Learning to Enhance Images From Human Feedback

Carly Atwell
Department of Computer Engineering
University of Utah
carly.atwell@utah.edu

Victor Petrov
Kahlert School of Computing
University of Utah
victor.petrov@utah.edu

Abstract

Image enhancement takes time and skills that are costly, so in this paper we propose a way to train a reward model to be able to automate the process of editing images. The model is trained on human preference data over different sets of images with various filters applied to them. We cover some past work in the area of automated image enhancement and detail our methods, experimental design, and results, and demonstrate some success on models trained with synthetic implicit preference data for performing image enhancements that mimic user behavior.

1 Problem Statement

We want to develop an agent that performs basic image enhancement by applying a set of filters to an input image. Image editing and enhancement is applicable in many domains. We see its use quite obviously on social media and websites, whether for artistic reasons or marketing motivations, but also for more practical uses where higher contrast or better visual accessibility may be necessary. Editing images can be a time consuming process, however, and it requires aptitude with elements of design and skills with image editing tools. While an in-depth knowledge of design and image components may not be everyone's forte, any person could look at some images and tell you which ones look better than the others. This presents an opportunity to automate the process of image enhancement, training a machine learning model based on feedback from what humans think looks good.

For this project, we are more concerned with the application in general image enhancement for social media or photography, etc. On a broader scale, this concept of being able to automate image editing with a specific human preference could be applied to other domains like medical imaging, or underwater photography as we will see in the related work section. For whichever domain or application is chosen, the same framework could be applied by simply altering how pairwise preferences are chosen by the human and what kind of images the model is trained on.

2 Related Work

Formulations of image enhancement as an RL task are numerous. In an early work Farhang and Hamid (2003) formulated image enhancement as a weighted sum between an input image and the output of that input filtered on a preconfigured set of filters. At each step, the agent can take an action of adjusting one of these weights, and receives a reward by direct input from a human overseer. This means, however, that a considerable amount of input is necessary from the user, and the filtering is only done in a single-step, where filters cannot be repeatedly applied. The output is also a weighted combination of filters, rather than a direct sequential application of these filters. The policy's state definition is also composed not of the full pixel data of the image, but rather a series of crude metrics computed on the current output, such as the intensity of noise and edge sharpness, limiting its utility.

Class Project for CS 6960: Human-AI Alignment at the University of Utah, Fall 2022.

Kosugi and Yamasaki (2020) highlight a more advanced GAN-based approach, where the generator is an RL algorithm rather than a traditional neural network, leveraging this formulation to avoid problems with purely neural-network based approached that suffer from certain artifacts such as generated by CNNs and autodecoders, which require limiting to only certain filters or generating superresolution images with careful downscaling to avoid, which is undesirable. Their RL-based approach integrated with Adobe Lightroom, where the action output of the policy is the strengths of a series of preselected enhancements in the software. As with the prior work, however, this only permits a single step of transformations, as the filters are only configured and applied once, and have a predefined order to their application, as the action taken configures all the parameters of the set of filters at once. We want to apply any number of filters, one at a time, without restrictions to their order.

Sun et al (2022) use image enhancement in the context of underwater photographs, aiming to remove color distortion, light scattering, and other undesirable phenomenon for processing the image data. Their work avoids the problem of a single-step transformation, instead learning an action sequence over the entire set of possible single-step enhancements, as our work seeks to do. However, their model uses heuristics to determine the reward signal, with no basis in human input.

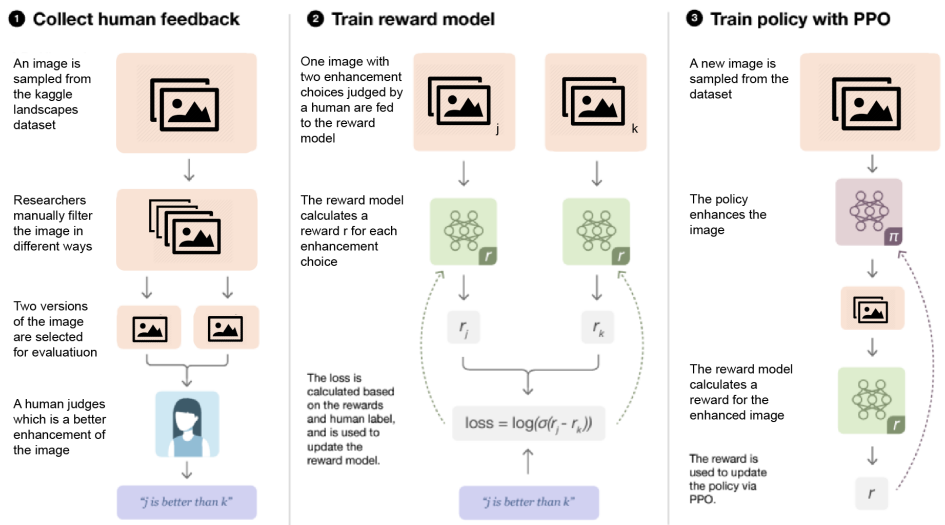
This work also highlights various areas where image enhancement is useful beyond generally increasing subjective aesthetic quality. Similarly, Minh et al (2021) demonstrate using an RL agent to automatically process image data for deep learning, learning to mimic data scientists' intuitive transformations of images to make them better suited for training. Thus, while our context is purely within general image aesthetics, it can be retrained for any given context and provide value across a variety of domains.

Lv et al (2023) train a policy to improve image aesthetics, where the agent selects as an action an enhancement and configuration of its properties as an action, and thus also allows a sequence of enhancements. The process is user-guided, where a user must at each step modify the selected set of transformations if they are unhappy with the output. This second step can be expensive and taxing on the user, and thus differs from our preference-based approach which requires less user interaction in order to train the policy, and does not require online user action for policy training by instead training the reward rather than direct policy from the user's inputs.

Our work therefore combines and expands on prior literature by specifically: Leveraging human input through pairwise preferences to train a reward model approximating human desires, across a trajectory of actions. Training a policy that iteratively selects and applies filters, to arbitrary length, in arbitrary order.

The prior works highlighted do not build their models satisfying both of these desires.

3 Methodology



We are using the framework presented in 'Learning to summarize from human feedback' to train an RL agent to enhance images with a reward model based on human feedback in the form of pairwise preferences over different sets of filters on images, using the LHQ1024 JPG landscape photo database. It builds off of past work in RL image enhancement by adding the human-in-the-loop component of pairwise preferences to train the reward model.

We use filters from Pillow's Image Enhance Module, including options to increment or decrement brightness, contrast, color balance, and sharpness. We created a simple GUI for applying filters to images to generate our training data and another for collecting the human preference data. This preference information is then used to train a reward model which is in turn used to train the policy. The final policy returns whichever filter action it estimates to give the highest reward based on the image state, then iterates with more filters until the highest reward is the stop action or we reach the max number of possible filters.

3.1 Reward and Policy Models

We started with creating our own convolutional nets for both the reward model and the policy. The results it produced were far from ideal, so we switched to modifying a known pretrained model, Alexnet. From Alexnet, we kept the feature extractor and froze its weights so that it wouldn't be modified by gradients, then tacked on a linear network that transformed to the 9 actions.

4 Experiments

For the experiment, we collect human preference data two different ways to train the reward model and compare the results. In the first part, we set up the experiment by manually filtering a set of images and generating the pairs of images to be judged and chosen between by a third person. 230 images were manually filtered and preference generated by a third party for this experiment.

For the second method, we generate synthetic preference data from a user's filtering of a set of images. Assuming they applied some chain of filters $\{a, b, c\}$, there is an implicit preference of this filtered image over any other potential chain of filters, $\{a, c, b\}, \{c, a, b\}, \{d, e\}$, etc. Therefore we can generate these other permutations and create a synthetic preference of the user's filter chain over all others. This removes the need for any additional effort on the part of the user in selecting explicit preferences, using purely their data from previously filtering images, allowing such a system to be seamlessly integrated into existing photo editing applications. It also allows for generating a significantly larger amount of preference data from a small set of filtered images.

For generating implicit preferences, if there are k possible filters, with a maximum chain length of N, that is $\sum_{i=0}^{N} k^i$ possible permutations, which would quickly be too much data for training and likely generate garbage results. Therefore, we instead simply do a sampling of x random permutations of each possible length for our synthetic preferences, where x may be tuned to create more/less implicit data.

Further, we noticed that generating max-length permutations tends to cause a less desirable distribution over output actions, as it likely creates many long filter chains which are selected against in the preferences, thus placing too much weight on the 'STOP' action. We therefore also ran experiments training the model to only generate permutations of length 0 to $N = |T| + 2$, where $|T|$ is the length of the human-created filter trajectory. Thus it sees some chains that are longer, and creates preferences against them, but is not dominated by significantly longer filter data.

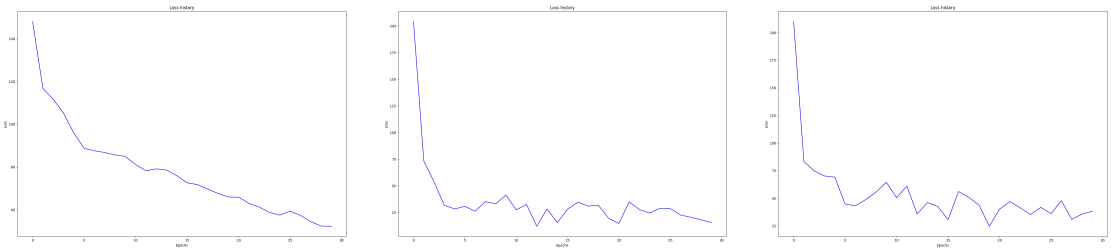
Using $x = 3$, we generated 9229 preferences for the regular implicit case, and 4102 preferences for the length-limited case.

The null hypothesis is that there is no significant difference in how these two forms of human feedback collection affect the performance of our image enhancement policy. The alternative hypothesis is that these two methods will produce different models, with one performing better or worse than the other on new images to filter. Performance is measured subjectively by a human evaluator on new test images.

4.1 Results

Our reward models were generally able to train well. For our models, we tracked the sum Binary Cross Entropy loss for each epoch, as well as produced a final accuracy metric by feeding in the ground truth data into the network after training, and tallying the proportion of preferences it was able to correctly match with the label.

For our models, the *explicit* achieved a final accuracy of 88.70%, *implicit* got 97.28%, and *implicit length-limited* got 95.20%, representing a good fit to the training inputs.



Explicit Implicit Implicit Length-Limited

Figure 1: Reward Network Training Loss Per Epoch

The policy networks were then trained using the reward networks to stand-in for expected human evaluation of the policy actions.

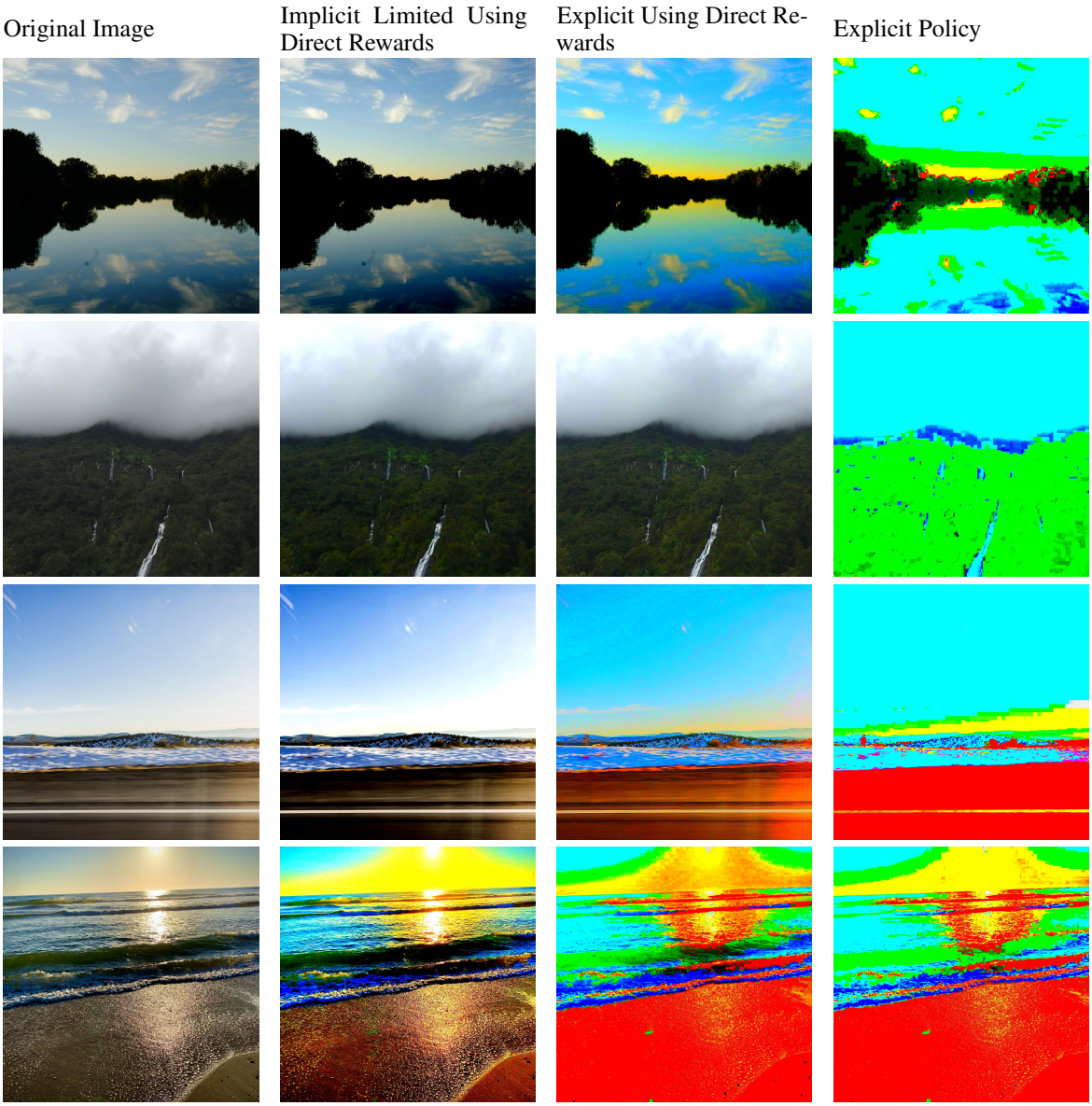
The policy network training ultimately produced sub-par results, with a network often getting stuck in a minimum learning to take only a single filter action repeatedly. For the explicit case, it would often for instance take the action for *increase color balance*, likely as a product of the human preference data rarely taking the *decrease color balance* action, causing the network to associate increasing color balance with nearly universally being a good action and thus the safest policy choice. For the implicit case it was slightly more robust from having trained on a better formed distribution of possible filter chains, however this would often train it to simply take the *stop* action immediately as a relatively safe choice against many possible poor rewards from taking other actions. This is technically the beginnings of a good policy, as we want to avoid overfiltering, but is too safe and therefore produces no useful results. We introduced the length-limited variant to try to combat this issue, but it still produced the same results.

However, given the reward network also outputs 9 values for each action like the policy network, we can just use it as a policy network directly as well to observe results. The reasoning here is that the highest activation in the reward output represents the best single-step filtering action that the network believes could be taken. This is technically a 'greedy' approach as it lacks any future planning, but this somewhat mimics how a human would filter an image as well, in taking a single action they think is best, evaluating the new image, and repeating the process, and could therefore be a good model for filtering behavior. Trying this direct approach netted considerably better results than with the policy network.

The explicit case still outputs poor results with overfiltering, with the network often just rewarding one of the actions as always being best, which explains some of the behavior of the policy network.

For the implicit models, the non-length limited only occasionally applies a filter, but usually almost immediately takes the *stop* action, likely influenced by the long synthetic filter chains being preferred against and thus making stopping early almost universally better. With the length-limited case, however, we get very nice results, with the network now learning a more useful relationship, leveraging all the possible filters to do enhancement, but also avoiding overfiltering and stopping before image data becomes significantly muddled.

The implicit policy and non-length-limited reward direct results are not shown, given they are identical to the original image (no filters applied).



The first column of edited images shows some examples of how the implicit model performed using the direct rewards. Notice on image 1 how it increased the sharpness to make the image less blurry and did a little bit of bumping contrast, overall looks nice but not blown out. On image 2, it pushed the color balance up to make it less gray and pop more, but didn't overdo it leading to a good result! On image 3 similarly made things pop more with better color balance but didn't overdo it. On image 4 definitely overdid the color balance a little bit but still reflects 'decent' result in that unlike the bad policy results, it isn't just mindlessly overblowing, it's still trying to keep some good image structure and it stops before the image completely deteriorates. So the effect is harsh, but doesn't destroy the image. This is one of the worst examples, which shows that even at it's worst this method did not get terrible results.

In the next column we can see some examples of the Explicit model using direct rewards. In images 1, 3, and 4, it's clear how this method ends up overdoing the color balance, contrast, and sharpness. Especially in image 4 we see one of the worst results. These aren't quite as bad as what we'll see with the last method, but still definitely not achieving the desired results.

Lastly, the explicit policy overfilters to the extreme as seen in these images. The images are hardly recognizable anymore and definitely not "enhanced".

5 Limitations & Future Work

5.1 Tuning Hyperparameters

Further tuning hyperparameters, such as the lengths of the filter chains generated for the implicit data, or applying heuristics to the types of chains generated, etc, could potentially net better results. Exploring other potential CNNs and architectures, such as ResNet to replace AlexNet, might also produce better results.

5.2 Types of Human Feedback

With limited time and resources to dedicate to this project, we haven't been able to iterate over more than two forms of human preference feedback for image enhancement. An interesting avenue to continue on with this project would be to try out other forms of feedback like ranked preferences or generating preferences over one type of filter or enhancement at a time.

5.3 Types of Filters

Future work could also involve just including more filter options using other libraries with image editing tools such as instafilter, pilgram, OpenCV, and scikit. Continuous actions spaces could also be explored for applying filters.

5.4 Image Domains

Another possible future direction could be to explore using human preference data to train reward models for different classes of images or different applications. Perhaps the image enhancement desired for x-ray images is different than the type of enhancement for social media images. We used a dataset of landscape images, but it would be interesting to compare the results with portraits or macro photography. A large appeal of this human preference based image enhancement learning is that it should be able to be trained for any domain, so some exploration into more image domains would be valuable.

6 Conclusions

Using the trained policy we started with failed to give useful image enhancement results. However, using the reward model directly as a policy model instead improved performance. We also found that using just the explicit preference data collected from human judges had worse effects on images overall than using implicit preferences, specifically when those implicit preferences were generated with a max-length of filter sequences that they could not exceed.

References

[1] Satoshi Kosugi and Toshihiko Yamasaki. "Unpaired Image Enhancement Featuring Reinforcement-Learning-Controlled Image Editing Software". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 11296–11303. DOI: 10.1609/aaai.v34i07.6790. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6790.

[2] Pei Lv et al. "User-Guided Personalized Image Aesthetic Assessment Based on Deep Reinforcement Learning". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 736–749. DOI: 10.1109/TMM.2021.3130752.

[3] Tran Ngoc Minh et al. *Automated Image Data Preprocessing with Deep Reinforcement Learning*. 2021. arXiv: 1806.05886 [cs.CV].

[4] F. Sahba and H.R. Tizhoosh. "Filter fusion for image enhancement using reinforcement learning". In: *CCECE 2003 - Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436)*. Vol. 2. 2003, 847–850 vol.2. DOI: 10.1109/CCECE.2003.1226027.

[5] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. "Aligning Latent and Image Spaces to Connect the Unconnectable". In: *arXiv preprint arXiv:2104.06954* (2021).

[6] Nisan Stiennon et al. *Learning to summarize from human feedback*. 2022. arXiv: 2009.01325 [cs.CL].

[7] Shixin Sun et al. "Underwater Image Enhancement With Reinforcement Learning". In: *IEEE Journal of Oceanic Engineering* (2022), pp. 1–13. DOI: 10.1109/JOE.2022.3152519.