and peak signal-to-noise ratio to the colorized images. With our concept we show the limitations of using traditional metrics for image colorization and which metrics are most suitable for colorization processes.**

***Impact Statement*—The impact of this research lies in its contribution to the field of image colorization and its evaluation. The results have important implications for improving the quality of automatic colorization algorithms and for the development of more effective evaluation methods. By showing the limitations of traditional metrics and the importance of human feedback, this study gives a broad overview in the field and leads to the creation of more visually appealing and realistic colorized images. Ultimately, this could benefit a range of applications, from historical archival to artistic expression and entertainment.**

***Index Terms*—Image Colorization, Metrics, User Test for Image Colorization**

I. Introduction

OVER the last decade machine learning has set new standards in the field of image related tasks. The advent of generative adversarial networks (GAN) provided a huge step for image-to-image translation. Tasks such as image generation (Gregor et al. [7], Qiao et al. [19]), image inpainting (Bertalmio et al. [3], Yu et al. [24], Guillemot et al. [9]) and super resolution (You et al. [23]) developed rapidly. Colorization of images is one of those image-to-image translation tasks, which among others can be solved by generative adversarial networks. Though original methodologies still seem to achieve better results as mentioned by Isola et al. [14]. Usually computer vision tasks rely on a strong ground truth, which indicates that there is only one correct answer. In the task of colorization it is not desired to limit the prediction to one correct answer, as the image should not be replicated. Ultimately the task is to create one possible color image which will withstand a human evaluation. In machine learning the human is typically used as a baseline model, but in colorization the estimation of the human is the objective of the model. To put it simply: A human cannot distinguish a colorized image from a non-colorized image. Due to the size of state-of-the-art image datasets and the resulting time expenditure, it is not possible to use human evaluation. Therefore, the common metrics in the field of colorization are investigated. Further we compare and discuss the results of which metrics best reflect the human assessment.

This paper is submitted on the 12 March 2023.

Mario Amann is with the Ravensurg Weingarten University (RWU) (e-mail: mario.amann@rwu.de).

Felix Hamburger is with the Ravensurg Weingarten University (RWU) (e-mail: felix.hamburger@rwu.de).

Tobias Schmähling is with the Ravensurg Weingarten University (RWU) (e-mail: tobias.schmaehling@rwu.de).

II. Problematic

The problem of image colorization is complex due to the subjective nature of color and the lack of ground truth information. Some images have the property of ambiguity which means that they still make sense if certain elements of the images are colorized differently. This can be seen in Figure 1, where it subjectively does not matter if the color of the carpet is either blue or green.

Fig. 1. Ambiguity of image colorization with greyscale image (left), ground truth (middle) and colorized image (right).

As shown in Figure 2, it is possible for a colorized image to look even more convincing than the ground truth, as it adds a new dimension of visual information and appeals to human perception in a unique way.

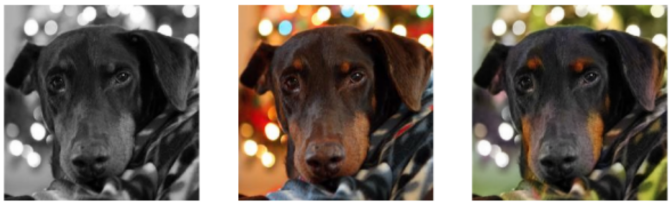


Fig. 2. Problematic of image colorization with greyscale image (left), ground truth (middle) and colorized image (right).

Accordingly, the objective is to produce colorized images that look appealing and accurately convey the information contained in the grayscale image, rather than to match the

colors exactly. Therefore, the success of colorization algorithms should be evaluated based on how well they achieve this goal, rather than on how closely they match a ground-truth reference.

III. Previous Work

A. Colorization Models

There are several approaches for coloring images. The network from Hariharan et al. [10] introduced the concept of hypercolumns on pixel level. While usually features are represented as the output of the last layer, Hariharan et al. represent a features pixel as a vector of all convolutional units.
Zhang et al. [26] (ECCV16) showed that the root mean squared error (RMSE) lacks robustness due to the multimodal and inborn uncertainty of the colorization problem. When elements of an image can take on different colors, the network learns a mean between those colors which makes the generated images rather brownish with a low saturation. To avoid this effect Zhang et al. [26] converts the regression task into a classification task by dividing the possible color-values into bins. The problem arises that the colors of a natural image are not equally distributed. To compensate for this, they balance each pixel according to its frequency. In the end, the probability distribution of the colors is adjusted.
The work of Zhang et al. [27] (Siggraph17) uses an interactive colorization approach. Interactive colorization means, that the user can inject desired colors and determine which colors some pixels should have. Siggraph17 simulates these user inputs during training. The advantage of this is that it solves much of the ambiguity problem and makes it possible to use regression error detection. The network can still be used as an automatic colorizer.
Additionally there exist generative colorization approaches. Deshpane et al. [6] used a variational autoencoder to learn a low-dimensional embedding variable and a mixture density network to learn their modal conditional model. Ardizzone et al. [2] used invertible neural networks and extended them by adding conditional inputs to the core building blocks.
Cao et al. [4] used a conditional generative adversarial network which generates high-quality colorized images.
Guadarrama et al. [8] (PixColor) and Royer et al. [20] (PIC) are autoregressive approaches, same as the colorization transformer (Coltran) introduced by Kumar et al. [16].
Coltran is composed of three model parts where self-attention is the foundation for the architecture. The coltran core is an autoregressive colorizer which is based on axial transformer with conditional transformer layers and an auxiliary parallel head.
The second part is a color upsampler and the third part is a spatial upsampler which also use attention layers.

B. Colorization Metrics

While there are many papers which introduce and compare different methods for image colorization (Zeger et al. [25], Anwar et al. [1]), to our knowledge only few tackle the comparison of image quality metrics in its entierty. To our knowledge only few papers such as Hore et al. [13] and Sara et al. [21] try to evaluate different metrics based on perceptive image quality. It is a common trend in the literature that papers often do not compare the metrics used in their work with each other or with existing benchmarks, which can make it challenging to compare the performance of different colorization models. Our paper is addressing this issue and establishing a set of standard metrics for image colorization. While we do not introduce new metrics to the community, we want to give an understanding which types of metrics suit image colorization tasks and which metrics do not meet the standards.

IV. Experimental Setup

In the following we will describe how we conduct our experiment. Initially we select different colorization methods to colorize images. Finally, we use a set of metrics and human annotations to provide a comprehensive assessment of the colorization quality, allowing us to understand the strengths and limitations of each approach.

A. Colorization Models

We have selected four different pretrained networks, each with a different approach.

VGG16: The network implemented by the paper of Hariharan et al. [10] uses hypercolumns at pixel level.
ECCV16: The network of Zhang et al. [26] this uses a classification and rebalance methods.
Siggraph17: The network of Zhang et al. [27] this is actually an interactive model, but is used as an automic model.
Coltran: The colorization transformer (Coltran), which was published by Manoj Kumar et al. [16], is an autoregressive colorizer based on axial transformer.

The focus of our comparison is not on the performance of the networks themselves, but rather on the comparison of the metrics used to evaluate image colorization with the help of human feedback. The output of the four networks is therefore helpful in this comparison, as our goal is to understand the connection between the metrics and human perception of the colorization quality.

B. Dataset

The used colorization algorithms were all trained on the ImageNet dataset [5].
For some of the metrics, classification labels of the images are necessary, thus it is suitable to use the ImageNet dataset to evaluate the metrics. The labels for the ImageNet dataset are only available for the training and validation dataset. To fully investigate the errors on a single image we need the label for a single image. Therefore, we use the validation dataset instead of the test dataset. This is valid because the performance of the models is not the target of our comparison.
Additionally we remove black and white images from the dataset.

C. User Test

For our user test we have been inspired by the perceptual realism test from Zhang et al. [26] and have extended it with an additional approach. In the first test, the survey participant is shown an image for one second. After that, the participant must decide whether the image shown to him is from the ground truth set or is one of the four models colorized image. Each session begins with two sample images so that the task of the experiment is clear to the participants. The experiment is carried out with 100 images. In order to ensure the comparability of the models and to keep human error as low as possible, we decided to assign the same number of images of each model to each participant. Each participant therefore gets 20% images of the ground truth set and 20% of each of the model sets. To ensure that each image is used

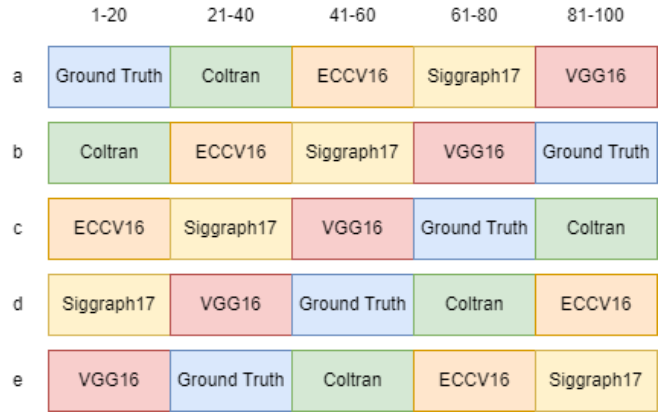


Fig. 3. Allocation of the images of the first user test.

in each colorization, the allocation is made as seen in Figure 3. With this allocation method five participants are needed to assess each image with every colorization method once.

The second test is more close to the test from Zhang et al. [26]. Instead of just one image, the participant is now shown two images side by side. One is from the set of ground truth and the other is from the set of model outputs. The participants see the two images simultaneously for one second, then decide which image belongs to the set of ground truth. To ensure the comparability of the participants, each participant gets the same number of images from the different models. As in the first test, there is a kind of shift which images are assigned to which participant. To ensure that each image is used in each colorization, the allocation is made as seen in Figure 4. With this allocation four participants are needed so that each image of the models is presented to one participant.

D. Metrics

Class Accuracy: The class accuracy metric (ACC) is a widely used metric for colorization that was also used by Zhang et al. [26]. For the class accuracy metric (ACC) the colorized

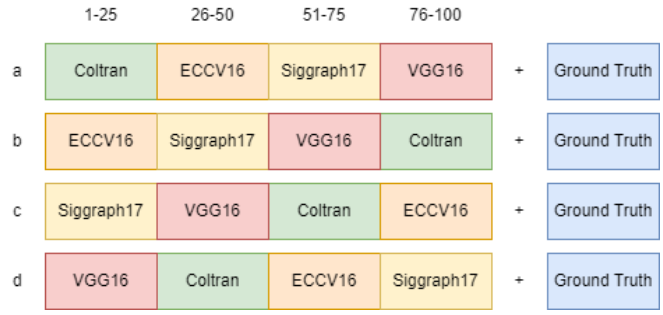


Fig. 4. Allocation of the images of the second user test.

images are fed into a pretrained classification network[1] which was trained on the ImageNet Dataset to predict classes. This accuracy of the model will be compared with the accuracy of the prediction with the ground truth images. This shows if the colorized images still hold their semantic information. The higher the class accuracy, the closer the created images are to the original.

Pixel Distance: The pixel distance (PixelDist) is used to quantify how closely the predicted image ($\hat{I}$) matches the ground truth (I). We calculate it as the mean absolute error (MAE)

$$\text{PixelDist} = \frac{1}{HW}\sum_{x=0}^{W}\sum_{y=0}^{H}|I_{x,y} - \hat{I}_{x,y}| \tag{1}$$

with H being the height and W being the width of the image.

Convolutional Distance: The convolutional distance (ConvDist) compares the output matrix M and $\hat{M}$ of the last convolutional layer from the classification network.

$$\text{ConvDist} = \frac{1}{CHW}\sum_{c=0}^{C}\sum_{x=0}^{W}\sum_{y=0}^{H}|M_{c,x,y} - \hat{M}_{c,x,y}| \tag{2}$$

with H being the height, W being the width and C being the channel of the matrix M.

Logit Distance: The logit distance (LogitDist) is the measure of the dissimilarity between two probability distributions. We use it to evaluate the difference between the predicted vector of the colorized images $\hat{Y}_{col}$ and the predicted vector of the ground truth images $\hat{Y}_{gt}$.

$$\text{LogitDist} = \frac{1}{N}\sum_{y=0}^{N}|\hat{Y}_{col} - \hat{Y}_{gt}| \tag{3}$$

with N being the number of classes.

[1]pretrained VGG16 classification network. This differs from the VGG16 used in image colorization.

Fréchet Inception Distance Metric (FID): The FID was first introduced by Heusel et al. [12] which measures the difference between the Gaussian with the mean (m_{col}, C_{col}) calculated from the probability of the model generated colorized images and the Gaussian with the mean (m_{gt}, C_{gt}) calculated from the probability of the ground truth images.

$$d^2((m_{col}, C_{col}), (m_{gt}, C_{gt})) = \\ \|(m_{col} - m_{gt})\|_2^2 + Tr(C_{col} + C_{gt} - 2(C_{col}C_{gt})^{\frac{1}{2}}) \quad (4)$$

We implemented the FID with the python library[2] from Maximilian Seitzer [22].

Peak Signal-to-Noise Ratio (PSNR): PSNR [15] is a commonly used metric for measuring the quality of reconstructed or compressed images. PSNR expresses a ratio between the maximum possible value for a pixel and the power of distorting noise.

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad (5)$$

with MAX_I being the maximum possible pixel value of the image and MSE being the mean squared error between the original image and the colorized image. A higher PSNR value indicates better image quality, as it means that the reconstructed or compressed image has a lower level of noise relative to the original image.

Structural Similarity Index Measure (SSIM): SSIM [11] is an attempt to predict the preceived quality of a image.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (6)$$

The SSIM is calculated by comparing the mean, standard deviation, and covariance of the pixel values in local windows of the images, and applying non-linear transformations to these values to reflect the perceptual sensitivity to luminance and contrast changes. The resulting SSIM index ranges from -1 to 1, where 1 indicates perfect similarity and -1 indicates complete dissimilarity.

V. Results

A. User Test

The results of the user tests are listed in Table I. The first user test shows that 80% of the images were correctly recognized as real images. The colorized images which fooled the participants the most were produced by Coltran and Siggraph17 with 57% each. VGG16 performed worst with 30%.
The results of the first user test have to be treated carefully since 80% of the shown images for a participant are colorized images from the models. This fact is not known by the participants. Therefore, it can be assumed that a user tries to generate the same number of votes for both classes. With this knowledge, the recognition rate of the ground truth is surprisingly low.
The second user test shows convincing results with Coltran at 34% and Siggraph17 at 31% considering that 50% would be the optimal result where a user can no longer distinguish between real images and colorized images.
The outcome of the second user test confirms the results of the first user test.

Network	User Test 1 labeled as real (%)	User Test 2 labeled as real(%)
Ground Truth	80	-
Coltran [16]	**57**	**34**
Siggraph17 [27]	**57**	31
ECCV16 [26]	47	28
VGG16 [10]	30	24

TABLE I
Results of the user test

Figure 5 shows the three images which users thought to be real, as well as the three images which users always labeled as fake. It is noticeable that the images of the networks differ. Generally, Coltran has more confidence in predicting colorful images. While Siggraph17 and VGG16 seem to be more confident if the scene has lower saturated colors.
It can be seen that there are even real images which subjectively look fake. If there is no semantic meaning in the real image, a user could have problems to determine if colors are realistic for the given image. Another problem could be that real images recorded different as expected or with poor quality (e.g. a bad illumination or different colors due to weather conditions).
Moreover, it is obvious that there are some colorized images by the networks which are unrealistic.
To show statistical significance of the user tests, we provide the null hypothesis which would state that the proportion of correct responses for AI generated images is equal to the proportion of correct responses for real images. This null hypothesis assumes that any differences in the proportion of correct responses are due to chance or random variation. The alternative hypothesis we provide claims that there is a significant difference in the ability of users to correctly distinguish between colorized and real images.

H_0: The users take random guesses if the images are real of fake, hence we expect a mean of the total uservotes for one image divided by two.

H_1: The users can either distinguish between real and fake images or can not distinguish between real and fake images.

Utilizing the chi-square test of independence χ^2 introduced by McHugh et al. [18] we calculate the χ^2-value with the following formula:

$$\chi^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i} \quad (7)$$

We can not provide guarantee that the colorized images will

[2]https://github.com/mseitzer/pytorch-fid

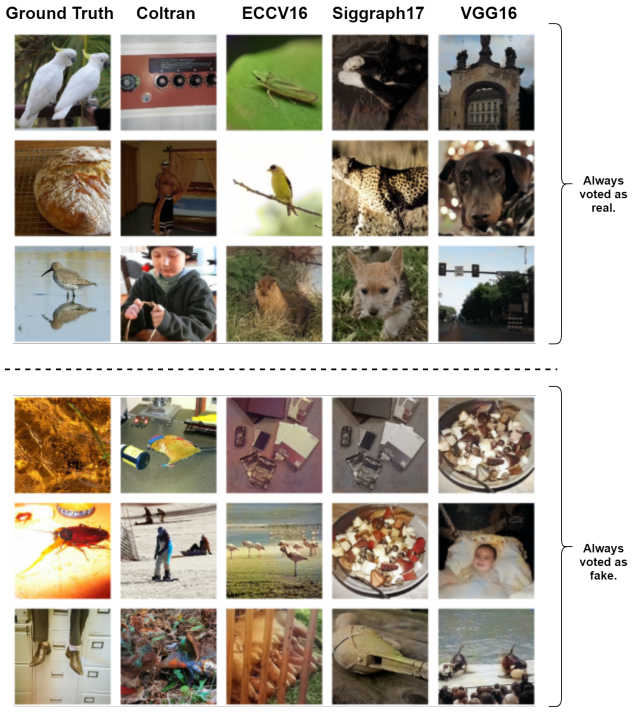


Fig. 5. Top three images which where either always voted as real or recognized as fake.

get less votes than the real images. Therefore we consider a two tailed test to cover both outcomes.
Considering a level of significance of 0.1 [17] and considering both a right-tail (0.05) and a left-tail (0.05) critical value we can reject the null hypothesis if for the first user test the χ^2-value is in between 448.20 and 552.07. For the second test the null hypothesis can be rejected if the χ^2-value is in between 316.09 and 404.18. By plotting the values within the χ^2 distribution[3] in Figure 6 we can see that the null hypothesis can neither be rejected for the first user test nor for the second user test.

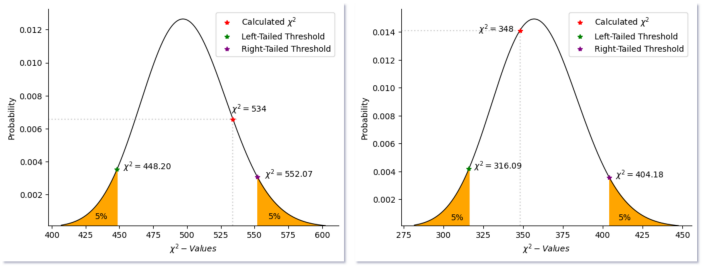


Fig. 6. Chi squared distribution for user test 1 (left) and user test 2 (right)

The predefined level of significance could not be achieved for both user tests, nevertheless this does not exclude statistical significance. It rather indicates that the chance this data could be observed randomly is higher than we would accept. In the following parts we will evaluate these results, with remark to the results from the test for statistical significance.

[3] $f(x;k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$

Due to the results of the statistical significance we will only consider the first user test in the following sections, since it came closest to the predefined value of significance.

B. Metrics

After we concluded the user test, the metrics described in chapter IV-D are applied to the whole validation ImageNet dataset described in the chapter IV-B. Investigating the class accuracies in Table II shows that the networks generally perform slightly better than gray scale images, but much better than images, which received a color channel from a random image. The ranking of the results almost coincides with the user tests. Applying this metric, it can be said, that all networks preserve the semantic meaning.

	Class ACC
Ground Truth	0.69
Coltran [16]	0.56
Siggraph17 [27]	**0.58**
ECCV16 [26]	0.57
VGG16 [10]	0.56
Gray	0.55
Random	0.45

TABLE II
CLASS ACCURACY BASED ON A CLASSIFICATION MODEL PRETRAINED ON THE IMAGENET DATASET.

We obtained the fréchet inception distance metric on the whole validation dataset described in chapter IV-B. This is important to know due to the fact that the numbers of data has an influence on the results. The results can be seen in table III.
Coltran achieves the best FID-score with 4.55, followed by Siggraph17 with 5.73. The worst results of the models are again obtained by VGG16 with 9.72. The results of the FID are consistent with the user test.
It can be seen that the grayscale images and images that received a color channel from a random image have poor results.

The SSIM and PSNR metrics show interesting results where Siggraph17 does seem to outperform all other networks. It can be seen that the PSNR metric for all networks and even for the gray and random set are close together and do not carry any information about the quality.

	FID	SSIM	PSNR
Coltran [16]	**4.55**	0.62	30.14
Siggraph17 [27]	5.73	**0.72**	**31.01**
ECCV16 [26]	7.29	0.64	30.02
VGG16 [10]	9.72	0.64	30.33
Gray	21.46	0.58	30.90
Random	16.64	0.45	29.50

TABLE III
RESULTS OF FID, SSIM AND PSNR.

In Table IV we calculate the different distance metrics mentioned in chapter IV-D. The two networks Siggraph17 and VGG16 achieve the smallest pixel distance with 0.20 and

0.23 on the dataset. We already expected this result, since both networks minimize pixel distance during training. At pixel distance, the grayscale images achieve better results than Coltran. This shows that a low pixel distance does not lead to quality images. Generally, the Siggraph17 achieves the best results over pixel distance, convolutional distance and logit distance.
It can be seen, that VGG16 only performs in pixel distance and get worse if the distance metrics are getting more abstract. With Coltran it is the exact opposite compared to the other models.

	Distance		
Network	PixelDist	ConvDist	LogitDist
Coltran [16]	0.27	0.25	0.89
Siggraph17 [27]	**0.20**	**0.22**	**0.82**
ECCV16 [26]	0.24	0.23	0.85
VGG16 [10]	0.23	0.59	1.27
Gray	0.25	0.23	0.93
Random	0.35	0.31	1.22

TABLE IV
THE DIFFERENT DISTANCE METRICS, PIXEL DISTANCE, CONVOLUTION DISTANCE AND DISTANCE FROM THE OUTPUT VECTOR.

C. Metric Correlation

We calculated the correlation between the distance metrics and the user test for the images used in the user test. This is not applicable with the FID, because this metric does not output a value for a single image. In table V it can be seen, that there are only slight correlations between the results of the user test and the distance metrics which also depend on the respective network. The strongest correlation is found with the Coltran network between the user test and the convolutional distance and the logit distance of 0.27 and 0.35. It can be seen that VGG16 and Siggraph17, which minimizes the pixel distance during training, have a higher correlation with pixel distance than with convolutional or logit distance.

Network		Correlation to ACC User Test 1
Coltran [16]	Pixel Distance	0.06
	Conv Distance	0.27
	Logit Distance	**0.35**
Siggraph17 [27]	Pixel Distance	**0.14**
	Conv Distance	0.12
	Logit Distance	0.07
ECCV16 [26]	Pixel Distance	0.17
	Conv Distance	**0.18**
	Logit Distance	0.17
VGG16 [10]	Pixel Distance	**0.15**
	Conv Distance	-0.03
	Logit Distance	-0.03

TABLE V
CORRELATION BETWEEN THE USER TESTS AND DISTANCE METRICS.

In addition, we sorted the images according to the distance metrics. The results can be seen in the diagram in Figure 7. The y-axis of the graph shows how well these images performed in the first user test. To smooth the graph, we took the average of 10 images, otherwise the noise would have been too large. A perfect metric that correlates with the user test should show a monotonic decreasing curve. For most networks, the metrics show a falling curve, which means that on average the images perform similar to the user test. If one takes the 10 images from Coltran or Siggraph17 that have the smallest convolutional distance, these images were labeled as real by about 75%. If, on the other hand, one takes the 10 images with the largest convolutional distance, the rate is only about 35%. This trend can also be seen with logit and pixel distance as well. It can also be stated that it depends on the particular network. With the VGG16 network there is only a small trend with the user test in pixel distance, the other distance metrics show no trend at all.

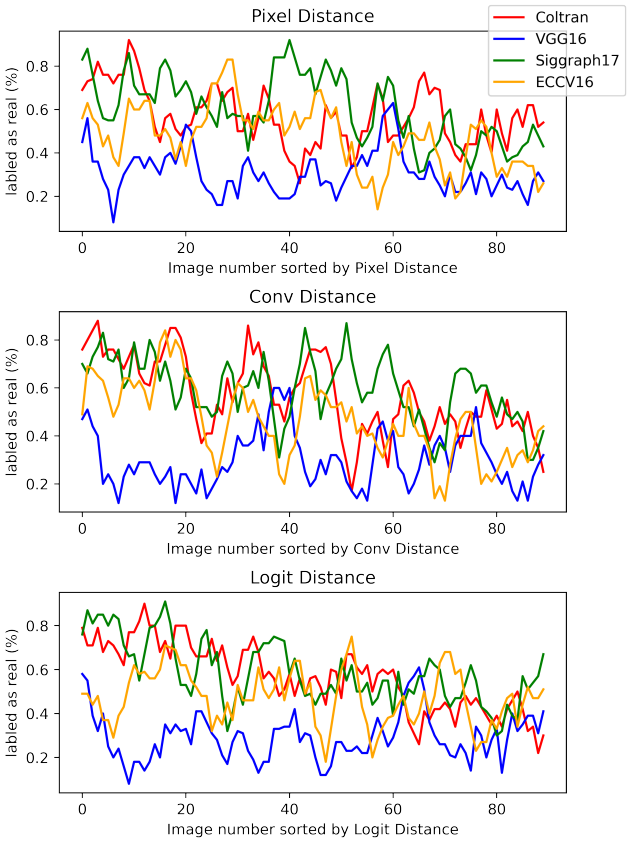


Fig. 7. The pictures are sorted according to the respective metric. The 10 pictures (5 before and 5 after) are considered together, as they performed on average in the user test 1.

VI. CONCLUSION

Metrics for colorization were applied to the colorized images of four models and compared with user tests.
The FID metric is most consistent with the results of the user tests. However, the problem is that the FID cannot be applied to individual images. Therefore, it makes most sense to use the FID at the end to evaluate the network.
If you want to evaluate a single colorized image, the convolutional, the logit distance metric or the SSIM are the most suitable.
It has been shown that metrics that minimize pixel distance perform better on pixel distance than on convolutional or logit distance, but low pixel distance is not an indication of whether the image has been colorized well.

We have shown that the performance of metrics in colorization methods depend on the type of network.
Due to the small number of users in the user test, we cannot make a clear statement about individual images, but we can make an overall statement about all images and identify a trend. This tendency would be even more evident with a larger target group, which would increase the significance of the results.

REFERENCES

[1] Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image colorization: A survey and dataset. *arXiv preprint arXiv:2008.10774*, 2020.

[2] Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[4] Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. Unsupervised diverse colorization via generative adversarial networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 151–166. Springer, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6837–6845, 2017.

[7] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1462–1471, Lille, France, 07–09 Jul 2015. PMLR.

[8] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. Pixcolor: Pixel recursive colorization, 2017.

[9] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2013.

[10] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization, 2014.

[11] Mohammed Hassan and Chakravarthy Bhagvati. Structural similarity measure for color images. *International Journal of Computer Applications*, 43(14):7–12, 2012.

[12] 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. *Advances in neural information processing systems*, 30, 2017.

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[15] Don H Johnson. Signal-to-noise ratio. *Scholarpedia*, 1(12):2088, 2006.

[16] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *International Conference on Learning Representations*, 2021.

[17] Sanford Labovitz. Criteria for selecting a significance level: A note on the sacredness of. 05. *The American Sociologist*, pages 220–222, 1968.

[18] Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

[19] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Amelie Royer, Alexander Kolesnikov, and Christoph H Lampert. Probabilistic image colorization. *arXiv preprint arXiv:1705.04258*, 2017.

[21] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.

[22] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

[23] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE transactions on medical imaging*, 39(1):188–203, 2019.

[24] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[25] Ivana Žeger, Sonja Grgic, Josip Vuković, and Gordan Šišul. Grayscale image colorization methods: Overview and evaluation. *IEEE Access*, 9:113326–113346, 2021.

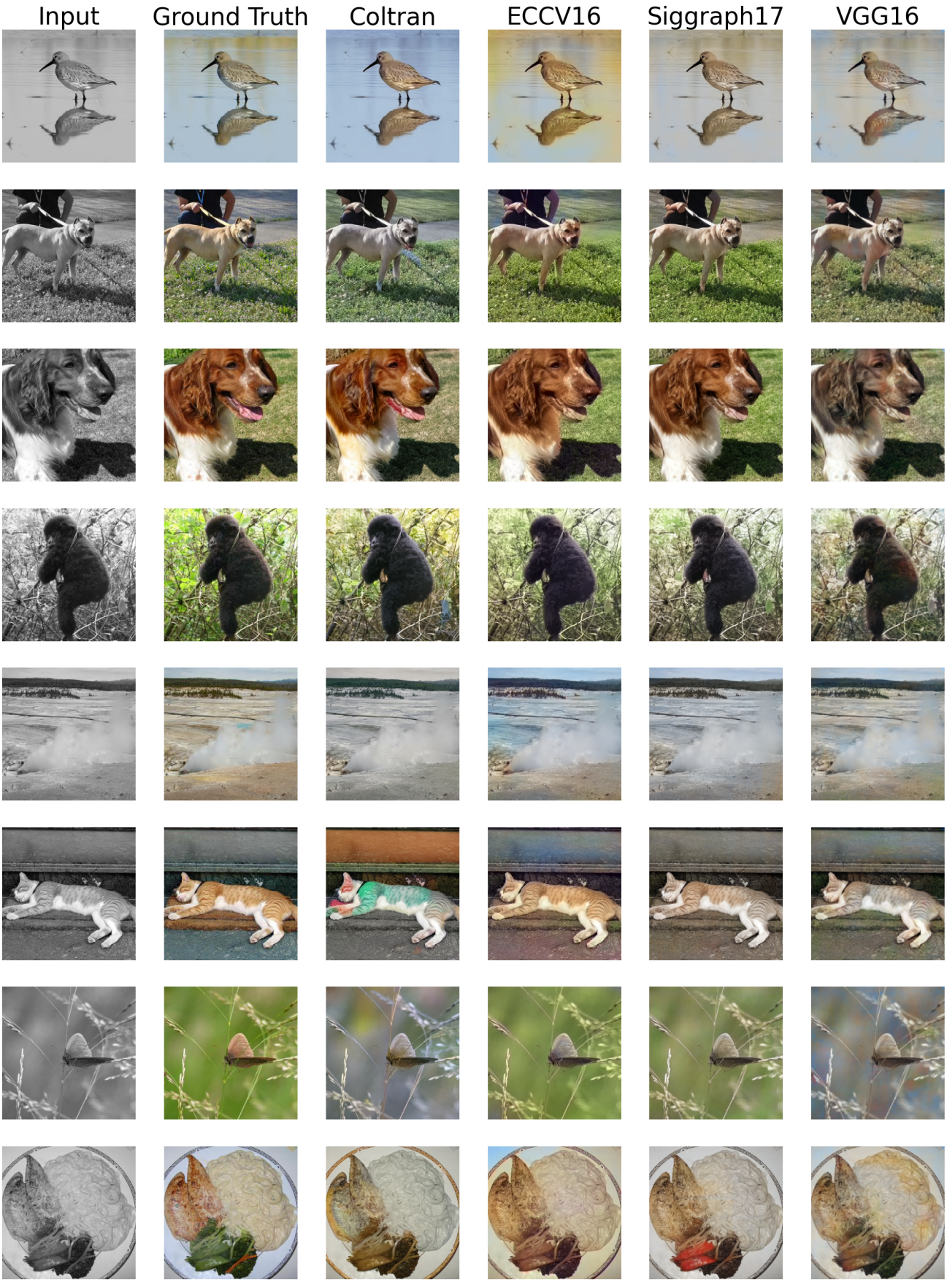
[26] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

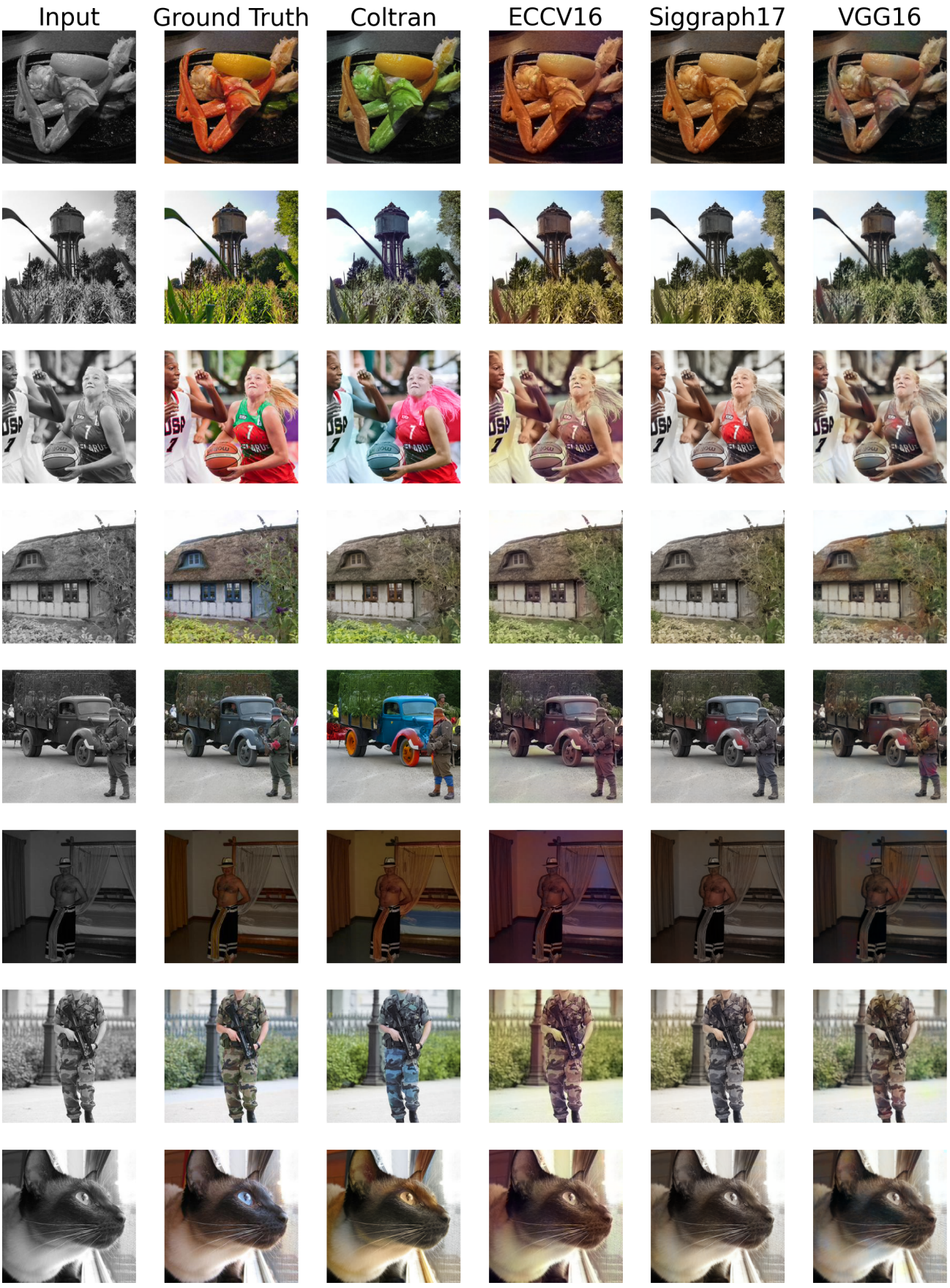
[27] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017.

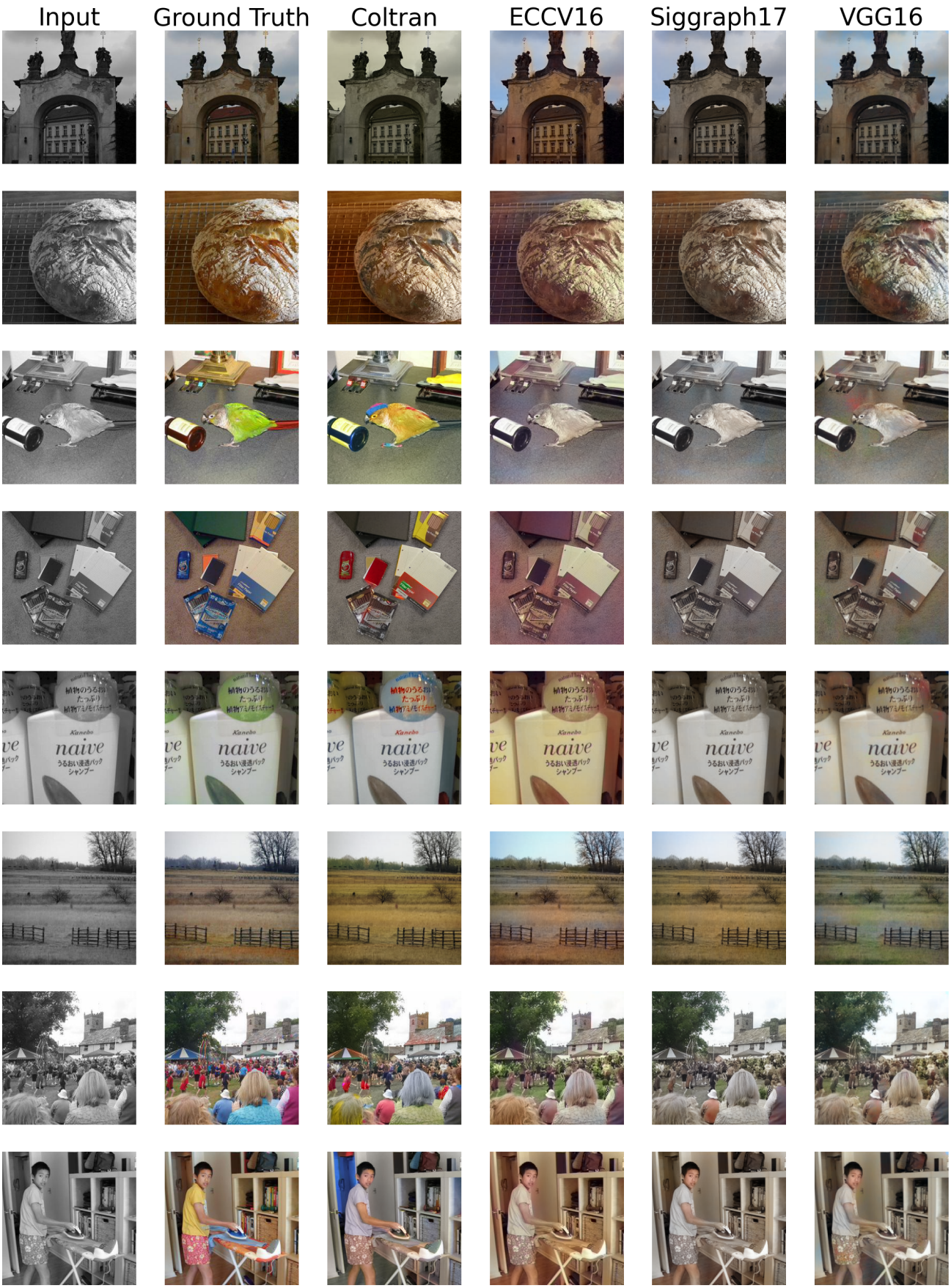
ACKNOWLEDGMENT

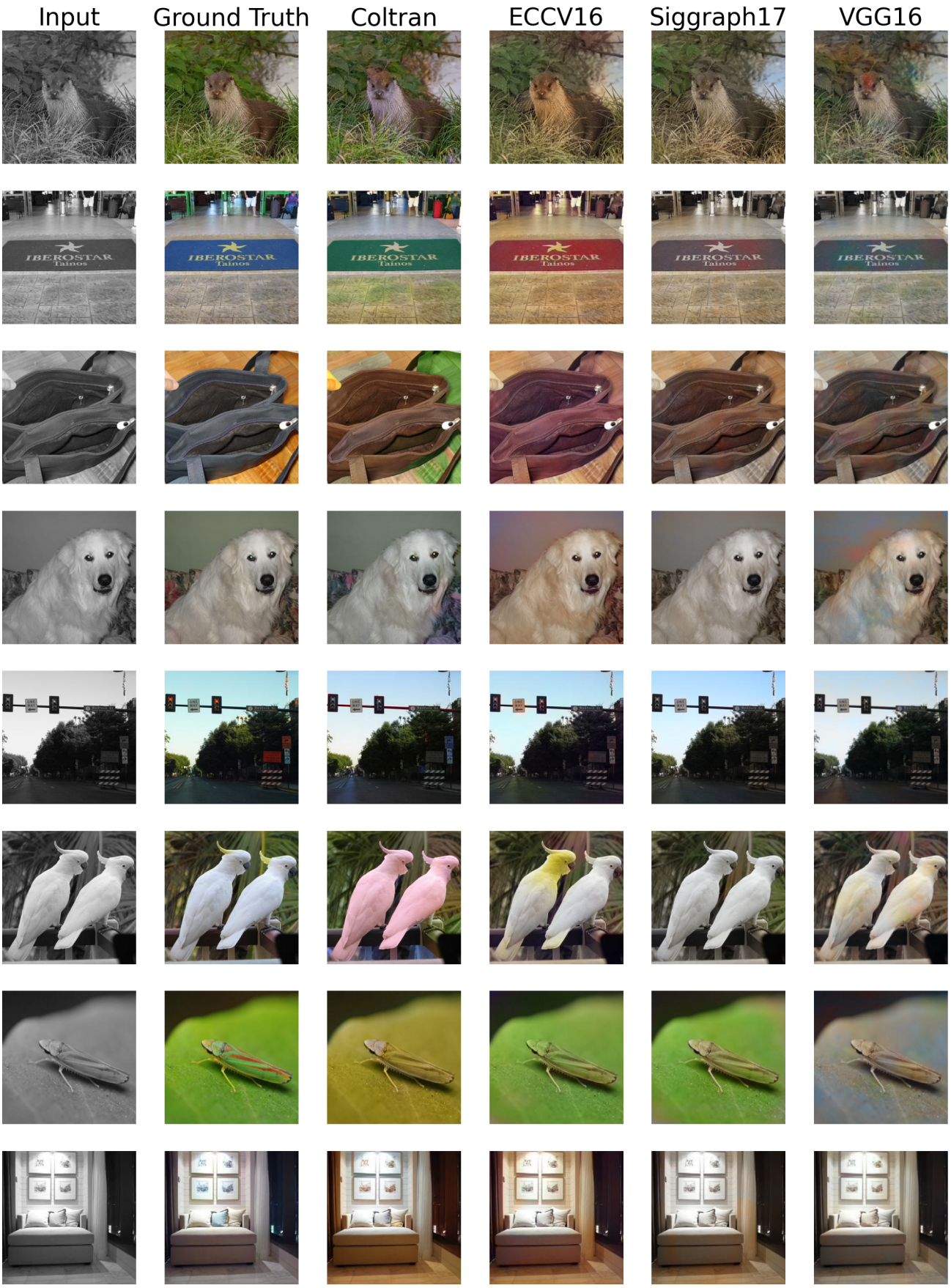
We want to thank Prof. Dr. rer. nat. Markus Schneider from the Ravensburg-Weingarten University of Applied Sciences for the assistance and his valuable time during our research. Furthermore our thanks goes to all of the participants of our user test. Lastly we want to thank the authors of the networks we used.

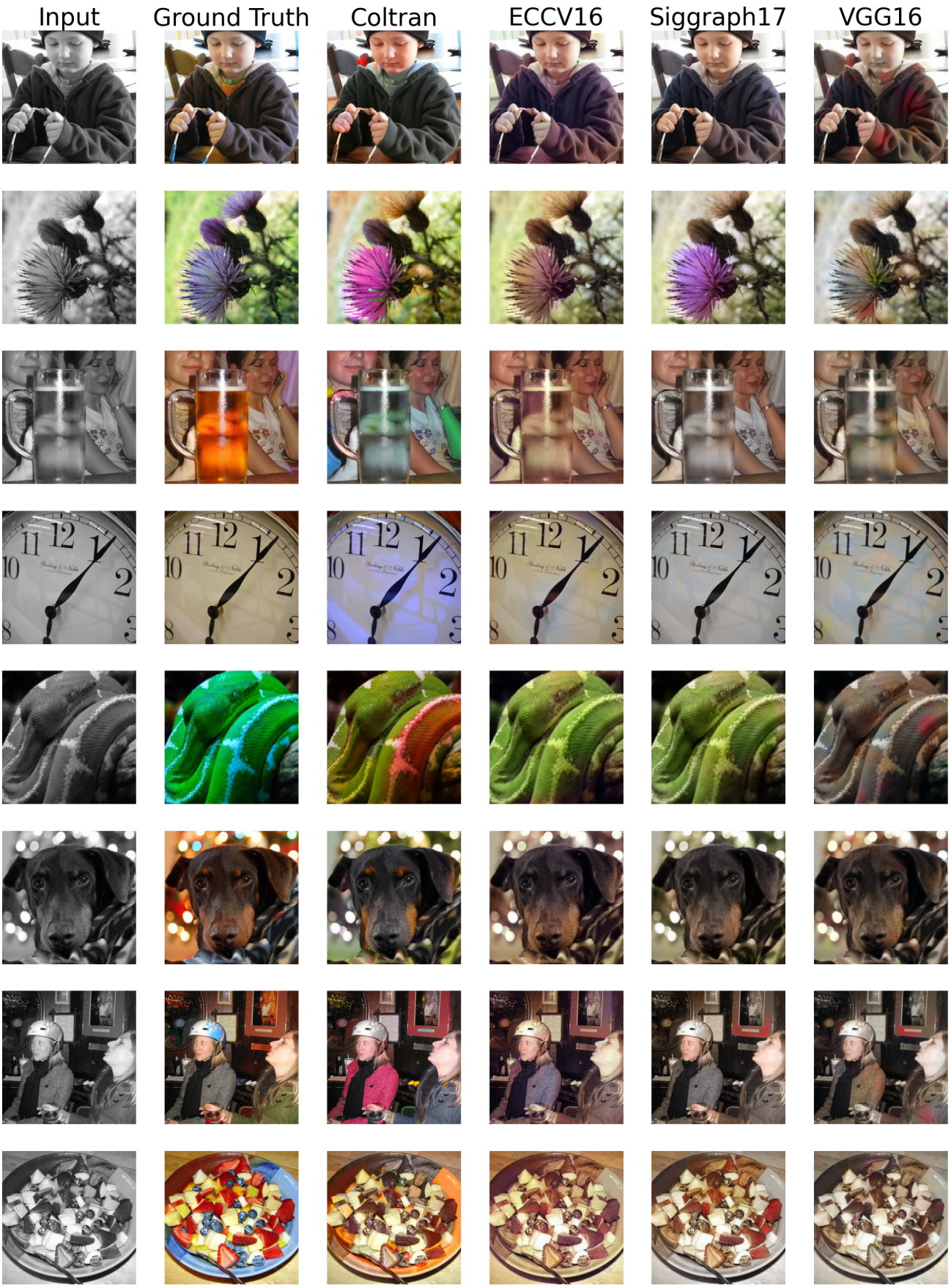
APPENDIX

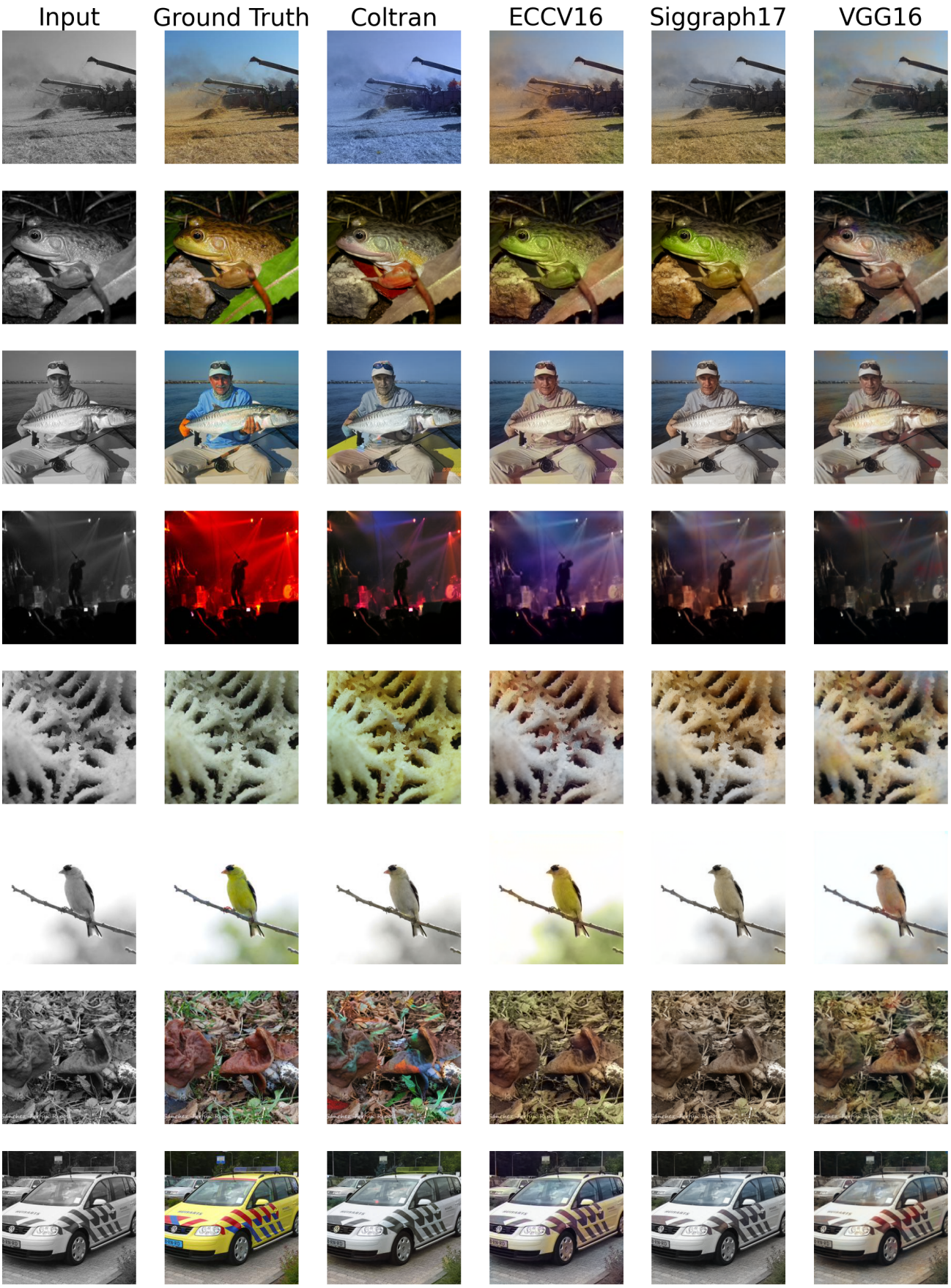


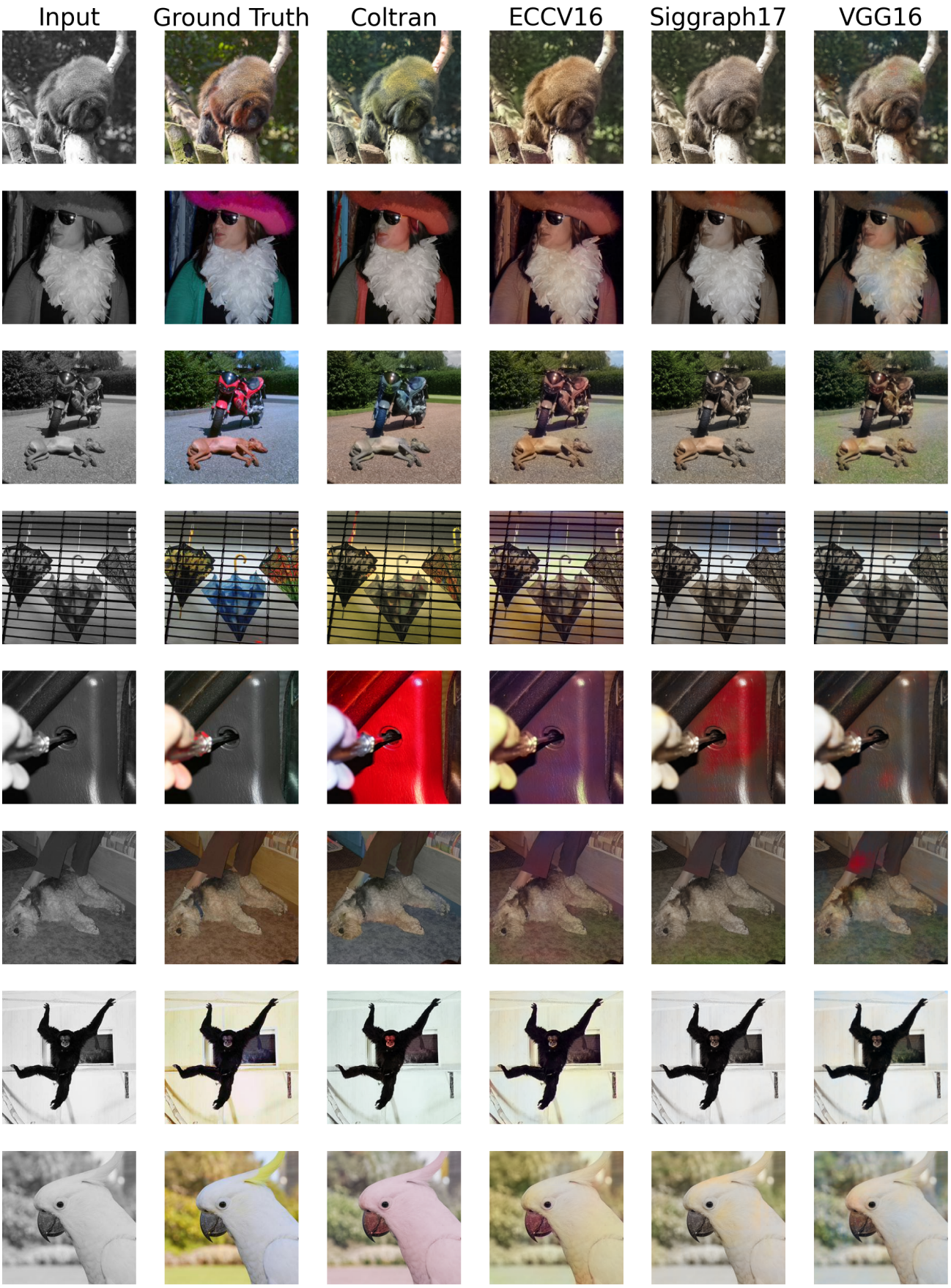

Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16

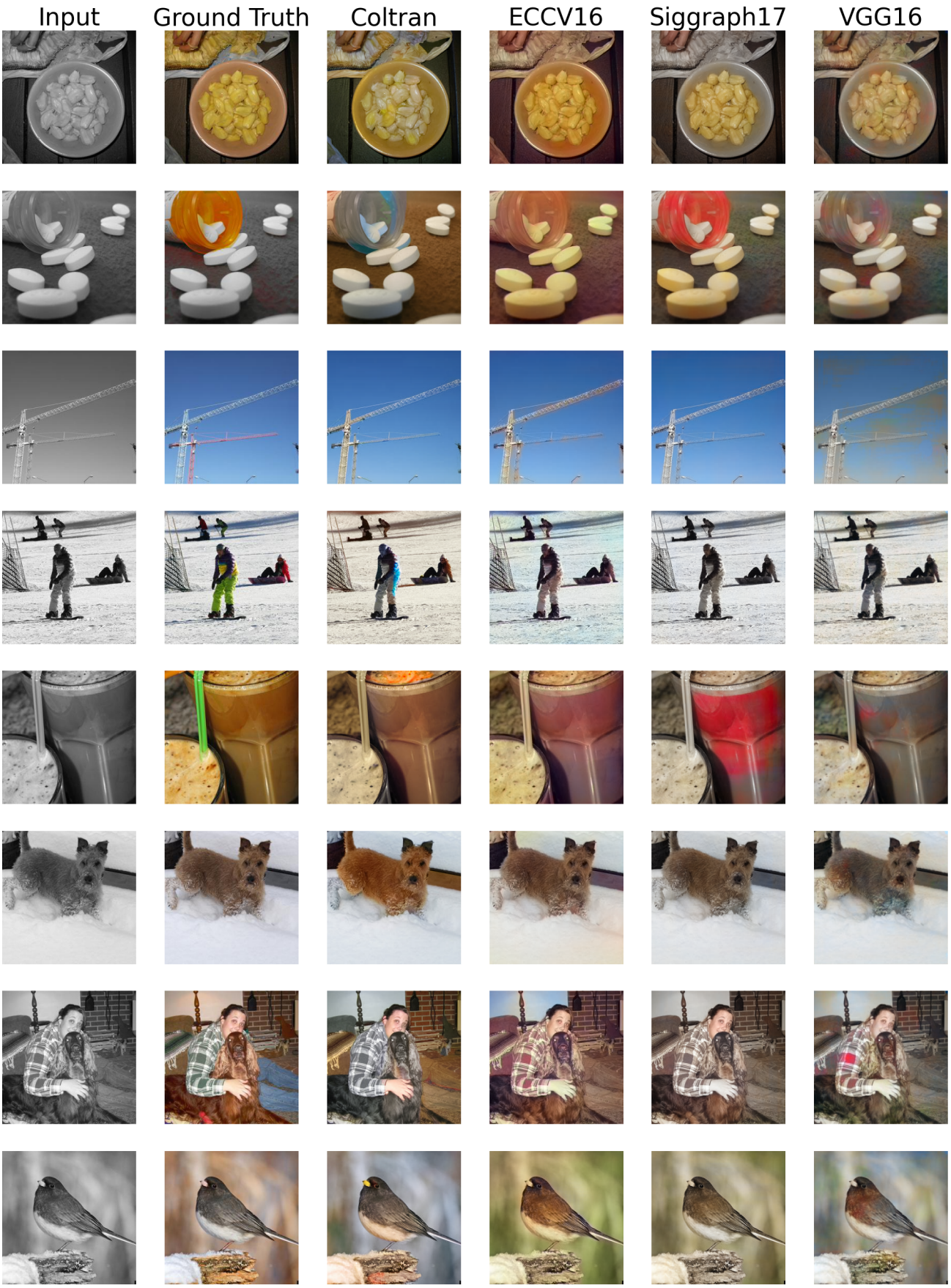

Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16

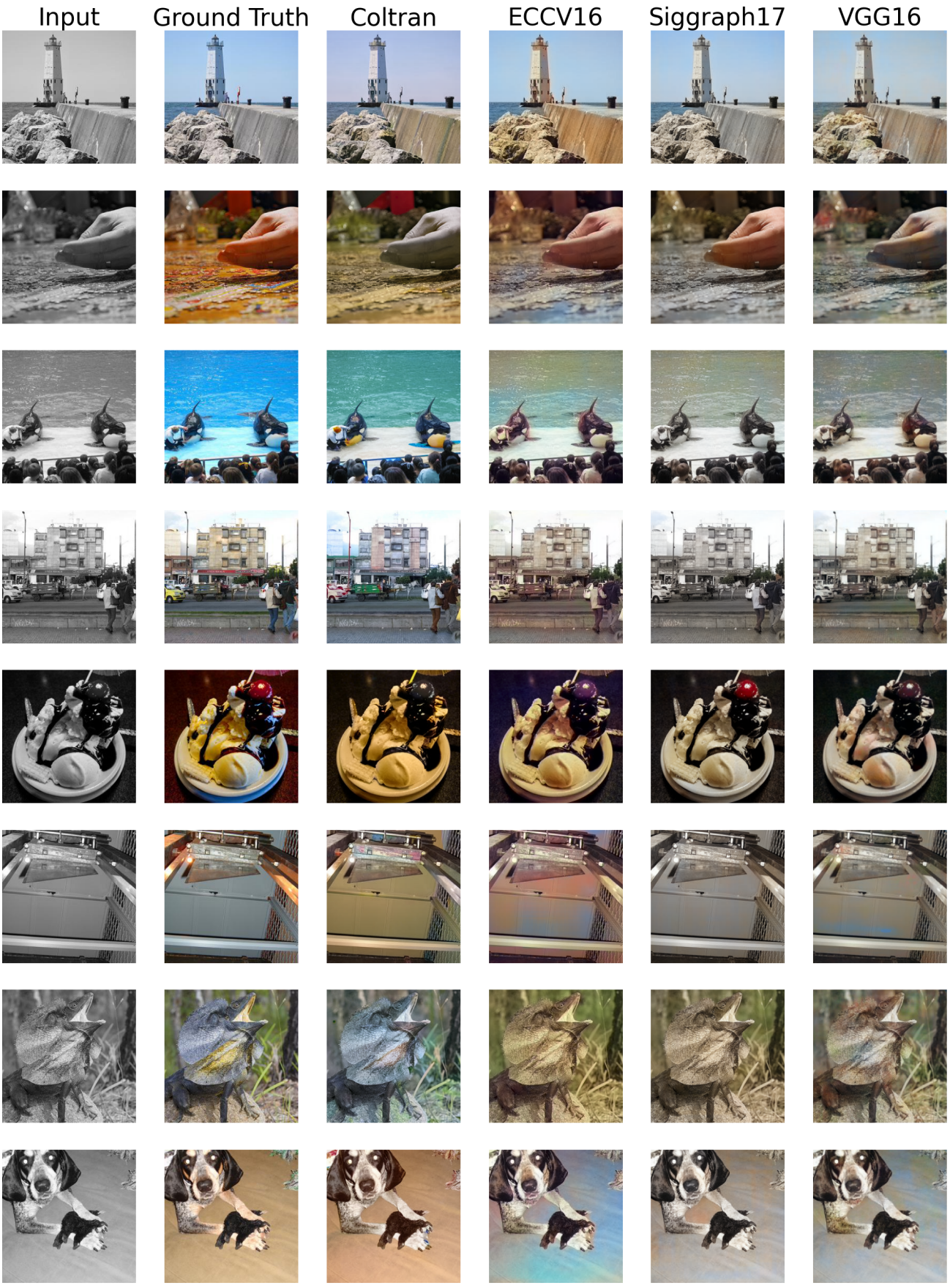

Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16

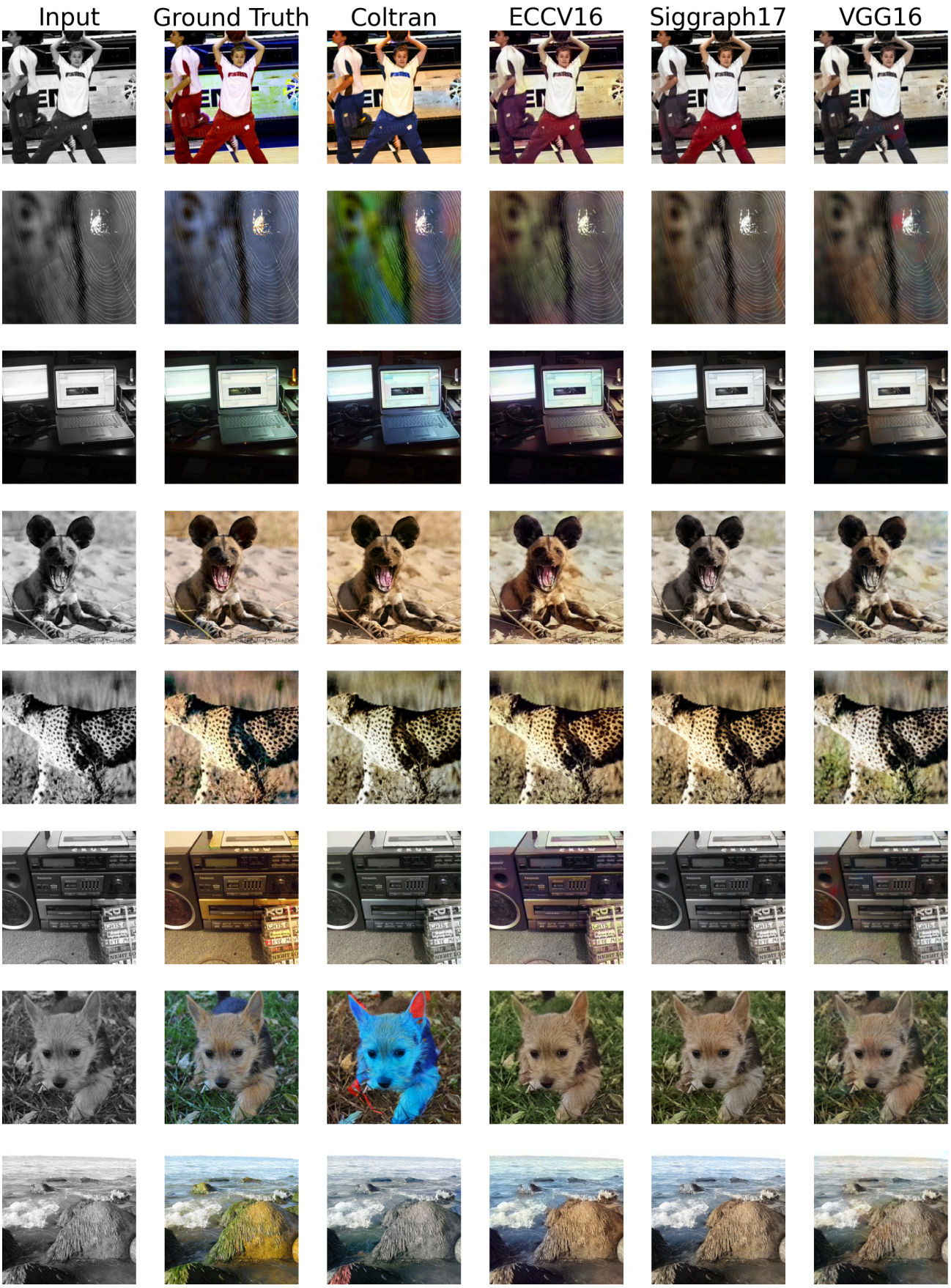

Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16

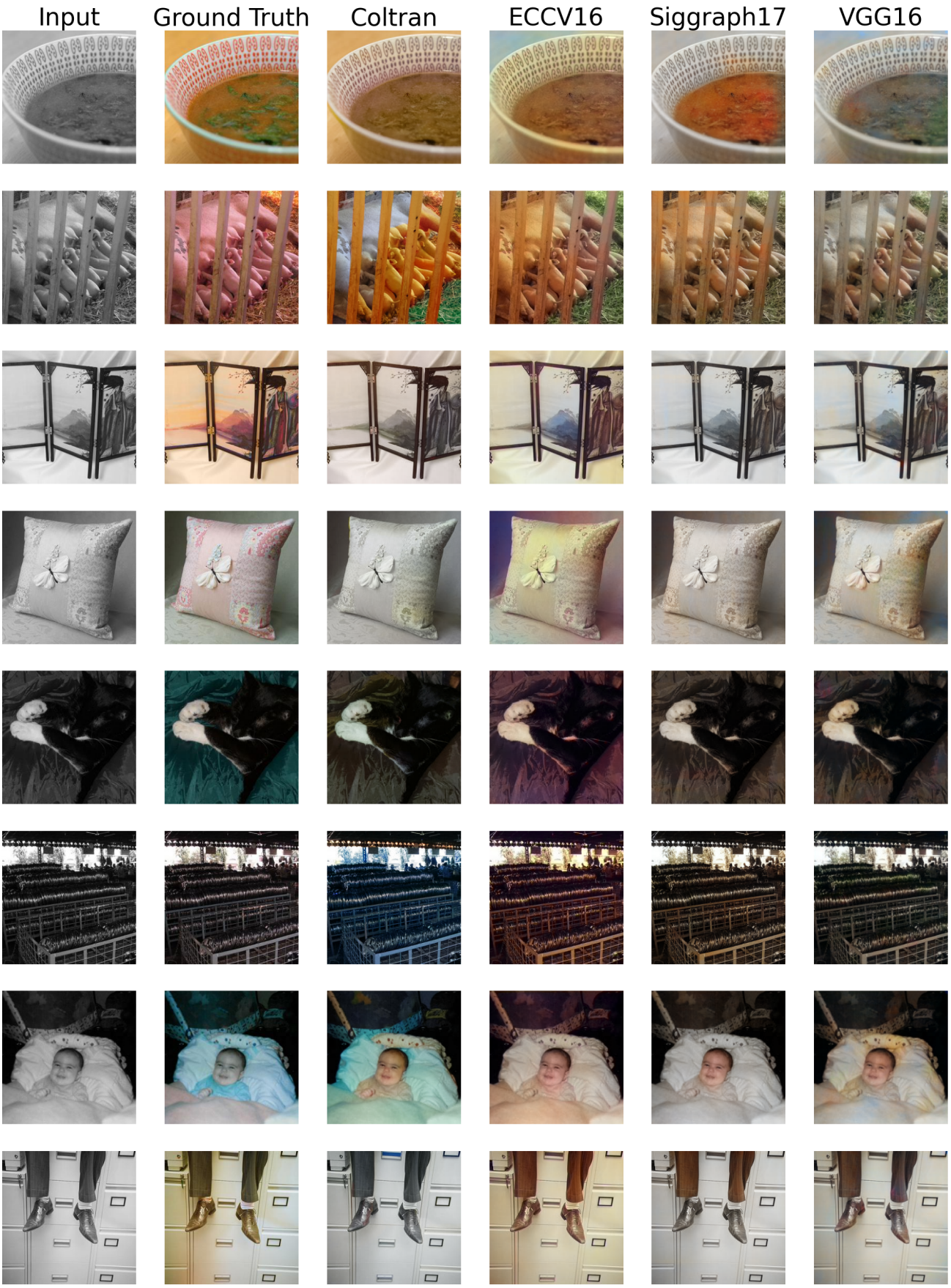

Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16


Input
Ground Truth
Coltran
ECCV16
Siggraph17
VGG16