Landmarks Detection for Manga Face

WANG YIWEN 5120FG52

Fig. 1. Blue dots painted are landmarks detected on images for validation.

Facial Landmark detection for human faces has been researched for long. Impressive progress has been made with neural-network based methods and large-scale datasets. However, it is still a challenging and largely unexplored problem in Manga domain. Compared to natural face images, Manga are much more diverse. They contain a much wider style variation in geometry. Moreover, datasets that are necessary to train neural networks are unavailable. We propose a method for Manga augmentation. In addition, we use a classification module to help understanding global feature.

Additional Key Words and Phrases: facial landmark detection , neural networks, Manga image augmentation, geometry aware style transfer

ACM Reference Format:
Wang Yiwen 5120FG52. 2021. Landmarks Detection for Manga Face. *ACM Trans. Graph.* 1, 1 (July 2021), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Manga are comics or graphic novels originating from Japan. The term manga is used in Japan to refer to both comics and cartooning. People of all ages read manga, and the medium includes works in a broad range of genres. Since the 1950s, manga has become an increasingly major part of the Japanese publishing industry. By 1995, the manga market in Japan was valued at ¥586.4 billion, with annual sales of 1.9 billion manga books and manga magazines in Japan [Paul 2004]. With such a big volume and market, however, there are almost no digital methods to help researching on the style and trend of Mangas.

Despite the variation in brackground, dressing, and characters' postures, the faces in Manga have patterns which allow us to apply Manga analysis, synthesis and stylization on. However, most algorithms that work well for natural faces fail when applied to Manga. The differences between Manga and face photos span two domains: texture appearance differences, and geometric differences.

In this article, we concentrate on detecting the landmarks for Manga faces with deep learning. Facial landmark detection aims to localize a set of predefined landmarks such as eye corners or mouth corners in a face depiction. It is the basis of any Manga analysis, and a fundamental problem in computer vision. Detecting facial features in Manga allows us to model the geometric style of an artist and use it, for instance, in geometry-aware style transfer. Using neural networks, impressive progress has been made on facial landmark detection in recent years on natural face images. However, moving to Manga this becomes much more of a challenge. Compared to natural face images, Manga are much more diverse. They contain a much wider variation in geometry and are more complex to analyze. Facial features in Manga are often exaggerated in ways that lead to the deviation from the implicit humanly attributes. Moreover, large scale datasets that are necessary to train neural networks are unavailable.

We propose to use image augmentation of Manga. We utilize manually labeled Manga landmarks datasets, and transform their content. This transformation is performed using various geometric variations and Delaunay-Triangulation based face swaping. Our augmentation method enables training deep neural networks despite the deficiency of annotated data of Manga. The network we train in this work is based on the ECpTp (Estimation-Correction-Tuning) framework for facial landmark detection [Jordan Yaniv 2019], which combines the advantages of the global robustness of a data-driven method, the outlier correction capability of a model-driven method, and non-parametric optimization of landmark mean-shift. We enhance the frameworks performance on Manga by using multitask for image classification. This addition enlarges the model's receptive field, which is necessary due to the importance of global congnation for local detection. Finally, to evaluate our landmark detection, Normalized Mean Error is computed to compare with other methods.

In conclusions, our main contributions are:

- Face-swap based on Delaunay-Triangulation to generate more labeled Manga faces.
- Thin-Plate-Spline for data augmentation.
- Add global feature module into the fully convolutional network.
- Model performance better in Normalized Mean Error than existed methods.

Author's address: Wang Yiwen 5120FG52.

0730-0301/2021/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

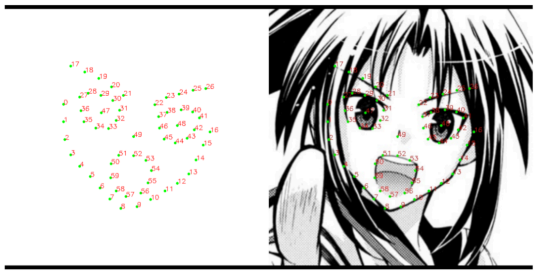

Fig. 2. Facial landmarks in model from [Stricker et al. 2018]

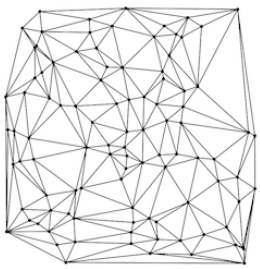

Fig. 3. A Delaunay triangulation of a random set of 100 points in a plane.

2 RELATED WORK

Availability of manga images which can be used for research and shared publicly are scarce because of copyright issues. However Matsui et al. [Matsui et al. 2015] released "Manga109", a dataset of 109 manga volumes for research purposes.

[Chu and Li 2019] used this dataset to train a CNN for detecting the face bounding boxes on manga images. To achieve this, they manually labeled the faces in 66 volumes. They have used 50 pages per title and published a subset of this dataset, which can be freely downloaded. This subset contains face bounding boxes from 24 different volumes. We have used these face bounding box information in order to extract the faces from the manga pages.

The similar pipeline of Manga faces' landmarks detection has been completed before by [Stricker et al. 2018]. They provide a new landmark annotation model for manga faces, and a deep learning approach to detect these landmarks using the "Deep Alignment Network", a multi-stage architecture where the first stage makes an initial estimation which gets refined in further stages.

Our work follows Estimation-Correction-Tuning (ECT) framework presented by [Zhang et al. 2016]. This framework combines the advantages of the global robustness of data-driven method (deep CNN), outlier correction capability of model-driven method (PDM) and non-parametric optimization of regularized landmark Mean-Shift (RLMS). We adapt the ECT framework for the Manga faces domain by using image augmentation in the training procedure.

To reduce the dependency between the different facial features [Jordan Yaniv 2019] uses a featurebased correction step, namely CpTp, and incorporating a Spatial Transformer Network (STN) component into the network to increase the accuracy of a wide range of tasks (e.g. classification), as the network learns invariance to geometric warping. Besides, this work also applys Thin-plate-spline on real photo faces for geometric augmentation. However, it fails to deform real photos when the shift and scale of landmarks are too large.

3 DATASET

Landmarks selection in Manga is different with canonical human faces. For example, noses in Manga faces are mostly simplified compared to human features. The nostrils are not drawn and the nose dorsum line is sometimes only implied. Besides, many characters have round shaped eyes which is much larger than human's proportion. Therefore, [Stricker et al. 2018] decided to simplify their landmarks for the nose compared to the iBUG model and assign each eye more landmarks. They followed the landmarks positioning proposed by the iBUG model for the chin contour, brows and mouth. To sum up, their landmark model consists of the following 60 landmarks with fixed index order as Fig. 2 shows:

- 5 landmarks for each eyebrow.
- 10 landmarks for each eye.
- 1 landmark for each pupil.
- 10 landmarks for the mouth.
- 17 landmarks for the chin contour.

Each image has been labeled by at least one participant. Most images were labeled twice. If the distance between two landmarks was greater than 2 pixels, then these landmarks were manually corrected and compared again. Finally 608 usable pairs of images and landmarks are achieved.

4 DATA AUGMENTATION

Because landmarks detection is a data driven problem, the amount and variability of dataset could be crucial. Despite routines for data augmentation like scaling, rotation, we did two more innovation for Manga faces, Delauney triangulation based face swapping and Thin-plate-spline. Notice that when doing image deformation, we should keep knowing where the landmarks go.

4.1 Delauney triangulation based face swapping

In mathematics and computational geometry, a Delaunay triangulation (also known as a Delone triangulation) for a given set P of discrete points in a general position is a triangulation $DT(P)$ such that no point in P is inside the circumcircle of any triangle in $DT(P)$. As Fig. 3. shows.

Simply speaking, Delaunay-Triangulation is a method to separate a plane given landmarks. The intuition of swapping two faces based on Delaunay-Triangulation, is that the landmarks' indexes order for face model are fixed, so the order of triangles. And we suppose the corresponding triangles of two different faces are similar. The combinition of different faces will hopefully enlarge the dataset. Besides, the process of swapping keep the positions of landmarks being known.

Two faces are thus separated into triangles according to their landmarks. Then the corresponding triangles are swapped. Finally the colors are unified by Seamless Cloning. Seamless clone is based on Poisson Blending to make the color look fitted. Part. The part

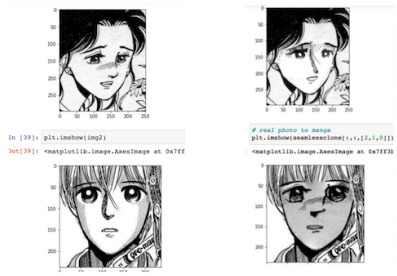

Fig. 4. A good example:The left column is original Manga face pair. The right swapped.

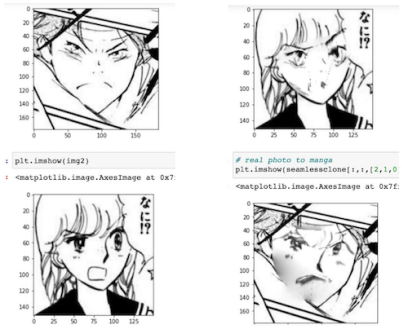

Fig. 5. A bad example:The left column is original Manga face pair. The right swapped.

Face Swapping in appendix illustrates the practical process [Canu [n.d.]].

The whole process is non-deep learning, which could be conducted by open-cv. Applying the method on Manga faces and we can see some pictures generated look good as Fig. 4. Some bad as Fig. 5. So we cherry-picked the good ones to enlarge the dataset into 1000 pictures.

4.2 Thin-plate-spline

To use Thin-plate-spline (TPS) to deform faces, we randomly shift each landmarks to a reasonable new position and accordingly spline the picture. The name Thin Plate Spline refers to a physical analogy involving the bending of a thin sheet of metal. Just as the metal has rigidity, the TPS fit resists bending also, implying a penalty involving the smoothness of the fitted surface.

The TPS arises from consideration of the integral of the square of the second derivative—this forms its smoothness measure. In the case where x is two dimensional, for interpolation, the TPS fits a mapping function $f(x)$ between corresponding point-sets y_i and x_i that minimizes the following energy function:

$$E_{tps}(f) = \sum_{i=1}^{K} ||y_i - f(x_i)||^2 \tag{1}$$

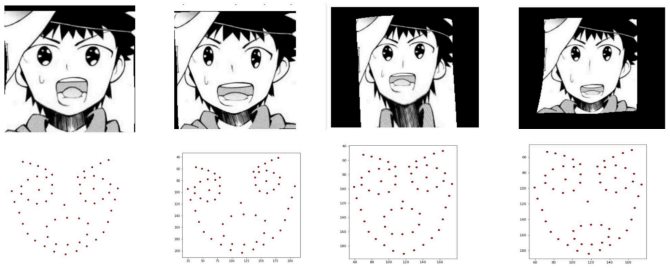

Fig. 6. Manga face after TPS

The smoothing variant, correspondingly, uses a tuning parameter λ to control the rigidity of the deformation, balancing the aforementioned criterion with the measure of goodness of fit, thus minimizing:

$$E_{smooth}(f) = (1) + \lambda \iint [(\frac{\partial^2 f}{\partial x_1^2})^2 + 2(\frac{\partial^2 f}{\partial x_1 \partial x_2})^2 + (\frac{\partial^2 f}{\partial x_2^2})^2] dx_1 dx_2 \tag{2}$$

For this variational problem, it can be shown that there exists a unique minimizer f.

The process of TPS is also non-deep learning, which is conducted by menpo API. As the Manga boy's face shows in Fig. 6., not only the whole scale changes, but also the proportion of features.

[Jordan Yaniv 2019] do almost the same work on drawing portraits, but it fails to apply on Mangas later. Because when deforming real photos, Thin Plate Spline can not change features in big scales as photos to Mangas. That's why we give up augmenting real photos for Manga dataset.

5 NETWORK

Here we simpy adopt the network used in [Jordan Yaniv 2019] and [Zhang et al. 2016] as a baseline. Besides, we utilize the swapped faces to add a multitask of classification, to help the model better understand the global features [Iizuka et al. 2016].

5.1 Baseline

For the pipeline named ECT [Zhang et al. 2016], E refers to the estimation network, as Fig. 7. shows. The design of the PrimaryNet and FusionNet is originated from [Pfister et al. 2015]. The PrimaryNet can not learn the spatial dependencies of landmarks very well. To address this, conv3 and conv7 are first concatenated and then fed into the fusion net. Finally a deconvolution network is used to upsample the small heatmaps into the original size. Dilated convolution filters are used inside to enlarge the receptive field.

For input, the ideal heatmap of the i-th landmark for image I is defined as a single-channel image M^i with the same resolution as I, the value at position z is defined as $M_z^i = \mathcal{N}(z; x_i^*, \sigma^2)$, where x_i^* is the ground truth location of the i-th landmark, and σ is the variance of a blurring kernel creating the map.

The training dataset consists of pairs $D = \{(I, 1^*)\}$, where $1^* = [x_1 ..., x_n]^T \in \mathbb{R}^{n \times 2}$ is the ground truth positions of n landmarks

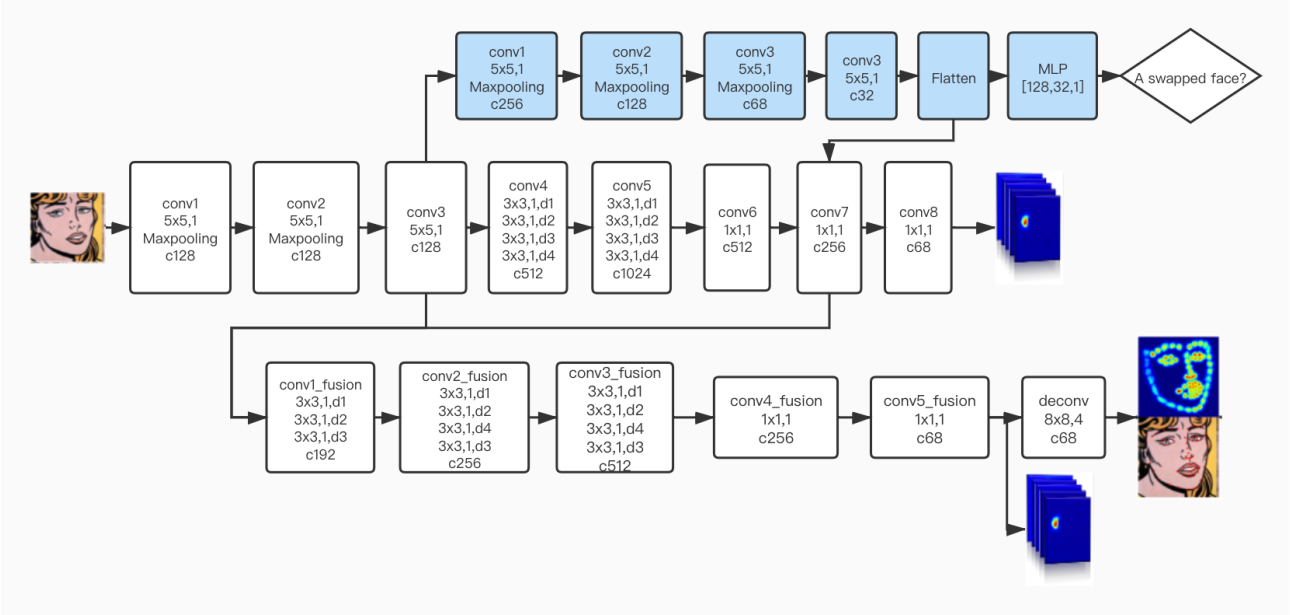

Fig. 7. Estimation network with the global feature module with blue background

embedded in image I. The objective of the regressor becomes estimating the network weights λ that minimizing the following L2 loss function:

$$\mathcal{L}(\lambda) = \sum_{(I,1^*)\in D} \sum_{i} ||M^i - \phi^i(I,\lambda)||^2 \tag{3}$$

where $\phi^i(I,\lambda)$ is the output of the i-th channel of the heatmap network, on the input image I.

C stands for correction and T stands for tuning. In the correction step, a more accurate initial shape is computed by correcting outlier landmarks using a pretrained point distribution model. Finally the landmark locations are finetuned based on weighted regularized mean shift.

For what the index p means, Manga faces have significantly variability in the facial features geometric location and proportion. Using one global constraint for all facial features will not allow much feature deviations from the canonical face shape. So [Jordan Yaniv 2019] created a feature based correction approach, i.e. A separate PDM model for each facial features.

In conclusion, E is data-driven, but C and T are model driven.

5.2 Global feature module

As the blue background blocks in Fig. 7. shows, our innovation here is to insert a classification subnetwork to do multitask of classifying whether a picture is original or face swapped. The flattened one channel are repeated and concatenated into the estimation network to help it better get global features. The index g for E represents the estimation network with the classification module.

And the overall loss becomes

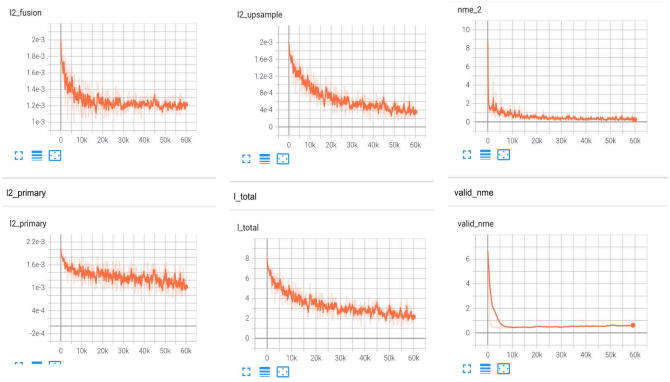

Fig. 8. Loss and evaluation metrics during training

$$\mathcal{L}(\lambda,\gamma) = \alpha \sum_{(I,1^*)\in D} \sum_{i} ||M^i - \phi^i(I,\lambda)||^2 + \beta \mathrm{CE}(f(I) - y^*) \tag{4}$$

where γ denotes the learnable parameters in the classification network. And α, β denotes the weights to balance the regressor and classifier. CE stands for cross entropy. $f(I)$ is classification output based on the imput image I. And y^* is the ground truth telling whether the picture is original or swapped.

6 EXPERIMENTS

We implement the network using Tensorflow framework. The model takes an input of 256 × 256 face image and outputs a set of 60 heatmaps with the same resolution. During training, we randomly flip the input image horizontally and crop a 248 × 248 arbitrary

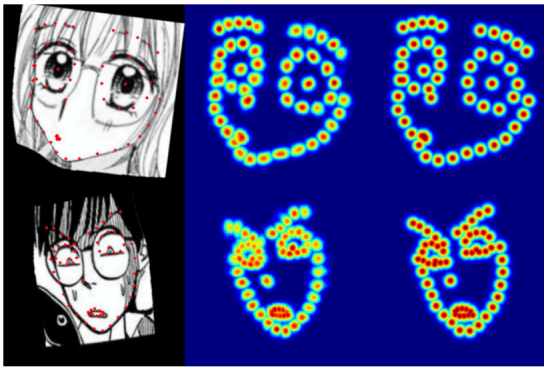

Fig. 9. Images in training with corresponding heatmaps

Table 1. NME of different models

NME	
DNA[Stricker et al. 2018]	0.039
ECpTp[Jordan Yaniv 2019]	0.033
EgCpTp	0.030

subimage from it. Then, we rotate it with a random angle from $-30°$ to $30°$ before rescaling it back to original size. For training the neural networks, we use batch size of 6, and weight decay of 10^{-8}, learning rate 10^{-4}. We use Adam optimization with $\beta 1 = 0.9, \beta 2 = 0.999, \epsilon = 10^{-8}$. The network is trained for 200 epochs. Network weights are initialized using Xavier weight initialization. Menpo Project is used in data augmnetation and landmark correction. For w-RLMS we use the code by [Zhang et al. 2016]. More implementation detailed can be found in [MyRepo]

The training process is shown in Fig. 8., in which $l_{fusion}, l_{primary}$, and $l_{upsample}$ represent the MSE of fusion net,primary net, and upsample net respectively. l_{total} represents the overall loss, including the classification cross entropy loss. NME is short for Normalized mean error.

$$NME = \frac{1}{n}\sum_{i=1}^{n}\frac{||x_i - x_i^*||^2}{d} \tag{5}$$

Where n is for the total number of landmarks, i for landmarks' indexes, x_i for predicted landmark's position, x_i^* for the ground truth landmark's position, d for inter pupil distance. The loss and NME for both train and validation drop down properly as Fig. [?]. shows. And Fig. 9. shows the model is predicting the heatmaps.

NME of different models for facial landmarks detection on Manga are computed for comparison, as listed in table 1.

7 CONCLUSIONS

This work provides a model to detect the landmarks for Manga faces, which applies innovative data augmentation methods to enlarge the limited dataset, and a classification module to help better understanding the global features. Besides, there are some points to improve on:

- Geometric style can be analyzed by calculating the distribution of different Mangas' face landmarks. And thus to analyze how artists' taste for human face changes over history. However, the dataset we possess now is too limited to get statistic features. The mean and variance are severely influenced by certain images. If time allowed, Manga images with tags such as time and authors can be crawled and analyzed with our model.
- More Evaluation metrics, such as MAPE, CED, AUC and ect. should be computed and compared with existed works.
- It could be interesting to include face landmarks into Manga generation as conditions, and generate characters with certain styles.

REFERENCES

Sergio Canu. [n.d.]. Face-swapping, open-cv with Python. [EB/OL]. https://pysource.com/2019/05/28/face-swapping-explained-in-8-steps-opencv-with-python/ Accessed June 25, 2021.

Wei-Ta Chu and Wei-Wei Li. 2019. Manga Face Detection based on Deep Neural Networks Fusing Global and Local Information. *Pattern Recognition* vol. 86 (2019). https://www.cs.ccu.edu.tw/~wtchu/projects/MangaFace/

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. Let There Be Color! Joint End-to-End Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Trans. Graph.* 35, 4, Article 110 (July 2016), 11 pages. https://doi.org/10.1145/2897824.2925974

Ariel Shamir. Jordan Yaniv, Yael Newman. 2019. The Face of Art: Landmark Detection and Geometric Style in Portraits. (2019).

Yusuke Matsui, Kota Ito, Yuji Aramaki, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2015. Sketch-based Manga Retrieval using Manga109 Dataset. *CoRR* abs/1510.04389 (2015). arXiv:1510.04389 http://arxiv.org/abs/1510.04389

Gravett Paul. 2004. Manga: Sixty Years of Japanese Comics. *New York: Harper Design* (2004).

Tomas Pfister, James Charles, and Andrew Zisserman. 2015. Flowing ConvNets for Human Pose Estimation in Videos. *CoRR* abs/1506.02897 (2015). arXiv:1506.02897 http://arxiv.org/abs/1506.02897

Marco Stricker, Olivier Augereau, Koichi Kise, and Motoi Iwata. 2018. Facial Landmark Detection for Manga Images. *CoRR* abs/1811.03214 (2018). arXiv:1811.03214 http://arxiv.org/abs/1811.03214

Hongwen Zhang, Qi Li, and Zhenan Sun. 2016. Combining Data-driven and Model-driven Methods for Robust Facial Landmark Detection. *CoRR* abs/1611.10152 (2016). arXiv:1611.10152 http://arxiv.org/abs/1611.10152

A PROCESS OF FACE SWAPPING

Fig. 10. Find landmark points of both images.[Canu [n.d.]]

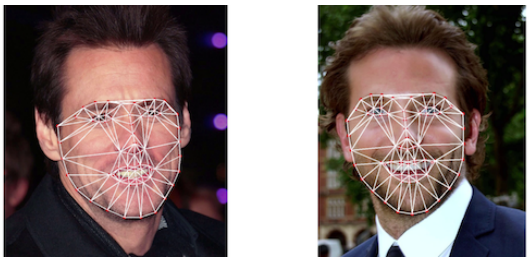

Fig. 11. Triangulation destination image.[Canu [n.d.]]

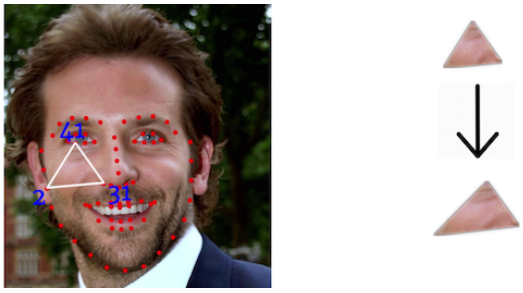

Fig. 12. Extract and warp triangles.[Canu [n.d.]]

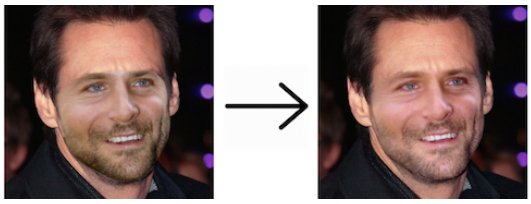

Fig. 13. Replace the face on the destination image.[Canu [n.d.]]

B DIFFERENCE IN REPO

Thanks to the base work [face-of-art]. And my revised repo is [here]

The main changes are:

- face_of_manga/deep_heatmaps_model_fusion_net.py L298-328 for global feature module.
- face_of_manga/pts60_deformation_functions.py for Manga faces' Thin Plate Spline
- face_of_manga/face_swap_menpo_landmarks_success.ipynb for face swapping.

Other changes are for the data IO and trivial. For details you can check the commits in the repo mentioned above.