LEARNABLE MASKS FOR POSE-GUIDED VIEW SYNTHESIS

Mohamed Ilyes Lakhal[1], *Oswald Lanz*[2], *Andrea Cavallaro*[1]

[1]Centre for Intelligent Sensing, Queen Mary University of London, UK
[2]Technologies of Vision, Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

Pose-guided human view synthesis uses a target pose to generate the appearance of a new view of a person. The input view and the target pose can be processed separately with UNet architectures that combine the results in a late fusion stage. UNet architectures link their encoder and decoder with skip connections that preserve the location of spatial features by injecting input information in the decoding process. However, direct skip connections may transfer irrelevant information to the decoder. We overcome this limitation with learnable mask for skip connections that encourage the decoder to use only relevant information from the encoder. We show that adding the proposed mask to UNet architectures improves the performance of view synthesis with only a slight increase in inference time.

Index Terms— View synthesis; UNet; Mask Skip Connection.

1. INTRODUCTION

Pose-guided human image synthesis is the process of generating new images of a person given a set of 2D keypoints representing the target pose (Figure 1). Because of asymmetries in human poses, this problem is particularly challenging. Methods for view synthesis include Pose Guided Person Generation Network (PG^2) [1], View-Disentangled Generator (VDG) [2], Pose-Normalized GAN (PN-GAN) [3], Deformable GAN (Def-GAN) [4], and Vari-UNet [5].

Table 1 summarizes recent models for pose-guided image synthesis that use UNet or ResNet architectures [6]. The models may use reconstruction, $\mathcal{L}_r$, adversarial $\mathcal{L}_a$, variational [7] $\mathcal{L}_v$ and perceptual [8], $\mathcal{L}_p$, losses. An in-depth analysis on how to combine these losses is presented in [9]. Most view-synthesis methods [1, 2, 4, 5] use UNet [10], a U-shaped encoder-decoder Convolutional Neural Network (CNN) with the ability to retain spatial information [11]. The encoder is composed of multi-layer convolutional blocks that map the input view to a lower-dimensional feature map, which is then upsampled by the decoder with a deconvolution [12].

Architectures for pose-guided human image synthesis use early or late fusion. *Early fusion* [1, 3] combines the input view and the target pose in a feature representation for the encoder. PG^2 [1] is a two-stage model that converts the target pose into a heatmap, synthesizes a coarse image in the target pose and then refines it with adversarial training by optimizing the weights of the network. PN-GAN [3] learns identity-sensitive and view-invariant features by reformulating the problem on a pre-defined set of poses for person re-identification.

Late fusion [2, 5, 4] keeps the processing of the input view (and input pose) and the target pose separate at the encoder and combines the results in the decoder. VDG [2] processes input view (and pose) and the target pose in two separated encoders and then fuses their results to reconstruct a coarse target view with the decoder, which is learnt using a loss function based on Structural Similarity (SSIM) [13]. Vari-UNet [5] uses a variational inference constraint. Def-GAN [4] extends the UNet architecture to learn a set of affine transformations that are applied to the convolutional block of the encoder using pre-defined masks.

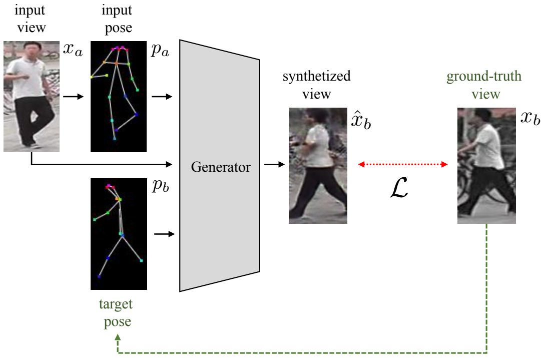

Fig. 1: Pose-guided view synthesis. A generator processes the input view and the target pose to implicitly learn the multi-view mapping through a full reconstruction loss, $\mathcal{L}$.

Masks have been used to filter information in MaskConnect [14] and Modular-GAN [15]. MaskConnect [14], which allows for flexible residual connections between the previous layers, is preferable for image recognition than fixed (hand-designed) residual connections such as those in ResNet [6] or ResNeXt [16]. Modular-GAN [15] defines the mask as convolutional operation in the transformer module for face synthesis, a less complex task than person synthesis.

As pre-defined masks [4] may eliminate important information for the generation, we aim to learn how to mask only irrelevant information. To this end, we include in the UNet architecture a mask module that learns to establish selective skip connections. A selective skip connection blocks (*masks*) input features from being accessed in the decoding stage if they corrupt the target view. The mask consists of a parameter tensor and an operator that combines the tensor with the encoder feature map. The proposed module is end-to-end trainable and can be plugged into existing late fusion architectures to allow selective information passing.

2. MASKED SKIP CONNECTIONS

In this section we present the proposed mask UNet (mUNet) model and motivate the need for learnable mask skip connections. We also investigate two strategies to combine the learnt mask with the encoder feature map in the skip connection operation, namely a dense

Table 1: Pose-guided view synthesis methods: main design choices and losses used ($\mathcal{L}_r$: reconstruction loss; $\mathcal{L}_a$: adversarial loss; $\mathcal{L}_v$: variational loss [7]; $\mathcal{L}_p$: perceptual loss [8]).

Method	**Losses**				**Architecture**	**Fusion**
	$\mathcal{L}_r$	$\mathcal{L}_a$	$\mathcal{L}_v$	$\mathcal{L}_p$		
PN-GAN [3]	✓	✓			ResNet	Early
PG^2 [1]	✓	✓			UNet	
VDG [2]	✓	✓				Late
Def-GAN [4]	✓	✓		✓		
Vari-UNet [5]	✓		✓	✓		

masking, where each pixel is considered separately from its neighbors, and a sparse masking, where the encoder feature map is masked using a grid of cells (see Figure 2).

Given an input view of a person, x_a, the corresponding body pose in that view, p_a, and a target pose, p_b, the goal is to design and train a generator, f_θ, that produces $\hat{x}_b$, the appearance of a new view of the person. The target view, $\hat{x}_b$, shall be as similar as possible to the ground-truth image, x_b, from which the target pose, p_b, was extracted.

We decompose the generator in three parts, namely an input encoder, E_I, a pose encoder, E_P, and a decoder, D. The input encoder, E_I, maps x_a and p_a to a lower dimension feature vector $v_I = E_I(x_a \oplus p_a)$. The pose encoder, E_P, processes p_b such that $v_P = E_P(p_b)$, which will guide the decoder in the synthesis. Finally, v_I and v_P are concatenated and fed to the decoder, D, to produce the target view $\hat{x}_b = D(v_I \oplus v_P)$, where $\oplus$ indicates the concatenation operation.

Let N be the number of layers of the encoder and the decoder, and w, h and c be the width, height and the number of channels (depth) of the feature map $f_E^i \in \mathbb{R}^{h\times w\times c}$ of the encoder at layer i and $f_D^{N-i} \in \mathbb{R}^{h\times w\times c}$ be the corresponding feature map at layer $N-i$ of the decoder. In the UNet architecture [10] a skip connection concatenates the feature f_E^i of the encoder with the feature f_D^{N-i} of the decoder, such that:

$$f_S^{N-i} = f_E^i \oplus f_D^{N-i}. \tag{1}$$

We add in each layer i of the encoder CNN a learnable mask, $\mathbf{M}^i \in \mathbb{R}^{w\times h\times d}$ where $d \leq c$. $\mathbf{M}^i$ is a weight matrix that is applied to the encoder feature map, f_E^i, through $\psi\colon \mathbb{R}^{h\times w\times c} \times \mathbb{R}^{h\times w\times d} \rightarrow \mathbb{R}^{h\times w\times c}$ such that the skip connection at each layer becomes:

$$f_{S'}^{N-i} = \psi(f_E^i, \mathbf{M}^i) \oplus f_D^{N-i}, \tag{2}$$

where $f_{S'}^{N-i}$ is the proposed mUNet model (see Figure 3). The resulting values of the weight matrix $\mathbf{M}^i$ after optimisation are those that minimize the loss of training. The binary mask is obtained by combining the weight matrix with the input feature map f_E^i.

We consider two variants of the mUNet model, which has a mask for each channel of the encoder: the mUNet$_o$ model, which has a mask with only one channel; and the mUNet$_g$ model, which uses one sparse mask (similar to mUNet$_o$) for all the feature channels to block the pixels used in the encoder.

Let ϕ be a function such that for a given input x and y:

$$\phi(x, y) = \mathbb{1}_{[\sigma(x\odot y)\geq 0.5]}, \tag{3}$$

where the function $\mathbb{1}$ returns 1 if its condition holds and 0 otherwise, σ is the sigmoid function, and $\odot$ is the element-wise product.

For skip connections with dense masks, each pixel in the feature map is independent of its neighbors and therefore the spatial relationship between pixels is learnt from the data. The resulting masked feature map is:

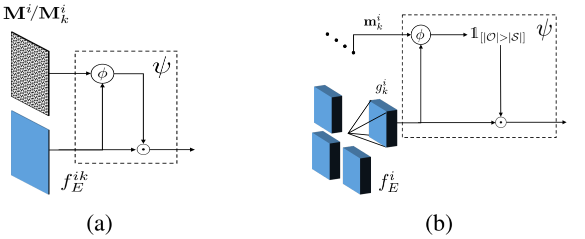
$\mathbf{M}^i/\mathbf{M}^i_k$
ϕ
ψ
f_E^{ik}
(a)
$\mathbf{m}^i_k$
ϕ
$\mathbb{1}_{[|\mathcal{O}|>|\mathcal{S}|]}$
ψ
g^i_k
f_E^i
(b)

Fig. 2: The proposed mask skip connection and its variants: (a) dense mask (the mUNet$_o$ model uses one mask $\mathbf{M}^i$ for all the feature map channels) and (b) sparse mask.

$$\psi(f_E^i, \mathbf{M}^i) = f_E^i \odot \phi(f_E^i, \mathbf{M}^i). \tag{4}$$

For skip connections with sparse masks, pixels are locally related using cells that are treated separately, whereas the pixels inside a cell are all either filtered out or passed on. We divide the encoder convolutional feature map f_E^i into $\lfloor \frac{h}{h_g} \rfloor \lfloor \frac{w}{w_g} \rfloor$ cells, with each cell, g_k^i, having size $h_g \times w_g \times c$, and

$$f_E^i = \left\{ g_k^i \in \mathbb{R}^{h_g\times w_g\times c} \middle| k \in \lfloor \frac{h}{h_g} \rfloor \lfloor \frac{w}{w_g} \rfloor \right\}, \tag{5}$$

where h_g (w_g) is the height (width) of the grid. We multiply the k-th value of the mask $\mathbf{m}^i \in \mathbb{R}^{\lfloor \frac{h}{h_g} \rfloor \lfloor \frac{w}{w_g} \rfloor}$ with the cell g_k^i thus generating an $h_g \times w_g \times c$ binary matrix, b_k^i, defined as:

$$b_k^i = \phi(g_k^i, \mathbf{m}_k^i). \tag{6}$$

We use the dominant binary value in each cell to mask the information through the grid as:

$$\psi(g_k^i, \mathbf{m}_k^i) = \mathbb{1}_{[|\mathcal{O}|>|\mathcal{S}|]} g_k^i, \tag{7}$$

where $|\mathcal{O}|$ and $|\mathcal{S}|$ are the total number of 1s and 0s, respectively, in b_k^i. If the number of 1s inside a cell is greater than the number of 0s, we pass the information of all the pixels of that particular cell.

3. VALIDATION

In this section we evaluate different masks for mUNet and compare the robustness of different architectures with and without the proposed learnable mask.

We use as datasets for the evaluation Market-1501 [17] and the *In-shop Clothes Retrieval Benchmark* subset of DeepFashion [18]. As train and test splits we use $263,631$ and $12,000$ pairs, respectively, for Market-1501 and $101,268$ and $8,670$ pairs, respectively, for DeepFashion [2]. We apply mUNet, mUNet$_o$ and mUNet$_g$ on the VDG [2] and define No-mask as the core VDG model (without masks). The image size for Market-1501 is 128×62 and for DeepFashion is 256×256 pixels. To account for this difference in resolution, we use a mUNet architecture with $N = 5$ layers for Market-1501 and with $N = 6$ layers for DeepFashion. To quantify the impact of the mask for VDG [2] and Def-GAN [4], we use the training parameters presented in the original papers. We further analyze the input pose effect to the generation by removing it from *image branch* encoder (the one receiving the input image and input pose) of the VDG model. We refer to this model as VDG$_{np}$ and we train it with the same parameters as VDG. Unlike the Stage I, the

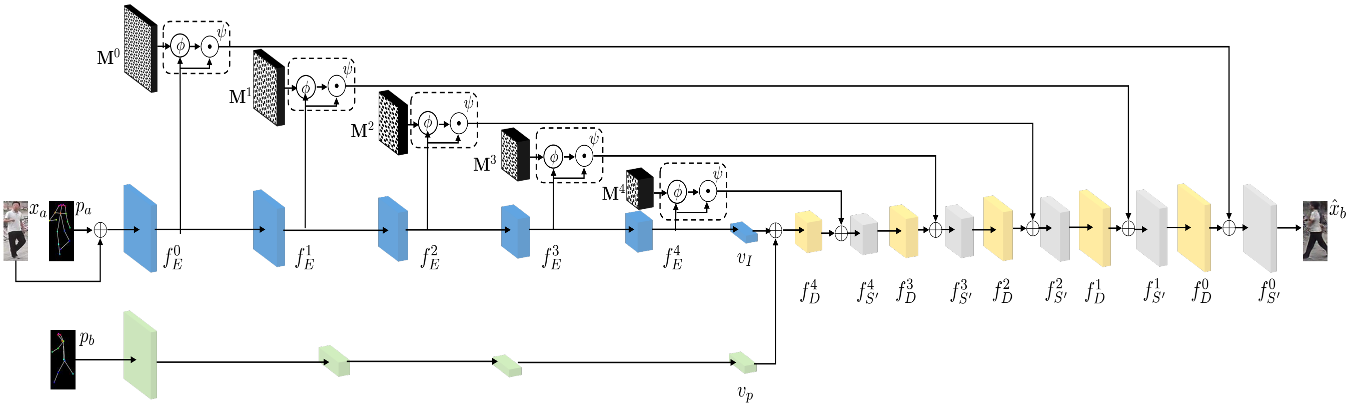

Fig. 3: The proposed mUNet architecture for pose-guided view synthesis.

Stage II in VDG (VDG$_{np}$) does not use the fully connected layer. The mUNet of the VDG model is trained for 30k (50k) steps in Stage I and 14k (30k) steps in Market-1501 (DeepFashion). Note that the mask for VDG is only applied in Stage I, since the Stage II is for refinement only. This explains our choice of keeping the same architecture and hyper-parameters in Stage II. The mUNet of Def-GAN is trained for 90 epochs in each dataset. For model selection we use a one-stage generator for all the models. We use mask $\mathcal{L}_1$ loss along with adversarial training [2]. All the models are implemented in TensorFlow [19]. As performance measures, we use SSIM (Structural Similarity [13]), IS (Inception Score [20]), and their mask versions, mask-SSIM and mask-IS, which keep only the person of interest before computing the scores [1] for Market-1501 to remove the (irrelevant) impact of the background on the measurement of the quality of the synthetized person.

Table 2 shows the results of the methods under analysis. The dense masks mUNet and mUNet$_o$ improve over No-mask. mUNet$_g$ performed worse than the original model, No-mask, because its masking is too strong and suppresses important information from the encoder. mUNet slightly outperformed mUNet$_o$ as having a mask for each channel of the encoder feature map gives more flexibility in masking between pixels.

To evaluate why the sparse mask method, mUNet$_g$, performs poorly, Figure 4 shows the average number of non-zero pixels in the Market-1501 test set. We extract the first convolutional block of the encoder, which is of shape $dim(Conv_1) = 128 \times 64 \times 128$, for each channel $k \in \{1, \ldots, 128\}$ and we count the number of non-zero pixels of 128×64, *i.e.*, $|\mathcal{O}^c_{Conv_1}|$. For No-mask we use the original skip connections so all the channels have an average pixel of $\approx$ 8192 pixels. For dense mask methods about half of the pixels are masked. For the grid mask, $40 \sim 60\%$ pixels in each channel are masked and no patterns can be noticed in the learned mask as shown in Figure 5, which displays the learned mask of the first convolutional block. In each row we randomly select one feature channel. For the dense masks, the masked pixels are uniformly distributed across each frame even if we only keep 50% of the pixels in each channel, whereas for the grid mask, a whole cell that may carry important spatial information may be masked, thus leading to worse results.

Table 3 shows the results for Stage I on Market-1501 and DeepFashion to evaluate the influence of the input pose of VDG (we remove it from the *image branch* of the VDG model) and also compares with state-of-the-art methods mUNet of the models VDG, VDG$_{np}$ and Def-GAN. The use of the mask improves most scores and VDG$_{np}$ improves over IS. Images produced by the mask version of the models seem more realistic and have fewer artifacts. The mask UNet of Def-GAN gets a drop in 4 out of 6 of the scores used. The model uses deformable convolution to transform the encoder feature map using affine transformations and when applying our mask we might have discarded some of these transformed feature information that would instead be useful.

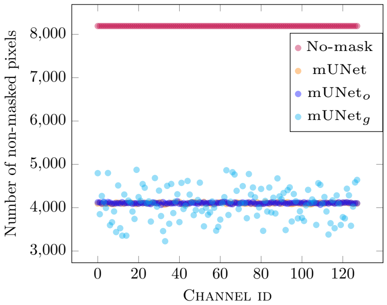

Fig. 4: Comparison of total number of non-masked pixels on all the channels of the first convolutional layer using the method under analysis. Note that the number of pixels for mUNet is similar to that of mUNet$_o$ and therefore the two lines overalp in the plot. KEY - No-mask: original model; mUNet: masked skip connection model; mUNet$_o$: mUNet with one mask for all the feature map channels; mUNet$_g$: grid mask model.

Table 2: Influence of different masks on Market-1501. KEY - mUNet: masked skip connection model; mUNet$_o$: mUNet with one mask for all the feature map channels; mUNet$_g$: grid mask model.

Method	**SSIM**	**IS**	**mask-SSIM**	**mask-IS**
No-mask	.222	3.294	.763	3.140
mUnet	.226	3.473	.769	3.294
mUNet$_o$	.227	3.412	.769	3.199
mUNet$_g$	.217	3.046	.760	3.183

Figure 6 shows that the masked version, mUNet, produces in general better quality images and the mask version of both VDG and VDG$_{np}$ (mUNet column) produces sharper body limbs. Figure 7

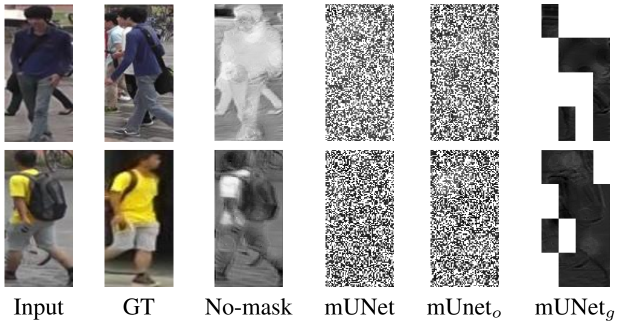

Fig. 5: Sample mask learned on the first convolution layer. First column: input view. Second column: target (ground-truth) image. First row: 120^{th} feature map; second row: 20^{th} feature map. White correspond to the masked pixels for mUNet models. KEY – mUnet: masked skip connection model; mUNet$_o$: mUNet with one mask for all the feature map channels; mUNet$_g$: grid mask model.

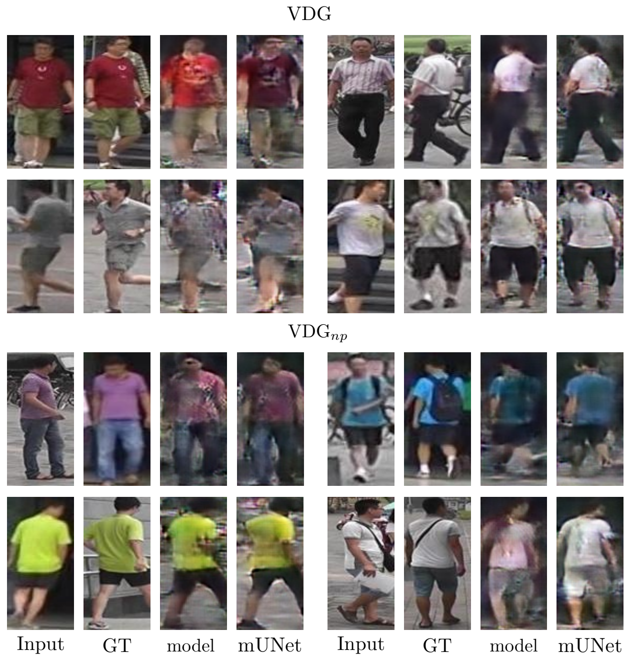

Fig. 6: Comparison of three models with their masked skip connection version on Market-1501 and DeepFashion. KEY – GT: ground-truth image; model: original model; mUNet: masked skip connection version of the VDG (VDG$_{np}$) model.

shows instead two examples of challenging cases where mUNet for VDG and VDG$_{np}$ produces more artifacts or missing body parts than the original architecture. For VDG, mUNet was unable to synthesize the face from side to frontal view and for VDG$_{np}$, mUNet generated shorts instead of the long black pants in the input view.

To conclude, including the mask module only moderately affect the execution of the overall architecture as the average running time of 1000 test samples is 33.6, 34.2, 34.4 and 189.8 ms for No-mask, mUNet$_o$, mUNet and mUNet$_g$, respectively. The grid mask is considerably slower than other solutions because of its loops inside each feature. mUNet is slightly slower than mUNet$_o$, but both running times are very similar to those of No-mask.

Table 3: Performance on Market-1501 and DeepFashion of the Stage I of VDG with and without the input pose along with the mask version (first four rows) and with and without mUNet (last six rows). KEY – mSSIM: mask-SSIM; mIS: mask-IS.

	Market-1501				DeepFashion	
Method	**SSIM**	**IS**	**mSSIM**	**mIS**	**SSIM**	**IS**
VDG [2]	.274	3.407	.799	2.733	.691	2.773
+ mUNet	.268	2.673	.803	2.743	.691	2.878
VDG$_{np}$ [2]	.263	2.560	.797	2.705	.686	2.887
+ mUNet	.268	2.573	.801	2.753	.689	2.897
PG^2 [1]	.252	4.015	.771	3.555	.641	3.187
PDIG [21]	.099	3.483	.614	3.491	.614	3.228
VDG [2]	.238	4.007	.775	3.354	.708	3.003
+ mUNet	.244	4.257	.777	3.452	.707	3.075
VDG$_{np}$ [2]	.264	3.470	.773	3.220	.710	2.938
+ mUNet	.254	4.093	.779	3.541	.706	2.997
Def-GAN [4]	.290	2.990	.798	3.544	.665	3.420
+ mUNet	.274	2.765	.805	3.347	.617	3.879

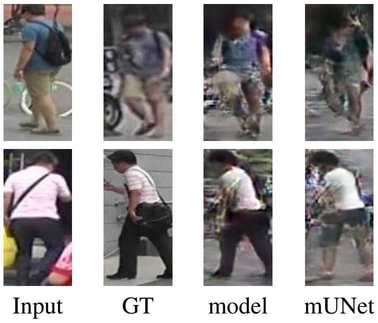

Fig. 7: Examples of challenging view synthesis cases for VDG (first row) and VDG$_{np}$ (second row). KEY – Input: input view; GT: ground-truth image; model: original model; mUNet: masked skip connection version of the model.

4. CONCLUSION

We proposed a module that learns to keep only the relevant information of a convolutional feature map in a generator network. The masked encoder feature map is passed to the corresponding layer of the decoder and both feature are combined via skip connections. As future work, we will incorporate priors in the learnable mask module.

5. REFERENCES

[1] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, California, United States, December 2017.

[2] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro, "Pose guided human image synthesis by view disentanglement and enhanced weighting loss," in *European Conference on Computer Vision Workshop (ECCVW)*, Munich, Germany, September 2018.

[3] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue, "Pose-normalized image generation for person re-identification," in

European Conference on Computer Vision (ECCV), Munich, Germany, September 2018.

[4] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe, "Deformable GANs for pose-based human image generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States, June 2018.

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer, "A Variational U-Net for Conditional Appearance and Shape Generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States, June 2018.

[6] Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas Valley, Nevada, United States, June 2016.

[7] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, Calgary, Canada, April 2014.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, October 2016.

[9] Alexey Dosovitskiy and Thomas Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain, December 2016.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, Munich, Germany, 2015, Springer International Publishing.

[11] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, December 2018.

[12] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus, "Deconvolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, California, United States, June 2010.

[13] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[14] Karim Ahmed and Lorenzo Torresani, "Maskconnect: Connectivity learning by gradient descent," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[15] Bo Zhao, Bo Chang, Zequn Jie, and Leonid Sigal, "Modular generative adversarial networks," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[16] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, United States, July 2017.

[17] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, Las Condes, Chile, December 2015.

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas Valley, Nevada, United States, June 2016.

[19] Martín Abadi and Ashish Agarwal et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," in *CoRR*, 2016.

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain, December 2016.

[21] Ma Liqian, Sun Qianru, Georgoulis Stamatios, Van Gool Luc, Schiele Bernt, and Fritz Mario, "Disentangled person image generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States, June 2018.