Learning pose-aware human representations for conditional person image translation

Hsuan-I Ho[1], Chi-Ching Hsu[2], and Juan-Ting Lin[3]

[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Department of Mechanical and Process Engineering, ETH Zurich, Switzerland
[3]Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland

***Abstract*—Most existing pose-guided image generation approaches rely on pre-computed pose inputs and paired ground truth to transfer appearance details from the source image to the conditional pose. We alternatively model this problem in an image translation setting which does not require auxiliary pose inputs to translate human appearances. With the goal of learning representative pose and appearance details, we propose an end-to-end *Multi-Objective Multi-Identity Network*. Our model explicitly encodes semantic pose information while capturing the corresponding appearance details in a multi-task learning scheme. Moreover, our newly designed *skeleton patch verification* loss mitigates the constraint of using paired ground truth for training. Both qualitative and quantitative results from our experiments confirm the effectiveness of our proposed method.**

I. Introduction

In recent years, research attention has been attracted to learn robust human representations for a variety of computer vision tasks like pedestrian re-identification [1] or person image generation [2]. In this project, we focus on the task of conditional person image generation, which aims at transferring the appearance details (e.g. clothes, hairstyles, accessories...) from a source image to a target pose as depicted in Fig. 1a.

While recent works have shown impressive performance in light of generative adversarial networks (GANs) [3] based image generation models, these methods highly rely on paired (images with the same identity) ground truth and auxiliary pose maps from off-the-shelf pose estimation models (Fig. 1b). Such a training scheme results in the inflexibility of utilizing pairwise data only, the instability for relying on the quality of generated pose maps, and also extra computations for generating the pose maps even in the inference stage.

On the other hand, other approaches also attempt to model this problem in a more practical yet challenging image-to-image translation setting [4], which learns an implicit image translation function (e.g., the changing of colors or styles) given only RGB source images and target images. However, they cannot transfer human appearances well as shown in Fig. 1c, since the learned human representation is easily affected by irrelevant background or contents. Thus, it would be a challenging task to adapt human appearance details in the source image while keeping its pose fixed as the target image without any guidance from the auxiliary pose map.

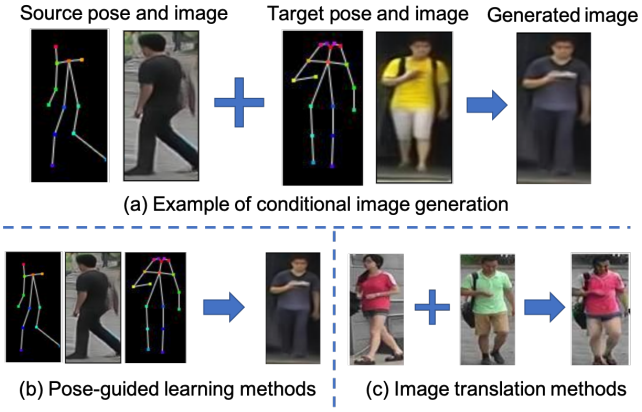

Figure 1: (a) Human image generation task conditioned on a target pose. (b) Example of pose-guided training methods which rely on paired training images and auxiliary pose inputs. (c) Unsupervised image translation methods which do not necessarily transfer human appearance details to the target pose.

To overcome the aforementioned challenges we propose *Multi-Objective Multi-Identity Network* (MOMI-Net), which not only relaxes the requirement of using auxiliary pose maps in a multi-task learning scenario, but it is also capable of transferring human appearance details to any unpaired identities. Our network consists of a pose distillation branch plus an image translation branch, which can be optimized jointly by sharing information with each other. This also provides a better generalization to the conditional image translation task. We uniquely adopt the technique of knowledge distillation [5] to ensure the pose distillation branch would encode semantically meaningful human pose features. Further, the image translation branch learns pose-aware appearance details by manipulating the obtained features for image reconstruction. It is also worth noting that with the aid of our newly proposed *skeleton patch verification* (SPV) loss, our model can be trained more efficiently by sampling

any unpaired training images.

In the following subsections, we provide an overview and issues of existing works, followed by a summary of our contributions before entering the details of our proposed method.

A. Related works

Pose-guided image generation: A vast number of approaches have been proposed to tackle with pose variances, irrelevant background, and occlusions using auxiliary pose landmarks [2], [6], [7], [8], [9], [10] or semantic parsing [11] as guidance during training. While pose-guided methods demonstrate the ability to extract appearance details effectively, they also introduce computational overhead for the pre-computed pose inputs or cumbersome components in their models. Meanwhile, their results also highly depend on the quality of pose inputs obtained from the off-the-shelf human pose detector, which once again limits their flexibility. In contrast, our network shows the ability to extract meaningful pose information only from the input RGB image without extra guidance.

Unsupervised image translation: Recent studies also utilize the idea of style transfer to synthesize images across viewpoints/poses. For instance, Zheng et al. [12] utilize MUNIT [13] to decompose images into structure and appearance codes and generate images by recombining them. Xiao et al. [14] disentangle pose-related features by reconstructing pseudo ground truth sampled from a well-constrained dataset. However, rather than explicitly translate images with respect to their poses, these methods only transfer them into another implicit style. Thus, the generated results cannot fully exhibit appearance details from the source images. While Wu et al. [15] recently manage to interpret the disentangled structural (pose) feature via geometry distillation with VAE [16], their assumption is still insufficient to model complicated human images with large pose variances. In comparison, our method explicitly preserves semantic information in the pose feature via knowledge distillation and is capable of adapting appearance details more accurately.

B. Contributions

- We resolve the issue of existing pose-guided learning methods, which require auxiliary pose inputs for both training and inference. Also compared with image translation methods, MOMI-Net can explicitly learn semantic pose features and more representative appearance details.
- Our newly proposed SPV loss mitigates the limitation of utilizing paired ground truth for training. Under our training scheme, our network can produce more realistic and diverse images by transferring the appearance information between arbitrary identities.
- The re-implementation of several state-of-the-art methods alongside experimental results have quantitatively and qualitatively confirmed the effectiveness of our method on the challenging dataset.

II. Methods

In this section, we introduce our MOMI-Net which performs conditional image translation from a source image to any target pose. The entire pipeline is shown in Fig. 2: Given an input image, our pose distillation branch first extracts semantic pose information from the image and encodes it into a pose feature z_p. Similarly, our image translation branch takes the image alongside its semantic pose information to return a pose-aware appearance feature z_a. Finally, the task of image translation is accomplished by recovering the images after swapping and recombination of z_p and z_a with other training samples.

In particular, a training tuple $x, x^+, x^- \in \mathbb{R}^{H\times W\times 3}$ is sampled from the dataset at each training step. x^+ denotes a positive sample that shares the same identity (similar appearance) with x while x^- is a negative sample with distinct appearance. Our goal is recovering x^+ using (z_p^+, z_a) and translating a new image with (z_p^-, z_a). We note that all samples in the dataset would have different poses, i.e, it is impossible to get a ground-truth image for the newly translated image. Properties of each component will be further discussed in the following subsections.

A. Pose distillation branch

As we have mentioned in the previous section, the pose distillation branch aims at extracting representative human pose feature $z_p \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times d}$ from the input image. However there is no guarantee that the encoder would preserve only pose-related information while leaving out unnecessary contents from the images. We therefore adopt the technique of knowledge distillation which is able to transfer the behavior of an existing teacher model to our student model.

To be more precise, the pose distillation branch consists of an encoder-decoder based pose detector F_p, which generates a 17-channel landmark heatmap $h \in \mathbb{R}^{H\times W\times 17}$. Unlike existing methods which only implicitly approximate the landmarks or features based on some geometric assumptions, knowledge distillation ensures our model also explicitly detect meaningful pose landmarks as the teacher model does. In consequence, we define the distillation loss by minimizing the difference between their output activations:

$$\mathcal{L}_{distill} = \|\psi(x) - F_p(x)\|_2, \tag{1}$$

where ψ denotes output activations from the pose estimation teacher model [17]. Instead of directly using the activations as human pose landmarks, we further re-project the coordinates of its max activations to another Gaussian-like heatmap by a fixed Gaussian convolution kernel centered at the landmark coordinates. As a result, we obtain a new Gaussian-like heatmap $p \in \mathbb{R}^{H\times W\times 17}$ which contains only semantic pose information for the later pose feature encoding $E_p : p \rightarrow z_p$.

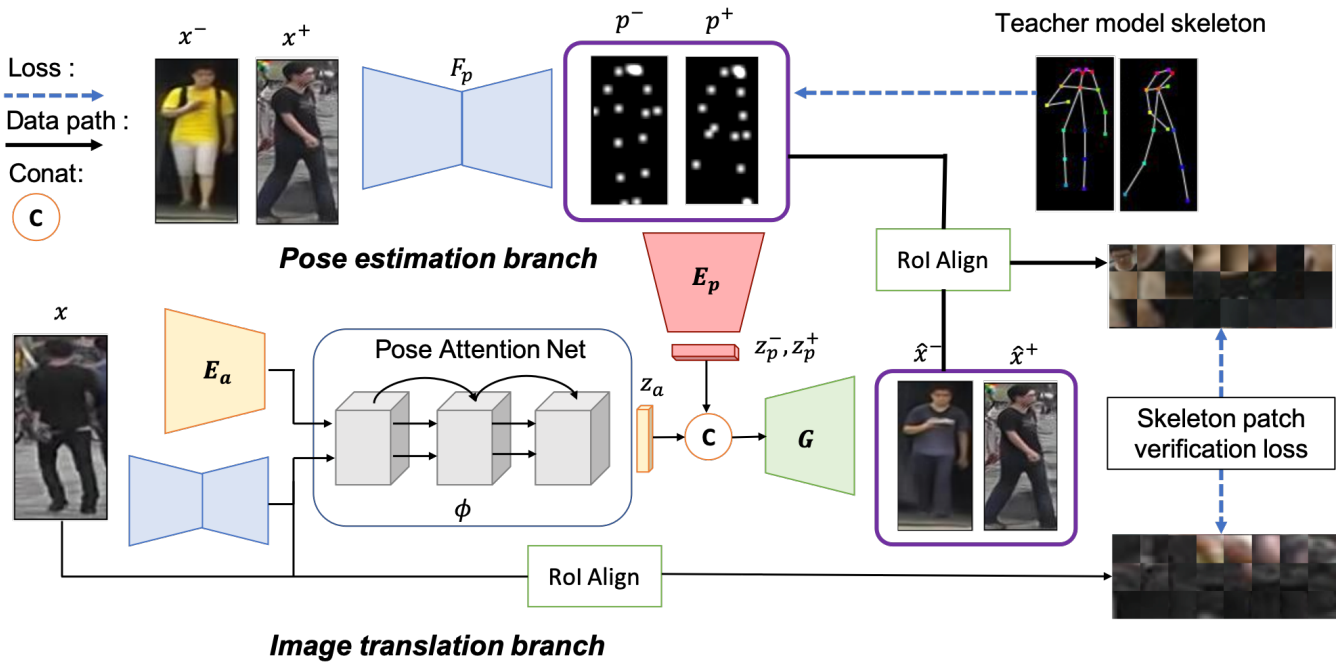

Figure 2: Our proposed MOMI-Net including a pose distillation branch and an image translation branch. We have x, x^+, x^- denote the source image, positive and negative samples respectively. z_p and z_a represent their pose and appearance features and they are recombined for generating images $\hat{x}^+$ and $\hat{x}^-$.

B. Image translation branch

Appearance feature encoder. After the pose distillation branch encodes the representative pose feature z_p from the input, our image translation branch would extract the remaining appearance details for image reconstruction and translation. With the recent success of attention mechanism [18], [19] in convolutional neural networks, we could also utilize pose information together with the input image to extract more accurate appearance features. To be detailed, our appearance feature extractor consists of a downsampling encoder E_a plus a series of pose attention blocks ϕ [10]. The final appearance representation is compute as follow:

$$z_a = \phi(E_a(x), z_p), z_a \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times d}. \quad (2)$$

Image generator. The image generator G is deployed to recover appearance details from input images conditioned on pose features. Rather than directly learning an image-to-image mapping function with paired ground truth like other pose-guided methods do, we train our image generator by manipulating the intermediate latent features (z_a, z_p) with different combinations. For instance, the input image x should be recovered by z_a and z_p itself, and meanwhile we also expect z_a and z_p^+ should recover x^+ since they share the same identity (appearance). Thus, we calculate the reconstruction loss of image translation branch:

$$\mathcal{L}_{recon} = \|x - G(z_a, z_p)\|_1 + \|x^+ - G(z_a, z_p^+)\|_1. \quad (3)$$

However, simply minimizing pixel-level L1 loss cannot guarantee the image perceptual quality and result in blurry images as observed in [20]. We consequently reformulate the objective of image translation branch as:

$$\mathcal{L}_{translate} = L_{recon} + \sum_l \lambda_l \|\psi_l(x) - \psi_l(\hat{x})\|_1, \quad (4)$$

where $\hat{x} = G(z_a, z_p)$ is the reconstruction of input x, and ψ_l is the feature map obtained from the l-th block of VGG-19 [21]. We empirically find out using $l = \{2, 5\}$ leads to better results.

Discriminator. Finally, the discriminator D (not shown in the figure) is deployed to distinguish whether the generated image and the input image belong to the same person, and meanwhile to ensure the generator would produce realistic images. That is, the image generator tries to fool the discriminator by recovering and translating sufficient appearance details which contains identity-related information. The adversarial loss of the discriminator D is thus defined as

$$\begin{aligned} \mathcal{L}_{adv} = & \mathbb{E}_{x \sim \mathcal{X}}[\log D(x, x^+)] + \\ & \mathbb{E}_{\hat{x} \sim \mathcal{X}'}[\log(1 - D(\hat{x}, x^+)] + \\ & \mathbb{E}_{\hat{x}^+ \sim \mathcal{X}'}[\log(1 - D(x, \hat{x}^+)], \end{aligned} \quad (5)$$

where $\mathcal{X}$ and $\mathcal{X}'$ represent the true data distribution and generated data distribution by the image generator G.

C. Skeleton patch verification

Besides learning representative pose and appearance features in a multi-task learning scenario, we further pro-

Figure 3: Qualitative comparisons of conditional image translation on Market-1501 dataset. The generated image should share appearance details in src while having the pose in tar. DG, PN and PG^2 represent the results of [12], [8] and [2], respectively. Our* denotes the variant of our method excluding perceptual loss during training.

pose *skeleton patch verification* (SPV) to jointly optimize our network by exploiting the information and knowledge learned from both branches. The key concept of SPV is to online match human local patches according to their pose landmarks. For instance, when transferring appearance details from the source image x to its negative sample x^-, there is actually no ground truth image to tell the network how $\hat{x}^- = G(z_a, z_p^-)$ should look like. However we could still expect their local appearances (e.g, color of the sleeve, trouser, or shoes) to be invariant after changing the pose. To this end, we apply RoI Align [22] to the source image x and translated images $\hat{x}^-$ according to their landmark coordinates on heatmap p. As depicted in Fig. 2, the $RoI(\cdot)$ operation would return a series of image crops given the corresponding bounding box coordinates. Thus we integrate this operation into MOMI-Net to extract significant human patches for the online verification process. Finally, we define our SPV loss as

$$\mathcal{L}_{patch} = \|RoI(x, p) - RoI(\hat{x}^-, p^-)\|_1. \quad (6)$$

Is is also worthwhile to emphasize that SPV loss is applicable to any unpaired images, which allows our network to generate more realistic images by viewing a diversity of training pairs.

D. Training of our model

Our multi-task network can be jointly train end-to-end by optimizing the following full loss function:

$$\mathcal{L} = \mathcal{L}_{distill} + \alpha\mathcal{L}_{translate} + \beta\mathcal{L}_{patch} + \gamma\mathcal{L}_{adv}, \quad (7)$$

where α, β, γ are the weighting hyperparameters for the translation branch, SPV loss and adversarial learning respectively. We found choosing $(\alpha, \beta, \gamma) = (1.0, 0.5, 0.5)$ leads to best image quality in our experiments. We also analyze the influence between γ and generated images in Fig. 4.

III. Experiments

A. Dataset and settings

Market-1501 [23] is a large scale multi-view dataset that have been commonly used for person image generation. It is originally a person re-identifiction (re-id) dataset containing 1501 identities in total 32668 images captured from 6 disjoint camera views. The images are split into train and test sets of 12936 and 19732 images with 750 and 751 identities.

However, existing works evaluate their methods using paired images only (i.e., the source and the target are the same identity), which is not able to fully evaluate the performance in the image translation setting. Therefore, we additionally divide 751 test identities into non-overlapped

source group (376 identities) and target group (375 identities). We randomly create 12000 source-to-target image pairs and ensure the identity of source and target are not identical.

B. Evaluation Metrics

In the quantitative experiments, the Inception Score (**IS**) [24] and Fréchet Inception Distance (**FID**) [25] are reported to measure the quality of generated images. To eliminate the influence of background in images, we further implement **mask-IS** by filtering out unnecessary background before computing the score.

We note that these metrics only indicate how realistic the generated images are but not how well they translate the appearances from source to target, therefore we introduce a new evaluation metrics *appearance preservation score* (**APS**). We finetune an ImageNet pre-trained ResNet-50 on the test set. This network returns the score of affinity $\in [0, 1]$ between two images, i.e., the score indicates how similar the generated image and the source image are and whether they have the same identity.

C. Methods to be compared

In our experiments, we quantitatively and qualitatively compare our method with both pose-guided methods (**PG^2** [2], **PN-GAN** [8]) and image translation methods (**DG-Net** [12]). We re-implemented their official source code and generated new images based on our selected 12000 pairs. Note that PG^2 and PN-GAN require additional pose inputs when generating results while DG-Net and our method need RGB images only.

We also perform an ablation study to analyze each component in our network. **Ours w/o** $\mathcal{L}_{patch}$ denotes removing of SPV loss Eq. (6), **Ours w/o per.** represents without using perceptual loss in Eq. (4) and **Ours w/o att.** is the variant of replacing pose attention blocks into normal residual blocks.

IV. Discussion

A. qualitative results

Fig. 3 shows the qualitative comparison of our model and state-of-the-art methods. It can be seen that our model produced favorable results even the person in the source image is blocked by the umbrella or shifting to the corner. In contrast, DG-Net simply copy-pasted all the content from target images and changed their color, which failed to extract correct appearance information from source images due to irrelevant background. Also compared to pose-guided methods, they cannot generally translate appearances to targets well since they only learn an image-to-image mapping from limited paired training data.

B. quantitative results

Table I summarizes the quantitative results of our proposed method against state-of-the-art methods. Our method achieved the highest on APS which once again confirmed our ability of image translation. Interestingly, although **Ours w/o per.** got the highest APS, its image quality is much worse. The ablation study thus verified the design choices in our full model, which is able to balance image quality and translation ability.

Figure 4: Analysis of the hyper-parameter γ.

Methods	Market-1501 test			
	IS ↑	mask-IS ↑	FID ↓	APS ↑
Real (source)	3.51	3.25	-	-
PG^2	2.95	**3.34**	131.56	0.8160
PN-GAN	3.17	3.21	**45.10**	0.7220
DG-Net †	3.62†	3.26†	17.01†	0.8880
Ours	2.44	2.98	57.50	0.9026
Ours w/o $\mathcal{L}_{patch}$	2.33	2.93	54.32	0.8677
Ours w/o per.	**3.33**	3.16	110.64	**0.9106**
Ours w/o att.	2.08	2.85	78.20	0.8619

Table I: Quantitative results and ablation studies on our proposed method. We note that DG-Net incorrectly copy-pastes all content from targets, thus their realism results is incompatible.

Another interesting observation is the realism metrics, where we did not get a desirable result corresponding to the qualitative one. We come up with several explanations: 1) Background has a larger impact on the score. We can see that most methods got improvement after masking out the background. However, evaluating the background quality makes no sense since our model only focuses on translating human appearances. 2) As pointed out in [26], IS score is biased. We surprisingly find out PG^2 and Ours w/o per. got the best IS scores but the actual image quality seems not necessarily better from human eyes. We conclude that FID could better evaluate the quality of the images and our model is still capable of producing comparable results. 3) In this project, we did not finetune our model for long epochs. Also, we did not search out optimal hyperparameters for training, but we believe there is still room for improvement.

V. Summary

In this project, we propose MOMI-Net and SPV loss to tackle the conditional image translation problem. We solve issues in existing works by learning representative features with mult-task learning. Experimental results confirmed the effectiveness of our proposed model.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[2] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2223–2232.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[6] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8857–8866.

[7] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3408–3416.

[8] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 650–667.

[9] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems (NIPS)*, 2018, pp. 1229–1240.

[10] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2347–2356.

[11] S. Song, W. Zhang, J. Liu, and T. Mei, "Unsupervised person image generation with semantic parsing transformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[14] F. Xiao, H. Liu, and Y. J. Lee, "Identity from here, pose from there: Self-supervised disentanglement and generation of objects using unlabeled videos," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7013–7022.

[15] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, "Disentangling content and style via unsupervised geometry distillation," *Proceedings of the International Conference on Learning Representations Workshops (ICLR Workshops)*, 2019.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[17] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, p. Proceedings of the European Conference on Computer Vision (ECCV).

[19] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 705–10 714.

[20] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4491–4500.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[23] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2234–2242.

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[26] S. Barratt and R. Sharma, "A note on the inception score," 2018.