Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI, the Arts and Creativity

Re:Draw - Context Aware Translation as a Controllable Method for Artistic Production

João Libório Cardoso[1] , **Francesco Banterle**[2] , **Paolo Cignoni**[2] and **Michael Wimmer**[1]
[1]TU Wien, Austria
[2]CNR-ISTI, Italy
{jaliborc, wimmer}@cg.tuwien.ac.at, {francesco.banterle, paolo.cignoni}@isti.cnr.it

Abstract

We introduce context-aware translation, a novel method that combines the benefits of inpainting and image-to-image translation, respecting simultaneously the original input and contextual relevance – where existing methods fall short. By doing so, our method opens new avenues for the controllable use of AI within artistic creation, from animation to digital art.

As an use case, we apply our method to redraw any hand-drawn animated character eyes based on any design specifications – eyes serve as a focal point that captures viewer attention and conveys a range of emotions; however, the labor-intensive nature of traditional animation often leads to compromises in the complexity and consistency of eye design. Furthermore, we remove the need for production data for training and introduce a new character recognition method that surpasses existing work by not requiring fine-tuning to specific productions. This proposed use case could help maintain consistency throughout production and unlock bolder and more detailed design choices without the production cost drawbacks. A user study shows context-aware translation is preferred over existing work 95.16% of the time.

1 Introduction

Alfred Yarbus's influential work [Yarbus, 1967] quantified a long-standing intuition: using an eye tracker, he noted that observers spend a surprisingly large fraction of time fixated on the eyes in a picture. The eyes of others are important to humans because they convey subtle information about a person's mental state (e.g., attention, intention, emotion) and physical state (e.g., age, health, fatigue). This significance has translated into the realm of hand-drawn animation, where the eye designs have often become increasingly complex and expressive to capture these nuances. However, this complexity comes at a cost. Despite its massive resurgence in the last decade [Masuda *et al.*, 2019], traditional animation has struggled to benefit from advances in computer graphics: techniques used in production remain largely the same, with productions relying on repetitive manual labor from a large workforce. As a result, the eyes, being the most time-consuming and intricate to draw, are often the first elements to be simplified, leading to compromises in both expression and artistic consistency. Our aim is to introduce a computational method that can alleviate some of these challenges without sacrificing the artistic integrity of the medium.

We further conducted a survey among 17 professional animators, of which 29% work at established studios, and the rest either freelanced or worked for smaller studios. We asked them multiple optional questions about time consumption, of which the full breakdown is shown in the author version[1] (**AV**). Character faces were reported to be the most complex part of animation, with 50% reporting it as the element they spend the most time on. Of the remaining animators, 75% voted for either anatomy or hair. Drawing was estimated to constitute the vast majority of the work (85%), and doing it at the highest level of detail was estimated to take 1.7 times the amount of work than on average, for a total of 66m of additional human effort per key frame from 1st key through coloring. Sadly, the fact that 53% still use paper drawings in their studios, despite 100% preferring to draw digitally, indicates that using computational tools during the early drawing stages might not be possible yet in practical terms.

1.1 Problem and Contributions

Existing deep learning methods present significant limitations within artistic applications. Inpainting, while capable of generating detailed art that fits within existing content, offers little control over the generated content, making it unsuitable for most precise artistic endeavors [Akita *et al.*, 2020]. Image-to-image translation, while being able to take artistic input, is constrained by only being applicable to entire images, as it does not take into account the context surrounding target regions.

We propose context-aware translation as the solution to these limitations. We then apply it in a novel pipeline that automates increasing consistency and amount of detail in the eyes of hand-drawn animation characters. It effectively mimics the work of cleanup animators, who redraw frames to fix mistakes and better match the character color guides – despite the misleading name, color guides, also known as model

[1]**AV:** https://jaliborc.github.io/re-draw

Figure 1: Teaser. As shown, the proposed context-aware translation is capable of automatically redrawing parts of images according to any provided design, without the need for fine-tuning. Unlike image-to-image translation [Liu *et al.*, 2019], which neglects surrounding context, our approach considers the entire frame. Unlike inpainting [LahIntheFutureland, 2023], which lacks artistic control by ignoring the original content, our method honors the artist's input. This facilitates the production of more consistent artwork and allows for more complex design choices.

sheets, depict all the information an artist would need to draw a character while remaining true to its intended design and the art style of the production (see **AV** for an example). We also tackle an additional problem this use-case raises: the lack of training datasets of anime production, which we address by proposing methods to negate the need for production data entirely, including a novel character recognition method.

In summary, our key contributions are: 1) Context-aware translation, a novel general deep-learning method that avoids the limitations of both inpainting and image-to-image translation – this includes 1.a) a dual discriminator structure and novel adversarial losses that enforce simultaneous respect for input content, translation requirements, and context constraints; and 1.b) a triple-reconstruction loss that yields greater generation capabilities than traditional loss. 2) A character design recognition network that outperforms existing work by using a production-style-aware latent space. 3) A novel pipeline that takes advantage of the aforementioned contributions to automatically increase the consistency and amount of detail in the eye region of characters, and without the need of production data during training.

Furthermore, we present an ablation study in Section 4 that scrutinizes the benefits of each of our novel components, contribution of each loss used, compares both our context-aware translation and style-aware clustering against existing work, and assesses the robustness and temporal coherence of our method. We also present a user study with 63 participants in Section 5 that tests three key properties: the absence of detectable artifacts, the enhancement of artwork detail, and the overall aesthetic preference when compared to existing methods, all of which our pipeline successfully validated.

2 Related Work

In this Section we describe the minimum animation production background necessary to frame our use-case, and analyze relevant deep-learning existing work.

Anime Production Background Limited animation production involves a precisely defined pipeline comprising of drawing, finishing, and compositing steps [Furansujin Connection, 2016], each performed by different artists. The process starts with frame planning and drawing, where paper is prevalent, followed by digital cleaning and coloring, and concludes with compositing. For computational methods to be effective in this domain, they must integrate seamlessly into the existing pipeline, not expect extensive per-frame manual intervention, and allow for artistic control.

Editing Content with Style The seminal work by Gatys *et al.* [2016] introduced style transfer using deep learning, sparking extensive research in image editing through style variation [Huang and Belongie, 2017; Karras *et al.*, 2019; Karras *et al.*, 2020; Karras *et al.*, 2021a; Karras *et al.*, 2021b]. Liu *et al.* [2019] introduced FUNIT, an unsupervised network capable of image-to-image translation from unseen domains. Subsequent works [Kim *et al.*, 2020; Saito *et al.*, 2020; Nizan and Tal, 2020; Li *et al.*, 2021] have built on this foundation, but invariably modify entire images and unintentionally compromise the consistency of character poses, expressions, and other elements. A crucial aspect of our work is to locally modify art, but only the design should be varied on the targeted elements; e.g., the eyes.

Artistic Methods In comics and illustration editing, most literature focuses on colorization (of illustrations or shaded manga drawings) and line extraction. Early efforts by Simo-Serra *et al.* [2016] and Li *et al.* [2017] demonstrated techniques for extracting main lines and patterns. These methods have seen enhancements through user interaction and adversarial training [Simo-Serra *et al.*, 2018b; Simo-Serra *et al.*, 2018a; Lee *et al.*, 2019]. Colorization techniques have evolved from user-assisted GAN-based methods [Ci *et al.*, 2018; Zhang *et al.*, 2018] to more sophisticated approaches that allow high-quality outputs with minimal training data [Silva *et al.*, 2019; Shimizu *et al.*, 2021]. Recent studies have also explored colorization using text tags and user inputs for specific parts, although these methods often sacrifice artistic control [Akita *et al.*, 2020; Kim *et al.*, 2019]. The training illustration dataset by Branwen *et al.* [2022] is a staple of many methods. Maejima *et al.* [2021] tackle animation colorization using a few-shot strategy and an ad-hoc sampling method for patches.

Recent work has also brought segmentation [Zhang *et al.*, 2020] and clustering [Nir *et al.*, 2022] to cartoon and illustration art. Nir *et al.* [2022] developed a self-supervised tech-

nique for style-specific analysis in animation, yet it requires separate training per production and is more suited for tracking characters and not their designs. In our work, we introduce a more flexible semantic clustering that decouples the style of the anime from the content, allowing it to generalize to unseen productions.

3 Method

Our approach for enhancing animated eyes takes as input the animation frames to be improved and a character color guide; see Fig. 2. We use an unsupervised convolutional network trained alongside classification networks, capable of telling designs apart, as its adversaries. Using such a model requires artists to manually associate regions to redraw and color guides manually, which is not practical. Even more problematic, to train this type of adversarial structure, one normally uses pairs of these images, labeled into different classes (character designs). In particular, to ensure our model is capable of generalizing to new designs, we need to train on a large enough variety of them. Yet, art direction is not easily available and generally not created in high enough quantities that would be needed for a robust training. Moreover, manually tagging and cropping this data would be extremely labor intensive and hard to replicate. To address these issues we propose a novel character design clustering method, and use it to automatically infer training data from random frames, thus solving the association problem. As such, Re:Draw does not require internal production data for training, instead it only only requires a set of random sampled frames from different productions; see Fig. 4.

Image in-painting has shown to be capable of completing missing regions, yet predictions based only on the surrounding of the area to be redrawn do not allow artists to finely control the output results using art or style direction examples. Image-to-image translation and style transfer are capable of using both of these inputs, yet existing work is incapable of generating art that fits and correctly matches within the actual context of the drawing: they can be very unreliable in preserving the artwork pose. For these reasons, we introduce context-aware translation. We make use of two adversarial discriminators built using partial convolutions, allowing them to weight images differently and independently, and a novel triple reconstruction loss based on the concept of the generation of image triplets.

3.1 Dataset Generation

We will now describe how we avoid the need for production data, by automatically clustering art by character design and then further splitting it into low- and high-levels of detail. This categorized dataset is required during the training phase of our context-aware translation model, which involves solving multiple adversarial classification tasks simultaneously.

Object Detection We first train an object-detection network – we use the well-established Faster R-CNN network [Ren *et al.*, 2015] – to identify character faces and details in them (such as eyes) and run it on the randomly sampled frames. This results in a dataset of character faces in a variety of poses, split by the sources they were sampled from.

Style-Aware Clustering Although re-identification of human faces is a long studied topic [Balaban, 2015], we found existing work to be ineffective at automatically identifying animated characters not seen during training. We attribute this to the fact that, while human faces have a consistent and predictable structure, animated characters are not restrained by the laws of reality and thus present a much higher variance: in a given production, characters might have a very similar look and feel, while in another production they might vary widely in structure and shape (see **AV**).

To address this issue, we improve upon the state of the art of character recognition by combining ideas from facial recognition and image-to-image translation. We propose a supervised network that, unlike existing work, maps character portraits to an art-style normalized Euclidean space, where distances between these portraits correspond to a measure of character design similarity within its production. It takes as input character portraits to be mapped and a collection of random portraits from the same production for normalization estimation.

As shown in Fig. 3, latent representations of both inputs are estimated: we compute the content representation using a ResNet [He *et al.*, 2016] encoder – which is well established for object recognition – and the production representation using a convolutional encoder. The latter is done using only the lightness in $l\alpha\beta$ color space [Reinhard *et al.*, 2001], as we found that the normalization input works better if it only contains the main shape information, so we use a color space to decorrelate it from color variation. This style-latent representation is then used to compute a set of affine transformations, with the goal of mapping the encoding from an absolute Euclidean-space representation of portraits to the style-normalized one. This mapping is done using Adaptive Instance normalization [Huang and Belongie, 2017] on the content-input latent representation. This finally results in 32 parameters per portrait thanks to the linear layers, which are then clustered using traditional hierarchical clustering, with unweighted pair group method, arithmetic mean and Euclidean distance. These methods and parameters were chosen by testing the rate of correct clustering across a validation dataset.

To train this network E and ensure the content-encoding output respects the desired intra and inter-class proprieties of the normalized Euclidean space, we use the option of Triplet Margin Loss [Balntas *et al.*, 2016; Hermans *et al.*, 2017] – that is, given a pair of portraits from the same design $\{p_1, p_2\}$ and one from another p_3 but from the same production $\mathbb{P}$, we minimize the distance from the first two, while maximizing the distance of the third (images shown in lowercase; functions and classes in uppercase):

$$\underset{E}{\operatorname{argmin}}\ max\big\{\|E(p_1,\mathbb{P}) - E(p_2,\mathbb{P})\|^2 - \|E(p_1,\mathbb{P}) - E(p_3,\mathbb{P})\|^2 + 1, 0\big\} \tag{1}$$

This means that, during training, character portraits must be provided in sets of three. We also ensure that the total training weight of each production style and of each character design within each style is the same, to further help with generalization. While it is technically possible to train E in conjunction with the context-aware translation, it is more computationally

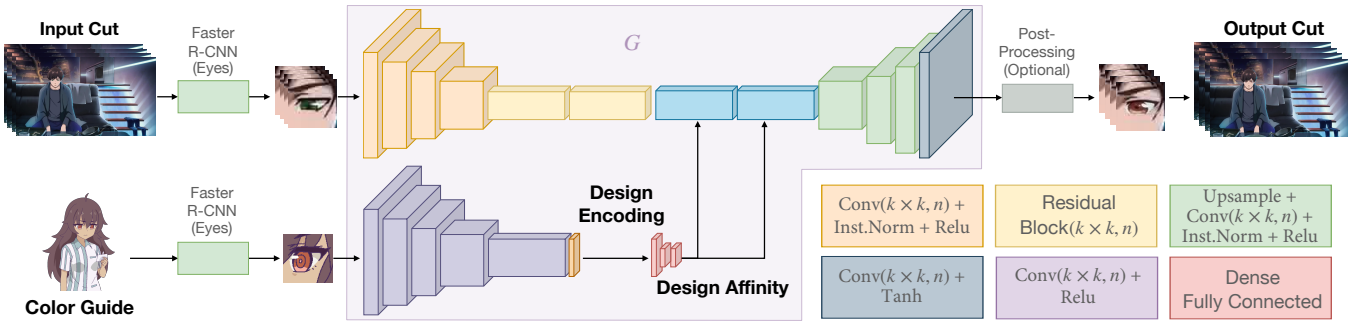


Figure 2: Context-Aware Inference. Eyes are detected in a sequence of frames and a color guide, then fed to our context-aware redrawer G, and the resulting styled eyes are post-processed into the original art. The whole process can run in real-time. See AV for full version.

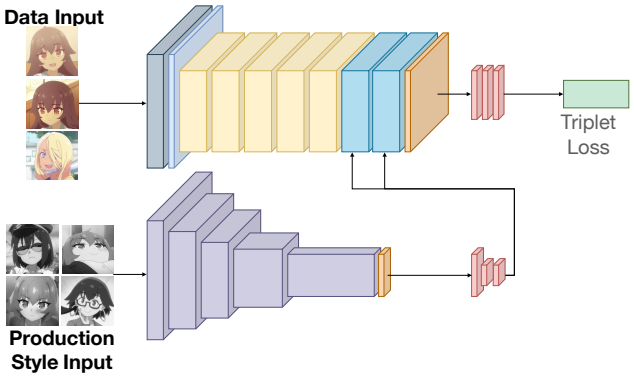


Figure 3: Style-Aware Encoder. We train a network capable of generalized character design recognition using triplet loss. Unlike traditional facial recognition, where only a set of labeled portraits is seen, we additionally input random unlabeled portraits to account for production style. See AV for full version.

efficient to train E first and then freeze it to train the remainder networks.

Level-of-Detail Split Having the portraits organized by character design, we extract the intended art details (eyes in our case study) from them using a Faster R-CNN network again, but now with the knowledge of their corresponding designs. We standardize all extracted art to the same size. Then, we exploit the fact that characters are often drawn with different levels of detail, depending on their prominence on screen, to discriminate between low and high details regions. So, after empirical observation, regions with less than 0.31% pixels were assumed to be low-detail and to be redrawn, while regions with more than 0.48% pixels were used as art direction examples (see **AV**).

3.2 Context-Aware Translation

Having generated the dataset, we can now train the redrawing model. We want to find a function $x^+ = g(x, \mathbb{S})$ that, given a low-detail content image x and color guide $\mathbb{S}$ (equivalent to style images in a style transfer context), is capable of outputting a higher detail version x^+ of x. This leaves us with two conflicting goals: we want the translated artwork x^+ to match the provided design $\mathbb{S}$ and its level of detail, but to still fit within the original drawing of x.

We define an image-to-image network G, to be trained as a context-aware redrawer, with the purpose of approximating g. As illustrated in Fig. 2, it is composed of a convolutional encoder-decoder structure with an additional style encoder. The latter matches exactly the encoder described in Section 3.1 and is used to compute a set of affine transformations that control the Adaptive Instance normalization in the decoder.

Triple Reconstruction Loss Let l be a low-detail image and h a high-detail one, each sampled from different designs $\mathbb{L}$ and $\mathbb{H}$, respectively. Our approach is to train the redrawer as an image translation problem such that $t = G(l, h)$ outputs the result of applying design $\mathbb{H}$ to l.

To help G learn a translation model and ensure it maintains the local structure of l, a second output $\hat{l} = G(l, l)$ is frequently used as part of a reconstruction loss [Huang *et al.*, 2018]. However, this is not appropriate for our problem, as we are not interested in the network producing low detail images. We propose a novel reconstruction loss that analyses a total of three generated images:

$$\mathcal{L}_R = [\|h - G(h,h)\| + \|F(l) - F(t)\| + \|F(l) - F(\hat{l})\|]_1^1, \quad (2)$$

where $F(x)$ is a low-pass image filter applied on the lightness of the image x, implemented by converting to frequency space using fast Fourier transform and remove any frequencies above a set threshold (0.06, as shown in Fig 5). The basis is that, by removing high frequencies and color changes, differences between low-detail and high-detail images are ignored as well, allowing us to create a reconstruction loss in low-detail images.

Adversarial Discriminators We address our aforementioned conflicting goals by using two independent image multi-task classifiers instead: a *quality discriminator* Q judging whether the output is high detail and matches the intended design, and a *context discriminator* C judging whether it fits within the original artwork and its own design, irrespective of detail-level.

To achieve this purpose, as we show in Section 4, we need to train each discriminator differently, despite sharing many commonalities: both have the same partial convolutional structure and are trained using hinge loss with R1 regularization [Mescheder *et al.*, 2018], to prevent over-fitting and mode collapse. This results in the following losses for

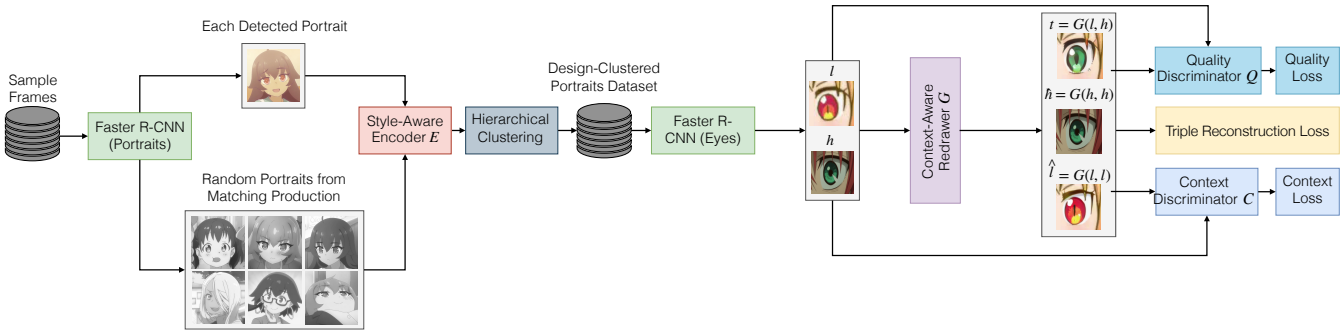


Figure 4: Context-Aware Training. Detected portraits are clustered using our style-aware encoder (left), with additional portraits as style guides. Content and style images are extracted from portraits of different designs (center). These are used to generate 3 redrawings from different combinations of these input pairs, which are judged by multiple losses, including two multi-class discriminators (right).

penalizing wrong classifications of positive $\mathcal{L}_P$ and negative $\mathcal{L}_N$ examples:

$$\begin{aligned} \mathcal{L}_P(x, \mathbb{S}) &= [\max(0, 1 - D(x)_{\mathbb{S}}) + \gamma \|\nabla D(x)_{\mathbb{S}}\|^2]_1 \\ \mathcal{L}_N(x, \mathbb{S}) &= [\max(0, 1 + D(x)_{\mathbb{S}})]_1 \end{aligned}, \tag{3}$$

where $D \in \{Q, C\}$ can be any one of the discriminators, $D(x)_{\mathbb{S}}$ is the discriminator's score of image input x for design class $\mathbb{S}$, $\nabla D(x)_{\mathbb{S}}$ its derivative used for R1 regularization and $\gamma = 10$ (R1 standard weight). That is, D should converge to $[0, 1]$ for positive entries and to $[-1, 0]$ otherwise. The discriminators are then given different input masks to weight disparate regions of images differently: the quality discriminator Q focuses on the interior of the redrawn region, while the context discriminator C focuses on the opposite, including an outer border that is not redrawn. They meet and oppose each other in the intersection of their two regions. Finally, they are trained to judge the training image pairs $\{l, h\}$ and the generated triplets $\{t, \hat{l}, \hat{h}\}$ such that Q looks for high detail output, while C tries to tell real and generated art apart:

$$\begin{aligned} &\underset{Q}{\operatorname{argmin}}\, \mathcal{L}_P(h, \mathbb{H}) + \frac{\mathcal{L}_N(l, \mathbb{H}) + \mathcal{L}_N(t, \mathbb{H})}{2} \\ &\underset{C}{\operatorname{argmin}}\, \mathcal{L}_P(h, \mathbb{H}) + \mathcal{L}_N(t, \mathbb{H}) + \mathcal{L}_P(l, \mathbb{L}) + \mathcal{L}_N(\hat{l}, \mathbb{L}) \end{aligned} \tag{4}$$

That is, Q attempts to learn to identify real high-detail images as positive examples, and low-detail or generated art as negative ones; while C attempts to identify real as positive and generated as negative, independently of detail. Images are always judged for the design class they are supposed to belong to. Then, to train the redrawer network G using these discriminators, we use hinge loss with a latent feature loss to regularize the adversarial training. Let D^F be the latent features computed by a discriminator D in a hidden layer, and

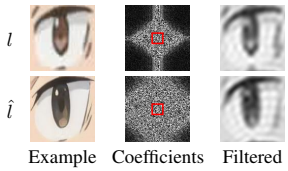


Figure 5: **Low-Pass Filter**. $\hat{l}$ is computed during training from enhancing l. Right shows the result of our low pass filter.

s a sampled image (either l or h) from the given design class $\mathbb{S}$. The adversarial loss function of each discriminator D becomes:

$$\mathcal{L}_D(x, \mathbb{S}) = [1 - D(x)_{\mathbb{S}}]_1^1 + [D^F(x)_{\mathbb{S}} - D^F(s)_{\mathbb{S}}]_1^1 \tag{5}$$

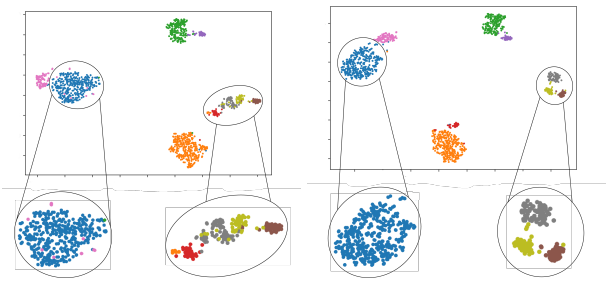
The use of hinge loss, R1 regularization and feature matching loss have been used in different forms in image-to-image translation problems [Liu *et al.*, 2019; Saito *et al.*, 2020]. Just as with training the discriminators themselves, our contribution is how these are then used to train a redrawer capable of addressing our problem. We combine our novel reconstruction loss with two adversarial losses to verify discriminator conditions and another two to ensure reconstruction persistence, where $\mathcal{L}_Q$ and $\mathcal{L}_C$ are the adversarial functions of each discriminator, defined in Equation 5:

$$\underset{G}{\operatorname{argmin}}\, \mathcal{L}_R + \mathcal{L}_Q(t, \mathbb{H}) + \mathcal{L}_Q(\hat{l}, \mathbb{L}) + \mathcal{L}_C(t, \mathbb{L}) + \mathcal{L}_C(\hat{h}, \mathbb{H}) \tag{6}$$

Post-Processing To further ensure image regions generated fit within the original image, we apply a few image processing operations: after re-sampling the network output to the original resolution, we apply color transfer [Reinhard *et al.*, 2001] and place it into the original image using Poisson image editing [Pérez *et al.*, 2003].

4 Ablation Experiments

Clustering We compared our style-aware clustering approach with FaceNet [Schroff *et al.*, 2015], trained on the same labeled animated character faces. We statistically analyzed how effective the latent representations learned from either method work on a validation dataset of production styles not seen during training: we measured the ratio of the average squared norm distance between each point of the same character, and the average squared norm distance between the mean points of each character. The lower this value, the better is the latent space representation in principle. Our method measures a ratio of $1.212e^{-4}$ in the validation data, which outperforms FaceNet's $1.494e^{-4}$ ratio. Interestingly, our method performs similarly to a FaceNet network trained on the validation dataset, whose ratio was $1.209e^{-4}$. This shows our method presents better generalization to unseen data. Furthermore, we show in Fig. 6 how well split the character faces from validation production styles are. The colors



(a) FaceNet (b) Ours

Figure 6: Clustering Validation. 2D visualization of characters faces from a production not seen during training [Coolkyousinnjya and Dragon Life Improvement Committee, 2017], generated using T-Distributed Stochastic Neighbor Embedding [Van der Maaten and Hinton, 2008] on the latent spaces learned using FaceNet [Schroff *et al.*, 2015] and our proposed encoder network. Colors correspond to ground truth labels. As shown, character differentiation is clearer in our learned space. See AV.

of the points represent their ground-truth labels, and there is a visible improvement with our method.

Redrawing We evaluated our redrawing approach against neural cross-domain image translation with content and style inputs. While multiple variations of these networks exist [Saito *et al.*, 2020; Ojha *et al.*, 2021], they mostly compete in translation ability and do not address our problems. Thus, we chose as a baseline for comparison Liu *et al.* [2019] method (FUNIT). We progressively introduced each of our novelties to show the importance of each one (Fig 7): we introduced our style-aware clustering by training FUNIT only on the few manually labeled faces versus the generated dataset. We then added our double-discriminator method, while maintaining FUNIT's traditional reconstruction loss. Finally, we introduced the proposed triplet reconstruction, followed by the post-processing step. FUNIT fails to respect pose/expression and context. Without our large-scale dataset, it often fails to generate realistic art and cannot generalize to designs not seen during training. Our dataset and double discriminator approach solve all of these issues, but reduce the ability of the network to generate highly detailed art. Our novel triplet reconstruction fixes that.

Our method is also very stable and demonstrates temporal coherence (see **AV**) despite not specifically incorporating it in the loss. We believe the reason for this consistency is its emphasis on preserving the intended pose and context of the drawing: removing the regularizer from the adversarial losses results in mode-collapse, and removing the quality loss results in a poor auto-encoder as expected (see **AV**). Yet, removing the reconstruction loss does not result in unpredictable output as expected, just an inability to learn useful transformations, and removing the context loss does not result in output reminiscent of FUNIT. Our explanation is that the network, by following the context loss, is being incentivized to exhibit spatial and temporal consistency in its output. This stability is what enables our less explicit form of reconstruction loss to direct the training toward useful results.

We critically evaluate the performance of existing techniques – namely style-transfer, image-to-image translation, and text-to-image diffusion – against our proposed context-aware translation. Figs. 1 and 8 clearly illustrate the limitations of these existing approaches, thereby reinforcing why our method is a more apt solution for this particular type of problem. Furthermore, tests shown in the **AV** provide compelling evidence of our method's robustness and versatility in handling a variety of challenging scenarios, further substantiating its suitability for this application.

The limitations of the network we found boiled down to two cases: uncommon occlusions and strong rotations. As most occlusion to anime eyes is hair and the vast majority of are drawn up-right, the network can generate artifacts outside of those expected conditions: the shape of occluders below the eyeline might be distorted as if was a skin tattoo, and eyes will be drawn upright if a character is reversed (see **AV**). Adding a network capable of estimating eye rotation to the pipeline would automate this process and further improve the generated dataset. But outside of these two unusual spaces, artifacts we did find were created by Poisson-blending, not our models, which leads us to conclude post-processing is the current main limitation of the method.

5 User Study

To validate our work, we conducted a study with 63 volunteer participants, predominantly self-identifying as anime experts, comprised of three tasks with images from anime productions, all shown in randomized order. The user study and our statistical analysis is provided in the **AV**.

Realness (T1) Participants assessed eight images – four originals and four with our technique – to spot drawing issues, not knowing the mix. Statistical analysis using the χ^2 test with Yates correction [1988] showed no significant difference in error detection between original and modified, indicating our generated eyes do not have evident artifacts and are indistinguishable from real productions.

Level of Detail (T2) Participants selected for higher detail among eight pairs of images unedited and enhanced by our method. Using the [David, 1988] multiple comparison test, we found a statistical significant preference for our enhanced images, indicating our method effectively produces images with higher detail from original artwork.

Preference (T3) Participants chose the better-looking image from pairs made to cross-compare 3 methods: our method, FUNIT [Liu *et al.*, 2019] and FUNIT enhanced with our clustering. Our method was preferred in 95.16% of cases. Using the same statistical analysis as in **T2**, we found images generated by our method are preferred to the ones generated by previous work, even when it is enhanced with our dataset.

6 Conclusions

We have presented context-aware translation: a novel unsupervised image-to-image network, trained with two adversarial classification networks. We built these classifiers using partial convolutions, allowing them to weight generated images differently and independently. We introduced novel loss functions for these discriminators and a novel reconstruction

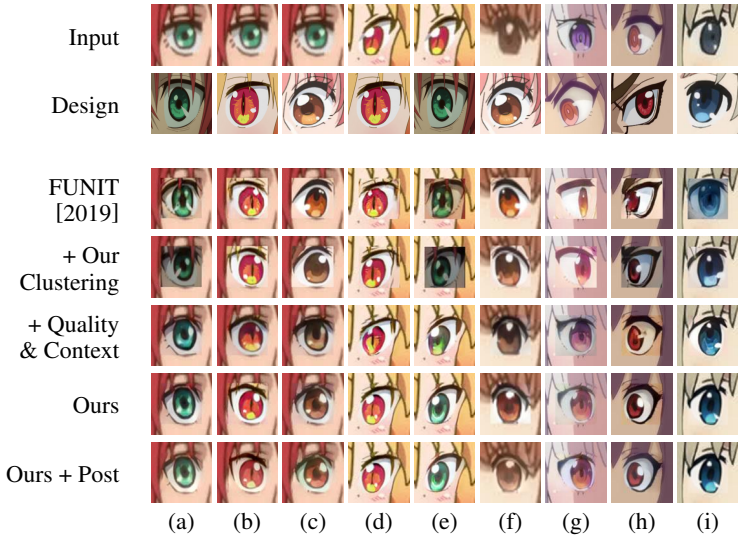


Figure 7: Ablation of Contributions. We compared our method with traditional image translation by progressively introducing our contributions (rows 4 to 6). The synthetic data extrapolated using our style-aware clustering prevents over-fitting (a,e) and allows generalization to unseen designs (g,h,i). Further introducing our dual discriminators allows redrawn areas to maintain artwork fit, but sacrifices detail (a,b,c,e,i) and ability to redesign shape or color (g,h). Finally, introducing the triplet reconstruction loss brings that expressiveness back. See AV.

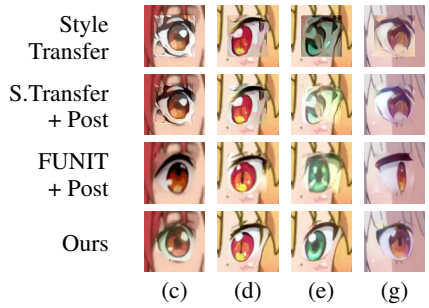


Figure 8: Ablation of Related Work. We compared our method with style transfer and image-to-image translation with and without post-processing. See AV.

loss based on image triplets to achieve context-aware translation. We proposed a deep-learning approach with a novel character design clustering to automatically collect training data from animation frames and input data during inference.

We have shown the method is capable of automatically redrawing the eyes of an anime character according to a provided character design direction. In this way, the shape and color of the iris can be changed, features, like reflexes and shades, can be added and the level of detail can be increased under the precise artist's control. We have made the method easily replicable, by removing the need for internal production data or labeled data. Given the general nature of our method, we expect it to be usable or extendable to other elements in and outside of animation. Only our frequency threshold value and the quality-split criteria could be specific to our use case of animation, but such is the case of many meta parameters in deep learning methods.

The obtained results also indicate the model might be presenting emergent behavior of precise image segmentation, despite never having seen segmented data, which we plan to explore in future work. As the models run nearly in real-time, Re:Draw could be used interactively as artists draw or color line-art. Only post-processing, which is significantly more computationally intensive, prevents the entire method from being run interactively. Given blending is also the main culprit behind artifacts, the focus of future work will be to further improve the method, likely within the reconstruction loss, to remove the need for post-processing altogether.

We have substantiated the need for such a style-driven enhancement with a professional user survey that reported the impact in the time of high quality drawing of the face details. Finally, we have validated our approach and results with ablation and a user study showing that our style-normalized latent space pushes the state of the art regarding the identification of non-photorealistic imagery, our approach is preferred over traditional image-to-image translation 95.16% of the time and the images generated are not discernible with respect to images drawn by artists with traditional techniques.

Acknowledgments

We would like to especially thank Jarret Martin for teaching us about the animation pipeline needs, his continuous feedback throughout the course of our research and help running the industry survey. Our gratitude also goes to Tonari Animation and OtakuVS for providing production data and other materials from their works, Otachan and Second Self, which we utilized whenever possible to illustrate our research.

Finally, we'd like to thank some of the first author's students for their contributions during the early stages of this project, in particular Yanic Thurner, Felix Kugler, and especially Dominik Hanko, who worked more closely with this topic.

References

[Akita *et al.*, 2020] Kenta Akita, Yuki Morimoto, and Reiji Tsuruno. Deep-Eyes: Fully Automatic Anime Character Colorization with Painting of Details on Empty Pupils. In Alexander Wilkie and Francesco Banterle, editors, *Eurographics 2020 - Short Papers*. The Eurographics Association, 2020.

[Balaban, 2015] Stephen Balaban. Deep learning and face recognition: the state of the art. *Biometric and surveillance technology for human and activity identification XII*, 9457:68–75, 2015.

[Balntas *et al.*, 2016] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

[Branwen *et al.*, 2022] Gwern Branwen, Anonymous, and Danbooru community. Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. https://www.gwern.net/Danbooru2021, 1 2022. Accessed: DATE.

[Ci *et al.*, 2018] Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1536–1544, New York, NY, USA, 2018. Association for Computing Machinery.

[Coolkyousinnjya and Dragon Life Improvement Committee, 2017] Coolkyousinnjya and Dragon Life Improvement Committee. Miss kobayashi's dragon maid, 2017. [Online; accessed 4-August-2023].

[David, 1988] H. A. David. *The Method of Paired Comparisons, 2nd ed.* Oxford University Press, 1988.

[Furansujin Connection, 2016] Furansujin Connection. Les étapes de fabrication. https://web.archive.org/web/20220401075414/http://www.furansujinconnection.com/les-etapes-de-fabrication/., 2016. [Online; accessed 2022-March-21].

[Gatys *et al.*, 2016] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2414–2423. IEEE Computer Society, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Huang and Belongie, 2017] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1510–1519. IEEE Computer Society, 2017.

[Huang *et al.*, 2018] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 9 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020.

[Karras *et al.*, 2021a] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 852–863, 2021.

[Karras *et al.*, 2021b] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021.

[Kim *et al.*, 2019] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9055–9064. IEEE, 2019.

[Kim *et al.*, 2020] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020.

[LahIntheFutureland, 2023] LahIntheFutureland. Mysteriousv4 sdxl 1.0. https://civitai.com/models/118441/lah-mysterious-or-sdxl, 2023. [Online; accessed 19-October-2023].

[Lee *et al.*, 2019] Gayoung Lee, Dohyun Kim, Youngjoon Yoo, Dongyoon Han, Jung-Woo Ha, and Jaehyuk Chang. Unpaired sketch-to-line translation via synthesis of sketches. In *SIGGRAPH Asia 2019 Technical Briefs*, SA '19, page 45–48, New York, NY, USA, 2019. Association for Computing Machinery.

[Li *et al.*, 2017] Chengze Li, Xueting Liu, and Tien-Tsin Wong. Deep extraction of manga structural lines. *ACM Trans. Graph.*, 36(4), 7 2017.

[Li *et al.*, 2021] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Transactions on Multimedia*, pages 1–1, 2021.

[Liu *et al.*, 2019] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019.

[Maejima *et al.*, 2021] Akinobu Maejima, Hiroyuki Kubo, Seitaro Shinagawa, Takuya Funatomi, Tatsuo Yotsukura, Satoshi Nakamura, and Yasuhiro Mukaigawa. *Anime Character Colorization Using Few-Shot Learning*. Association for Computing Machinery, New York, NY, USA, 2021.

[Masuda *et al.*, 2019] Hiromichi Masuda, Tadashi Sudo, Kazuo Rikukawa, Yuji Mori, Naofumi Ito, Yasuo Kameyama, and Megumi Onouchi. Anime Industry Report. Technical report, The Association of Japanese Animations, 2019.

[Mescheder *et al.*, 2018] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[Nir *et al.*, 2022] Oron Nir, Gal Rapoport, and Ariel Shamir. CAST: Character labeling in Animation using Self-supervision by Tracking. *Computer Graphics Forum*, 2022.

[Nizan and Tal, 2020] Ori Nizan and Ayellet Tal. Breaking the cycle - colleagues are all you need. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7857–7866. Computer Vision Foundation / IEEE, 2020.

[Ojha *et al.*, 2021] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021.

[Pérez *et al.*, 2003] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.

[Reinhard *et al.*, 2001] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA, 2015. MIT Press.

[Saito *et al.*, 2020] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. COCO-FUNIT: few-shot unsupervised image translation with a content conditioned style encoder. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2020.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Shimizu *et al.*, 2021] Yugo Shimizu, Ryosuke Furuta, Delong Ouyang, Yukinobu Taniguchi, Ryota Hinami, and Shonosuke Ishiwatari. Painting style-aware manga colorization based on generative adversarial networks. In *2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021*, pages 1739–1743. IEEE, 2021.

[Siegel and Castellan, 1988] Sidney Siegel and N. John Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGrall-Hill International, 1988.

[Silva *et al.*, 2019] Felipe Coelho Silva, Paulo André Lima de Castro, Hélio Ricardo Júnior, and Ernesto Cordeiro Marujo. Mangan: Assisting colorization of manga characters concept art using conditional GAN. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pages 3257–3261. IEEE, 2019.

[Simo-Serra *et al.*, 2016] Edgar Simo-Serra, Satoshi Iizuka, Kazuma Sasaki, and Hiroshi Ishikawa. Learning to simplify: Fully convolutional networks for rough sketch cleanup. *ACM Trans. Graph.*, 35(4), 7 2016.

[Simo-Serra *et al.*, 2018a] Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Mastering sketching: Adversarial augmentation for structured prediction. *ACM Trans. Graph.*, 37(1), 1 2018.

[Simo-Serra *et al.*, 2018b] Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Real-time data-driven interactive rough sketch inking. *ACM Trans. Graph.*, 37(4), 7 2018.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Yarbus, 1967] Alfred L. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967. Translated from Russian.

[Zhang *et al.*, 2018] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage sketch colorization. *ACM Trans. Graph.*, 37(6), 12 2018.

[Zhang *et al.*, 2020] Lvmin Zhang, Yi JI, and Chunping Liu. Danbooregion: An illustration region dataset. In *European Conference on Computer Vision (ECCV)*, 2020.