AutoToon: Automatic Geometric Warping for Face Cartoon Generation

Julia Gong
Stanford University
jxgong@stanford.edu

Yannick Hold-Geoffroy
Adobe Research
holdgeof@adobe.com

Jingwan Lu
Adobe Research
jlu@adobe.com

(a) Input Photo (b) Cartoon (Ours) (c) Warping Field (Ours) (d) Caricature (Ours + CartoonGAN [6]) (e) Caricature (WarpGAN [28])

Figure 1: Example images from our test set (a), exaggerated cartoons (b) and overlaid warping fields (c) generated by our model (AutoToon), our model's cartoons stylized with CartoonGAN [6] to create caricatures (d), as compared to WarpGAN [28] caricatures (e).

Abstract

Caricature, a type of exaggerated artistic portrait, amplifies the distinctive, yet nuanced traits of human faces. This task is typically left to artists, as it has proven difficult to capture subjects' unique characteristics well using automated methods. Recent development of deep end-to-end methods has achieved promising results in capturing style and higher-level exaggerations. However, a key part of caricatures, face warping, has remained challenging for these systems. In this work, we propose AutoToon, the first supervised deep learning method that yields high-quality warps for the warping component of caricatures. Completely disentangled from style, it can be paired with any stylization method to create diverse caricatures. In contrast to prior art, we leverage an SENet and spatial transformer module and train directly on artist warping fields, applying losses both prior to and after warping. As shown by our user studies, we achieve appealing exaggerations that amplify distinguishing features of the face while preserving facial detail.

1. Introduction

Every human face is slightly different. While most people can identify faces familiar to them, it requires the more trained eye of a caricature artist to pick up on the most distinctive features that characterize an individual's face. In fact, caricature is a specific form of portraiture in which artists exaggerate the most visually salient characteristics of their subjects that distinguish them from others. Amplifying these defining features lets artists create more distilled portrayals of their subjects, and studies have shown that this skillful exaggeration can allow viewers to identify a subject's identity more easily from a caricature than from a normal photograph [27].

With the rise of applying computer vision techniques to tackle creative tasks, an interesting problem that has emerged is automatic caricature generation. Similar to how an artist might approach caricatures, the computer vision analogy to caricature generation can be decomposed into two steps: 1) applying a geometric warp to the face that exaggerates salient features, and 2) stylizing the warped image for an artistic effect. The complete disentanglement of these two steps allows them to be independently learned and applied, leading to greater flexibility and higher quality of generated caricatures.

Early work in caricature generation mostly relied on rules-based methods [1, 2, 11, 20, 21, 24]. More recently, with the rise of deep learning for artistic tasks such as sketch synthesis, image-to-image translation, and style transfer [10, 16, 35, 38, 41], caricature generation has been re-introduced as an image-to-image translation problem first by Cao et al. [4] and then Shi et al. [28]. While these systems do achieve geometric exaggeration and artistic stylization, the exaggerations still have room for improvement. They often either do not precisely target the most salient facial features due to the constrained set of warping handles, or the warping is not disentangled completely from the artistic stylization, resulting in weaker standalone warps and less flexibility for combining different warps and styles.

In comparing the difficulty of these two stages of caricature generation, it is noteworthy that the computer vision community has seen much progress in general image stylization and style transfer in recent years, such as [10, 18]. However, effective geometric warping, especially applied to faces, has more room for improvement. In fact, there is less room for error in pure geometric warping; not only are our eyes highly attuned to faces [32], but viewers are also more sensitive to the quality of unstylized, warped faces than that of stylized caricatures, since the resulting images are photo-realistic. Thus, in this work, recognizing there are numerous high-quality methods that can perform stylization in the caricature generation pipeline, we focus on the more difficult stage: geometric warping of distinguishing characteristics to create a high-quality, warped version of the original photograph, the result of which we term a *cartoon*.

Specifically, we aim to create an automated, end-to-end pipeline (AutoToon) that geometrically warps images of faces to generate cartoons, which are then used to create caricatures via existing stylization techniques. Our model learns a smooth warping field of pixel displacements that is applied to the input image, which can be scaled in magnitude to increase the exaggeration. By virtue of learning a warping field rather than performing image-to-image translation, our model preserves facial details more effectively and generates higher quality images for a given portrait.

Finally, to accompany our model, we also introduce the AutoToon dataset, a paired dataset of human facial portrait photos and their corresponding geometrically warped cartoons by trained artists. We hope that proposing a model for higher-quality face warping will accelerate the progress in creating end-to-end systems for caricature generation and other face-related cartoonization tasks.

Qualitative evaluation via user studies and artist appraisal of cartoons produced by AutoToon show that the generated cartoons from our approach exaggerate facial features more effectively than state-of-the-art warping methods. A summary of our contributions is as follows:

- To our knowledge, AutoToon is the first supervised deep learning face cartoon generation model. It
  - automatically exaggerates salient facial features well in a caricature-like manner and can be scaled to control warping extent,
  - is completely disentangled from stylization, and thus can be paired with any stylization method,
  - is trained on less data, and preserves image details more effectively than previous methods.
- A paired dataset, the AutoToon dataset, which also includes artist warping fields for photorealistic facial exaggeration and cartoon generation.

2. Related Work

Human faces have received a lot of attention in the literature over the years. Many approaches were developed to either model [29], interact [12] or generate them [19]. In this section, we review the work relevant to caricature generation and face warping.

2.1. Learned Warping

Multiple works have learned and applied spatial transforms on images. First, parametric approaches such as the spatial transformer [17] have been proposed to estimate global transform parameters. Flow-based approaches such as [26] further this idea by learning a dense deformation field over the whole image. DeepWarp [9] proposes to apply this to gaze manipulation. Recently, Zhao et al. [40] uses this dense flow estimation to remove geometric distortion from close-range portrait images. Cole et al. [8] also warp portrait images using spline interpolation on pre-detected landmarks while preserving identity. Similar to our loss functions, Zhang et al. [37] use smoothness, local, and global alignment terms for parallax-tolerant image stitching. Given the efficacy of flow estimation in these related application domains, our work on AutoToon aims to integrate this work with caricature generation by using dense flow estimation and the differentiable warping module from [17] to predict warping fields for generating cartoons.

2.2. Caricature Generation

One goal of caricature generation is to detect and amplify the unique features of a given face. Traditional techniques typically approached this by amplifying the difference from the mean, either by explicitly detecting and warping landmarks [3, 11, 22, 25] or using data-driven methods to estimate unique face features [23, 36, 39]. Early work largely relied on rules-based methods [1, 2, 20, 21], which limited caricature diversity. More recently, deep learning techniques have also been applied. For instance, Wu et al. [33] model the subject face in 3D to improve how natural the caricature expression looks using a neural network.

Newer techniques for caricature generation are data-driven. There exist some readily available datasets of annotated caricatures, such as WebCaricature [15], comprised of 6042 caricatures and 5974 photographs from 252 different identities. Despite these efforts, the limited amount of data available is still a major challenge. Thus, most of the work on this topic has taken inspiration from the recent generative image-to-image translation literature trained on unpaired images [7, 14, 41] and focuses on learning from unpaired portraits and caricatures [4, 34, 35]. Wu et al. [34] proposed to improve this image-to-image translation approach [16] by adding a geometric motion module.

Closer to our work, the first deep learning approach to caricature generation, CariGAN [4], proposed to train a Generative Adversarial Network (GAN) using unpaired images to learn the image-to-caricature translation. Building on previous work on style transfer and learned warping, Shi et al. [28] then proposed a method that uses the GAN framework to jointly train style and warping end-to-end. However, while unpaired learning can leverage more data, they introduce highly varied exaggerations from artists with divergent styles, even for the same subject, making learning consistent exaggerations difficult. They also frequently have varying scales, poses, and low input-output correspondence, resulting in models learning very high-level features that may not be the most specific distinguishing features of a given face. The exaggerations learned by these models are relatively coarse as well due to the use of sparse warping points. Thus, in our work, we instead take a *paired* supervised learning approach based on the work of two artists to balance this tradeoff, electing to learn specific artist styles well rather than an average of all styles. We also leverage the differentiable warping module from [17] to generate denser warping fields for more detailed exaggerations.

In contrast to previous work, we focus purely on the warping step of caricature generation to create high-quality warps while completely disentangling geometry and style.

3. Problem Formulation and Warping Model

In caricature generation, the task is to generate an exaggerated and stylized caricature for a given input portrait. Our new method, AutoToon, tackles the exaggeration portion of this pipeline. Given a normalized RGB portrait image $X_{in} \in \mathbb{R}^{H\times W\times 3}$, our task is to apply an artist-like facial exaggeration to X_{in} to generate a cartoon image $\hat{X}_{toon}$. $\hat{X}_{toon}$ is then the input image to any stylization network to complete the caricature generation task.

3.1. Warping and Linear Interpolation

To discuss our method, we first need to formalize our definition of warping fields and grid sampling, which are key to our approach.

To perform the facial exaggeration for X_{in}, our network learns a flow field, which we call a warping field. The learned warping field $\hat{F} \in \mathbb{R}^{H\times W\times 2}$ is applied to X_{in} to obtain $\hat{X}_{toon}$. The first channel of dimension $W \times H$ is a grid of scalar values representing the per-pixel displacement of X_{in} in the x direction, while the second channel encodes the same for the y direction.

To perform exaggeration, this warping field is applied to X_{in} via the differentiable warping module taken from Spatial Transformer Networks [17]. The module performs bilinear interpolation to displace the pixels of X_{in} according to the learned displacements $\hat{F}$, or $\mathit{Warp}(X_{in}, \hat{F})$. We call *Warp* the Warping Module, as shown in Figure 2.

4. Proposed Method

4.1. Dataset

101 portrait images of frontal-facing people (non-celebrities) were collected from Flickr. The people selected covered a broad range of age groups, sexes, races, and face shapes. These images were then warped via Adobe Photoshop by two caricature artists with similar styles to generate the ground-truth artist cartoons. This paired dataset of 101 images (X_{in}, X_{toon}) was split into 90 training and 11 validation images. The test set, without ground truth labels, was collected from various subjects and public sources. Sample images from the training set are shown in Figure 3.

An additional component of the dataset that we provide are the estimated artist warping fields $F_{32} \in \mathbb{R}^{32\times 32\times 2}$ that, after bilinear upsampling to size $H \times W \times 2$, correspond to each artist caricature. We discuss this choice to select 32 $\times$ 32 as the warping field spatial size choice in the next section. To obtain these, we performed gradient descent optimization on the warping field for each X_{toon} with $L1$ loss through the differentiable Warping Module to obtain the artist warping fields that correspond as closely as possible to each X_{toon}. To be precise, we solved the optimization

$$\underset{F_{32}}{\operatorname{argmin}} \, ||X_{toon} - \mathit{Warp}(X_{in}, \mathit{Upsample}(F_{32}))||_1 \, . \quad (1)$$

4.2. Model Architecture

AutoToon, our proposed method to tackle cartoon generation, is outlined in Figure 2. The exaggeration network of AutoToon is comprised of two components: the Perceiver Network and Warping Module. The Perceiver Network is a truncated Squeeze-and-Excitation Network (SENet50) [13] with weights pretrained on the VGGFace2 Dataset [5], chosen due to its state-of-the-art facial recogni-

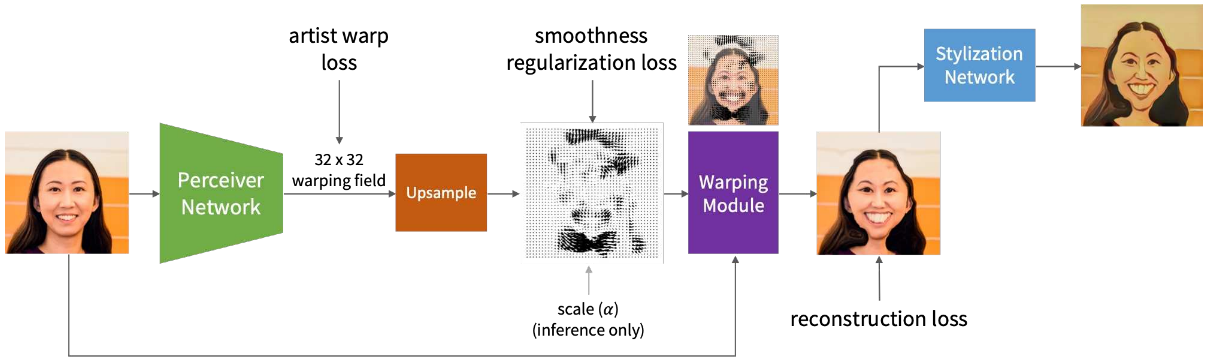


Figure 2: AutoToon model architecture and training losses. Given an input image, the Perceiver Network generates a 32×32 warping field. The warping field is upsampled via bilinear interpolation to obtain pixel-wise displacements, which is used to warp the input image into the resulting cartoon. The cartoon can then be stylized using any desired stylization network, such as CartoonGAN [6], used here. At inference time, a scaling factor α can be applied to the warping field to manipulate warping intensity.

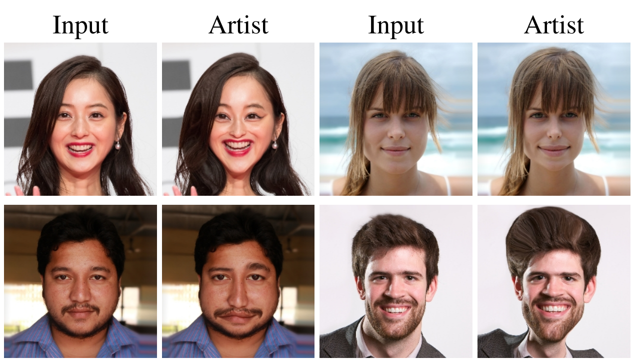


Figure 3: Four example pairs of input images and artist-warped cartoons from the training dataset. Photos by Dick Thomas Johnson, Shannon Luk, Possible, and Chuck Grimmett; modified.

tion performance. In particular, we modify it by only keeping the original layers up to and including the second bottleneck block, followed by an adaptive average pooling layer with output size $32 \times 32 \times 2$. The purpose of truncating the network is to reduce network capacity and prevent overfitting to the small dataset. The Perceiver Network takes input image X_{in} and outputs the warping field $\hat{F}_{32} \in \mathbb{R}^{32\times32\times2}$. $\hat{F}_{32}$ is then upsampled via bilinear upsampling to obtain $\hat{F}$, the per-pixel displacement. The Warping Module applies the warping field $\hat{F}$ to X_{in} to obtain $\hat{X}_{toon}$. In inference, the warping field can also be multiplied by a scaling factor α to control the intensity of the warp, as shown in Figure 7.

The choice to upsample a 32×32 warping field was motivated by two primary reasons. First, upsampling allows for an inherent smoothing of the warps, which intuitively creates smoother cartoons. Second, in keeping with powers of 2, a 64×64 warping field would have been too granular, and a 16×16 warping field was found to yield less exaggerated cartoons (see supplementary materials for details).

4.3. Loss Functions

We propose three loss functions to train AutoToon: the reconstruction loss, artist warping loss, and smoothness regularization loss.

The reconstruction loss $\mathcal{L}_{recon}$ penalizes the $L1$ distance between the artist cartoon X_{toon} and the generated cartoon $\hat{X}_{toon}$. In addition to this supervision on the model output, we also supervise the warping fields themselves with the artist warping fields. The artist warping loss $\mathcal{L}_{warp}$ penalizes the $L1$ distance between the artist warping field F_{32} obtained with (1) and the estimated warping field $\hat{F}_{32}$.

Finally, we use a cosine similarity regularization loss $\mathcal{L}_{reg}$ to encourage the warping field to be smooth and have fewer sudden changes in contour. This can be described as

$$\mathcal{L}_{reg} = \sum_{i,j\in\hat{\mathbf{F}}} \left(2 - \frac{\langle \hat{\mathbf{F}}_{i,j-1}, \hat{\mathbf{F}}_{i,j} \rangle}{\|\hat{\mathbf{F}}_{i,j-1}\| \|\hat{\mathbf{F}}_{i,j}\|} - \frac{\langle \hat{\mathbf{F}}_{i-1,j}, \hat{\mathbf{F}}_{i,j} \rangle}{\|\hat{\mathbf{F}}_{i-1,j}\| \|\hat{\mathbf{F}}_{i,j}\|} \right), \tag{2}$$

where $\langle \hat{\mathbf{F}}_{i,j-1}, \hat{\mathbf{F}}_{i,j} \rangle$ denotes the dot product of the upsampled warping field $\hat{\mathbf{F}}$ at pixel indices $i, j-1$ and i, j.

Thus, the loss function used to train our model is

$$\mathcal{L}_{autotoon} = \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{warp} + \lambda_3 \mathcal{L}_{reg}. \tag{3}$$

5. Experiments and Discussion

5.1. Training Details

We use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and with learning rate decay 0.95. With a batch size of 16, each minibatch consists of a randomly selected and aligned input-cartoon pair with the corresponding artist warp. Two types of online data augmentation are applied to the input images: random horizontal flips, as well as color jitter (brightness, contrast, and saturation jitter each uniformly sampled from the range $[0.9, 1.1]$ and hue jitter

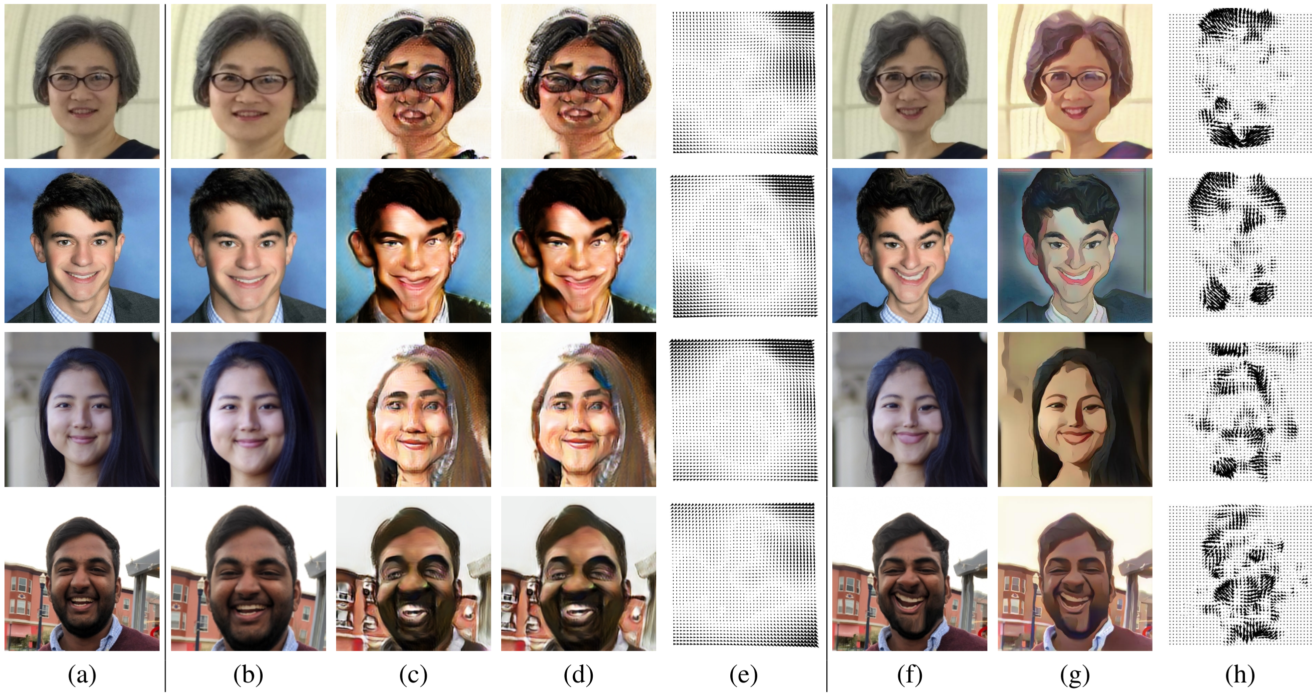


Figure 4: Comparison of our method to WarpGAN [28] to visualize disentanglement of geometry from style. From left to right: input images (a) from our test set, (b) the result of passing the image through WarpGAN's warping module without performing stylization, (c) stylizing using WarpGAN's encoder and decoder, but without the warping module, (d) the final output of WarpGAN, and (e) the visualized WarpGAN warping fields. Then, we have exaggerated cartoons (f) generated by our model, our model's cartoons stylized with CartoonGAN [6] to create caricatures (g), and our visualized warping fields (h). See supplementary materials for more comparisons.

uniformly sampled from the range $[-0.05, 0.05]$ as specified by the PyTorch color jitter API). We empirically set $\lambda_1 = 1$, $\lambda_2 = 0.7$, and $\lambda_3 = 1e$-6. All experiments were conducted with PyTorch version 1.1 on Tesla V100 GPUs.

5.2. Ablation Study

We train three additional variations of our model to analyze the contribution of each loss function to the system performance, as shown in Figure 5. Without the artist warp loss, the warps are much weaker and constrained to detailed features, and they do not dramatically alter the face shape. Without the reconstruction loss, the warps are larger in scope, but twist the face dramatically to the point where it unnaturally distorts the face. Without the proposed cosine similarity regularization loss, the warping field is less smooth and introduces some implausible asymmetries, artifacts, and inconsistencies in the facial warping.

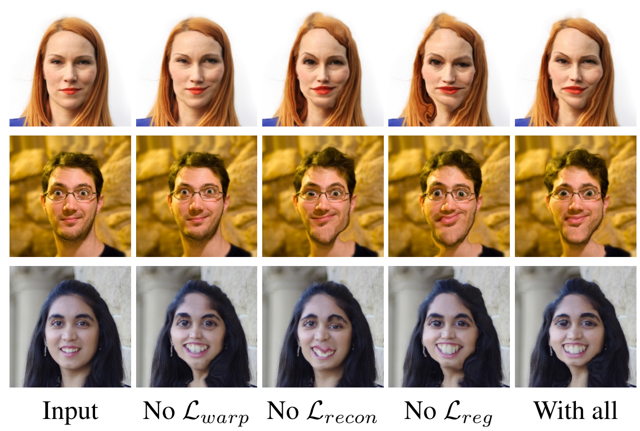


Figure 5: Cartoons of model variations without each proposed loss on images from validation (first two) and test set (last). Photos 1 and 2 by Pirátská strana and Frédéric de Villamil; modified.

5.3. Warping Quality User Study

We conducted two user studies to assess the quality of the warps learned by AutoToon. Since our contribution is purely the warping component of the caricature generation framework, we evaluated the quality of our warps against the performance of the warping module in the state-of-the-art, WarpGAN [28]. For each of 24 images, we asked 14 trained artists to provide ratings of cartoons generated by each network from 1 (worst) to 10 (best) for exaggeration quality, or the exaggeration of the subject's most prominent features. We also asked 37 casual observers to select between AutoToon cartoons and WarpGAN cartoons for the more "visually convincing" cartoon for the subject. To ensure earnest responses, the participants were in a controlled

Score	WarpGAN [28]	AutoToon
Exaggeration	3.2	**4.5** ($p < 0.01$)
User Preference	30.1%	**69.9%** ($p < 0.0001$)

Table 1: Results (averages) of user studies for artists and casual observers. Artists rated images from 1 (worst) to 10 (best). Casual observers chose the image with more convincing exaggeration; the proportion of user selections for each model are shown here.

setting and attentive to the task rather than randomly crowd-sourced. These results are shown in Table 1.

AutoToon consistently performs higher for both casual observers and artists ($p < 0.0001$ from 1-sample proportion test, $p < 0.01$ from 2-sample t-test respectively), making it a strong warping module for cartoon generation. We hypothesize that $\sim 30\%$ of users preferred WarpGAN cartoons because these images are often warped so weakly that they nearly exactly match the original image (see Figure 4), creating such a stark contrast to AutoToon's that users perceived AutoToon's as distorted. The artists also provided feedback that they would have liked to see even more symmetry and less distortion in AutoToon warps, but that they preferred this to WarpGAN warps that did not alter specific facial features and only mildly stretched the image. We leave these improvements to future work.

5.4. Disentanglement of Geometry and Style

Disentangling warping and stylization is valuable for providing greater flexibility in combining warped images with different styles, as well as potential uses where only pure warping is desired to create photorealistic deformation. It also encourages preservation of details, as we will discuss shortly. A strong warping module is thus an important contribution to a complete caricature pipeline. AutoToon only performs geometric warping, so its output is photorealistic and can be separately stylized by any stylization method.

To evaluate AutoToon's warping quality, we can compare it to the warping module of WarpGAN [28] by evaluating the extent of disentanglement. In Figure 4, we examine WarpGAN's output image with only warping from its warping module, only stylization, both warping and stylization, and the corresponding warping field. We compare the warping-only cartoon to the output cartoon of AutoToon, the result of applying stylization to this output to create the final caricature, and the warping field learned by AutoToon.

We find that WarpGAN's cartoon images (b) do not significantly deviate from the input images (a), only providing relatively coarse and somewhat weak warping. The geometric differences between the stylized images (c) and the final caricatures (d) are also minimal. We can confirm that the warps are not that strong and do not provide a clear signal of distinguishing facial characteristics between identities by

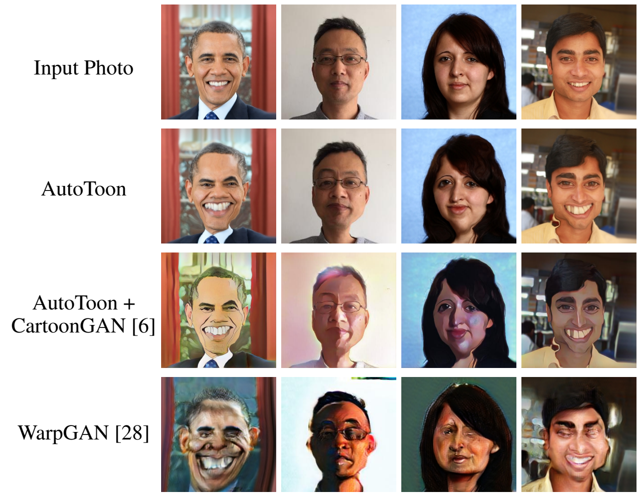


Figure 6: Comparison of AutoToon to WarpGAN [28] w.r.t. facial detail preservation (first two test, last two validation images). Photos 3 and 4 by Robby Schulze and Possible; modified.

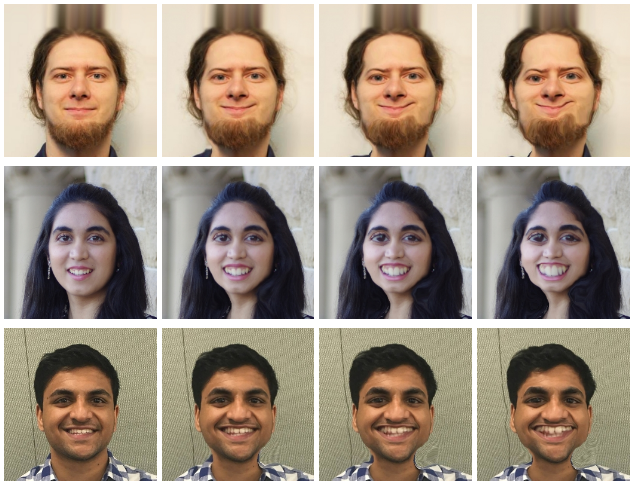
looking at the warping fields (e), which have very general shapes. Thus, WarpGAN's stylization network carries the majority of the geometric contribution to the final caricature in looking at the difference between the inputs (a) and stylized images (c).

In contrast, applying stylization (g) to AutoToon's outputs does not significantly alter the geometry of the cartoons (f), and geometric differences between (a) and the cartoons (f) are large, so the vast majority of the geometric contribution to the final caricature comes from AutoToon. Note also the strength and specificity of the warps learned by AutoToon in (h). Not only are the warps larger in magnitude and localized around facial features, but they are also clearly different for each identity on the level of facial features.

5.5. Preservation of Facial Detail

Caricatures need not sacrifice visual quality of the input image when exaggerating salient facial characteristics. However, due to the incomplete disentanglement of geometry and style in WarpGAN, there exists an inherent tradeoff between stylization and facial detail preservation. As shown in Figure 6, WarpGAN's style is inseparable from its warping, creating inconsistencies or sacrificing details of the eyes, lowering the caricature quality. On the other hand, AutoToon exaggerates yet still preserves the overall quality and consistency of facial features in a way that is faithful to the original image, especially with respect to details such as the eyes, ears, and teeth. This is especially noteworthy because of the difficulty of convincingly preserving facial detail in a photorealistic image due to the lack of stylization that could potentially compensate for any warping artifacts.

It is also interesting to note that while AutoToon preserves facial plausibility, it is also in "toon" with facial



Input $\alpha = 1$ $\alpha = 1.5$ $\alpha = 2$

Figure 7: Result of scaling the warping field of various examples from the test set with scaling factor α.

(a) (b) (c) (d)

Figure 8: Input images (a) from test (first, third) and validation set (second), unwarped, stylized images generated by WarpGAN [28] (b), warped caricatures generated by WarpGAN (c), and (d), result of applying the warping fields generated by AutoToon to (b).

asymmetries. For example, in Figure 7, the second subject's left eye (from their perspective) is slightly smaller than their right; with increases in the scaling factor α, this asymmetry is amplified. We also see similar amplifications for the crooked smile of subject 4 in Figure 4 and the smirks of subjects 1 and 4 in Figure 10. This sort of exaggeration of asymmetry is crucial for creating caricatures because they often mark distinguishing features in individuals' faces.

5.6. AutoToon Warp Transfer

To illustrate the efficacy of AutoToon warps, we show in Figure 8 the effect of applying AutoToon warps to stylized WarpGAN test images, in comparison to the end-to-end WarpGAN caricatures. The resulting images have stronger warps that enhance the prominent features of the subjects.

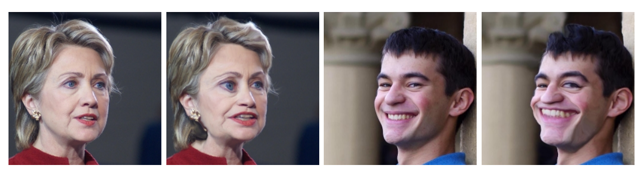
The warping quality of AutoToon warps can also be observed through manipulating the scaling factor α, which scales the magnitude of the warping field used to generate the cartoons as shown in Figure 7. Larger scale factors create more intense exaggerations, but still remain plausible and maintain the overall warping quality.

5.7. Facial Feature-Specific Warping

Despite the small dataset size, AutoToon has learned a diverse range of warping styles, and in particular, specific facial feature-level exaggerations that are distinct for different individuals. Examples of different learned facial feature warps are shown in Figure 10. Many other examples exist, including the curved smile of the second individual in Figure 4. In contrast to previous work that utilizes sparse warping, this more granular level of amplification helps to bring out more nuanced features of an individual's face beyond a rough exaggeration of face shape.

5.8. Face Pose Generalization

Though only trained and validated on frontal-facing images, AutoToon performs relatively robustly on images in the test set with subjects that deviate from the frontal pose, shown in Figure 9. This suggests that the Perceiver Network has successfully captured face features that are robust to changes in angle and position.



Input Cartoon Input Cartoon

Figure 9: Model generalization to non-frontal test set images.

5.9. Visualization of Network Attention

In order to get a sense of the features used by our method to generate cartoons, we employ guided backpropagation [31] that we couple with smoothgrad [30] for a more stable analysis. We visualize the result of this analysis on 4 different images from our validation set in Figure 11.

5.10. Limitations

Some limitations of AutoToon are illustrated in Figure 12. Compared to the ground-truth image, the model incorrectly enlarges the eyes in (a), likely because bulging of eyes is very common in the dataset. The chin in (a) and mouth and eyebrows of (b) are not as successfully warped and introduce some distortion and warping artifacts.

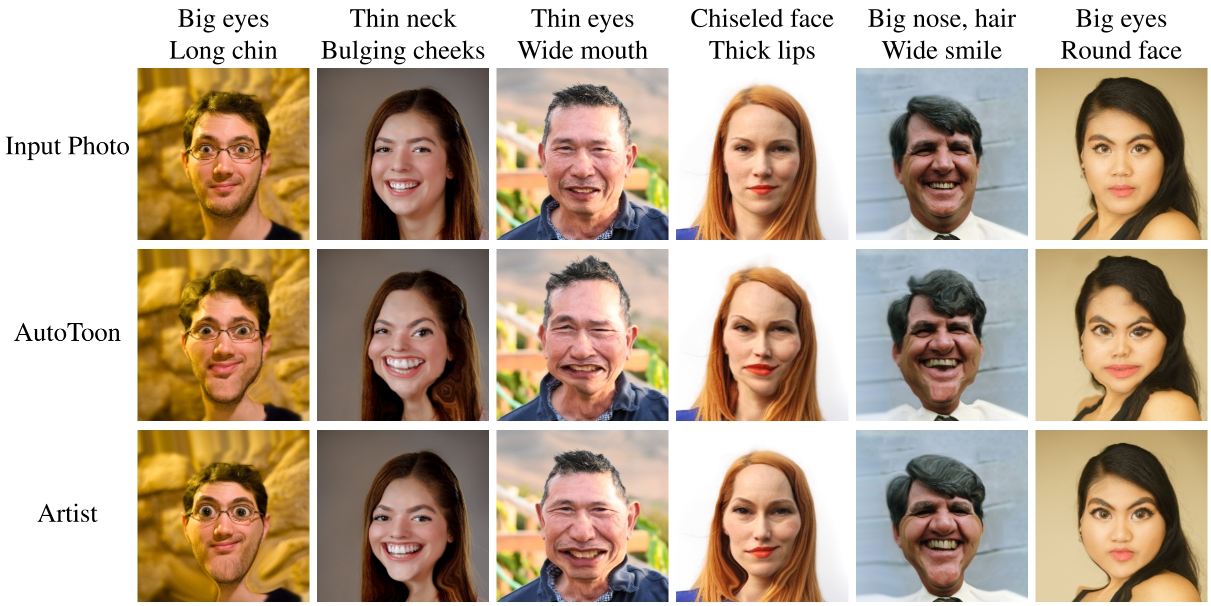


Figure 10: Examples of different detailed, face feature-specific exaggerations on the validation set learned by AutoToon as compared to artist cartoons. Shown are the input images, cartoons generated by our model, and the corresponding artist cartoons for the same subject. See supplementary materials for more results. Photos 2, 5, 6 by Jacob Seedenburg, Community Archives, and Aaron Stidwell; modified.

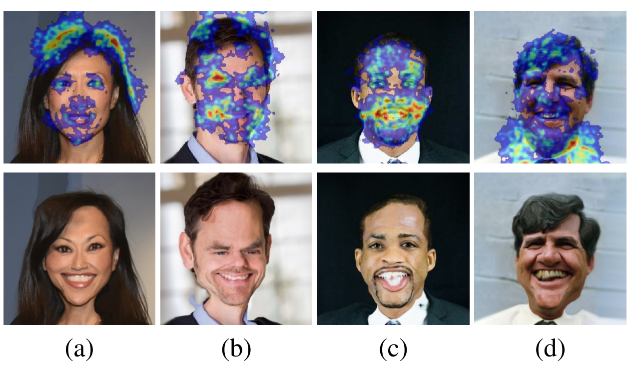


Figure 11: First row: network attention with smoothed guided backpropagation [30, 31] jet-overlaid on validation images (increasing from blue to red). Second row: generated cartoons. Our model focuses on specific features for each face, such as (a) hair and eyes, (b) eyes and smile dimples, (c) mouth, and (d) chin and neck. Photos 1 and 2 by Maryland GovPics and Si1very; modified.

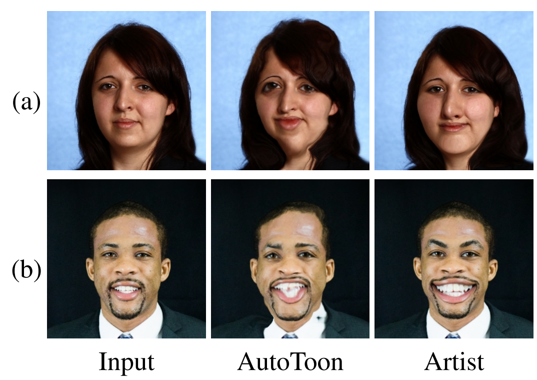


Figure 12: Model limitations illustrated by examples from the validation set, consistent with artist comments from the user study. Photo 2 by Vince Crabeo; modified.

6. Conclusion

In this paper, we present AutoToon, the first supervised deep learning method for cartoonization, or the warping step of facial caricature generation. Our warping method yields high-quality warps that outperform the state-of-the-art. Our model is also disentangled entirely from style, allowing it to be paired with any stylization network, including existing caricature generation models, to create diverse caricatures. Unlike previous work, it leverages the power of the SENet and differentiable warping module, and also learns directly from artist warping fields. In addition to creating convincing exaggerations that are subject- and facial feature-specific, it also preserves facial detail faithful to the original image and generalizes to non-frontal portrait images. We evaluated these caricatures qualitatively in comparison to prior art with respect to geometry and style disentanglement, facial detail preservation, and warping quality and feature-level specificity, and quantitatively showed through our user study and artist ratings that AutoToon outperforms state-of-the-art networks in geometric warping. Future directions of interest include further smoothing of the warping field to avoid pixel collision, identity preservation, and few-shot learning to adapt to different artist styles.

References

[1] E. Akleman. Making caricatures with morphing. In *ACM SIGGRAPH 97 Visual Proceedings: The Art and Interdisciplinary Programs of SIGGRAPH '97*, SIGGRAPH '97, pages 145–, New York, NY, USA, 1997. ACM.

[2] E. Akleman, J. Palmer, and R. Logan. Making extreme caricatures with a new interactive 2d deformation technique with simplicial complexes. *Proceedings of Visual 2000*, 01 2000.

[3] S. E. Brennan. The dynamic exaggeration of faces by computer. *Leonardo*, 18(3):170–178, 1985.

[4] K. Cao, J. Liao, and L. Yuan. Carigans: Unpaired photo-to-caricature translation, 2018.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] Y. Chen, Y.-K. Lai, and Y.-J. Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[8] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2017.

[9] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*, pages 311–326. Springer, 2016.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[11] B. Gooch, E. Reinhard, and A. Gooch. Human facial illustrations: Creation and psychophysical evaluation. *ACM Transactions on Graphics (TOG)*, 23(1):27–44, 2004.

[12] X. Han, K. Hou, D. Du, Y. Qiu, Y. Yu, K. Zhou, and S. Cui. Caricatureshop: Personalized and photorealistic caricature sketching. *arXiv preprint arXiv:1807.09064*, 2018.

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[15] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin. Webcaricature: a benchmark for caricature recognition. In *British Machine Vision Conference*, 2018.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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.

[17] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[18] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[19] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[20] N. K. H. Le, Y. P. Why, and G. Ashraf. Shape stylized face caricatures. In K.-T. Lee, W.-H. Tsai, H.-Y. M. Liao, T. Chen, J.-W. Hsieh, and C.-C. Tseng, editors, *Advances in Multimedia Modeling*, pages 536–547, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[21] L. Liang, H. Chen, Y.-Q. Xu, and H.-Y. Shum. Example-based caricature generation with exaggeration. In *Pacific Conference on Computer Graphics and Applications*, pages 386 – 393, 02 2002.

[22] P.-Y. C. W.-H. Liao and T.-Y. Li. Automatic caricature generation by analyzing facial features. In *Proceeding of 2004 Asia Conference on Computer Vision (ACCV2004), Korea*, volume 2, 2004.

[23] J. Liu, Y. Chen, and W. Gao. Mapping learning in eigenspace for harmonious caricature generation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 683–686. ACM, 2006.

[24] Z. Mo, J. Lewis, and U. Neumann. Improved automatic caricature by feature normalization and exaggeration. *SIGGRAPH Sketches*, 08 2004.

[25] Z. Mo, J. P. Lewis, and U. Neumann. Improved automatic caricature by feature normalization and exaggeration. In *Siggraph Sketches*, page 57, 2004.

[26] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1164–1172, 2015.

[27] G. Rhodes, S. Brennan, and S. Carey. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, 19:473–497, 1987.

[28] D. D. Shi, Yichun and A. K. Jain. Warpgan: Automatic caricature generation. In *Conference on Computer Vision and Pattern Recognition*, 2019.

[29] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Josa a*, 4(3):519–524, 1987.

[30] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[32] D. Y. Tsao and M. S. Livingstone. Mechanisms of face perception. *Annual Review of Neuroscience*, 31(1):411–437, 2008. PMID: 18558862.

[33] Q. Wu, J. Zhang, Y.-K. Lai, J. Zheng, and J. Cai. Alive caricature from 2d to 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7336–7345, 2018.

[34] R. Wu, X. Tao, X. Gu, X. Shen, and J. Jia. Attribute-driven spontaneous motion in unpaired image translation. *arXiv preprint arXiv:1907.01452*, 2019.

[35] W. Wu, K. Cao, C. Y. Li, C. Qian, and C. C. Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. *ArXiv*, abs/1904.09571, 2019.

[36] W. Yang, M. Toyoura, J. Xu, F. Ohnuma, and X. Mao. Example-based caricature generation with exaggeration control. *The Visual Computer*, 32(3):383–392, 2016.

[37] F. Zhang and F. Liu. Parallax-tolerant image stitching. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 3262–3269, Washington, DC, USA, 2014. IEEE Computer Society.

[38] S. Zhang, R. Ji, J. Hu, Y. Gao, and C.-W. Lin. Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1163–1169. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[39] Y. Zhang, W. Dong, C. Ma, X. Mei, K. Li, F. Huang, B.-G. Hu, and O. Deussen. Data-driven synthesis of cartoon faces using different styles. *IEEE Transactions on image processing*, 26(1):464–478, 2016.

[40] Y. Zhao, Z. Huang, T. Li, W. Chen, C. LeGendre, X. Ren, J. Xing, A. Shapiro, and H. Li. Learning perspective undistortion of portraits. *arXiv preprint arXiv:1905.07515*, 2019.

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.