ACM/EG Expressive Symposium
Eurographics Workshop on Intelligent Cinematography and Editing (2025), pp. 1–4
D. Berio and A. Bruckert (Editors)

Towards Automated 2D Character Animation

H. Mailee[1] and R. K. Dos Anjos [2]

[1]The University of Manchester, United Kingdom
[2]University of Leeds, United Kingdom

Figure 1: ***Eyes and Mouths Detected by Our Three Fine-Tuned Models.*** *From left to right, the images show inferences from our three trained models: YOLOX-l, YOLOX-Tiny, and Faster R-CNN. YOLOX-Tiny outperforms the other two, while YOLOX-l produces false positives, and Faster R-CNN fails to detect correctly.*

Abstract

Automating facial expression changes in comics and 2D animation presents several challenges, as facial structures can vary widely, and audiences are susceptible to the subtlest changes. Building on extensive research in human face image manipulation, landmark-guided image editing offers a promising solution, providing precise control and yielding satisfactory results. This study addresses the challenges hindering the advancement of landmark-based methods for cartoon characters and proposes the use of object detection models —specifically YOLOX and Faster R-CNN— to detect initial facial regions. These detections serve as a foundation for expanding landmark annotations, enabling more effective expression manipulation to animate expressive characters. The codes and trained models are publicly available here.

CCS Concepts

• ***Computing methodologies*** → ***Interest point and salient region detections;*** *Object detection;*

1. Introduction

In the comic and 2D animation industry, production is heavily dependent on artists delivering the respective keyframes. These keyframes are consecutive drawings, either of the same scene with minor differences or characters in new settings and emotions. We will refer to the keyframes for comics as 'strips' in our text for clarity. Focusing on characters' faces, the changes in expression are the most significant alterations, particularly in scenes depicting characters' interactions. While the changes are minimal in these scenes, artists are required to draw the frames nonetheless, a stage that has the potential for automation to alleviate this burden.

Quality for the audience is of utmost importance in the final automation process, as they are sensitive to the subtle nuances of facial expressions for emotion recognition. Concurrently, the artists require partial control on the process to determine the proper target keyframe with the exaggeration exhibited in cartoons. All these

Figure 2: *Results of YOLOX-Tiny detections on iCartoonFace & Monsters Inc. samples.*

necessitate employing a channel that provides artists with enough supervision to accurately edit character faces, leading us to the field of guided image manipulation.

Recent advances in guided image editing have been substantial, especially for human faces. These edits include but are not limited to age, hairstyle, skin tone and makeup. The dimensionality of these edits can also include 2D and 3D, with 3D models focusing on mesh reconstruction of the human image and applying the edits. The guidance is also diverse, and models can edit human images with textual prompts, another human face, an art style, or landmarks.

While these significant findings exist for human faces, and some apply to stylized characters, exclusive studies on the cartoon domain are still sparse. There are many limitations when it comes to this particular domain, most of which stem from the diversity of the cartoon characters. These faces differ not only in art styles but also in their facial features. Noting eyes, for instance, one can see that the number of eyes and their shape and location diverges from the regular face and from one character to the other.

This diversity prevents the use of human-based models since those depend on facial similarity. In 3D editing, 3D Morphable Models (3DMMs) [BV23] are widely used as the base mesh for human faces, and the changes are then applied to these meshes, rendering the mesh back to an image. In addition to a lack of a standard shared mesh between all cartoon characters, rendering them back to the image requires non-photorealistic rendering, which brings forth a substantial research field.

Focusing on the 2D editing models, we narrow down the practical approaches for the cartoon domain based on the guides used for editing. Finding the perfect text prompt that can alter a cartoon's face can be more time-consuming than drawing the keyframe. Coming up with another face or expression that matches the artist's vision is not an optimal solution, specifically for exaggerated expressions. While drawing a sketch is plausible, it still requires a longer time than a method that can pinpoint the target areas for modification. Mask-based methods are more efficient than sketch-based methods, but they lack a fine-grained control for details. These limitations direct us to use 2D landmarks, enabling less erroneous editing by focusing on minimal points and small local areas.

With limited landmark datasets for characters and few studies on their detection in unconventional faces, we take the first step in identifying the eye and mouth areas for use in future landmark detection. By training two leading object detection models, YOLO and Faster R-CNN, we determine the most accurate model for detecting the targeted areas after fine-tuning. The trained models can serve as an initial step for landmark detection, facilitating landmark-guided image editing for stylized faces.

2. Related Work

2.1. Landmark-Guided Image Editing

Most landmark-guided image editing models use a landmark image as a constraint to generate a target face image with the desired expression while preserving the identity of the source image. The process involves extracting landmarks from the driving image and using a conditional generator to create the target image with the specified expression.

In the early works, such as those by Hu et al. [HWY*18] and Wu et al. [WZL*18], landmarks are integrated with GANs to generate face rotations and facilitate re-enactments, respectively. This approach has been widely adopted in subsequent research, with a comprehensive summary provided by Nickabadi et al. [NFFM22].

Most recently, Ma et al. [MLW*24] introduced expression-aware landmarks, enhancing realism in animated portraits. Using MediaPipe [LTN*19] as the backbone for keypoint extraction, their method is limited to faces that can be detected and analysed by the API. However, it demonstrates promising results for human-like stylized faces.

(a) *YOLOX-l* **(b)** *YOLOX-Tiny* **(c)** *Faster R-CNN*

Figure 3: *Superiority of Faster R-CNN for instances closer to the training domain.*

2.2. Stylized Face Datasets

We study the available datasets to understand the variety of stylised face literature. Most large-scale datasets, like iCartoonFace [ZZR*20] (∼400k images) and Danbooru [Ano21] (∼970k images), focus solely on face detection and recognition. A dataset with richer annotation is CCDaS [QPN*23], which includes both body boxes and faces for ∼140k images. Though these datasets include a wide range of varieties in terms of metadata, resources (such as cartoons, comics, merchandise), and characters, none include annotations solely for face landmarks, which is the focus of this work.

On the contrary, feature and landmark detection has received significant attention in the relatively close field of Caricature and Art Faces. For caricature, we have WebCaricature [HLS*17] and CaricatureFace [CGPZ21] and art faces are adequately covered by ArtFacePoints [SMC22] and Artistic-Faces [YNS19]. Even so, these datasets cannot be used directly due to their differences with the domains of cartoons and manga.

One important seminal public dataset in this field is Manga109 [MIA*17], which consists of 109 manga volumes totalling 21,142 pages. This work has been expanded by further annotations over the years, notably a landmark annotation by Stricker et al. [SAKI18] for 2,105 faces of the dataset. Lastly, StylizedFacePoints [CMP24] is an admirable addition to the collection of landmark annotations containing 4,086 faces with 98 landmarks per face. Regrettably, the dataset was not accessible when the experiments were conducted.

3. Experiments

3.1. Data Pre-Processing & Models

Using annotated landmarks on the Manga109 [SAKI18] dataset, we conduct experiments to detect eyes and mouths on faces. As the first step, the landmarks must be transformed into bounding boxes, following the default setting of object detection models. Having all the landmarks around the eye and mouth, we find the tightest bounding box by determining these landmarks' maximum and minimum coordination. We extend this box by an extra 15% to increase coverage and prevent losing data.

For our experiments, we use two renowned and widely adopted object detection models that have previously been applied to our target dataset, Manga109: YOLOX [GLW*21] and Faster R-CNN [Ren15]. YOLOX has been used for face and body detection on the same dataset in works such as [Shi21] and [TYS22], while Faster R-CNN has been reported to detect eye regions as a pre-processing stage for redrawing [CBCW24]. Using these two as our benchmarks, we fine-tune three models for eye and mouth detection: YOLOX-l (large), YOLOX-Tiny, and Faster R-CNN with a ResNet50 backbone. As these models are all initially trained on the COCO dataset—which does not include the 'eye' or 'mouth' categories explicitly—the comparison with non-refined backbones is impractical, since the models would classify regions into the existing COCO classes.

All experiments were conducted on two Tesla V100-SXM2-16GB GPUs, using their default implementations and instructions [Tor16, Meg25], with learning rates and preferred epochs set as suggested by the authors. To preview the data labels and export samples in different detection formats, we utilised the FiftyOne API [MC20], and kept the training, testing, and validation sets consistent across all experiments. The hyper-parameters, training logs, and fine-tuned models are available on the GitHub page.

	mAP	AP^{50}_{eye}	AP^{50}_{mouth}	AP^{75}_{eye}	AP^{75}_{mouth}	$AP^{50:95}_{eye}$	$AP^{50:95}_{mouth}$
YOLOX-l	0.48	0.97	0.51	0.76	0.30	0.66	0.30
YOLOX-Tiny	0.57	0.95	0.84	0.71	0.54	0.62	0.52
Faster R-CNN	0.53	0.94	0.87	0.64	0.53	0.57	0.50

Table 1: *AP comparison of all bounding boxes detected by fine-tuned models.*

3.2. Results

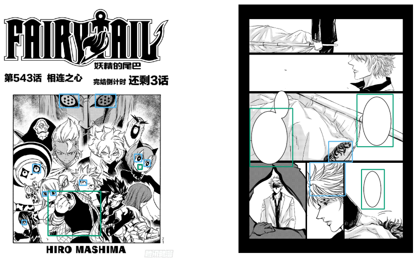
We summarise the results of our experiments after fine-tuning the three models in Table 1 and Table 2. Table 1 compares performance across models using variations of the Average Precision (AP) metric, which is widely used in object detection tasks. For improved visualisation, we colour the cells relative to the values in their respective columns, with darker shades indicating better performance. As shown in the table, YOLOX-Tiny consistently outperforms the other two models. Although YOLOX-l and Faster R-CNN are designed to perform better with larger datasets—given their number of parameters and model complexity—YOLOX-Tiny shares a similar structure with YOLOX-l but is optimised with fewer parameters for smaller datasets, resulting in its superior performance (Figure 2).

Being strong models by themselves, YOLOX-l and Faster R-CNN show their strengths in particular cases. For instances closely aligned with the training dataset's domain, Faster R-CNN achieves higher confidence in the predictions, demonstrating its ability to learn the training domain while highlighting its limitations in generalizing to unseen instances (Figure 3).

Table 2 compares the number of false positives (FP), false negatives (FN), and the FP/FN ratio. YOLOX-l exhibits lower average precision in the tests due to a higher number of false positives. Although increasing the number of detections can improve the chances of correct identifications and assist with the initial dataset annotation, the inconsistency of the results suggests that the model requires human supervision (Figure 4). YOLOX-Tiny achieves the best FP/FN ratio, with the lowest number of FPs and a fairly acceptable number of FNs.

	Eye			***Mouth***		
	FP	*FN*	*FP/FN*	*FP*	*FN*	*FP/FN*
YOLOX-l	1372	10	137.20	822	22	37.36
YOLOX-Tiny	**138**	19	**7.26**	**133**	29	**4.59**
Faster R-CNN	189	17	11.12	214	18	11.89

Table 2: *FP, FN and FP/FN comparison of bounding boxes detected by fine-tuned models per category.*



(a) *Success case* **(b)** *Failure case*

Figure 4: *YOLOX-l success and failure cases with the increased number of bounding boxes (the blue and green boxes represent eye and mouth detections respectively).*

4. Limitations

The most significant limitation is the scarcity of annotated datasets and the similarity exhibited in most of them. Although the trained model performs well in detecting eyes and achieves acceptable results for mouth detection, these results are all based on a limited dataset of only 2000 images of black-and-white manga faces. Furthermore, upon examination, we found that none of the faces contained occlusions, which are crucial factors in object detection.

5. Conclusion & Future Work

Despite the limitations, the satisfactory results motivate us to pursue this problem further. For future work, we plan to explore other approaches used in object detection, such as domain adaptation. These methods would allow us to leverage existing datasets without the need for additional annotation. If annotated data is required, the trained models presented in this paper can significantly aid in dataset collection by providing initial detection. All these efforts contribute to the broader goal of enabling guided face-image editing for stylized characters, ultimately equipping artists with a valuable tool to assist their work.

References

[Ano21] ANONYMOUS AND DANBOORU COMMUNITY AND GWERN BRANWEN: Danbooru2020: A large-scale crowdsourced & tagged anime illustration dataset. https://gwern.net/Danbooru2020, January 2021. Accessed: 01-02-2025. URL: https://gwern.net/Danbooru2020. 3

[BV23] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 157–164. 2

[CBCW24] CARDOSO J. L., BANTERLE F., CIGNONI P., WIMMER M.: Re: Draw–context aware translation as a controllable method for artistic production. *arXiv preprint arXiv:2401.03499* (2024). 3

[CGPZ21] CAI H., GUO Y., PENG Z., ZHANG J.: Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model. *Graphical Models 115* (2021), 101103. 3

[CMP24] CHENG S., MA C., PAN Y.: Stylizedfacepoint: Facial landmark detection for stylized characters. In *Proceedings of the 32nd ACM International Conference on Multimedia* (2024), pp. 8072–8080. 3

[GLW*21] GE Z., LIU S., WANG F., LI Z., SUN J.: Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* (2021). 3

[HLS*17] HUO J., LI W., SHI Y., GAO Y., YIN H.: Webcaricature: a benchmark for caricature recognition. *arXiv preprint arXiv:1703.03230* (2017). 3

[HWY*18] HU Y., WU X., YU B., HE R., SUN Z.: Pose-guided photorealistic face rotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8398–8406. 2

[LTN*19] LUGARESI C., TANG J., NASH H., MCCLANAHAN C., ...: Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv ...* (2019). Publisher: arxiv.org. URL: https://arxiv.org/abs/1906.08172. 2

[MC20] MOORE B. E., CORSO J. J.: Fiftyone. *GitHub. Note: https://github.com/voxel51/fiftyone* (2020). 3

[Meg25] Megvii-basedetection/yolox: Yolox is a high-performance anchor-free yolo, exceeding yolov3 v5 with megengine, onnx, tensorrt, ncnn, and openvino supported. documentation: https://yolox.readthedocs.io/. https://github.com/Megvii-BaseDetection/YOLOX, 2025. Accessed: 2025-02-03. 3

[MIA*17] MATSUI Y., ITO K., ARAMAKI Y., FUJIMOTO A., OGAWA T., YAMASAKI T., AIZAWA K.: Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications 76*, 20 (2017), 21811–21838. doi:10.1007/s11042-016-4020-z. 3

[MLW*24] MA Y., LIU H., WANG H., PAN H., HE Y., YUAN J., ZENG A., CAI C., SHUM H.-Y., LIU W., CHEN Q.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers* (New York, NY, USA, 2024), SA '24, Association for Computing Machinery. URL: https://doi.org/10.1145/3680528.3687587, doi:10.1145/3680528.3687587. 2

[NFFM22] NICKABADI A., FARD M. S., FARID N. M., MOHAMMADBAGHERI N.: A comprehensive survey on semantic facial attribute editing using generative adversarial networks. *arXiv preprint arXiv:2205.10587* (2022). 2

[QPN*23] QI Z., PAN D., NIU T., YING Z., SHI P.: CCDaS: a benchmark dataset for cartoon character detection in application scenarios. In *International Forum on Digital TV and Wireless Multimedia Communications* (2023), Springer, pp. 369–381. 3

[Ren15] REN S.: Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015). 3

[SAKI18] STRICKER M., AUGEREAU O., KISE K., IWATA M.: Facial landmark detection for manga images. *arXiv preprint arXiv:1811.03214* (2018). 3

[Shi21] SHINYA Y.: Usb: Universal-scale object detection benchmark. *arXiv preprint arXiv:2103.14027* (2021). 3

[SMC22] SINDEL A., MAIER A., CHRISTLEIN V.: Artfacepoints: High-resolution facial landmark detection in paintings and prints. In *European Conference on Computer Vision* (2022), Springer, pp. 298–313. 3

[Tor16] TORCHVISION MAINTAINERS AND CONTRIBUTORS: Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 3

[TYS22] TOPAL B. B., YURET D., SEZGIN T. M.: Domain-adaptive self-supervised pre-training for face & body detection in drawings. *arXiv preprint arXiv:2211.10641* (2022). 3

[WZL*18] WU W., ZHANG Y., LI C., QIAN C., LOY C. C.: Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 603–619. 2

[YNS19] YANIV J., NEWMAN Y., SHAMIR A.: The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics (TOG) 38*, 4 (2019), 1–15. 3

[ZZR*20] ZHENG Y., ZHAO Y., REN M., YAN H., LU X., LIU J., LI J.: Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM international conference on multimedia* (2020), pp. 2264–2272. 3