Cloud2Sketch: Augmenting Clouds with Imaginary Sketches

Zhaoyi Wan
University of Rochester
United States
i@wanzy.me

Dejia Xu
University of Texas at Austin,
Snap Inc.
United States
dejia@utexas.edu

Zhangyang Wang
University of Texas at Austin
United States
atlaswang@utexas.edu

Jian Wang*
NYC Research Lab, Snap Inc.
United States
jwang4@snapchat.com

Jiebo Luo*
University of Rochester
United States
jluo@cs.rochester.edu

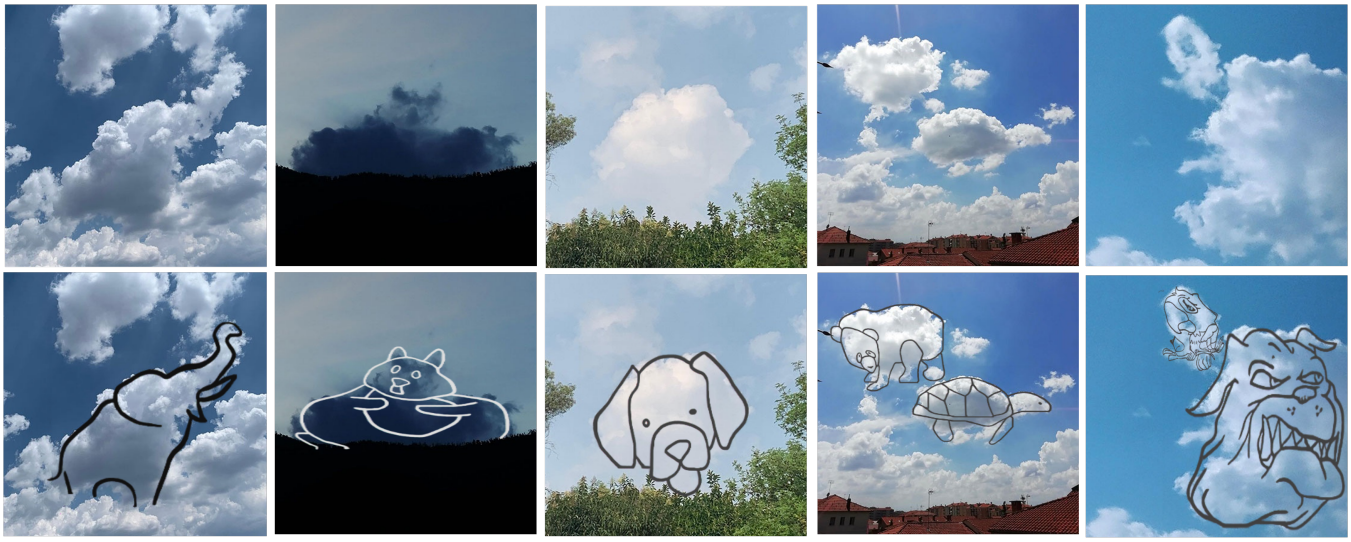

Figure 1: Cloud augmentation with imagined animal sketches. We propose a novel pipeline to draw sketches on cloud images so that the imagination of clouds as animals are visualized. The sketch drawings are naturally aligned with cloud contours.

ABSTRACT

Have you ever looked up at the sky and imagined what the clouds look like? In this work, we present an interesting task that augments clouds in the sky with imagined sketches. Different from generic image-to-sketch translation tasks, unique challenges are introduced: real-world clouds have different levels of similarity to something; sketch generation without sketch retrieval could lead to something unrecognizable; a retrieved sketch from some dataset cannot be directly used because of the mismatch of the shape; an optimal sketch imagination is subjective. We propose Cloud2Sketch, a novel self-supervised pipeline to tackle the aforementioned challenges. First, we pre-process cloud images with a cloud detector and a thresholding algorithm to obtain cloud contours. Then, cloud contours are passed through a retrieval module to retrieve sketches with similar geometrical shapes. Finally, we adopt a novel sketch translation model with built-in free-form deformation for aligning the sketches to cloud contours. To facilitate training, an icon-based sketch collection named Sketchy Zoo is proposed. Extensive experiments validate the effectiveness of our method both qualitatively and quantitatively. Our code and data are publicly available[1].

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; **Computer vision**.

KEYWORDS

image-to-sketch generation, cloud augmentation, shape alignment, sketch synthesis

*Co-corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547810

[1] https://wanzy.me/research/cloud2sketch

ACM Reference Format:
Zhaoyi Wan, Dejia Xu, Zhangyang Wang, Jian Wang, and Jiebo Luo. 2022. Cloud2Sketch: Augmenting Clouds with Imaginary Sketches. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3547810

1 INTRODUCTION

From time to time, we look into the sky and associate clouds with imaginary objects. In the sky, no two clouds are ever the same, while through our mind's eyes, we make imaginations and associations with similar animals.

It is human nature and creativity to imagine random natural shapes, such as clouds, as familiar shapes, such as animals [10]. Humans recall in their minds a collection of previously seen objects [6] and project them onto seen clouds. We attempt to bring this interesting imagination process into reality in this work. As shown in Fig. 1, we propose a new application: augmenting cloud images with imaginary sketch drawings that fit the shape of clouds.

Techniques for augmenting natural images have an important place in augmented reality applications, such as entertainment applications [32], education techniques [23], and human-computer interaction [14]. Due to its ubiquitous ability to represent visual objects, sketch is a natural media for these kinds of applications. As a result, sketch retrieval [43, 45, 46] and synthesis [41] are drawing an increasing amount of research attention. Meanwhile, our application is different from the existing sketch-related research tasks. In comparison to classical sketch retrieval, cloud augmentation requires a precise alignment of geometrical shapes. As for sketch synthesis, which exactly depicts natural scenes or objects with schematic drawing, imagined sketches are drawn from incomplete cues. An automatic system for achieving cloud augmentation is not only a reflection of human imagination but may also surprise humans with diverse results as shown in Fig. 2. On the other hand, different from a generic image-to-sketch translation task, unique challenges are introduced in cloud image augmentation with imaginary sketches.

For the task itself, the shape features of clouds are diverse and possibly vague. The imagination should focus on the overall shapes but neglect noisy or trivial shape details. Therefore, we make associations with animal sketches according to their contour shapes. On the other hand, even the outer shape of clouds does not necessarily correspond to a meaningful object (specifically animals in this work). Therefore, it is important for a valid cloud augmentation algorithm to add interior strokes inside the cloud area to ensure a vivid drawing.

From the dataset perspective, existing datasets cannot be simply transferred into our application because: (1) Sky detection / segmentation datasets regard clouds as continuous regions without distinguishing instances; (2) Sketch retrieval is normally defined as a search process according to semantic similarities while our application aims at geometrical correspondences; (3) The style of dominant human-hand sketch datasets is not expressive enough for aesthetic applications. From the model perspective, there are also challenges that typical image-to-image translation methods can hardly address. Cloud contours and sketch drawings represented in raster images are sparse in comparison with natural images, making

Figure 2: Cloud shapes are so diverse that some shapes invoke ambiguous imagination. Therefore, a computational algorithm may produce multiple results for the same cloud shape, which may inspire humans with different imagination. The finding is verified by the user study in Sec. 5.

it difficult to extract meaningful features. Moreover, the alignment of a sketch from imagination towards cloud contours is agnostic in two respects. First, sketches from imagination are not necessarily topologically equivalent to the targeted cloud shape in our application. It might make an ideal alignment impossible. Second, imagination-based augmentation is subjective and cloud shape is represented as incomplete contours. Therefore, there usually exist more than one optimal sketch drawings for a given cloud.

In this paper, we propose a self-supervised pipeline to tackle these challenges. Inspired by the human process of imagining and sketching up objects, we propose to break down cloud augmentation in images into three separate stages to which the algorithm can be flexibly applied and adjusted. First, a natural image with clouds is pre-processed to obtain the contours of the clouds. An input image is transformed into an edge map in grayscale, to which a detector is applied to detect possible cloud regions. Inside the detected bounding boxes, we assume that there are mostly sky and cloud pixels. Thus, we introduce and develop a segmentation algorithm based on thresholding to parse cloud contours. Second, a contour-to-sketch retrieval model is trained to retrieve sketches geometrically similar to the contours detected in the first stage. Finally, the alignment between the sketch and contour is achieved by a novel generation model built with free-form deformation (FFD). As shown in Fig 1, from a given cloud image, the proposed pipeline manages to perform an association to an animal sketch and a smooth alignment.

To facilitate training of our proposed pipeline, we collect an artist-drawn sketch collection named Sketchy Zoo. It contains 3464 mostly

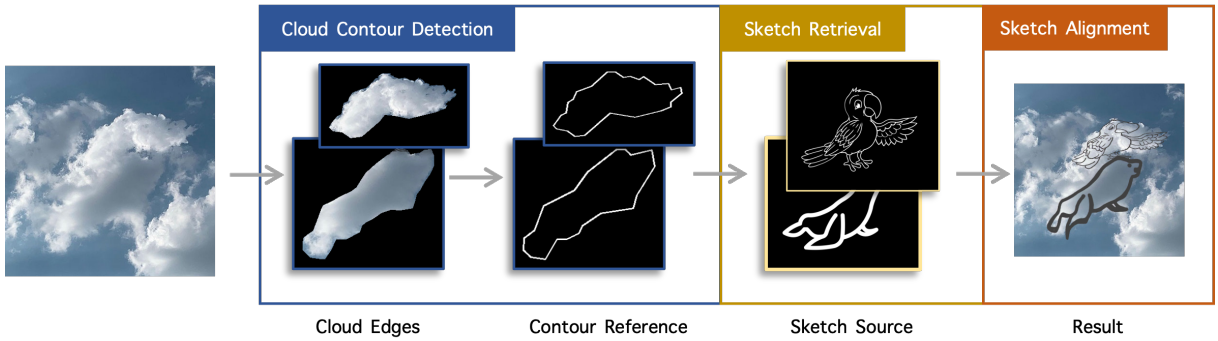

Figure 3: Our proposed pipeline for augmenting clouds with imaginary sketches. We propose to decompose the task into three stages, 1) cloud contour detection to obtain cloud contours as shape features, 2) sketch retrieval to search for geometrically similar sketch candidates, and 3) sketch alignment to smoothly align retrieved sketches and produce drawings on cloud.

animal sketches that are originally designed for aesthetic purposes by professionals. The proposed pipeline is trained with Sketchy Zoo and validated by extensive experiments both qualitatively and quantitatively. We also conduct a user study to collect comments and feedback. It provides an interesting comparison between the computational algorithm and human imagination in Sec 5.

In summary, our contributions are several folds:

(1) We propose a novel task, augmenting clouds in images with imaginary sketches, to bring human imagination of clouds into reality.
(2) A novel and effective pipeline is conducted to tackle the proposed task. It can be trained in a self-supervised manner without human labeling.
(3) We collect a dataset of an artist-drawn sketch collection for aesthetic applications. Using the collected sketches, we validate the proposed pipeline with extensive experiments and a user study.

2 RELATED WORKS

2.1 Imagination for machine vision

Pioneers are exploring simulating human imagination with artificial intelligence [15, 24]. Mahadevan [31] presents a new challenge to integrate imagination into machines. Hamrick [12], Wang et al. [42] propose to align learning systems with human imagination. While these explorations are still at their early stages, there are works that aim at bringing specific imagination into reality with methods in the current realm of computer vision and deep learning. Similar to our direction, Song et al. [37] proposes to implement the mental image of face pareidolia that perceives illusory faces do not actually exist with a computational algorithm. Our proposed application task belongs to this direction of pareidolia but aims at a totally different application and uses different methods.

2.2 Sketch Synthesis

Sketch synthesis has been widely studied due to the ability to represent objects and scenes abstractly. Human faces [40] and scenes [44] are the most common sources for sketch synthesis. Li et al. [25] learns to produce boundary-like drawings that capture the outline of the visual scene and can work well despite the imperfect alignment of the annotation and the actual ground truth. More recently, Chan et al. [4] adopts CLIP network [35] to generate informative drawings which is aware of geometry and semantics. In vanilla image-to-sketch translation, sufficient information is given by the natural images and models filter redundant details out while keeping the schema strokes. In contrast, our application requires models to hallucinate details that can not be perceived from input images.

With the rapid development of sketch synthesis techniques, many datasets [8, 11, 18, 25, 33, 36] have been proposed to facilitate the training of learning-based methods. However, these existing datasets are mostly designed for recognition of sketches [36], boundary detection [33], and image synthesizing from sketches [8, 11]. Although different levels of abstractions are provided, they usually contain badly drawn sketches. To this end, we propose an icon-based sketch collection named Sketchy Zoo, which contains sketch images drawn and curated by professional painters.

2.3 Spatial Transformation Networks

Spatial transformation networks are first introduced as Spatial Transformer Network (STN) [22] to extract invariant features against spatial transformation to help digital classification. STN is further extended to learn other kinds of transformations, such as projective or TPS [2]. Free-form Deformation (FFD) is recently introduced to model more partial shape alignment problems in [13]. In our scenarios, the deformation is more complex and flexible that we integrate FFD and a sketch generator to achieve both spatial transformation and detail preservation.

3 METHODOLOGY

In this section, we present our proposed method. It draws inspiration from how humans imagine and sketch objects [8]. A human usually (1) builds observation of the shape of the target object [9], specifically a cloud in our scenario, in mind, (2) makes association from the observation against his *visual corpus* [6], and (3) projects the association together with the object to form an "imagination" [3]. Analogous to the human process, the overall architecture of our proposed method consists of three main components: cloud contour detection, association search in a pre-collected sketch collection,

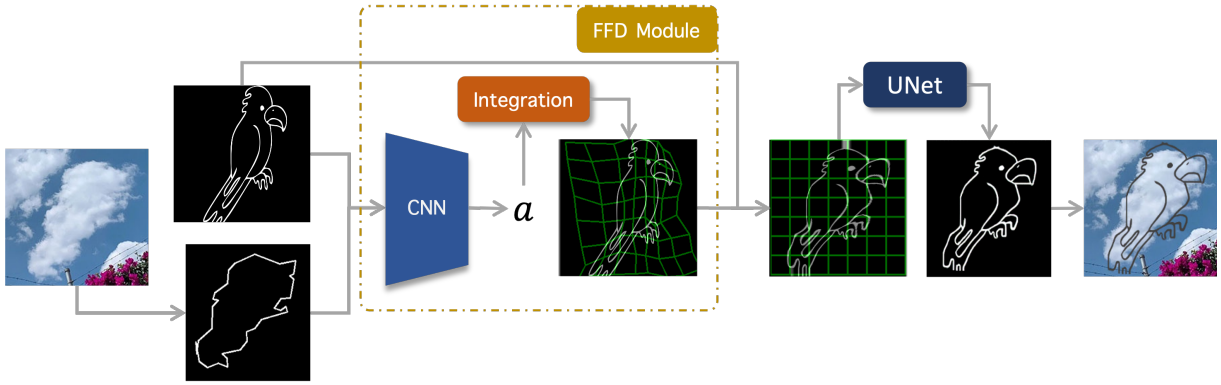

Figure 4: Agnostic sketch alignment using the FFD Generator. The FFD Generator estimates a in Eq. 2 to form an FFD from contours of clouds to obtain course generation. A UNet follows the deformation module to refine the generated sketch.

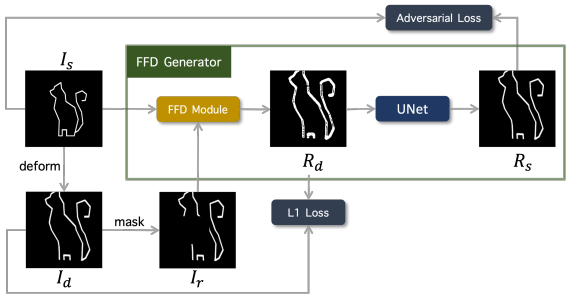

Figure 5: The training process of the FFD generator. It follows a self-supervised manner, where a sketch and its augmented view form a data pair.

and sketch alignment. The pipeline is shown in Fig. 3. In the following subsections, we introduce each of the components in detail.

3.1 Agnostic Sketch Alignment

The core idea is that the cloud itself *shapes* an animal sketch in our corpus, but in an agnostic spatial form. The task of agnostic sketch alignment is to align a sketch drawing *source* towards the (possibly partial) cloud shape *target* while keeping its gist and style. It is achieved by a network with two sequential components illustrated in Fig. 4. First, a spatial deformation is estimated from the concatenation of source and target to obtain a coarse alignment. Second, the deformed source sketch is refined by a UNet [21] model to recover the style and eliminate artifacts caused by the deformation.

FFD Module Inspired by STN [22] and ALIGNet [13], we utilize a lightweight CNN to estimate spatial transformation. Different from the conventional STN based on affine transformation, our FFD Generator is built with a more flexible FFD module. The estimation CNN can be a regular down-sampling network, e.g., ResNet18 [17] in our experiments, that yields a sampling grid at a lower resolution. The grid is interpolated to a higher resolution before warping is performed.

Formulation of FFD estimation is supposed to encourage smooth deformation and restrict the violation of monotonicity. Let us consider a 1D sequence $S = \{s_1, s_2 ... s_L\}$ with length L for simplicity. An identical sampling grid (sequence for 1D) $\mathcal{P}_{\mathbb{I}} = \{1, 2 ... L\}$ can be represented by a basis and accumulated shifts:

$$\mathcal{P}_{\mathbb{I}}^k = \begin{cases} \mathcal{P}_{\mathbb{I}}^{k-1} + \Delta\mathcal{P}_{\mathbb{I}}^k, & \text{if } L \geq k \geq 1 \\ 0, & \text{otherwise} \end{cases}$$
$$\Delta\mathcal{P}_{\mathbb{I}}^k = \frac{1}{L}. \tag{1}$$

Similarly, FFD in general cases can be formulated by a sequence of accumulating sampling shifts $a = \{a_1 ... a_L\}$:

$$\mathcal{P} = \{a_0 + \Sigma_{j=1}^{k} a_j\}_{k=1}^{L}. \tag{2}$$

We train a CNN to estimate a and make a_0 a constant $1/L$. With cumulative summation and regulation, the FFD estimation roughly preserves axial monotonicity while keeping differentiable. The CNN is initialized to produce identical deformation that simply yields output the same as input at the start of training.

Self-supervised Training As mentioned above, the FFD generator aims to reconstruct a sketch given a partial and distorted input. It is essentially hard to achieve since there is no existing paired data for training. To this end, we propose a simple yet effective self-supervised training strategy as shown in Fig. 5. To start with, a sketch I_s randomly sampled from our sketch corpus serves as *source* for training. It is then geometrically distorted and masked out as another view I_r to mock a partial *target*. The distortion consists of shear transformation and piece-wise affine transformation [34] after which we obtain I_d before masking. The mask is generated from a uniform distribution in a region occupying 30% of the image in height. The FFD generator estimates a free-form deformation that warps the sampled source referring to the partial target:

$$f_e(I_s, I_r)(I_s) = R_d \xrightarrow{shape} I_d,$$
$$f_g(R_d) = R_s \xrightarrow{style} I_s, \tag{3}$$

where $f_e(\cdot)$ and $f_g(\cdot)$ are the FFD module and refine generator, respectively. We enforce pixel-wise alignment between the augmented view I_d and the output of the FFD Module R_d:

$$\mathcal{L}_p = |I_d - R_d|. \tag{4}$$

On the contrary, the output of the refining UNet, R_s, is not well-aligned with the input I_s but is distorted due to the non-rigid deformations. Neither can it be supervised directly via the deformed result I_d, because I_d suffers from unrealistic shapes. To overcome the obstacle, we propose to regularize via adversarial training. We empirically adopt a Markovian discriminator [21] since it penalizes structure at the scale of local image patches. The adversarial loss is formulated as follows:

$$\mathcal{L}_{\text{adv}} = \log D(I_s) + \log(1 - D(R_s)), \tag{5}$$

where R_s is considered as negative samples and I_s is seen as positive samples. We further adopt a regularizor on the estimated deformation a_i according to Eq. 2:

$$\mathcal{L}_r = \sum ||a_i - a_0||. \tag{6}$$

This design pulls a_i towards a_0, limiting the magnitude that the estimated deformation differs from the identical free form deformation. The overall loss function for our FFD Generator can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_p + \lambda_a \mathcal{L}_a + \lambda_r \mathcal{L}_r, \tag{7}$$

where we set $\lambda_a = 1$ and $\lambda_r = 0.01$ for all experiments.

3.2 Sketch Retrieval

The sketch retrieval module is the key to our imagination. By incorporating a diverse sketch corpus, we are able to locate the most relevant sketches every time we extract the boundaries from clouds. We then warp the reference sketches to align with the cloud edges and the reference sketches are baselines of our imagination.

We implement our sketch retrieval module as a lightweight ResNet. The training strategy is illustrated in Fig. 6. We train the module in a contrastive manner using triplet loss

$$\mathcal{L}(A, P, N) = \max\left(\|\mathrm{f}(A) - \mathrm{f}(P)\|^2 - \|\mathrm{f}(A) - \mathrm{f}(N)\|^2 + \alpha, 0\right), \tag{8}$$

where A, P, N are anchor, positive, negative input, respectively.

The module is capable of learning translation-invariant representations of sketch images. To this end, we generate the positive samples from the anchor inputs using various data augmentation and distortion strategies. The negative samples are selected randomly, with a 50% possibility of having the same class as the positive samples, and a 50% possibility of having different a different class.

3.3 Cloud Edge Detection

For the source of augmentation, we look for clear cloud edges. However, not all forms of clouds are suitable for augmenting. They have different structures under different regions and conditions of the atmosphere. In Appendix, we show cloud genera in different physical forms [1], where we mainly consider cumuliform and cumulonimbiform clouds as detection targets.

Several cloud detection/segmentation datasets have been proposed in the literature [7, 38], while none of them achieves the separation of cloud layers and instances that are critical for our application. To avoid laborious data labeling, we train our detector with synthetic data on edge maps. We apply an off-the-shelf edge detection algorithm [39] on a popular object detection dataset,

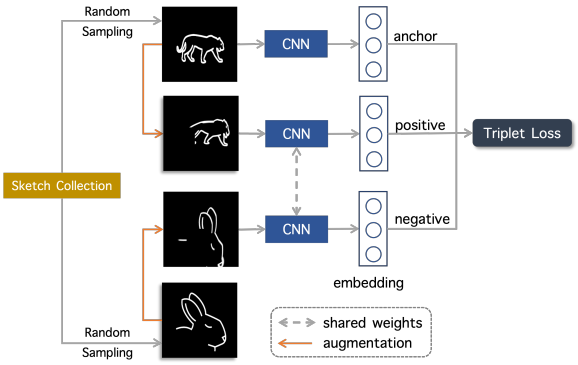

Figure 6: Contrastive training of contour-based sketch retrieval. We randomly sample an anchor sketch and a negative sketch. The anchor in its distorted view is regarded as a positive sample.

COCO [28], to obtain our training data with labels provided by COCO. During inference, the same algorithm is applied to the sky area of the input image, after which our detector [27] trained on COCO edges detects possible cloud regions to be augmented. Examples of training data and detection results are shown in Fig. **??**. When detecting edges from sky images, we focus on cloud areas obtained by an off-the-shelf semantic segmentation model [5].

Inside bounding boxes of clouds, we use thresholding techniques inspired by [26] to locate cloud pixels. As shown in Fig. 13 in the Appendix, the histograms of red and blue channels of an RGB image are informative for distinguishing cloud regions. An ideal feature space for cloud segmentation is supposed to be bi-modal where the foreground can be separated from the background. Therefore, we use the normalized B/R ratio as feature of clouds images:

$$\begin{aligned} \gamma_N &= (\gamma - 1)/(\gamma + 1), \\ \gamma &= b/r, \\ \Leftrightarrow \gamma_N &= (b - r)/(b + r) \end{aligned} \tag{9}$$

where r and b are red and blue values in a raster image. We can perceive from Fig. 13 in the Appendix that the feature image of γ_N maintains more consistent contrast and is more robust to noise. The thresholds shown in Fig. 13 are chosen by maximum entropy from histograms. According to [30], the fixed threshold is not accurate for cumuliform. Thus, the watershed algorithm is applied in our algorithm. Then from the precise regions of clouds detected, we can easily obtain their contours by finding connected components.

4 DATASETS AND EXPERIMENTS

In this section, we present our collected sketch collection for training and test images for evaluation of our pipeline. Based on collected data, we train our proposed pipeline and conduct extensive experiments to validate it.

4.1 Datasets

We collect two datasets to train and validate the proposed pipeline: A sketch collection of high-quality sketch images and a test set of cloud images.

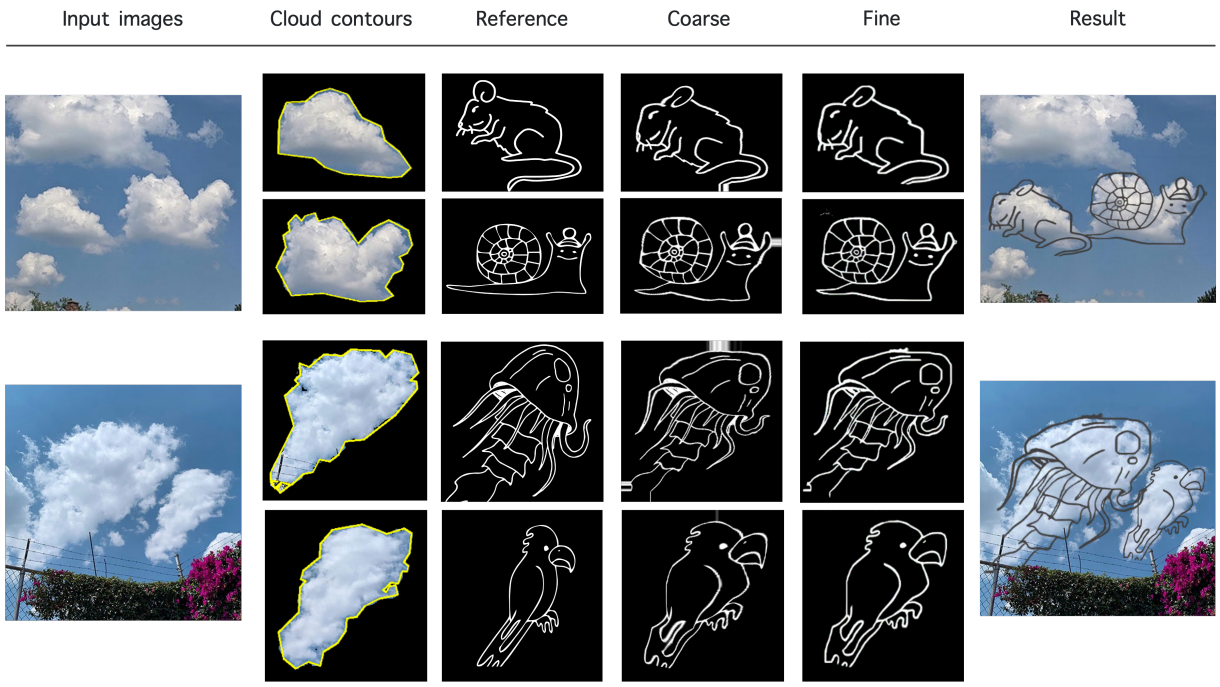

Figure 7: Qualitative presentation of augmented results by our proposed method. We also provide an extended application of our algorithm that augments other types of images (e.g., islands in the sea) in Appendix.

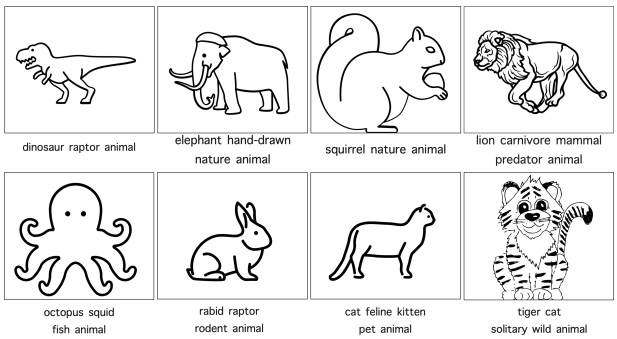

Figure 8: Examples of our collected sketch collection, named Sketchy Zoo. Sketchy Zoo consists of sketches created by professional designers for aesthetic purposes. Multiple tags are attached to each image for more general tasks.

Sketchy Zoo We collect a new sketch dataset consisting of 3464 images in total, all drawn and curated by professional painters. We refer to the dataset as *Sketchy Zoo*, since most of the sketches are cute animal icons with high quality and diversity. Initial data are crawled from an icon collection website[2]. We manually choose icon collections that are created with lines instead of stuffed colors. Our algorithm takes raster images as input, while vector graphics of collected sketches are also available.

[2]https://thenounproject.com/icons/

Sketchy Zoo is quite different from existing sketch datasets. Since it comes from designers for professional purposes, sketches in Sketchy Zoo are more vivid and expressive than those collected from user drawings. This quality advantage is substantially important for the success of sketch-based augmentation applications. Moreover, although class labeling is not requisite for our application, Sketchy Zoo provides tags describing the class and style of sketches labeled to each sketch. Therefore, it may open up new opportunities in image augmentation with sketches. Some examples of our datasets are shown in Fig. 8.

Cloud Images We collect ground-based cloud images to evaluate our proposed method. It contains 100 cloud images shot from the ground under diverse light conditions. The images are originally collected from Instagram by artists. We manually review and remove images without clear edges.

4.2 Qualitative Results

We use Sketchy Zoo to train our proposed method. Final augmenting results and intermediate results of each stage are shown in Fig. 1 and Fig. 7, respectively.

As stated in Sec. 3.3, three models are adopted sequentially to preprocess cloud images, DeeplabV3 [5] for sky region segmentation, DexiNed [39] turning RGB clouds into edges, and RetinaNet [27] for cloud contour detection on edge maps. Results of cloud instance segmentation and contour detection are shown in the second column in Fig. 7. It can be seen from the visualization that the detector locates salient clouds with clear edges that are possible for augmentation. The cloud segmentation algorithm presented in Sec. 3.3

Table 1: Quantitative results of average IOU and Moment Distance on our test set. A higher shape IOU indicates tighter alignment. Small Moment distance indicates similar shape properties. See Sec. 4.3 for details.

	Contour IOU	Moment L2 Distance
average	0.38	0.68
+ retrieval	0.39	0.55
+ FFD Generator	0.63	0.53

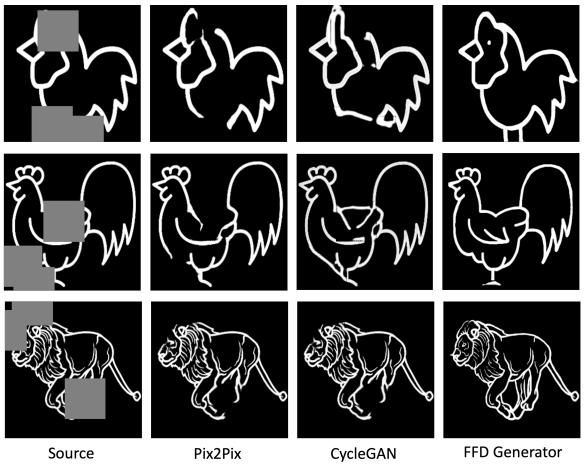

Figure 9: Comparison of the FFD Generator and the state-of-the-art image-to-image translation methods on sketch complement. Sketches from Sketchy Zoo are masked out as the source. Different from image inpainting, the gray regions are illustrated only for visualization, which is unknown to the models.

accurately locates clouds in different light conditions. The segmentation can be easily turned into a contour by finding the connected component with the maximum area inside each bounding box.

After pre-processing, the contour of target cloud regions serves as a reference for sketch retrieval in the corpus and the target for sketch alignment. The search in corpus manages to find sketches with similar geometrical shapes. The generated results demonstrate that the proposed FFD Generator smoothly aligns sketches with clouds and the gist of the sketches is faithfully preserved. We recommend seeing the Appendix for extended applications of the proposed pipeline as well.

4.3 Quantitative Evaluation

In addition to the qualitative results shown so far, we conduct our quantitative validation based on shape similarity to investigate the effect of sketch retrieval and alignment. In Tab. 1, we evaluate the augmentation with two metrics. Firstly, generated sketches are turned into contours, then IOU between pixels inside sketch contours and cloud contours can be computed. In addition, L2 distance between their Hu Moments [19] are introduced to measure the geometrical distance between generated sketch and cloud contours.

The average result of all images on our test set is shown in Tab. 1. In comparison to the average of total Sketchy Zoo, sketches from retrieval demonstrate significantly lower moment distance but similar contour IOU. The level of change is expected since Hu moments are invariant to translation, scale, and rotation. It verifies that our retrieval model retrieves sketches that maintain similar geometrical properties but is tolerant to differences in spatial forms. This spatial gap is fulfilled by FFD Generator. As Tab. 1 reads, FFD Generator elevates IOU between sketches and clouds, while maintaining and even improving the consistency in image moments. It indicates that FFD Generator effectively aligns sketches and preserves shape features.

4.4 FFD Generator

Since the effectiveness of our pipeline is verified by quantitative comparison, we conduct an ablation study focusing on FFD generators and possible substitutes.

We validate the effectiveness of our FFD generator with the sketch completion task to recover the original sketches from their masked-out views. We compare against state-of-the-art image-to-image translation methods, including Pix2Pix [21] and CycleGAN [47]. As shown in Fig. 9, sketches from Sketchy Zoo are masked out and fed with their deformed views into the networks. Training data of comparison methods are identical to that of our FFD generator. Although given the same input, Pix2Pix can only slightly extend lines at ends, without meaningful directions. CycleGAN produces more stable strokes. The line width is consistent and it tends to form closed shapes. However, it is still limited in building the geometrical relationship and thus, the result is incomplete. Benefiting from a different principle that estimates FFD from the partial source, FFD Generator faithfully completes the masked-out sketches. Note that different from image inpainting, the location of mask regions is unknown to models. To further investigate the comparison of completion capacity between Pix2Pix and CycleGAN, we present another interesting experiment in Appendix.

5 USER STUDY AND DISCUSSION

In this section, we conduct a user study to subjectively evaluate the proposed task and pipeline from the perspective of users. We invited 36 users with diverse backgrounds to participate in the study. Among them, 12 have advanced education in computer vision to balance the views of experts and common users. Each participant is asked to answer questions in a questionnaire anonymously and independently. Throughout the questionnaire, we avoid indicating that shown drawings are from human drawings or computational algorithms to prevent implication bias.

First of all, the participant is asked whether ever imagined clouds in the sky as animals and whether sketch drawings on clouds are interesting before any drawing is shown. The answer distribution is shown in Fig. 10(a). It shows the ubiquitousness and significance of the proposed task that all participants admit they have had this imagination, more or less. About 27% of participants think it is a frequent imagination ("often" or "normally"). They also demonstrate a promising interest in the task. When asked for an immediate score of this task without seeing actual drawings, 2/3 of participants are interested (those who score 4 or 5).

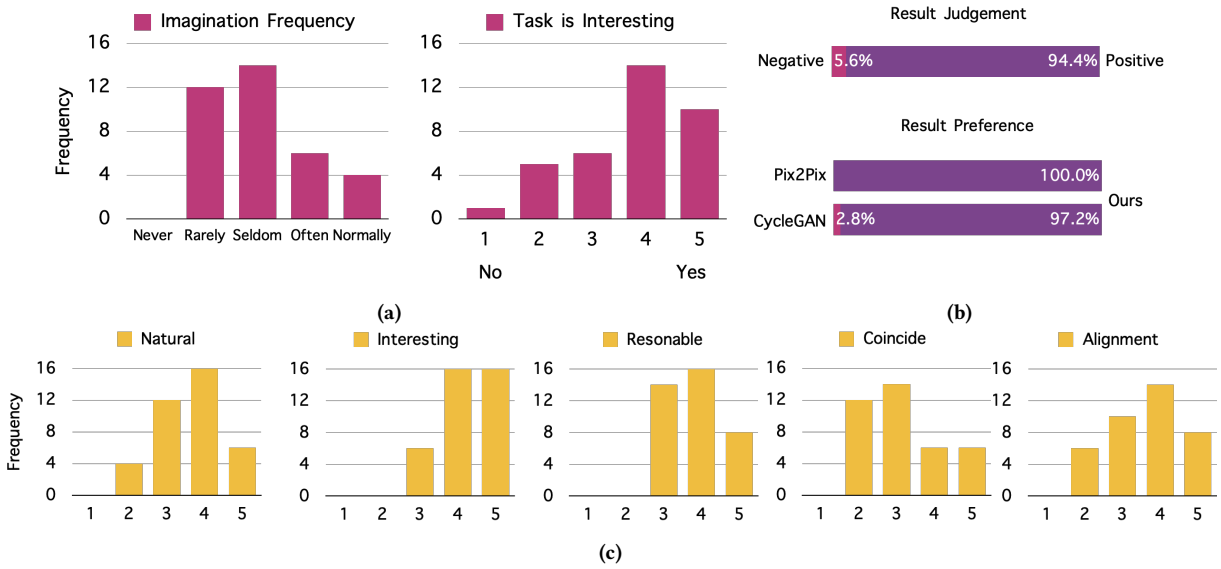

Figure 10: User study on the proposed task (a) and results of the proposed pipeline (b). Evaluation in different aspects is shown in (c). Participants score whether the algorithm results are *natural*, *interesting*, *reasonable* associations, *coincide* with human imagination, and precise *alignment*, respectively. Details and more results are provided in Sec. 5.

Then we display 10 cloud augmentation results generated by our proposed algorithm and ask participants to score them in 5 aspects: how **natural** the sketch drawing look; how **interesting** the sketch drawing look; how **reasonable** the association is; whether the imagination **coincide** with the participant's; how precise the sketches **align** with the cloud shape. We compute the average score from each participant in each dimension over all 10 shown results and illustrate the frequency in Fig. 10(c). The most concerted opinion is that the augmentations are "interesting", with the highest average score of 3.9. Meanwhile, participants hold contentious views to "coincide" scores. Many of them comment that some cloud shapes do not immediately invoke an explicit imagination so the chance of coinciding is not high. We think this phenomenon is interesting because, in the meantime, participants give higher scores in "interesting" and "reasonable", through which we perceive their agreement with the imagination produced by the algorithm. It verifies our anticipation that a computational algorithm not only reflects human imagination but may also inspire humans with different associations.

Due to the trade-off between the preciseness to coordinate with cloud shapes and smoothness to keep gist of sketches, approximately 47% of participants score the alignment as neutral or below. Moreover, cloud shapes that might be vague at boundaries are transformed into exact contours, which increases the uncertainty of alignment. In future work, the trade-off and uncertainty in alignment would be further investigated.

Following scores in detailed aspects, the questionnaire asks about the overall judgment of our results and preferences in comparison. As shown in Fig. 10(b), more than 94% of participants give a positive response to shown drawings. In comparison with the two baseline models, our proposed method overwhelmingly wins the votes. It is actually not surprising to us because of the challenges of the proposed task that can not be addressed by existing methods.

For the last questions, we ask participants to give comments on the task and show drawings. Regarding their feeling about the application, participants give positive feedback like "Very significant, demonstrating imagination", "relaxing and interesting", "an attractive technology", "a funny game I wish to join", "It attracts me, very interesting!", etc. On the suggestion of the task, it is interesting that participants comment diversely and even in contrast. For example, some suggest "avoid too complicated sketches" while another participant with a background in computer vision suggests "more details more interesting." We believe drawing with the same expressive power can be achieved by fewer strokes if the alignment can be more precise in future work.

6 CONCLUSION

In this paper, we present a novel application task, cloud augmentation with imaginary animal sketches, which brings imagination into reality. We design a self-supervised pipeline to address the unique challenges of this task. By decomposing the task into three stages, cloud contour detection, sketch retrieval, and sketch alignment, the proposed pipeline manages to augment clouds with vivid sketch drawings. Our extensive experiments validate its effectiveness and a concrete user study provides further guidance for this application.

ACKNOWLEDGMENTS

We thank Dr. Changqing Zou for his advice that helped to improve this work, especially on the technical direction during the early stage of this research. This research was funded in part by the New York State Center of Excellence in Data Science, an Empire State Development-designated Center of Excellence.

REFERENCES

[1] EC Barrett and Colin K Grant. 1976. *The identification of cloud types in LANDSAT MSS images.* Technical Report.

[2] Fred L. Bookstein. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence (TPAMI)* 11, 6 (1989), 567–585.

[3] Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. 2008. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences* 1124, 1 (2008), 1–38.

[4] Caroline Chan, Fredo Durand, and Phillip Isola. 2022. Learning to generate line drawings that convey geometry and semantics. *arXiv preprint arXiv:2203.12691* (2022).

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[6] Herbert James Clark. 1965. Recognition memory for random shapes as a function of complexity, association value, and delay. *Journal of Experimental Psychology* 69, 6 (1965), 590.

[7] Soumyabrata Dev, Florian M Savoy, Yee Hui Lee, and Stefan Winkler. 2017. Nighttime sky/cloud image segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 345–349.

[8] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.. In *International Conference on Learning Representations (ICLR)*.

[10] Aaron Gross and Anne Vallely. 2012. *Animals and the human imagination: a companion to animal studies.* Columbia University Press.

[11] David Ha and Douglas Eck. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477* (2017).

[12] Jessica B Hamrick. 2019. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences* 29 (2019), 8–16.

[13] Rana Hanocka, Noa Fish, Zhenhua Wang, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. 2018. Alignet: Partial-shape agnostic alignment via unsupervised learning. *ACM Transactions on Graphics (TOG)* 38, 1 (2018), 1–14.

[14] Matthias Harders and Gabor Szekely. 2003. Enhancing human-computer interaction in medical segmentation. *Proc. IEEE* 91, 9 (2003), 1430–1442.

[15] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. Neuroscience-inspired artificial intelligence. *Neuron* 95, 2 (2017), 245–258.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377* (2021).

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[18] Xiaodi Hou, Alan Yuille, and Christof Koch. 2013. Boundary detection benchmarking: Beyond f-measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2123–2130.

[19] Ming-Kuei Hu. 1962. Visual pattern recognition by moment invariants. *IRE transactions on information theory* 8, 2 (1962), 179–187.

[20] World international organization. 1987. *International Cloud Atlas Vol 2.* World Meteorological Organization.

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

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).

[23] Mehmet Kesim and Yasin Ozarslan. 2012. Augmented reality in education: current technologies and the potential for education. *Procedia-social and behavioral sciences* 47 (2012), 297–302.

[24] Michael R LaChat. 1986. Artificial intelligence and ethics: an exercise in the moral imagination. *Ai Magazine* 7, 2 (1986), 70–70.

[25] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. 2019. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1403–1412.

[26] Qingyong Li, Weitao Lu, and Jun Yang. 2011. A hybrid thresholding algorithm for cloud detection on ground-based color images. *Journal of atmospheric and oceanic technology* 28, 10 (2011), 1286–1296.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*. Springer, 740–755.

[29] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. 2019. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5830–5839.

[30] Charles N Long, Jeff M Sabburg, Josep Calbó, and David Pagès. 2006. Retrieving cloud characteristics from ground-based daytime color all-sky images. *Journal of Atmospheric and Oceanic Technology* 23, 5 (2006), 633–652.

[31] Sridhar Mahadevan. 2018. Imagination machines: A new challenge for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 32.

[32] Zahid Mahmood, Tauseef Ali, Nazeer Muhammad, Nargis Bibi, Imran Shahzad, and Shoaib Azmat. 2017. EAR: Enhanced augmented reality system for sports entertainment applications. *KSII Transactions on Internet and Information Systems (TIIS)* 11, 12 (2017), 6069–6091.

[33] D. Martin, C. Fowlkes, D. Tal, and J. Malik. 2001. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proc. 8th Int'l Conf. Computer Vision*, Vol. 2. 416–423.

[34] Florian A Potra, Xing Liu, Francoise Seillier-Moiseiwitsch, Anindya Roy, Yaming Hang, Mark R Marten, Babu Raman, and Carol Whisnant. 2006. Protein image alignment via piecewise affine transformations. *Journal of Computational Biology* 13, 3 (2006), 614–630.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.

[36] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.

[37] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Qian, Chen Change Loy, and Ran He. 2021. Everything's Talkin': Pareidolia Face Reenactment. *arXiv preprint arXiv:2104.03061* (2021).

[38] Qianqian Song, Zhihui Cui, and Pu Liu. 2020. An Efficient Solution for Semantic Segmentation of Three Ground-based Cloud Datasets. *Earth and Space Science* 7, 4 (2020), e2019EA001040.

[39] X. Soria, E. Riba, and A. Sappa. 2020. Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, Los Alamitos, CA, USA, 1912–1921.

[40] Lidan Wang, Vishwanath Sindagi, and Vishal Patel. 2018. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG)*. IEEE, 83–90.

[41] Xiaogang Wang and Xiaoou Tang. 2008. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* 31, 11 (2008), 1955–1967.

[42] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. 2018. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 7278–7286.

[43] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. 2018. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 8090–8098.

[44] Meijuan Ye, Shizhe Zhou, and Hongbo Fu. 2019. DeepShapeSketch: Generating hand drawing sketches from 3D objects. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[45] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. 2016. Sketchnet: Sketch classification with web images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1105–1113.

[46] Hua Zhang, Peng She, Yong Liu, Jianhou Gan, Xiaochun Cao, and Hassan Foroosh. 2019. Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval. *IEEE Transactions on Image Processing* 28, 9 (2019), 4486–4499.

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2223–2232.

A CLOUD GENUS

Clouds as the source of imagination in this work exist in diverse forms in different altitude levels or conditions of the atmospheres. There are many ways to identify and classify clouds in different orders [20]. From the ground-based perspective, we consider the physical forms of clouds to determine proper cloud types. In Fig. 11, we show cloud genus classified according to their physical forms [1]. To the end of clear cloud shapes and edges, we emphatically choose clouds in Cumuliform and Cumulonimbiform (Fig. 11d and Fig. 11e, respectively) to form our test data.

B SKETCH COMPLETION

In our paper, we verified that current image-to-image translation methods can hardly address the challenges posed by our proposed task. While in the early stage of our research, we investigated the potential capability of the state-of-the-art methods to complete images without knowing which parts are missing. In Fig. 14, Pix2Pix [21] and CycleGAN [47] demonstrate different completion results. Benefiting from the large receptive field of stacked convolutional layers and direct pixel supervision, Pix2Pix manages to reconstruct the synthetic color images. We expected that CycleGAN is limited in this task since the masked image can theoretically not be recovered from the ground truth, which is the foundation of CycleGAN. The observation coincides with Liu et al. [29] who propose to complete sketches with a multi-stage Pix2Pix model, named SketchGAN. However, their experiments are limited to a 10% mask that is insufficient for most applications. When extended to a larger area, the algorithm also demonstrates limited completion capacity, as shown in Fig. 15. Even with stacked prediction stages, SketchGAN can only complete small parts. It's still an open and challenging problem to complete a sketch without a reference (our FFD Generator achieves it with a reference from sketch retrieval). Recently, He et al. [16] propose masked autoencoders for self-supervised training where the agent task is to complete masked images and demonstrate promising results. It brings new chances to the problem of sketch completion, while sketch images are more sparse and thus more challenging.

C HU MOMENTS FOR SKETCH RETRIEVAL

In the quantitative comparison in our paper, we introduce Hu Moments Hu [19] as a measure of shape similarity. It is a natural idea to directly use Hu Moments as shape features for sketch retrieval in cloud2sketch, as Hu Moments are invariant to rotation and scale. In Fig. 16, we compare the retrieval results by Hu Moments and our proposed algorithm. Although Hu Moments describe features of shapes in images, the cloud contours maintain insufficient information and the retrieval from contour shapes requires more high-level geometrical correspondence that neural networks can capture.

D EXTENDED APPLICATION

The proposed cloud2sktech pipeline can be easily transferred to other augmentation sources with minor changes. In Fig. 17, we show augmenting islands on the see using our proposed algorithm. Although it is designed for cloud augmentation, it can be applied to other sources with minor changes: only the input channel of the thresholding algorithm is altered and no model is re-trained.

(a) Stratiform (b) Cirriform (c) Stratocumuliform (d) Cumuliform (e) Cumulonimbiform

Figure 11: Cloud genus according in different physical forms. We consider cumuliform and Cumulonimbiform as proper source for augmentation.

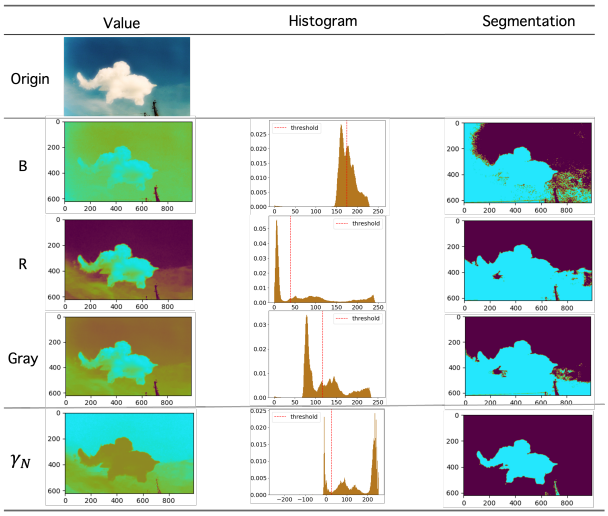

Figure 13: Normalized blue/red ratio for cloud segmentation, where $\gamma_N = (B-R)/(B+R)$. "B", "R", and "Gray" are blue, red, and grayscale channels, respectively.

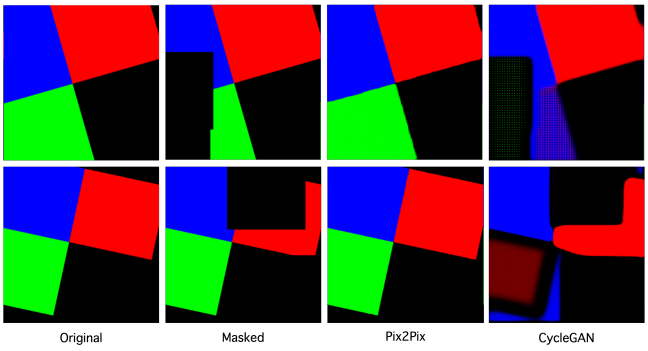

Figure 14: Comparison of image translation methods on a toy task to complete colored regions.

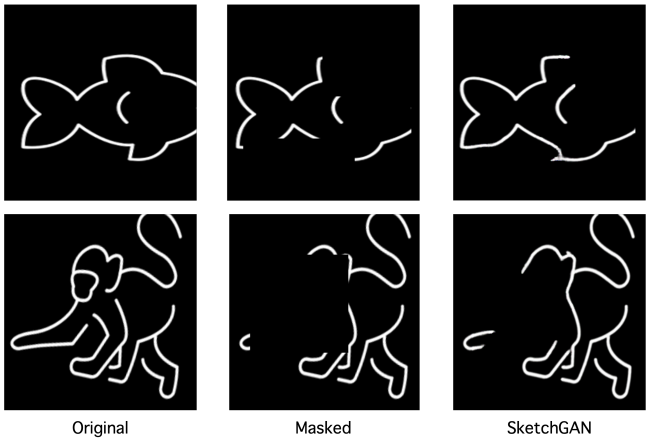

Figure 15: Sketch completion results by SketchGAN [29]. When applied to sketches with larger mask areas (30%), its completion capacity is limited.

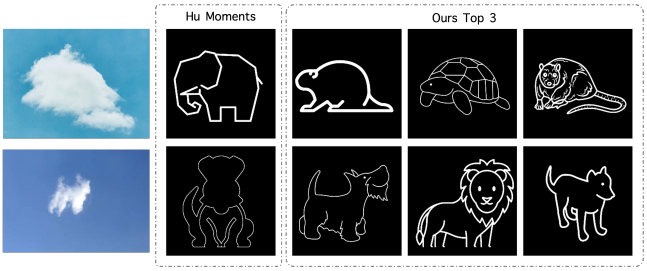

Figure 16: Comparison of sketch retrieval by Hu Moments [19] and ours.

Figure 17: Augmenting islands on the sea using our cloud2sketch algorithm. Only the input channel of cloud segmentation algorithm is changed and no model is re-trained.