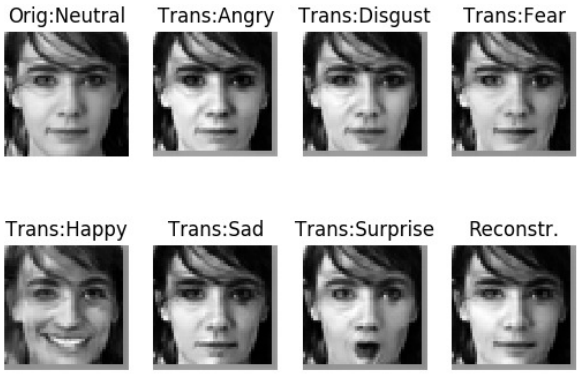
# Generating Realistic Facial Expressions through Conditional Cycle-Consistent Generative Adversarial Networks (CCycleGAN)

**Gino Tesei**
gino.tesei@gmail.com




## Abstract

Generative adversarial networks have been widely explored for generating photo-realistic images but their capabilities in multimodal image-to-image translations in a conditional generative model setting have been vaguely explored. Moreover, applying such capabilities of GANs in the context of facial expression generation conditioning on the emotion of facial expression and in absence of paired examples, to our knowledge, is almost a green field. Thus, the novelty of this study consists in experimenting the synthesis of conditional facial expressions and we present a novel approach (CCycleGAN) for learning to translate an image from a domain (e.g. the face images of a person) conditioned on a given emotion of facial expression (e.g. *joy*) to the same domain but conditioned on a different emotion of facial expression (e.g. *surprise*), in absence of paired examples. Our goal is to learn a mapping such that the distribution of generated images is indistinguishable from the distribution of real images using adversarial loss and cycle consistency loss. Qualitative results are presented, where paired training data does not exist, with a quantitative justification of optimal hyperparameters. The code for our model is available at `https://github.com/gtesei/ccyclegan`.

## 1 Introduction

The detection of human emotions has been long explored thanks to its applicability in various domains such as assisted living, health monitoring, real time crowd behavior tracking, and emotional security. Moreover, photo-realistic facial expression synthesis can be widely applied to face recognition, entertainment, virtual and augmented reality, computer graphics, data augmentation for emotion recognition, but such problem is much more challenging than the former, in part due to the scarcity

of large labeled paired datasets, i.e. where the same person is observed with different emotions of facial expressions.

## 2 Related Work

Generative modeling focus on observing data and learning a model to infer how this data was generated. Generative Adversarial Networks (GANs) [23, 3] achieved excellent results in image generation [19, 3], image editing [25] and representation learning [13, 16, 19]. Conditional GANs [14] extended GANs conditioning the generator or the discriminator on some extra information, e.g. class labels. Recent methods adopt the same idea for conditional image generation [11], translating visual concepts from characters to pixels [17], image inpainting [15], and future prediction [12].

### 2.1 Cycle Consistency

The idea to use transitivity to regularize structured data has been long used in different domain, e.g. in visual tracking enforcing forward-backward consistency [7, 20], in machine translation verifying and improving translations via back translation [22] and unsupervised machine translation [9, 1, 10], in monocular depth estimation supervising CNN training [2]. In the context of unpaired image-to-image translation, cycle consistent GANs proved to be effective in learning to translate an image from a source domain X to a target domain Y in the absence of paired examples [26, 27] but such domains never include facial expressions[1], and different facial expressions of the same person could be hardly modeled as different domains for unpaired image-to-image translation[2]. On the other hand, conditional difference adversarial autoencoder (CDAAE) [24] proved to be effective for facial expression synthesis but they were trained on paired datasets.

### 2.2 Facial Conditional Image-to-Image Translation

Recent methods of facial conditional image-to-image translation [11] and attribute editing on CelebA dataset [5] achieved impressive results, but they generate output images from input images with identical emotions of face expressions, while changing emotions of face expressions requires usually changing the shape of the face (e.g. *surprise* $\rightarrow$ *disgust*) along with a set of consistent changes of the facial expression not just related to color (e.g. *brown hair* $\rightarrow$ *blond hair*), texture (e.g. *old* $\rightarrow$ *not old*), the presence or absence of a given detail (e.g. *beard* $\rightarrow$ *no beard*).

## 3 Dataset

FER2013 [4], available online at Kaggle (accessed on 12 April 2019), consists of 28,709/7,178 train/test 48x48 pixel grayscale images of faces annotated with the emotion of facial expression as one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). It is a dataset of unpaired images and thanks to its resolution led to a good trade-off between accuracy and model complexity allowing us to iterate quickly many times. For data preprocessing we have normalized the input images from $[0, 255]$ to $[0, 1]$.

## 4 Methods

### 4.1 Conditional Cycle-Consistent GANs

Let's condition the GAN on the emotion of the facial expression, following the approach [14]. Our goal is to learn a mapping function between a domain $X$ and itself conditioned on the emotions of facial expressions $Y = \{0, 1, ..., k\}$. Hence, given an image $x_0 \in X$ annotated with $y_0 \in Y$ and given a desired emotion of facial expression $y_1$, we want to translate $x_0$ into $x_1$ having expression $y_1$, i.e. $x_1 = G(x_0|y_1)$, where G is the conditional mapping we want to learn or generator. Also, we split such generator into $G_{enc}$, the encoder responsible to encode a face image into its latent representation, i.e. $z_0 = G_{enc}(x_0)$, and $G_{dec}$, i.e. decoder responsible to perform the image-to-image translation given the desired facial expression label and the latent representation of the image, i.e.

[1] Examples of such domain pairs are *horse* $\rightarrow$ *zebra*, *winter Yosemite* $\rightarrow$ *summer Yosemite, apple* $\rightarrow$ *orange*, where shapes of objects are usually preserved.

[2] This is an experiment actually attempted in this project but there are several disadvantages in this approach. For example, if we have $n$ classes, the cardinality of the set of all possible image-to-image translations from a domain to another, i.e. $T_n = \{i \rightarrow j | 1 \leq i, j \leq n \wedge i \neq j\}$, becomes $|T_n| = n(n-1)$, that in our case (as $n = 7$) implies $|T_7| = 42$ distinct generators and 7 discriminators. This proliferation of generators and discriminators has the main downside of highly reducing parameters sharing (by a quadratic factor).

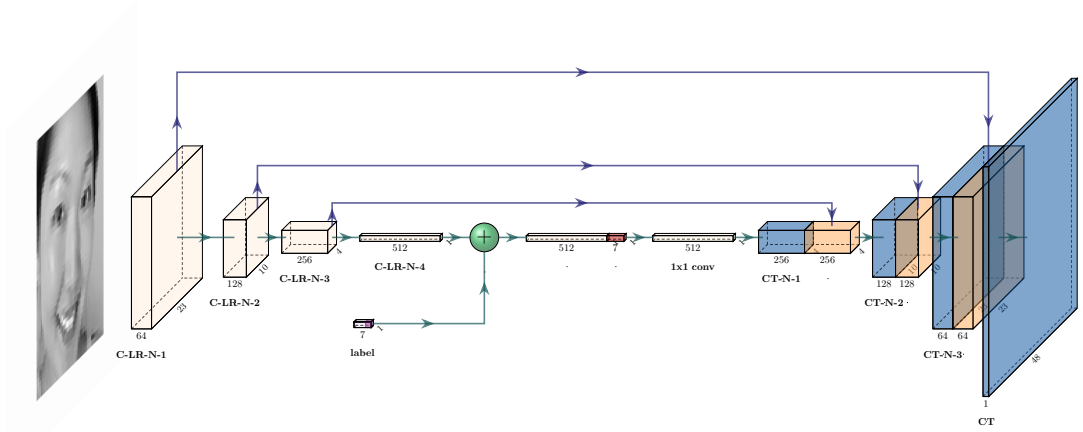



Figure 1: U-NET generator [18]. $G_{enc}$ has four convolutional layers, denoted as C-LR-N-#, composed by $4 \times 4$ unpadded convolutions (64/128/256/512 filters, strides 2), LeakyReLU and instance normalization [21], where the last $1 \times 1 \times 512$ feature map is the latent vector. $G_{dec}$ takes as inputs the latent vector and the one-hot label vector, reshaping the latter as a $1 \times 1 \times 7$ vector, concatenates it to the latent vector obtaining a $1 \times 1 \times 519$ vector and applying a $1 \times 1$ convolution to reduce the number channels to $512$, as it should be according to the U-NET scheme. Then three deconvolutional layers are applied, denoted as CT-N-#, composed by transposed convolution (256/128/64 filters, size 4, no stride, and no padding, except for the first one where it is used padding one), ReLU, instance normalization and concatenation with the correspondingly cropped feature map from the contracting path as described in [18].

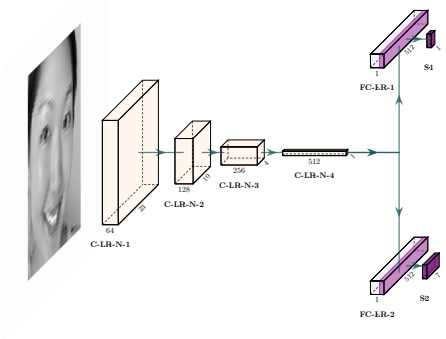
$x_1 = G_{dec}\left(z_0|y_1\right) = G_{dec}\left(G_{enc}\left(x\right)|y\right)$. During experimentation we found this split very beneficial compared to a monolithic design without encoder/decoder. In addition, we would like to introduce one adversarial discriminator $D(x|y)$ to distinguish true images conditioned on true facial expressions and translated images conditioned on desired facial expressions. During experimentation we found out this design, although conceptually correct, has the main disadvantage of back-propagating to the whole network both the error due to lack of realistic image generation and the error due to incorrect translation of facial expression, leading the network to learn one of these two sub-tasks but not both. Hence, we model the discriminator as two-tasks learning function, i.e. $D : X \rightarrow \{0,1\} \times Y$, denoting for conceptual convenience the first task as $D_{rf} : X \rightarrow \{0,1\}$ and as $D_{cl} : X \rightarrow Y$ the second one. To achieve parameters sharing, part of the network should be in common between these two tasks but the upper layers should be distinct. In section 4.2 the details of how to split the discriminator are reported. Regarding activations functions, the activation function of the last layer for $D_{rf}$ is sigmoid while for $D_{cl}$ could be either sigmoid or softmax. Although softmax may seem a good choice (and it was actually our first guess), during experimentation sigmoid proved to be a much better choice. Our explanation is that constraining the predicted probabilities of facial expressions to have sum one is conceptually incorrect. For example, although there are cases like *happy* vs. *sad* mutually exclusive, there are also fuzzier cases like *disgust* vs. *angry* not so clearly exclusive. Even for a human annotator for many of such cases it is not so clear whether the person is more disgusted than angry and, probably, the correct label should be both[3]. Applying adversarial losses, we can express the first objective term as a classical adversarial objective term, i.e. as

$$\mathcal{L}_{\mathcal{RF}}\left(G, D_{rf}, X, Y\right) = \mathbb{E}_{x \sim p(x)}\left[\log D_{rf}\left(x\right)\right] + \mathbb{E}_{x,y \sim p(x,y)}\left[\log\left(1 - D_{rf}\left(G\left(x|y\right)\right)\right)\right] \quad (1)$$

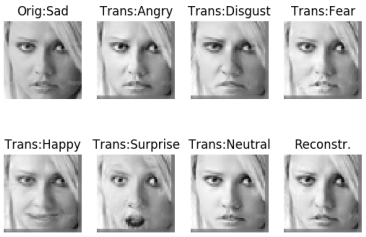
where $G$ tries to generate images $G_{dec}\left(G_{enc}\left(x\right)|y\right)$ that look similar to images from domain $X$, while $D_{rf}$ aims to distinguish between generated samples and real samples form domain $X$. $G_{enc}$ and $G_{dec}$ aim to minimize this objective against an adversary $D_{rf}$ that tries to maximize it, i.e., $\min_{G_{dec},G_{enc}} \max_{D_{cl}} \mathcal{L}_{\mathcal{RF}}\left(G, D_{rf}, X, Y\right)$. In the same way, we can express the second term

$$\mathcal{L}_{\mathcal{CL}}\left(G, D_{cl}, X, Y\right) = -\mathbb{E}_{x,y \sim p(x,y)}\left[l_{cl}\left(x, y\right)\right] - \mathbb{E}_{x \sim p(x), y \sim p(y)}\left[l_{cl}\left(G\left(x|y\right), y\right)\right], where \quad (2)$$

[3]FER2013 is not a multi-label dataset but probably it should have been. In real world people having complex cognitive states are usually holding different and maybe conflicting emotions at the same time.




(a) For the multi-task discriminator we use four convolutional layers, composed by $4 \times 4$ unpadded convolutions (64/128/256/512 filters, strides 2), LeakyReLU and instance normalization (except for the first layer). This is the part of the network shared by the two tasks. Then, for $D_{rf}$ and $D_{cl}$ we have two distinct fully connected blocks followed by LeakyReLU and sigmoid activations.




(b) Results with $\lambda_{cyc} = \lambda_{cl} = 1$, Adam solver, learning rate = 0.0002, $\beta_1 = 0.5, \beta_2 = 0.999$. On the top-left corner the original image with emotion of facial expression ("Orig:Sad"), next on the right the first translation ("Trans:Angry"), and so on. On the bottom left corner, the reconstructed image. Additional results are available at `https://github.com/gtesei/ccyclegan`

Figure 2: Multi-task discriminator architecture (a) and results with optimal hyperparameters (b).

$$l_{cf}(x, y) = \sum_{0 \leq i \leq k} -\mathbf{1}_{\{y_i = i\}} \log\left(D_{cl}(x)_i\right)$$

where $k+1$ is the number of class labels, $D_{cl}(x)_i$ is the i-th element of the output vector $D_{cl}(x)$. Hence, $G$ tries to generate images $G_{dec}(G_{enc}(x)|y)$ that look similar to images from domain $X$ with the desired facial expression $y \in Y$, while $D_{cl}$ aims to classify images with the correct facial expression label $y \in Y$. $G_{enc}$ and $G_{dec}$ aim to minimize this objective against an adversary $D_{cl}$ that tries to maximize it, i.e., $\min_{G_{dec}, G_{enc}} \max_{D_{cl}} \mathcal{L}_{\mathcal{CL}}(G, D_{cl}, X, Y)$. Also, following the approach [26], to further reduce the space of possible mapping functions, the learned mapping functions should be *cycle-consistent*, i.e. the image translation cycle should be able to bring $x$ back to the original image, i.e. $G_{dec}(G_{enc}(x_0)|y_0) \approx x_0$. This behavior can be incentivated by using a cycle consistency loss term:

$$\mathcal{L}_{cyc}(G, X, Y) = \mathbb{E}_{x,y \sim p(x,y)} \left\| G_{dec}(G_{enc}(x)|y) - x \right\|_1 \tag{3}$$

Hence, our full objective is:

$$\mathcal{L}(G, D, X, Y; \lambda_{cyc}, \lambda_{cl}) = \mathcal{L}_{\mathcal{RF}}(G, D_{rf}, X, Y) + \lambda_{cl}\mathcal{L}_{\mathcal{CL}}(G, D_{cl}, X, Y) + \lambda_{cyc}\mathcal{L}_{cyc}(G, X, Y) \tag{4}$$

where $\lambda_{cyc}, \lambda_{cl}$ controls the relative importance of the three objective terms. Hence, the general optimization problem can be formulated as:

$$G^*_{enc}, G^*_{dec} = \arg \min_{G_{enc}, G_{dec}} \max_{D_{rf}, D_{cl}} \mathcal{L}(G, D, X, Y; \lambda_{cyc}, \lambda_{cl}). \tag{5}$$

### 4.2 Network Architecture

We adopt U-NET generator [18] where for $G_{enc}$ encodes face images into latent vectors through four convolutional layers and $G_{dec}$ takes as inputs $1 \times 1 \times 512$ latent vectors and one-hot label vectors to decode them through three deconvolutional layers. Further details on fig. 1. For the multi-task discriminator we use four convolutional layers followed by fully connected layers and sigmoid activations both for $D_{rf}$ and $D_{cl}$. Further details on fig. 2a.

### 4.3 Training Details

For each train image $x_0 \in X$ annotated with facial expression $y_0 \in Y$, we extract the latent vector $z_0 = G_{enc}(x_0)$, and then generate all the possible 6 face images conditioning on the remaining 6 class labels, i.e. $G_0 = \{G_{dec}(z_0, y') | y' \in Y \wedge y' \neq y_0\}$ to train the discriminator and the generator. Notice that this procedure is different from [14] as we don't use any prior noise distribution. We use

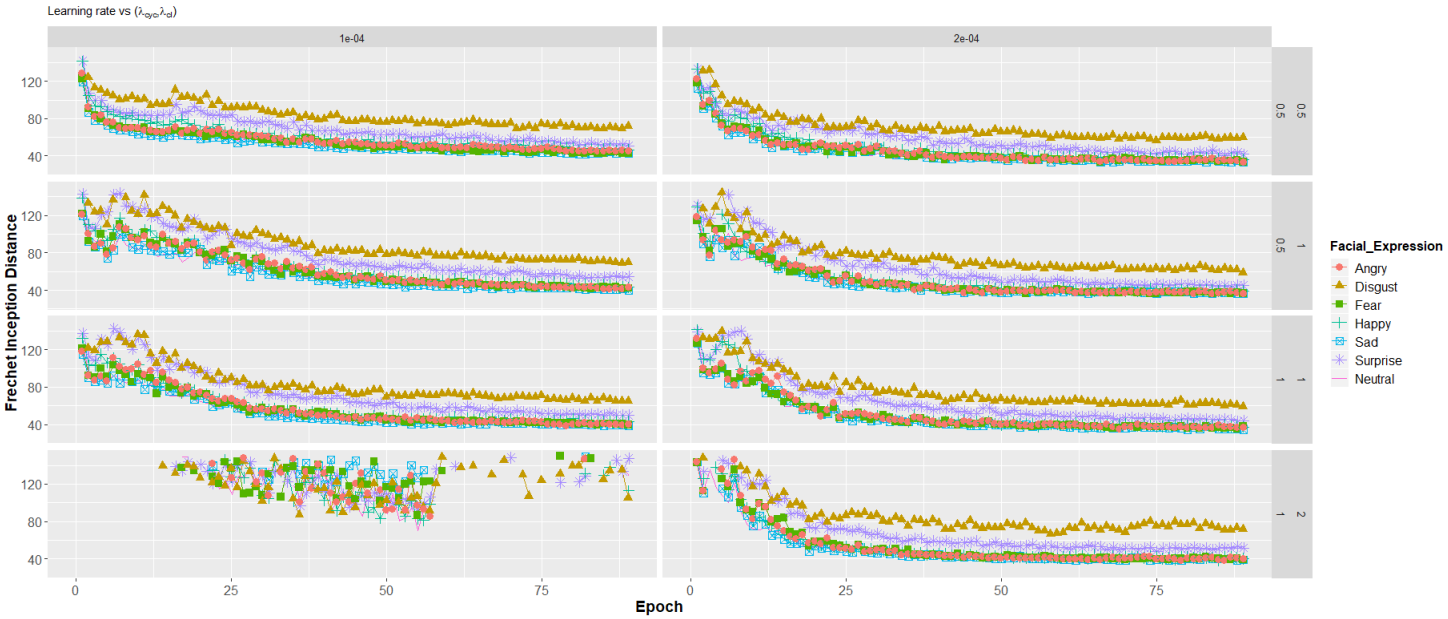



Figure 3: Frechet Inception Distance (FID) [6] is used to to find the best mix of learning rate (top x-axis), $\lambda_{cyc}, \lambda_{cl}$ (right y-axis) as measure of dissimilarity between the distributions of generated face images and real face images. Values have been smoothed adopting exponential smoothing average with bias correction ($\beta = 0.4$). We can see that *disgust* is the most difficult emotion to synthesis in all configurations (not surprisingly, it has the lowest class frequency of 1.5%). The three best configurations are the ones with lowest FID for a given epoch and class label, i.e. learning rate = 0.0002 (confirming [26]), $\lambda_{clyc} = 1 \wedge \lambda_{cl} = 1$ or $\lambda_{clyc} = 0.5 \wedge \lambda_{cl} = 1$ or $\lambda_{clyc} = 0.5 \wedge \lambda_{cl} = 0.5$. These configurations are confirmed by qualitative evaluation of generated facial images (see fig. 2b ).

mirroring as augmentation method and random shuffle of train data is applied both to generator and discriminator. Also, we use Adam solver [8] with batch size of 64 and to find the best mix of learning rate, $\lambda_{cyc}, \lambda_{cl}$ we use grid search on a restricted but convenient set of candidates as better explained in section 5.

## 5 Experiments

The model described by eq. 1, 2, 3, 4, 5 is the 26th successful tentative after 24 failures (see project repository for full list of experiments). We found out that monitoring the three losses of eq. 1, 2, 3 is not enough. For example, a typical problem we encountered is that to minimize the general optimization objective, the network could learn the identity transformation optimizing the losses $\mathcal{L}_{\mathcal{RF}}$ (eq. 1) and $\mathcal{L}_{cyc}$ (eq. 3) and sacrificing the loss $\mathcal{L}_{\mathcal{CL}}$ (eq. 2), reaching a better overall equilibrium than trying to do its job, i.e. performing the translation. To prevent this behavior, instead of increasing $\lambda_{cl}$, splitting the generator and the discriminator (section 4.1) and adopting the training procedure described in section 4.3 are key points. Following [26], we use Adam solver with $\beta_1 = 0.5, \beta_2 = 0.999$. For remaining hyperparameters, we use Frechet Inception Distance [6], applying an Inception-v3 network pretrained on ImageNet (converting grayscale images to RGB) to real and generated samples, to find optimal values of learning rate, $\lambda_{cyc}, \lambda_{cl}$ as better explained in fig. 3. The optimal hyperparameters found in this way are confirmed by qualitative evaluation of generated facial images (fig. 2b), with $\approx 150$ epochs.

## 6 Conclusion

We introduce CCycleGAN, a novel approach for the synthesis of realistic face images conditioning on the emotion of facial expression and in absence of paired examples. Qualitative results are presented and a quantitative justification is provided for optimal hyperparameters.

## References


[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. 2019. 2.1

[2] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. 2016. 2.1

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[4] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing*, pages 117–124. Springer Berlin Heidelberg, 2013. 3

[5] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. 2017. 2.2

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017. 3, 5

[7] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *Proc. 20th Int. Conf. Pattern Recognition*, pages 2756–2759, August 2010. 2.1

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. 4.3

[9] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. 2017. 2.1

[10] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. 2018. 2.1

[11] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. 2018. 2, 2.2

[12] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. 2015. 2

[13] Michael Mathieu, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representations using adversarial training. 2016. 2

[14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. 2014. 2, 4.1, 4.3

[15] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. 2016. 2

[16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015. 2

[17] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. 2016. 2

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015. 1, 4.2

[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. 2016. 2

[20] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science. Springer, Sept. 2010. 2.1

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. 2016. 1

[22] Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. 2016. 2.1

[23] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. 2016. 2

[24] Yuqian Zhou and Bertram Emil Shi. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. 2017. 2.1

[25] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. 2016. 2

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017. 2.1, 4.1, 3, 5

[27] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. 2017. 2.1