BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

MASTER'S THESIS

Brno, 2021 Bc. Adam Svorad

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF TELECOMMUNICATIONS
ÚSTAV TELEKOMUNIKACÍ

INCREASING QUALITY OF FACIAL IMAGES USING SEQUENCE OF IMAGES
ZVÝŠENÍ KVALITY V OBRAZU OBLIČEJE S POUŽITÍM SEKVENCE SNÍMKŮ

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR — **Bc. Adam Svorad**
AUTOR PRÁCE

SUPERVISOR — **doc. Ing. Radim Burget, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2021

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Master's Thesis

Master's study program **Communications and Informatics**

Department of Telecommunications

Student: Bc. Adam Svorad ***ID:*** 186195

Year of study: 2 ***Academic year:*** 2020/21

TITLE OF THESIS:

Increasing quality of facial images using sequence of images

RECOMMENDED LITERATURE:

[1] Joshi, Prateek. Artificial intelligence with python. Packt Publishing Ltd, 2017.

[2] Gowrishankar, S., and A. Veena. Introduction to Python Programming. CRC Press, 2018.

Date of project specification: 1.2.2021 ***Deadline for submission:*** 24.5.2021

Supervisor: doc. Ing. Radim Burget, Ph.D.

prof. Ing. Jiří Mišurec, CSc.
Chair of study program board

ABSTRACT

Master's thesis delves into the field of face super-resolution. It aims to review novel approaches to single-frame image sharpening and image editing in the theoretical part of the work. Practical part will focus on approaches to image reconstruction from a sequence of damaged images. Multiple multi-frame neural network models will be implemented and evaluated. As alternative option, a suite of image editing tools will be presented as well. These tools will utilize most modern image editing techniques to merge visual features of faces from multiple input images into a single output image. At the end of the thesis, all methods will be compared to each other.

KEYWORDS

convolutional neural network, face super-resolution, multi-frame super-resolution, single-frame sharpening, U-Net, GAN, StyleGAN

SVORAD, Adam. *Increasing Quality of Facial Images Using Sequence of Images*. Brno, 2021, 92 p. Master's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Telecommunications. Advised by doc. Ing. Radim Burget, Ph.D.

ABSTRAKT

Diplomová práca sa zameriava na oblasť zaostrovania obrázkov tvárí. V teoretickej časti práce budú prezentované moderné metódy zaostrovania obrázkov pomocou jediného obrázku a metódy editácie obrázkov. Praktická časť sa zameria na prístupy rekonštrukcie obrázkov zo sekvencie poškodených obrázkov. Viaceré modely neurónových sietí so vstupom pre viacero obrázkov budú zhotovené a vyhodnotené. Alternatívny prístup v podobe balíka nástrojov na editáciu obrázkov bude taktiež predstavený. Tieto nástroje budú využívať najmodernejšie prístupy k editácii obrázkov s cieľom spojiť vizuálne prvky tvárí zo vstupnej sekvencie obrázkov do jedného finálneho výstupu. V závere práce budú všetky metódy navzájom porovnané.

KĽÚČOVÉ SLOVÁ

konvolučné neurónové siete, superrozlíšenie tváre, multi-frame superrozlíšenie, single-frame zaostrenie, U-Net, GAN, StyleGAN

ROZŠÍRENÝ ABSTRAKT

Napriek súčasnej technicky pokročilej dobe mnohé bezpečnostné kamery stále zhotovujú snímky a video zábery v nízkom rozlíšení, a teda aj nízkej kvalite. Je to spôsobené buď obmedzenou veľkosťou lokálneho úložiska, alebo využitím zastaralej technológie pri výrobe bezpečnostnej kamery. V blízkej budúcnosti sa nevyrieši ani jeden z problémov, pretože by to vyžadovalo vynaloženie značných finančných prostriedkov. Avšak v prípade dopravnej nehody, lúpeže, vraždy alebo inej protizákonnej aktivity, záznamy z bezpečnostných kamier zhotovené v prinízkom rozlíšení neposlúžia v pátraní. Existujú prípady, keď sa páchatelia pozerajú priamo do šošovky kamery, avšak polícia ich nie je schopná identifikovať.

Vďaka nárastu výpočtovej sily sa dosiahlo v posledných dekádach mnoho významných pokrokov vo viacerých vedných oboroch. Prístupy založené na strojovom učení sa využívajú v takmer všetkých oblastiach výskumu a ich popularita neprestajne narastá. Oblasť spracovania obrazu nie je žiadnou výnimkou a techniky využívajúce neurónové siete sú predmetom stáleho výskumu. Avšak, žiadne významné objavy sa k dnešnému dňu neuskutočnili v oblasti zaostrovania obrazu. Existuje viacero prístupov, ktoré ešte neboli preskúmané, no napriek tomu, len malá skupina vedcov sa venuje tejto oblasti. Táto diplomová práca vníma potenciál nepreskúmaných prístupov a dáva si za ciel ich preskúmať a rozšíriť.

Modely neurónových sietí využívané pri zaostrovaní obrázkov napr. parkov, budov a iných obecných scén môžu vniesť do spracovaného obrázku istú úroveň kreativity. Rozmazané časti obrázku môžu byť plne prekreslené ‘z pamäte’ neurónovej siete, ktorá sa vytvorila pri jej trénovaní. Avšak, pri obrázkoch tvárí sa tento mechanizmus stáva dvojšečnou zbraňou. Na jednej strane sa rozmazané oblasti tváre môžu dokonale prekryť zaostrenými záplatami, na druhej strane môže dôjsť k zmene identity tváre. Ľudské oko je veľmi citlivé na detaily ľudskej tváre, a preto aj drobný nesprávny zásah pri rekonštrukcii obrazu môže viesť k významnej zmene identity a znehodnotiť celý výstup. Z toho dôvodu sa táto práca nezameriava na vyhotovenie maximálne ostrého obrázka ľudskej tváre z rozmazaného obrázka. Zameriava sa na rekonštrukciu rozmazaného obrázka s cieľom zachovania pôvodnej identity.

Viacero príbuzných prác ukázalo, že z jedného obrázka tváre sa len sťažka zhotoví kvalitnejšia alternatíva so zachovaním identity. Nie je totiž možné vygenerovať informáciu navyše z ničoho. Preto sa táto práca zameriava na rekonštrukciu tváre zo sekvencie obrázkov. Predpoklad je taký, že sekvencia obrázkov je vyhotovená jednou bezpečnostnou kamerou. Nie je to však nutná podmienka. Rozostup medzi obrázkami sú stovky milisekúnd, a teda každý obrázok zachytáva inú časť tváre. Navyše, jednotlivé obrázky zobrazujú tie isté časti tváre rôznou kvalitou. V tejto diplomovej práci nie je preto nutné generovať informáciu navyše z ničoho, aby sa zaostrila tvár človeka. Stačí, ak sa vyextrahuje informácia zo všetkých vstupných obrázkov a využije sa pri generovaní jedného výstupného ostrého obrázka.

V teoretickom úvode práca preskúmava najmodernejšie architektúry neurónových sietí v oblasti spracovania obrazu, konkrétne GAN a U-Net. Zároveň približuje projekt StyleGAN, ktorý umožňuje generovať plne syntetický obrázok tváre zo vstupného vektora, teda aj sekvencie náhodných čísel. Praktická časť sa zameriava na implementáciu klasickej U-Net architektúry, novo-predstavenej BiO-Net architektúry a navrhuje novú sieť Feature-Merge U-Net.

Klasická U-Net architektúra typicky pracuje len s jedným vstupným obrázkom. Vstupný obrázok sa dekóduje a opätovne zakóduje spolu s rysmi, ktoré sa sieť naučila počas trénovania. Tým, že táto práca má ako vstup sekvenciu obrázkov, všetky obrázky sú spojené do jedného bloku a posunuté priamo do modifikovanej U-Net architektúry. Idea je taká, že U-Net dekóduje rysy tváre z každého obrázku a počas kódovania ich automaticky spojí. Predbežné testovania však ukázalo, že U-Net nepracuje v tomto duchu, a preto bola navrhnutá nová modifikácia U-Net architektúry nazvaná Feature-Merge U-Net. Daná sieť spracováva vstupné obrázky individuálne v dekódovacej vetve. Keď sú rysy tváre vyextrahované, manuálne sa spoja a nasledujú do kódovacej vetvy.

Architektúra BiO-Net pracuje rovnako ako U-Net s tým rozdielom, že tie isté vstupy sú iteratívne spracovávané a spájané s predošlými výstupmi. Teoreticky by mala dávať lepšie výsledky oproti U-Net na úkor dlhšej doby spracovania a trénovania.

Veľkým nedostatkom neurónových sietí je, že sa nedajú natrénovať na príliš komplexnú úlohu, ak kritik nie je dosť sofistikovaný počas trénovania. Z toho dôvodu táto práca tiež skúma úplne iný prístup. Namiesto podsúvania všetkých obrázkov do neurónovej siete naraz sa komplexná úloha neurónovej siete rozpadne na viacero elementárnych problémov. Každý tento problém sa vyrieši individuálnou sieťou alebo nástrojom. V diplomovej práci bol tento prístup nazvaný ako rekonštrukcia obrázkov pomocou súboru nástrojov. Pri rekonštrukcii prostredného obrázku z viacerých vstupných obrázkov je treba nájsť obrázok s najväčšou kvalitou očí, nosu, úst a pod.. V ďalšom kroku sa zdrojový obrázok zarovná do pozície cieľového obrázka. Následne sa prekopíruje príslušná časť zdrojového obrázka vo vyššej kvalite do cieľového obrázka v nižšej kvalite. Takto sa postupuje, kým sa neprenesú všetky rysy tváre vo vyššej kvalite do výsledného obrázka. Na záver sa spustí jemné zaostrenie pomocou U-Net architektúry. Takto navrhnutý systém vykonáva zhodné kroky explicitne ako by mali vykonávať U-Net architektúry implicitne.

Záver práce porovnáva všetky testované prístupy pomocou objektívnych a subjektívnych metrík. Diplomová práca spochybňuje výpovednú hodnotu objektívnych metrík v oblasti zaostrovania obrázkov. Výsledky objektívnych a subjektívnych meraní sa totiž nezhodujú. K rovnakému záveru prišli aj mnohé iné práce v tomto obore.

Všetky navrhnuté a naimplementované U-Net architektúry pracujú rovnako slabo. Aj keď na svoj vstup dostanú viacero obrázkov, nie sú schopné vstupnú informáciu zúžitkovať. Práca ukazuje, že modely majú tendenciu ignorovať ostatné vstupy okrem prostredného. Možné vysvetlenie je také, že kritik nie je schopný natrénovať sieť tak, aby extrahovala a spájala rysy tvárí. Jednoduchšia cesta je totiž vziať prostredný obrázok a zaostriť ho pomocou naučených znalostí. Ďalším problémom je pretrénovanie. Siete majú veľkú tendenciu pamätať si tváre, ktoré videli a násilne tieto rysy vkladajú do nových rozmazaných vstupných obrázkov. Istá miera dokresľovania je akceptovateľná, avšak spomenuté architektúry len dokresľujú črty, ktoré danej tvári nepatria a menia jej identitu. Tento problém sa dá čiastočne riešiť väčšou trénovacou množinou. To je však len záplata na skutočný problém – nedostatočný kritik.

Ďaleko viac sa osvedčil rekonštrukčný systém, ktorý zaostruje obrázky pomocou viacerých elementárnych explicitných krokov. Veľkou jeho prednosťou je možnosť vrátiť sa ku ktorémukoľvek bloku a prerobiť ho pomocou úplne inej technológie alebo zvoliť úplne iný prístup. Navyše, ľahko sa testujú a opravujú bloky, ktoré majú plniť len jednu jednoduchú úlohu. Najväčšou limitáciou rekonštrukčného systému sú obrázky, s ktorými je schopný pracovať. Tento systém dokáže pracovať len s obrázkami tvárí vyhotovenými spredu. Navyše tieto tváre musia vyjadrovať rovnakú emóciu a byť približne rovnako osvetlené. Všetky opísané obmedzenia spočívajú v bloku, ktorý zarovnáva tváre do jednej polohy tak, aby sa dali črty priamo prekopírovať.

Najväčšiu perspektívu do budúcna má práve spomenutý rekonštrukčný systém. Jeho veľkou výhodou je, že sa dá rozložiť ma menšie úlohy, na ktorých sa dá pracovať paralelne. Navyše vstupy a výstupy každého bloku sa dajú exaktne popísať, na rozdiel od U-Net modelov, kde sa požaduje zaostrený vstupný obrázok. Veľký prínos do rekonštrukčného systému by predstavoval lepší blok zarovnania tvárí. Schopnosť presúvať črty tváre medzi obrázkami z rôznych uhlov pohľadu je nesmierne dôležitá. Aktuálne je táto úloha príliš zložitá, ale časom, s príchodom nových nástrojov založených na StyleGAN projekte sa výrazne zjednoduší. Ďalší veľký prínos nesie návrh systému schopného posúdiť kvalitu obrázka tváre z pohľadu človeka. Takýto blok by mal aj obecné využitie.

DECLARATION

I declare that I have written the Master's Thesis titled "Increasing Quality of Facial Images Using Sequence of Images" independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author I furthermore declare that, with respect to the creation of this Master's Thesis, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno ....

author's signature

ACKNOWLEDGEMENT

I would like to thank the advisor of my thesis, doc. Ing.Radim Burget Ph.D., for his valuable comments, inspiration and guidance.

Tato práce vznikla jako součást klíčové aktivity KA6 - Individuální výuka a zapojení studentů bakalářských a magisterských studijních programů do výzkumu v rámci projektu OP VVV Vytvoření double-degree doktorského studijního programu Elektronika a informační technologie a vytvoření doktorského studijního programu Informační bezpečnost, reg. č. CZ.02.2.69/0.0/0.0/16_018/0002575.

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Projekt je spolufinancován Evropskou unií.

Contents

List of Figures

List of Tables

Introduction

The increase in a computational power during the past few decades has paved the way for significant advances in many scientific fields. Machine learning based approaches started being utilized in almost all the domains of research and their popularity is persistently on rise. The field of image processing is no different and the techniques relying on neural networks are constantly being explored. Though, no significant progress has been made in the image sharpening field and numerous approaches have not been tested yet. There is still a lot of room for research, but few researchers focus on the development despite its practical application.

Despite current technologically progressed era, Closed-circuit Television cameras still record video sequences in a low resolution. It is caused either by limited memory storage (local or remote) or by use of an obsolete technology. Nevertheless, none of the two will be solved within short period of time since it would require an investment of considerably high resources. Though, in case of an accident, burglary, murder or any other illegal activity, recordings in such low resolution are of little or no use. There are even cases, when suspects look directly into camera lens, but the police are still unable to track down the identity.

Generally, models used for sharpening for example images of forests, buildings or other generic scenes can introduce certain level of creativity. Some areas of the images can be fully inpainted based on the knowledge the models have learnt during training. Though, this does not apply to images of faces. Human eye is sensitive to the shape of human faces. Even the slightest incorrect feature added to such an image can significantly change its meaning. That is why the primary focus will be on restoring the images of faces without letting the model become too creative and damage it. Instead, the model ought to use all possible input information to sharpen the image that fully complies with the identity in an input sequence of damaged images.

As of now, there is no working approach, which would provide a means of extracting more information from low resolution images in a video sequence and reconstruct them into a single high resolution image. Some related works been published in a domain of SF super-resolution and generic MF video reconstruction. Though, none of them have truly focused on human face sharpening and its reconstruction from multiple images. Moreover, these related works have not presented any groundbreaking approach so far and their results are poor or moderate. That is why this thesis will focus on a research in the described field and try to seek and implement a working solution.

Generally, the input for the system will be a video recording, which is just an image sequence. Each image will by capturing different parts of the face. It can

also be presumed, that each image will contain certain parts of the face relatively sharper than the other images. The goal of the thesis will be to develop a system able to reconstruct a face image from multiple similar damaged images.

The thesis is structured into a theoretical introduction, where related works in this field will be explored and used as inspiration for the further development. The second chapter describes the experiments carried out along with detailed dataset and models' description. Finally, in the third chapter, the achievements will be presented and the experiments will be evaluated.

1 Super-resolution

Super-resolution imaging (SR) is a set of techniques which enhances the resolution of an image. They can be divided into single-frame (SF) and multi-frame (MF) variants. while SF approaches attempt to increase the resolution of the image without producing blur using just a single source image, MF methods are based on sub-pixel shifts between multiple source images in a low resolution. The improved high-resolution image is thus created after fusing information from all source images.

1.1 Interpolation-based Approaches

An interpolation belongs to estimation methods in numerical analysis. It is a method of creating new data points lying within a range of known data points. In the field of image processing, interpolation is used to resample images to a higher resolution. Nearest neighbor, bilinear and bicubic interpolation methods are applied in practice the most [12].

Nearest-neighbor interpolation also called proximal interpolation is a multivariate interpolation in single or multiple dimensions. When a non-given point in some space is required, nearest-neighbor method picks value of the nearest point. It ignores values of all other given points in the space resulting in a piecewise-constant interpolant. This interpolation method is mostly used in real-time 3D rendering thanks to its simplicity and fast speed of operation [12].

Bilinear interpolation is based on linear interpolation on 2D grid. Interpolation is first performed in one direction and then again in the other. It is simple and belongs to the fastest resampling techniques in image processing. Bicubic interpolation interpolates data points on a two-dimensional regular grid. It can be computed using either cubic splines, cubic convolution or Lagrange polynomials. The interpolated area is much smoother and contains less interpolation artifacts compared to nearest-neighbor or bilinear interpolation. That is why it is preferred in image processing when speed is not a priority [12].

Generally speaking, interpolation methods compared to machine learning approaches in Super-resolution tasks suffer from smoothing and image information loss [13].

1.2 Machine-learning Approaches

Artificial Neural Networks (ANN) are computing systems whose architecture was loosely inspired by the biological neural networks which can be found in animal brains. ANN is a collection of connected computation units called neurons, which

vaguely model the neural networks in animal brains. AANs learn by processing examples. Every example consists of an input and known result. During training, the ANN produces a prediction by forwarding an input through its architecture. The difference between the desired result and the prediction is then evaluated as an error. In order to minimize the error value, network updates its parameters in a process called backpropagation. The training is carried out in multiple iterations and the successive adjustments are performed until the predictions are similar to desired results [13].

1.2.1 Generative Adversarial Networks

The original idea behind Generative Adversarial Networks (GANs) was to use generator to replicate real-world content from noise. The critic then stated how much the generated content resembled the real one. Though, as the generator was getting better, the critic started improving as well. Hence, this competition made both the networks improve until generator produced such good results that critic could not distinguish what was real and what was generated [14].

Formula 1.1 describes the whole process more formally. A minmax game is played between a generative network G and a discriminative (i.e. critic) network D. Noise sample $z \sim p(z)$ following normal or uniform distribution represents the input and $G(z)$ stands for generator's data output whose distribution p_g is expected to match distribution of the ground-truth data p_{data}. In the mean time, critic network D learns to recognize the real data sample $x \sim p_{data}(x)$ and generated data sample $G(z) \sim p_g(G(z))$.

$$\min_G \max_D E_{x \sim p_{data}}(log(D(x))) + E_{z \sim p_z}(log(1 - D(G(z)))) \tag{1.1}$$

In 2017 it was presented a famous modification of GANs, so called Wasserstein GANs [15] (WGANs). WGANs aim to solve stability problems of GANs and interpretability problems of their loss function during training. In essence, GANs are trying to learn the distribution of a real-world data by minimizing the difference in probability distribution. And that is done by generating adversarial data. The convergence can be interpreted as minimizing Jensen-Shannon divergence (JS) [15].

In [15] author shows the shortcomings of JS divergence when the two probability distributions being compared do not overlap and proposes to use Wasserstein distance instead [15].

In the field of image generation researches introduced Super-Resolution GANs (SRGANs), [16]. The motivation behind SRGANs is to restore finer textures from the picture when it is being upscaled. Their architecture relies on residual network

[17] instead of deep convolutional networks because residual networks can be considerably deeper and produce better results. Inside SRGANs there can be found 16 residual blocks and each block is built from 2 convolutional layers and a single skip connection. The output of residual block is passed through batch normalization and ReLU layer [16].

SRGANs generally produce outputs with unpleasant artifacts. In order to enhance the visual quality a group of researchers studied their architecture and a loss function. They came up with Enhanced Super-Resolution GANs (ESRGANs) and introduced them in the paper [18]. More specifically, they introduced so called Residual-in-Residual Dense Block without a batch normalizatiton as a replacement for a simple residual block. Moreover, they copied the idea from [19] to make the critic predict relative realness and not the absolute value. Finally, they also enhanced the feature loss by reading features before the activation which provides much stronger supervision for texture recovery and brightness consistency [18].

1.2.2 U-Nets

U-Nets are convolutional neural networks (CNNs) originally developed for biomedical image segmentation in 2015. Their architecture and behavior are precisely documented in [20]. U-Nets are the greatest rival of GAN architectures in the field of image sharpening or, more generally speaking, image upscaling.

U-Net network is always built from two parts – an encoder and a decoder. Encoder usually accepts high resolution input image with few channels and passes it through multiple successive layers. These layers mostly include convolution, which increases the number of channels and pooling operations, which decrease the resolution. The activations at the output of encoder contain features of an input image [20].

Once features are extracted, encoder is directly followed by a decoder. Decoder also uses convolutional layers, but, this time, to decrease number of channels. Pooling operations are replaced with upsampling operators, which increase the resolution [20].

An important feature of U-Nets is the use of skip connections tied directly from encoder to decoder. Features extracted from encoder are either merged or concatenated with features in corresponding decoder layer depending on an implementation. This allows the network to propagate encoder's context information to decoder's restoration layers. As a consequence, the resulting model is more or less symmetric and yields a U-shaped architecture [20].

Despite U-Nets' age, they still represent state-of-the-art deep learning based approach in various computer vision tasks such as segmentation, sharpening, image

denoising and inpainting. Only few better variants of this amazing architecture have been published in recent years. One of those is V-Net [21]. V-Nets aim to extract low-level features from the data as well as reduce the resolution at the end of each block. Their encoder consists of multiple stages that operate at different resolution. Every stage contains at least one convolutional layer using volumetric kernels of size 5x5x5 voxels. As the input data passes through the network, the resolution is reduced. This approach gives similar results as pooling layers. But its advantage is, that it leaves smaller memory footprint while training. Moreover, convolutions with an appropriate stride tend to reduce the size of the data[21].

Bi-directional O-shape network (BiO-Net) [1] represents another well designed variant of U-Nets published in July 2020. BiO-Net is a novel approach because unlike other variants, it does not increase its complexity and yet, it significantly outperforms other state-of-the-art methods [1].

BiO-Net reuses U-Net's building blocks in recurrent manner without adding extra parameters. It introduces backward skip connections, which pass decoder's features back to an encoder to further improve network's capabilities. Architecture is shown in Figure 1.1.

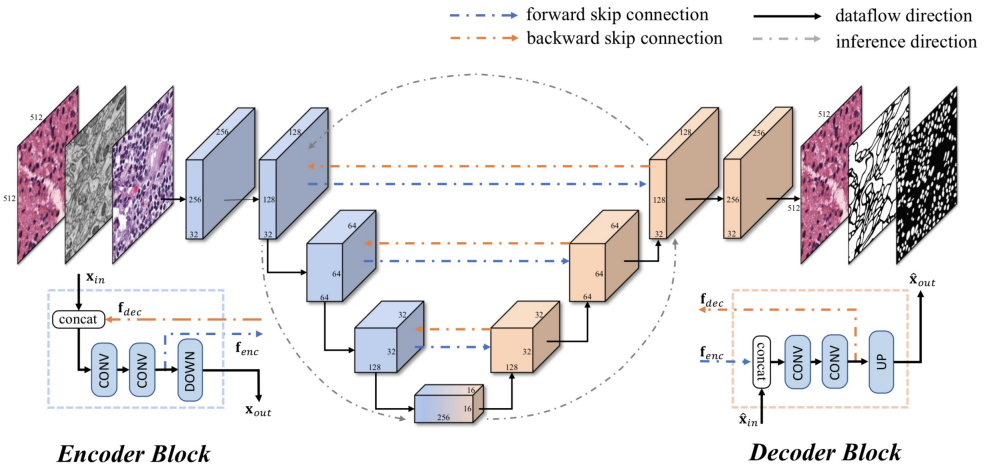

Fig. 1.1: BiO-Net architecture. Source [1].

The uniqueness of BiO-Net architecture is use of bi-directional skip connections which allow the decoder to evaluate the semantic features in the encoder and vice versa. As it can be seen from the Figure 1.1, forward skip connections link encoder and decoder at the same level and maintain the encoded low-level visual features. Backward skip connections pass high-level semantic features back to encoder which fuses them with its inputs. This way, high-level semantic features and low-level visual features are flexibly aggregated [1].

During- the first iteration BiO-Net behaves like a regular U-Net. Inputs are passed through encoder, which extracts the visual features. Since decoder does not contain any activations yet, encoded features remain intact. Features are then decoded into high-resolution outputs in the decoder. In the following iterations, the same inputs are provided to the encoder. Though, this time, decoder's activations from the previous cycle are merged with newly encoded visual features. Authors of BiO-Net claim, the more iterations the network performs, the better results are achieved. The only drawback is increasing computational time [1].

Recently, a great progress in medical image segmentation has been made thanks to introduction of UX-Nets [2]. UX-Nets propose a novel neural architecture search method (NAS) for image segmentation. UX-Nets search scale-wise feature aggregation strategies and also block-wise operators in given encoder-decoder network. Authors claim that UX-Nets greatly enhance the flexibility of a classical U-Net architecture which just aggregates features of encoder and decoder in an equivalent resolution. Moreover, relaxation of UX-Nets is thoroughly designed and enables its searching scheme to perform in efficient manner. UX-Nets with their novel approach to search of multi-level feature aggregation define current state-of-the-art method [2].

Multiple studies demonstrating the aggregation of multi-level features have been carried out so far. Even intuitively, merging low-level visual features and high-level semantic features extracted from different model's layers allows to capture more detailed information and enriches semantic representation. However, all the previous architectures designed their aggregation strategies manually. Fixed strategy may result in loss of useful information or involve useless information [2].

The comparison between different models and their aggregation strategies is presented in Figure 1.2. For example, original U-Net model performs feature fusion among layers at the same level. U-Net with Res or Dense blocks merges features at the same level and across 2 levels. Finally, deep aggregation counterparts perform merging between almost all the layers. On the other hand, UX-Nets allow each layer to select an optimum operation (e.g. dilated or traditional convolution) with proper receptive field. This process is shown in Figure 1.2.

As depicted in Figure 1.3, searching strategy of UX-Nets for feature aggregation is conducted to find more efficient merging method based on extracted features. Moreover, scale-wise aggregations and block-wise operators can be searched through relaxation in differentiable manner. The whole optimization process is driven automatically without a use of any pre-fixed set of receptive fields [2].

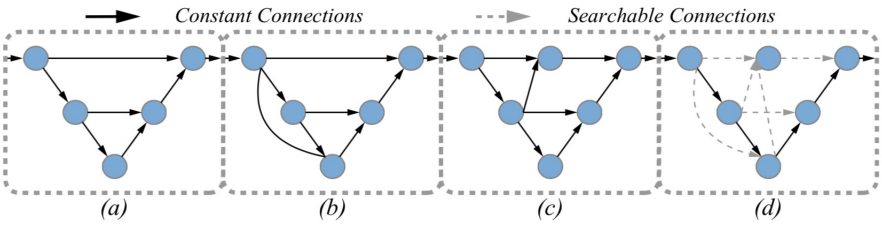

Fig. 1.2: Comparison of UX-Net and other architectures. Source [2].

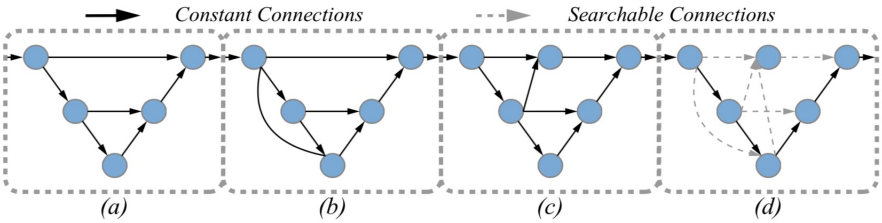

Fig. 1.3: UX-Net's multi-scale search architecture. Source [2].

1.2.3 StyleGAN

A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN) represents one of the best generative models at this time. It is able to synthetize realistic high-resolution facial images from noise. Many related works in the field of image processing, such as PULSE, InterFace or StyleGAN editor, use it as their backbone. Moreover, StyleGAN is currently a matter of extensive research and numerous works clarifying its internal operation will be soon released [3].

Generally speaking, in the field of image processing style-based generators can synthesize such outputs, whose look and feel, i.e. style, resembles the style of ground-truth images. Thus, they can produce fully synthetic images resembling real-world photos. Moreover, these generators are parametrized and expect some input tensor to guide the synthesis. Not to mention the fact, that they also take advantage of noise input in order to include variety in the generated predictions. The input tensor is usually referred to as a latent code. In case of StyleGAN the latent code represents the face features in some latent space. The mapping function between the real images and latent codes is not provided, since the purpose of generative models is to create new parametrized predictions and not to recreate the original images [3].

In case of StyleGAN, the latent code is normalized and passed through a mapping network comprised of fully connected layers as shown in Figure 1.4. That way, the input latent code z is mapped into a latent code w inside an intermediate latent

space W controlling the generator though multiple adaptive instance normalizations. The dimensions of vectors z and w are equal to 512. At the same time, Gaussian noise is injected into the synthesis network g to include more variety in the output predictions. Remaining parts of the system were omitted for brevity, but they match a typical GAN architecture [3].

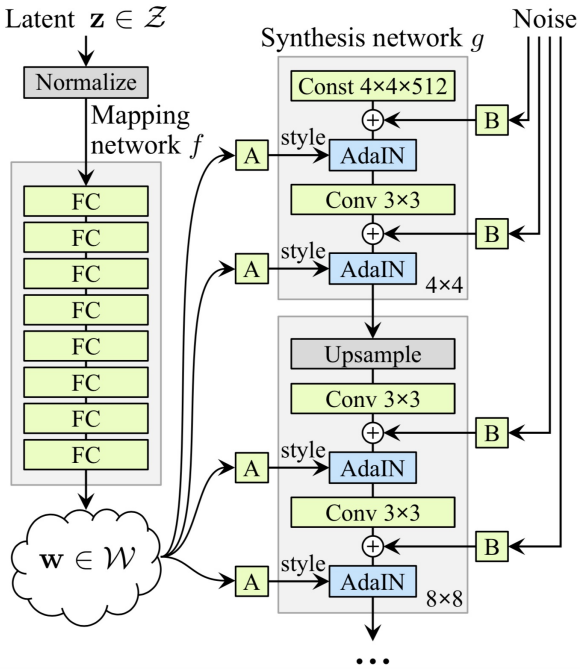

Fig. 1.4: System overview of StyleGAN. Source [3].

So as to show an effect of latent code on the outputs of StyleGAN, an idea of style mixing will be presented. In the Figure 1.5 *Source A* represents a latent code input into the StyleGAN at the beginning of image synthesis. And in the middle of the process, the latent code was switched to *Source B*. To be more specific from the implementation's point of view, two latent codes z_1 and z_2 were mapped into corresponding codes w_1 and w_2. w_1 was applied to synthesis network before given crossover point and w_2 after the crossover point. As it can be seen from the Figure 1.5, the predictions' faces resemble *Source A* images but their hair (and some other salient features) and copied from *Source B* images [3].

The limitation of StyleGAN for the image sharpening purposes is the fact that it does not provide any explicit mapping function between input real-world images and latent codes z. Thus, LR image cannot be simply fed it into the StyleGAN in a hope StyleGAN would synthesize an appropriate HR image. This missing piece is provided by other works such as PULSE [3].

Fig. 1.5: Effect of latent source on StyleGAN. Source [3].

1.2.4 StyleGAN Editor

Although StyleGAN can synthetize any generic facial image from input latent vector, it cannot perform the inverse process, i.e. generate latent vector from given facial image. This missing piece is provided by StyleGAN Editor. Formally speaking, StyleGAN Editor is an algorithm able to embed input image into StyleGAN's latent space. This embedding operation may find numerous applications in semantic image editing. Some of the editing operations, such as image morphing or style and expression transfer are also implemented by StyleGAN Editor [4].

Fig. 1.6: Examples of StyleGAN Editor's performance. Source [4].

Figure 1.6 shows an example of input images in the top row and results of embedding these images into StyleGAN's latent space using StyleGAN Editor. In order to study the implemented algorithm deeply, authors decided to not only use images of human faces as inputs, but images of animal faces and a car as well. As it can be seen in the figure, image of face was embedded with slight imperfections, thus the goal of StyleGAN Editor was successfully reached. When it comes to images of animal faces sharing the same overall structure with humans, embedded images are the same as input images, though of lower quality. The surprising fact is, that image of car could be embedded as well. Again, the quality is worse, but it clearly depicts the generative power of StyleGAN [4].

StyleGAN has multiple latent spaces such as initial latent space Z and intermediate latent space W as shown in Figure 1.4. Latent vectors $w \in W$ in created by passing latent vector $z \in Z$ through multiple fully connected layers. StyleGAN Editor aims to embed into latent space W. It produces 18 vectors w of 512 dimensions, one vector for each AdaIN block of StyleGAN. AdaIN blocks represent input blocks to generative network of StyleGAN and are shown in Figure 1.4 as well [4].

The embedding algorithm of StyleGAN Editor represents a basic optimization framework. An input is represented by two-dimensional image of three channels and pretrained StyleGAN generator. The output is a latent vector which, when passed through StyleGAN, matches the original input image. The algorithm starts with random latent vector. Each value of this vector is initialized independently following uniform distribution in range [-1,1]. Such latent vector is then fed into AdaIN layers of StyleGAN and generated image is compared with original input image using loss function. Loss function takes advantage of VGG 16 perceptual loss and MSE loss. Based on the loss, the value of latent vector is updated and algorithm performs another iteration. The whole process is over, when such latent vector is found which minimizes loss measuring similarity between the generated image and the original input image [4].

1.2.5 Image2StyleGAN++

Year later, authors of StyleGAN Editor released set of new algorithms able to perform more semantic editing operations of input images using StyleGAN and named it Image2StyleGAN++. New editing operations include image cross-over, image inpainting, style transfer, image reconstruction and feature transfer. On top of that Image2StyleGAN++ enhances original embedding algorithm which helps restore high frequency features of embedded images and greatly improves their quality [5].

Unlike StyleGAN Editor, Image2StyleGAN++ optimizes two variables – latent vector $w \in W$ and noise vector $n \in N$. While w encodes semantically meaningful

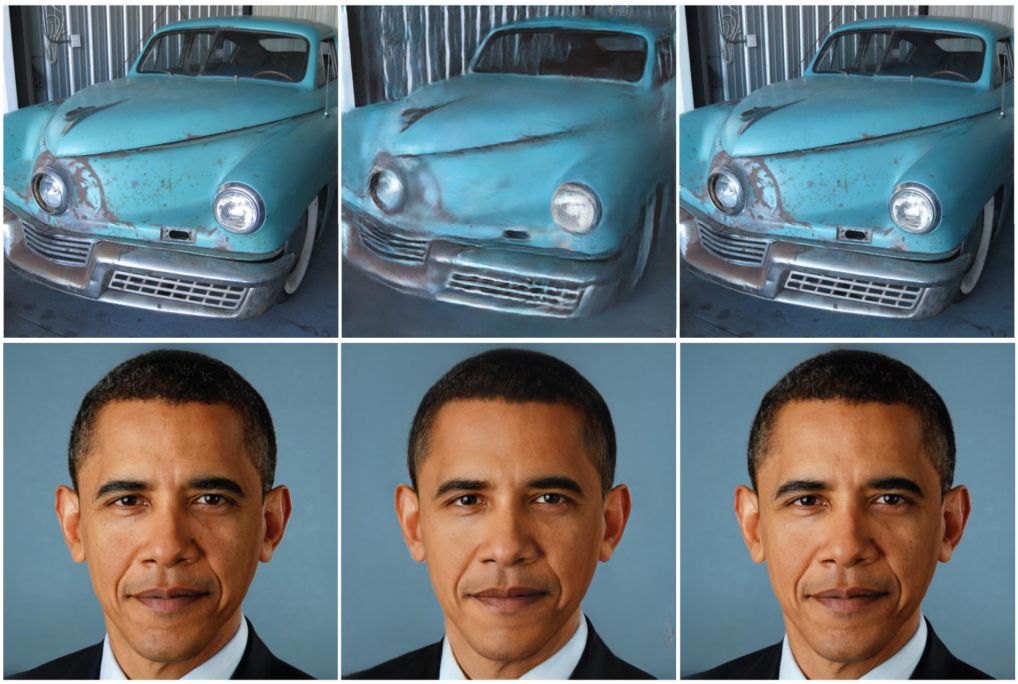

Fig. 1.7: Comparison of ground truth, StyleGAN Editor and Image2StyleGAN++ prediction. Source [5].

information, n stores high frequency component of input image in the Noise space. The actual optimization algorithm is outside the scope of this brief introduction. Rather, the results will be compared. Figure 1.7 shows comparison of embedded images using StyleGAN Editor and Image2StyleGAN++. The first column is an input image, the second column is StyleGAN Editor's prediction and the last column shows predictions of Image2StyleGAN++. As it can be seen, Image2StyleGAN greatly enhances the quality of embedded images. It works not only for facial images, but images of cars as well. Practically speaking, the quality of predictions increased by 20 – 40 dB in PSNR scale according to authors [5].

The idea of image cross-over operation is to copy parts of one image into the other image. Which parts should be copied are specified by a binary mask. In order to perform this operation, Image2StyleGAN++ uses regular embedding algorithm with some modifications. Instead of comparing the embedded image to a single input image, it compares just some parts of embedded image to one input image and remaining parts to the other input image. Input mask states which parts of embedded image should be compared to which input image. In order for embedded image not to contain sharp edges due to sharp edges in mask, mask is passed through Gaussian filter before used by the algorithm. The rest of the algorithm remains the same. Value of loss function obtained by evaluating embedded image is used to

update latent vector and algorithm steps into next iteration. The process is over when loss function is minimized below some specific threshold value [5].

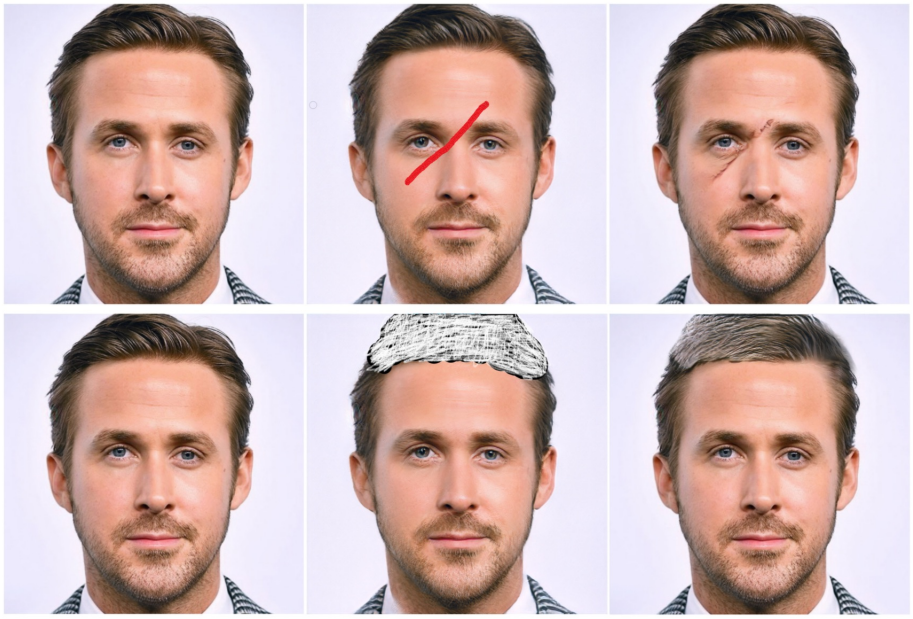

Fig. 1.8: Inpainting using scribbles with Image2StyleGAN++. Source [5].

Image2StyleGAN++ supports local editing using scribbles as well, i.e. generating missing parts of an image based on scribbles included in that image. This operation is again based on embedding optimization algorithm. Though, the explanation would be too lengthy. Some example results are presented in Figure 1.8 instead. The first column shows original input image, the second column shows local edits to input image, i.e. scribbles, and the third column shows image generated by Image2StyleGAN++. As it can be seen, red line was turned into a scar, some hair sketch was interpreted as regular hair and beard was fully removed as well. These examples depict the robustness of Image2StyleGAN++ algorithm [5].

1.2.6 PULSE

Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models (PULSE) represents current state of the art in the field of face hallucination. Term face hallucination is used rather than sharpening since it more precisely depicts PULSE's novel approach to this problem. While the vast majority of current super-resolution systems rely on CNNs trained on pairs of LR and HR images, PULSE represents a purely unsupervised approach. It explores the HR manifold of StyleGAN and chooses such HR image which downscales correctly to original

LR image. Unlike other systems which start with the LR image and gradually add texture details [6].

PULSE traverses the latent space of the StyleGAN, downscales the generated predictions and compares them to the input LR images. It operates in multiple iterations, e.g. 100 iterations as authors recommend, with a specific learning rate and gradually optimizes StyleGAN's predictions. Thanks to StyleGAN being so well trained, the PULSE's results always belong to a natural image manifold. Moreover, PULSE's algorithm ensures its outputs once downscaled match the LR inputs. A simplified diagram of PULSE system is presented in Figure 1.9. PULSE loads an input LR image I_init, converts it into latent code z_init and searches the latent space L. Once the optimization is successful, latent code $z - final$ represents such image I_final which belongs to natural image manifold and at the same time downscales correctly [6].

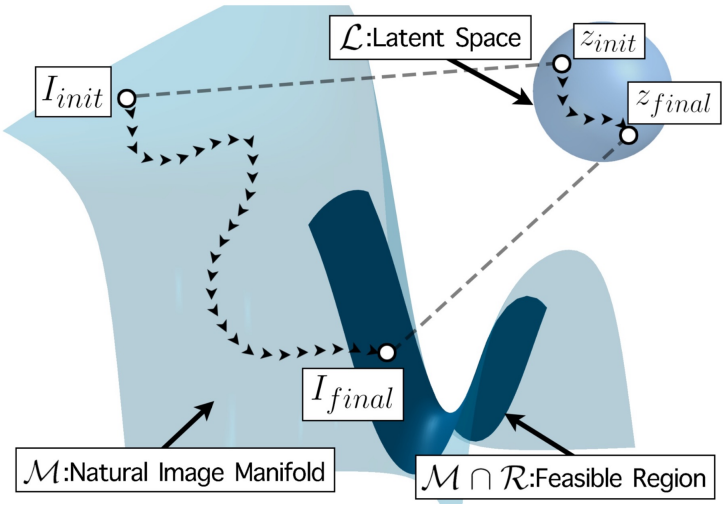

Fig. 1.9: System overview of PULSE. Source [6].

1.2.7 InterFaceGAN

InterFaceGAN network is a part of work trying to interpret StyleGAN's latent space and its latent vectors. It studies facial semantic properties encoded in StyleGAN's latent space. The main objective is to identify linear subspaces of StyleGAN, each subspace representing a single facial attribute, and realistically manipulate chosen attribute of given input image. As of now, InterFaceGAN has a precise control over gender, expression, age, pose and presence of eye glasses [7].

InterFaceGAN implementation paper employs GAN inversion approach to edit attributes of an input image. It trains encoder called InterFaceGAN network to

reverse StyleGAN's operation. Since InterFaceGAN paper aims control multiple subspaces of latent space of StyleGAN, it needs to provide that many trained InterFaceGAN networks [7].

Fig. 1.10: Example of InterFaceGAN operation. Source [7].

An example operation of InterFaceGAN is presented in Figure 1.10. Manipulation of all attributes can be observed here. In case of *Pose* image sequence, middle image represents the input to InterFaceGAN and images on sides represent the output. Input image is aligned and shows frontal face. Authors do not present the outputs when input image is unaligned or shows side of the face. In case of *Smile* image sequence, face without smile is the input and the two faces with smiles are outputs. InterFaceGAN can again produce multiple results based on different level of amplification of selected attribute. The same goes for images with glasses and age. Identity in the input image can be made look younger by using negative amplification of age attribute or made look older by using positive amplification of age attribute. Finally, InterFaceGAN has a control over gender as well. It can gradually turn female gender into male gender and vice versa. In case of *Gender* image sequence, image on left represents the input to InterFaceGAN, i.e. young female. Image on the right to the input image shows the same identity having few salient male features. Image on the right shows identity of male gender still resembling the previous image. As the figure shows, InterFaceGAN seized precise control over inspected five attributes [7].

1.2.8 PSFR-GAN

Most of the architectures presented so far are based on an encoder-decoder (U-shaped) structure and get trained to learn a direct black-box mapping from low-quality to high-quality images. These approaches represent current state of the art and but just few of them give satisfactory results for real-world low-quality images. Moreover, it is difficult to enhance their performance, because they literally operate as black boxes. No works have been published so far discussing the effect of hidden

layers of U-Net architectures on predictions. Up to this day, it is truly unknown what they internally do.

Progressive Semantic-Aware Style Transformation for Blind Face Restoration (PSFR-GAN) represents a novel architecture for SF super-resolution tasks. Unlike other U-shaped networks, PSFR-GAN approaches the face restoration process as semantic-aware multi-scale transformation. It uses semantic-aware style transfer process to progressively restore the features of different scales [8].

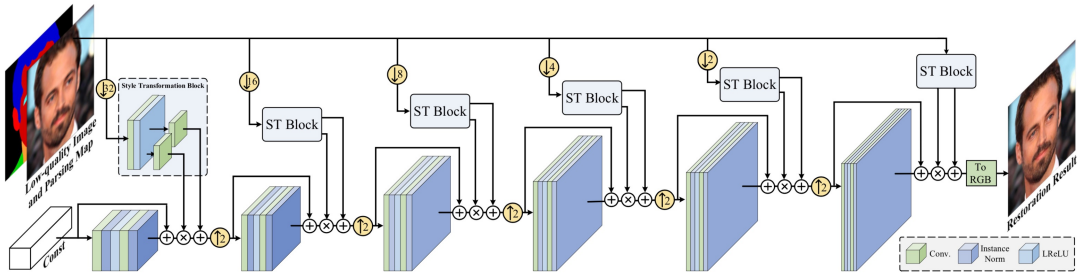

Fig. 1.11: PSFR-GAN network architecture. Source [8].

The Figure 1.11 shows the PSFR-GAN network architecture. It begins with learnt constant, i.e. latent code, and produces features at different scale using multiple upsampling layers. It also expects two other inputs – low-quality image and the corresponding face mask. The low-quality image provides information about the color and the mask provides information about the shape and semantics. As the input low-quality image progresses through the network's layers, the details are inserted in a coarse-to-fine manner. Another trick PSFR-GAN takes advantage of is a Semantic Aware Style Loss which greatly helps to enhance the restoration of the textures and limits the appearance of unwanted artifacts in the final prediction. Semantic Aware Style Loss will be explored in the following chapters [8].

1.3 Loss Functions in Super-resolution

Generally, in decision theory and mathematical optimization, a loss function, also known as cost function or fitness function is a function which maps value of a given variable onto a real number. This number stands for some "cost" related to the value. Optimization problems seek to minimize the "cost" of this loss function [13].

In machine learning models learn by means of loss function. It evaluates how precisely the model models given data. If model's predictions deviate from desired results, loss function yields a large number. The "cost" is high. Gradually, loss

function reduces its error value for predictions as the model learns to model data better. Improvement is driven by an optimization function [13].

Loss functions can be generally categorized into two main groups depending upon the task – classification and regression loss functions. In classification tasks, the model is trying to predict a category coming from a finite set of categorical values. On the other hand, in regression tasks, the model is trying to predict a value coming from a continuous range. Image segmentation domain mostly deals with regression tasks. That is why a closer look into them will be provided in the following subchapters [13].

1.3.1 Mean Squared error

Mean Squared Error (MSE) also known as quadratic loss or L2 loss computes the average of the squares of the errors. The error actually stands for the difference between the predicted and actual value. MSE is strictly positive, it does not take into account the direction of an error. It can be computed using Formula 1.2, where n stands for the number of predictions in vector $\hat{Y}$.Y represents the vector of actual values. Thanks to squaring operation, predictions which are less deviated from actual values yield low error value but, on the other hand, predictions far away from actual values get heavily penalized [13].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1.2}$$

In the field of image segmentation, MSE is usually used to compare red, green and blue (RGB) values of every single pixel of predicted image and ground-truth image. That is why MSE is also known as the pixel loss. Moreover, MSE is also used while comparing activations when predicted and target image are passed through another model used for evaluation. More on that later.

1.3.2 Perceptual Loss

In 2016 it was published a novel approach for evaluating predicted images and it was named perceptual loss [9]. It comes up with new ingenious loss function still used up to this day thanks to having an amazing image quality evaluation performance.

Perceptual loss is based on comparison of high-level semantic features extracted from a pretrained network of generated and ground-truth image. Since it compares features, it is many times referred as a feature loss. In this work, these two terms will be used interchangeably. Perceptual loss function can be used for training any feed-forward network dedicated to image transformation task. Compared to other optimization-based approaches, perceptual loss returns similar qualitative results

but it is more than three orders of magnitude faster. In the domain of SR, perceptual loss gives visually more pleasing results compared to a pixel loss. Thus, perceptual loss measures similarities between predicted and ground-truth image more robustly than a simple pixel loss [9].

Perceptual loss was tested on single-image SR task by its authors. As they correctly state, it is an ill-posed problem. There does not exist a single correct output image. Conversely, there are multiple high-resolution images which could be generated from the same low-resolution input image. Success in this task requires good semantic reasoning about the input picture. Fine details of a generated and visually ambiguous low-resolution input image must be inferred similarly, ideally equally. Theoretically, any high-capacity neural network could learn this kind of semantics reasoning implicitly, however in practice an explicit loss function guiding such training is required [9].

System overview of Perceptual loss function is shown in Figure 1.12. As a loss network authors used VGG16 architecture pretrained on image classification task using ImageNet dataset. As a generator network transforming low-resolution images to high-resolution images authors used some arbitrary model. During the training of a generator network, loss network remains fixed – its parameters do not get updated. Predicted high-resolution image and ground-truth high-resolution image are passed through the loss network and extracted high-level semantic features are compared using feature reconstruction loss function. Optionally, extracted low-level visual features can be compared as well [9].

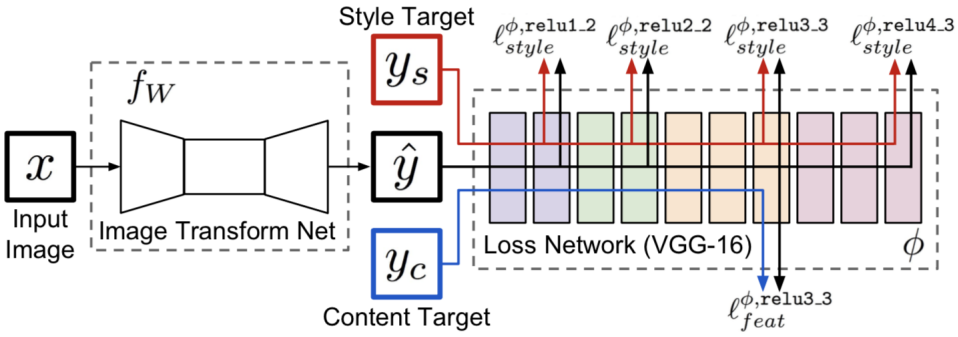

Fig. 1.12: System overview of Perceptual loss. Source [9].

Instead of encouraging the pixels of predicted image $\hat{y}$ to precisely match the pixels of the ground-truth image y, perceptual loss encourages them to have similar representations of features stated by the loss network ϕ. In the Formula 1.3 $\phi_j(x)$ represents activations of the jth layer of the network ϕ, when network process image x. j stands for a convolutional layer and $\phi_j(x)$ is a feature map of a shape $C_j \times H_j \times$

W_j. Feature reconstruction loss 1.3 then represents squared, normalized Euclidean distance between feature representations [9].

$$l_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} ||\phi_j(\hat{y}) - \phi_j(y)||_2^2 \tag{1.3}$$

As an example of perceptual loss performance, authors present Figure 1.13. It can be observed, that model trained with perceptual loss significantly outperforms bicubic interpolation and models trained with a simple pixel loss function. Moreover, as authors state, they did not beat Peak Signal-to-Noise Ratio (PSNR) or Structural Similarity Index Measure (SSIM) of other approaches, but instead they achieved much better visual results. This example greatly showcases incompetency of PSNR and SSIM metrics for SR tasks [9].

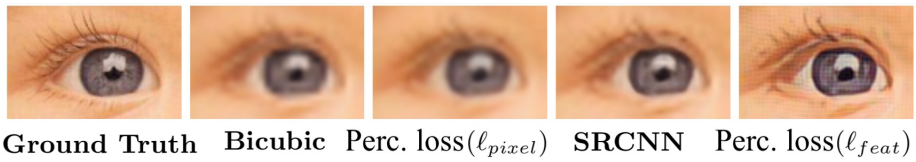

Fig. 1.13: Comparison of outputs of different SR methods. Source [9].

1.3.3 Gram Loss

Gram matrix of a set of vectors, also known as Gramian, is a Hermitian matrix of inner products. Typical application of a gram matrix is computation of linear independence of given vectors. If the Gram determinant is non-zero, then vectors in a given set are linearly independent [22].

In machine learning, gram loss is based on a computation of a Gram matrix and mainly finds an application in image style transfer tasks. An example of style transfer task is shown in Figure 1.14. The top left picture represents the content picture, the bottom left picture is the style picture and the result is presented on the right. Neural network takes in the content and style picture as inputs and with the help of gram loss trains itself to produce the output blended picture [13].

Gram loss function computes MSE between gram matrices of feature representations of generated and ground-truth image. Feature representations of the images are extracted by passing the images through a loss network presented in chapter Perceptual Loss [9].

More specifically, to compute Gram matrix of a given image, the image needs to be passed through a loss network. Loss network will generate activations at its all intermediate layers. Activations from any layer are then considered as feature

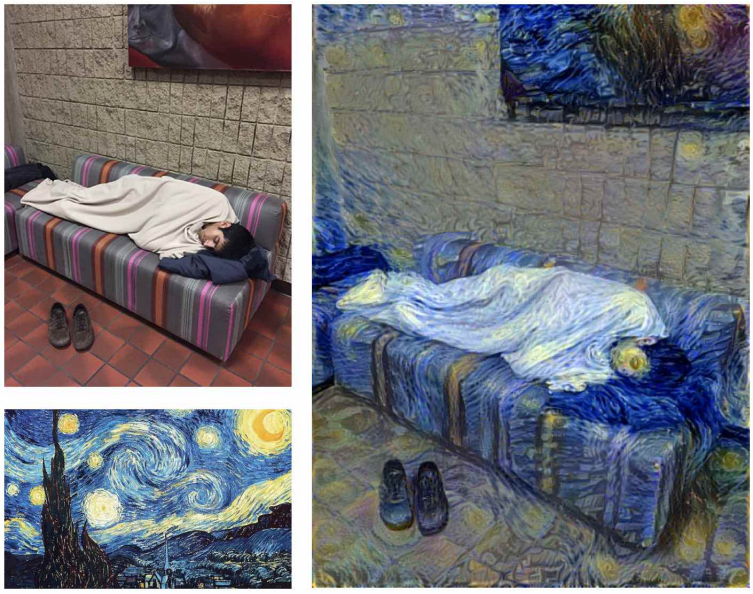

Fig. 1.14: Style transfer example. Source [10].

representations of a given input image. Usually, activations of multiple layers are taken into account, but for the sake of simplicity, consider just a single feature matrix. Feature matrix is then flattened and used for computation of a dot product. The dot product characterizes image's style, but completely loses information about image's spatial structure due to flattening. The result is the Gram matrix. When Gram matrix of content image and style image are compared using MSE and back-propagated, the network learns to regenerate content image with style of the style image [9].

1.3.4 SER-FIQ Loss

A novel approach to estimation of image quality was published recently. It was named SER-FIQ - Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness (SER-FIQ) [11]. This assessment mainly focuses on estimating the suitability of image for face recognition tasks. Hence, although this approach can state quality of an input image, the quality does not fully correspond to visual quality of the image. Rather, it states quality of an image from the point of view of neural network performing face recognition task. Still, the two kinds of quality are closely related [11].

SER-FIQ approach is based on determining variations in embeddings generated by random subnetworks of a face network. Image quality is then represented by the robustness of a sample representation. The whole concept is shown Figure 1.15.

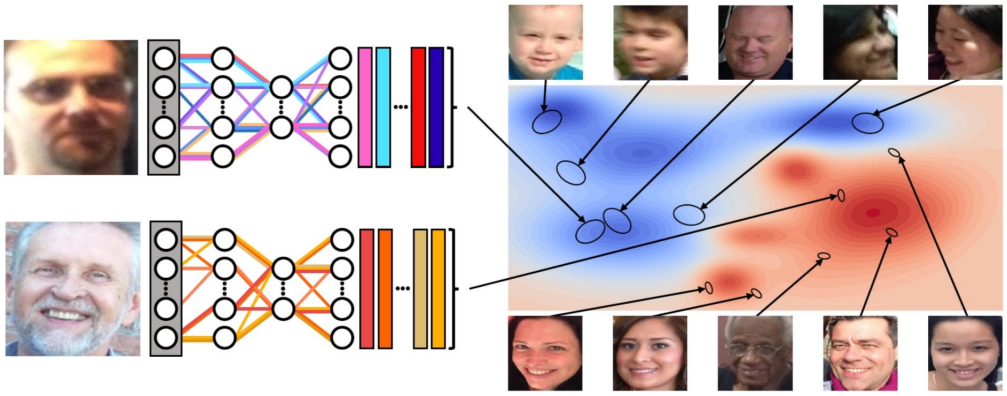

Fig. 1.15: System overview of SER-FIQ. Source [11].

As Figure 1.15 shows, high quality image (bottom left image) results in slight variations in stochastic embeddings and thus high robustness (red areas on the right). On contrary, low quality image (top left image) results in significant variations in stochastic embeddings returned by random subnetworks and hence indicates low robustness (blue areas on the right). In order to obtain random subnetworks of a face network, face network needs to be trained with dropout on face recognition task. Moreover, the dropout layer needs to be active even during inference. This way, by forwarding the same input image through the same face network m times, it will be obtained m different stochastic embeddings. The process is shown in Figure 1.16. The variations between these stochastic embeddings define the quality of the image [11].

More formally, SER-FIQ predicts quality $Q(I)$ of given image with face I using face network with dropout M trained on face recognition task. Model M needs to excel at extracting embeddings which are well identity-separated. In order to make the quality estimation of image I, m stochastic embeddings need to be generated by model M with the help of different dropout patterns. The value of m is always a trade-off between stability of quality measure and time complexity. Authors recommend using value $m = 100$. All stochastic embeddings are collected in a set $X(I)$ and the negative mean Euclidean distance between them is computed. The mean is then forwarded through the sigmoid layer which ensures that the quality will be in range <0, 1>. The whole formula is presented in Formula 1.4.

$$q(X(I)) = 2\sigma(-\frac{2}{m^2}\sum_{i<j} d(x_i, x_j)) \tag{1.4}$$

SER-FIQ's approach and use of Euclidean distance is backed by [23]. Authors in mentioned paper prove that repetitively applying dropout on network approximates

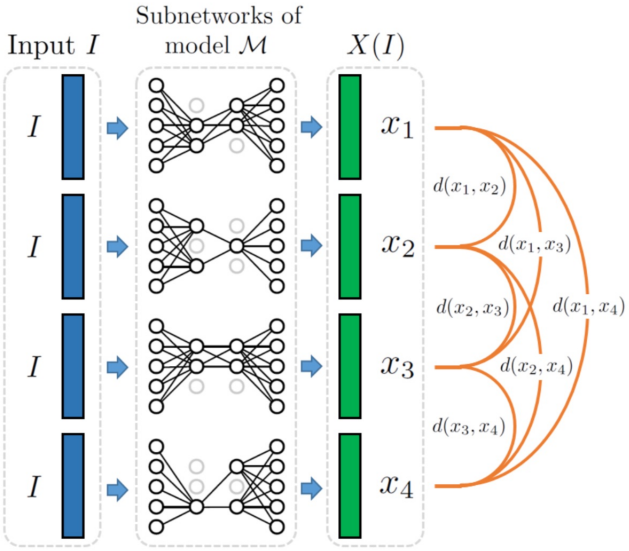

Fig. 1.16: Stochastic embeddings generation. Source [11].

uncertainty of Gaussian process [24]. More significant variation in the stochastic embeddings implies low robustness and thus, lower image quality [11].

An example of SER-FIQ's quality estimation is presented in Figure 1.17. As it can be seen, picture on left is evaluated by SER-FIQ as an image of high quality. Picture on the right is of poor quality and SER-FIQ also rates it poorly. Though, the three pictures in the center are also evaluated poorly despite having moderate visual quality. This is caused because SEF-FIQ states image suitability for face recognition model rather than visual quality. And since all the three pictures are ambiguous in terms of face recognition, they were rated poorly too. The reason why the images are ambiguous is because they either contain multiple faces in the same picture or because they do not capture full face from the front. Face recognition model would not be able recognize the identities with high level of confidence in either case [11].

Fig. 1.17: Example of quality evaluation by SER-FIQ. Source [11].

1.3.5 Semantic-Aware Style Loss

Simple gram matrix loss was first presented in super-resolution work [25] and has shown that gram matrix loss function is not only useful in style-transfer tasks, but it also positively affects texture recovery in super-resolution tasks. Authors of PSFR-GAN enhanced it and introduced a Semantic-Aware Style Loss l_{SS} helping to achieve better synthesis of texture details. It computes the gram matrix loss separately for every single semantic region of the visual features extracted from VGG19 model. More specifically, authors recommend reading layers `relu1_1, relu2_1, relu3_1, relu4_1` and `relu5_1`. Semantic-Aware Style Loss can be computed using formula 1.5:

$$l_{SS} = \sum_{i=1}^{5} \sum_{j=0}^{18} ||g(\phi_i(\hat{I_H}), M_j) - g(\phi_i(I_H), M_j)||_2 \tag{1.5}$$

where

- ϕ_i - i-th feature layer of VGG19,
- M_j - parsing mask with label j (e.g. background is M_0),
- $g()$ - gram matrix,
- $\hat{I_H}$ and I_H - predicted and label image

and gram matrix $g()$ can be computed using formula 1.6

$$g(\phi_i, M_j) = \frac{(\phi_i \odot M_j)^T (\phi_i \odot M_j)}{\sum M_j + \epsilon} \tag{1.6}$$

where ϵ avoids zero division, $\epsilon = 1e - 8$.

1.4 MLFDB Dataset

Multi-frame Labeled Faces Database[1] was primarily developed for multi-frame face superresolution tasks. It includes a wide range of ethnical groups, and age. Moreover, the images were taken in a different scale, angles, various lighting conditions and capture realistic background scenes. In total, the dataset provides exactly 12,200 training image sequences and 2,600 testing image sequences.

Each image sequence consists of seven images at resolution 32x32 representing the damaged images to be restored and a single label image at high resolution. The resolution of a label image varies in the dataset. It is either 64x64 or 128x128 or 256x256. Finally note that, the middle picture in the dataset matches the label picture.

[1]Source: http://splab.cz/mlfdb/

The dataset was built from around 300 youtube videos and includes roughly 7,000 unique identities. In order to ensure uniqueness between sequences, only such sequences were included, which differ from the other sequences by SSIM 0.7 or less. An example sequences are shown in Figure 1.18. These samples were randomly chosen from the training set and they clearly depict dataset's diversity. The first seven images to the left represent input damaged images and the image on the right represents a label image. Note that, since all images are displayed at a low resolution, the difference between input and label images seems subtle. Though, the figure is just misleading.

Fig. 1.18: MLFDB dataset examples.

1.5 Summary

Current state of the art in the field of image sharpening and superresolution has been presented in the previous chapters. These domains have not been truly explored yet for biometric purposes. Multiple works have tried to provide a solution for superresolution problem, though, their priority was the highest output image quality rather than preserving the original identity.

Nevertheless, promising novel approaches to image reconstruction have been explored in greater detail. U-Net and GAN networks still represent the leader in this field and that is the reason why many related works have focused on enhancing them. Multiple alternative networks have been presented recently and promise improvements in visual quality of sharpened images. It will definitely pay off to look

into them deeper and reimplement them in the practical part of the work. Note that, the greatest challenge to overcome will be to propose meaningful ways how to feed multiple images into these networks. They have all been originally designed for single image input.

StyleGAN and related works unarguably represent the most innovative approaches to face modeling, sharpening and editing at this time. StyleGAN on its own can generate fully synthetized images from input random vector. Other works, such as StyleGAN Editor offer an understanding of this vector and are able to perform an inverse operation – create latent vector from synthetized image. InterFaceGAN is built on top of StyleGAN Editor and it analyzes internal operation of StyleGAN. Thanks to that it can easily change attributes of the synthetized face. Finally, PULSE provides a simple way how to sharpen input image using StyleGAN. Instead of gradually increasing texture details of low-resolution image, PULSE returns fully synthetized high-resolution image. This synthetized image matches the original low-resolution input image when downscaled back to the original resolution. Mentioned works will be analyzed and taken advantage of in the practical section of this thesis as well. They can be utilized to build custom tools for image editing.

Finally, works such as SER-FIQ also present an opportunity to build custom critics for U-Net and GAN architectures. Moreover, SER-FIQ algorithm seems promising as metric as well since it is able to evaluate visual image quality. Though, it has not been put into practice yet. Its performance is unknown yet valuable to explore.

2 Implementation

Image reconstruction from a sequence of images still represents a domain in the modern science, which has not been fully explored yet. Few related works have been published in a domain of SF super-resolution and generic MF video reconstruction. Though, none of them have truly focused on human face sharpening and its reconstruction from multiple images. Moreover, these related works have not presented any ground-breaking approach so far and their results are usually poor.

The following chapters of this work will describe in detail multiple experiments that will be implemented and evaluated. Experiments will be split into three groups. The first group of experiments will present image sharpening approaches which rely on U-Net architecture and its most recent alternatives such as BiO-Net. The idea will be to propose new models and modify existing ones in such a way, that they accept all input images from given sequence at once. This approach has never been truly explored thus it is difficult to make any presumptions about its performance in real world.

The second group of experiments will present novel approaches to single image sharpening. New networks will be presented and compared with existing alternatives. Single-frame sharpening is generally a valuable tool required by multiple other tools presented in this thesis.

Last group of experiments will approach multi-frame sharpening from different point of view. It will introduce a suite of tools able to recognize, adjust and merge high-quality visual features of input images into a final image. The basic presumption is, that every image in the sequence contains compression artifacts at different locations. Moreover, different regions of each image have different quality. That is why the idea of finding high-quality regions and merging them seems promising.

Fastai library [26] will be used in the whole thesis mostly because it provides clear documentation, trainings and demo samples. The library is based on Pytorch library and aims to provide simpler, more practical and easier to use API. Though, fastai library does not contain implementations of the most recent or less known models. It will be still necessary to implement them using papers published by the authors. Some extra utility functions for masking and merging images will need to be implemented as well. Another limitation of fastai library is the absence of multi-frame dataloader, i.e. object being able to load multiple images from the disk as a sequence, perform some transformations on them and feed them into the model. It will be necessary to implement such dataloader which will fully comply with fastai API.

2.1 Multi-frame U-Net Based Sharpening

The following chapter with its subchapters will represent a group of experiments trying to sharpen the middle image from input image sequence. All images will be fed into U-Net based model at once. Although multiple networks will be presented, general system architecture will remain the same and it is presented in Figure 2.1.

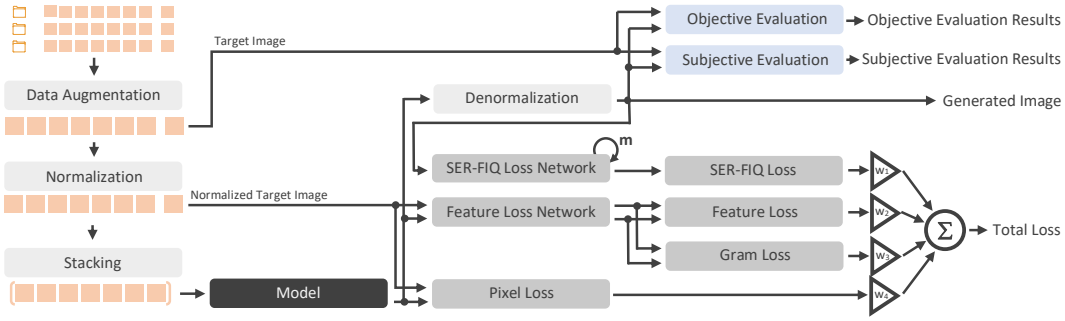

Fig. 2.1: System overview of U-Net based experiments.

The dataset will consist of sequences of damaged images and a single label image per sequence. Each image sequence is stored in a separated folder and before it is passed to the model, some preprocessing is applied to it. Pre-processing includes basic data augmentation techniques such as random zoom within a given range, vertical flip, random brightness adjustment and, most importantly, resizing to a desired resolution. All images are resized to a resolution 128x128 using bicubic interpolation and normalized using Imagenet parameters (mean and standard deviation values for RGB channels). Bicubic interpolation will be always performed before feeding images to the model. During the experiments it has proven that bicubic interpolation is better than upsampling using convolutional layers in terms of loss of information.

Once a sequence of seven images is randomly augmented and normalized, images are stacked along dimension 0 and put into a batch. Thus, in case of seven input images each consisting of 3 channels (RGB) and resolution 128x128, the batch dimensions would be $[bs, 21, 128, 128]$ where bs represents the number of sequences in the batch.

Model produces normalized images at its output. In order to evaluate them, images need to be denormalized using Imagenet parameters. Evaluation will comprise objective and subjective methods as well. For more information, refer to chapters Objective Evaluation Methods and Subjective Evaluation Method.

Model's normalized outputs are also used for computation of a total loss. Total loss is a weighted sum of partial losses. w_1, w_2, w_3, w_4 in the scheme represent the weights. Not all loss functions shown in the Figure 2.1 will be used for each model.

Particular loss functions used for a given model will be discussed in appropriate chapters. Generally, though, Feature loss described in chapter Perceptual Loss will be taken advantage of as well as, Pixel loss described in Mean Squared error and SER-FIQ loss discussed in SER-FIQ Loss. Note that, m in Figure 2.1 represents the number of generated stochastic SER-FIQ subnetworks through which the model's denormalized outputs will be passed.

2.1.1 Data Description

Two datasets will be utilized in the group of U-Net based experiments - MLFDB and CelebA dataset. There are already two reasons why. At the time of heavy experimentation, MLFDB had not been ready yet and the models implemented needed some dataset - CelebA dataset was at hand with images of faces already cropped out. The second reason why is that preliminary training on MLFDB has revealed some imperfections of the dataset. Thus, this thesis will also state what can be further improved about it in future.

CelebA[1] belongs to the most popular datasets in the field of face recognition. It contains around 200,000 SF aligned images of celebrities. Such size is big enough even for a heavy training of the final model. Huge limitation of CelebA dataset is the fact that it does not contain image sequences. This thesis solved the problem by data augmentation.

First of all, the whole CelebA dataset will not be used for the training. It is impractical since it would require lots of time spent copying the data and uncompressing it. Instead a smaller subset will be used. The subset was created by copying over the first 60,000 images from CelebA dataset and resizing and cropping them to resolution 128x128. These images represent the labels. The input damaged image sequences were created by applying seven different random transformations to each label image. The transformations included warping, rotation, brightness change, zooming and others. All transformed images were then resized to resolution 32x32. The final step introduced quality diversity in the sequences. Each resized image was saved at different quality from random range using JPEG compression. The whole dataset preparation pipeline is presented in Figure CelebA dataset transformation pipeline.

An example of an image sequence created from a single image is shown in Figure 2.3. The image on the right is the original label image. The greatest drawback of this approach is the fact, that it does not allow to create images from considerably different angles, e.g. picture of face vs. picture of the hair from behind. It is a huge limitation, but for now, this dataset suffices.

[1]Source: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

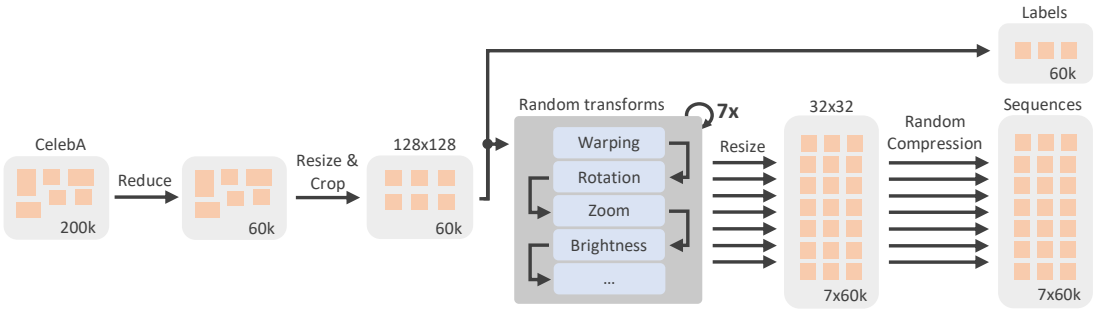

Fig. 2.2: CelebA dataset transformation pipeline.

Fig. 2.3: Example sequence of transformed CelebA dataset.

2.1.2 U-Net

The model proposed in this section is based on a simple U-Net architecture with ResNet34 encoder and it is presented in Figure 2.4. Although U-Net architecture is generic and supports various encoders, ResNet34 has proven to be give the best results during numerous experiments. Moreover, U-Nets with Residual blocks represents current state of the art.

Model's input is represented by a sequence of seven images in the same resolution cropped into a square. Before feeding them to the model, images are first randomly augmented, normalized, stacked along dimension 0 and put into a batch. Batch size of 64 is as much as could be reached while training on 15 GB of GPU memory.

ResNet34 encoder first passes the batch through initial basic layers. Beginning 2D convolutional layer of stride 2 reduces resolution to 64x64 pixels, BatchNorm layer than performs batch normalization, following ReLU introduces non-linearities and finally MaxPool layer further downsamples the batch to a resolution 32x32. Moreover, beginning 2D convolution works with large kernels 7x7. It is a recommended dimension for image resolution of 128 pixels.

Following layers are based on two basic building blocks, Res blocks and Downsampling Res blocks. While Res blocks only perform feature extraction, Downsampling Res blocks reduce the resolution and increase the number of output channels. Res block passes an input batch is through two series of 2D convolution, batch nor-

malization and ReLU activation. At the end, the output batch is summed with its original values. Downsampling Res blocks perform exactly same operation except for the first convolutional layer, which further reduces resolution thanks to stride 2 step and increases the number of output channels. All the convolutions use kernel size 3x3. Other related works have proven that such kernel size is a perfect for match for batch sizes these layers need to work with.

The output of MaxPool layer is sequentially passed through groups of Res blocks ended with a Downsampling Res block. Resolution of batch is gradually decreased to 4x4 and the number of channels is increased to 512 at the end of the encoder. As a last note to the encoder part, note that, the inputs of each Downsampling Res block are stored in the memory for the further use by the decoder.

Encoder is not directly followed by the decoder. Extracted low level visual features are first passed through two bottleneck layers. They sequentially increase the number of channels to 1024 and decrease it back to 512. Their purpose is solely to transform the representation of the same information, which suits the decoder more.

The decoder follows a fixed architecture. For each skip connection coming from encoder, i.e. for each operation changing dimensions of a batch in encoder, decoder provides a decoder layer group which consists of an Upsampling block followed by two convolutional layers separated and wrapped by ReLUs. Upsampling blocks are simply a sequential module of 2D convolution, Pixel Shuffle layer, Replication Pad, Average Pooling and ReLU. Furthermore, each output of decoder layer group is upsampled in the following layer and concatenated with encoder's features from corresponding encoder block. Note that, before concatenating encoder's features with decoder's feature, they are passed through BatchNorm layer which ensures that activations are within a similar range. The idea of directly feeding features from encoder to decoder greatly enhances the quality of generated pictures, though it has a negative impact on memory usage. The last concatenation of activations of the first convolutional layer and the last Upsampling block is not a part of U-Net architecture. It is just a small enhancement suggested by authors of fastai library.

Output of the U-Net model is a single normalized image. In order to denormalize it, an inverse operation to normalization needs to be performed using Imagenet parameters.

Loss function comparing generated and target image is rather complex. It includes a Pixel loss, along with Feature loss and Gram loss. Pixel loss, as explained in Mean Squared error, simply compares color of the pixels of the two images using MSE. Feature loss is based on a pretrained VGG16 model presented in Figure 2.5. Generated and target image are passed through the network individually and their activations are compared using MSE. These activations are also used for Gram ma-

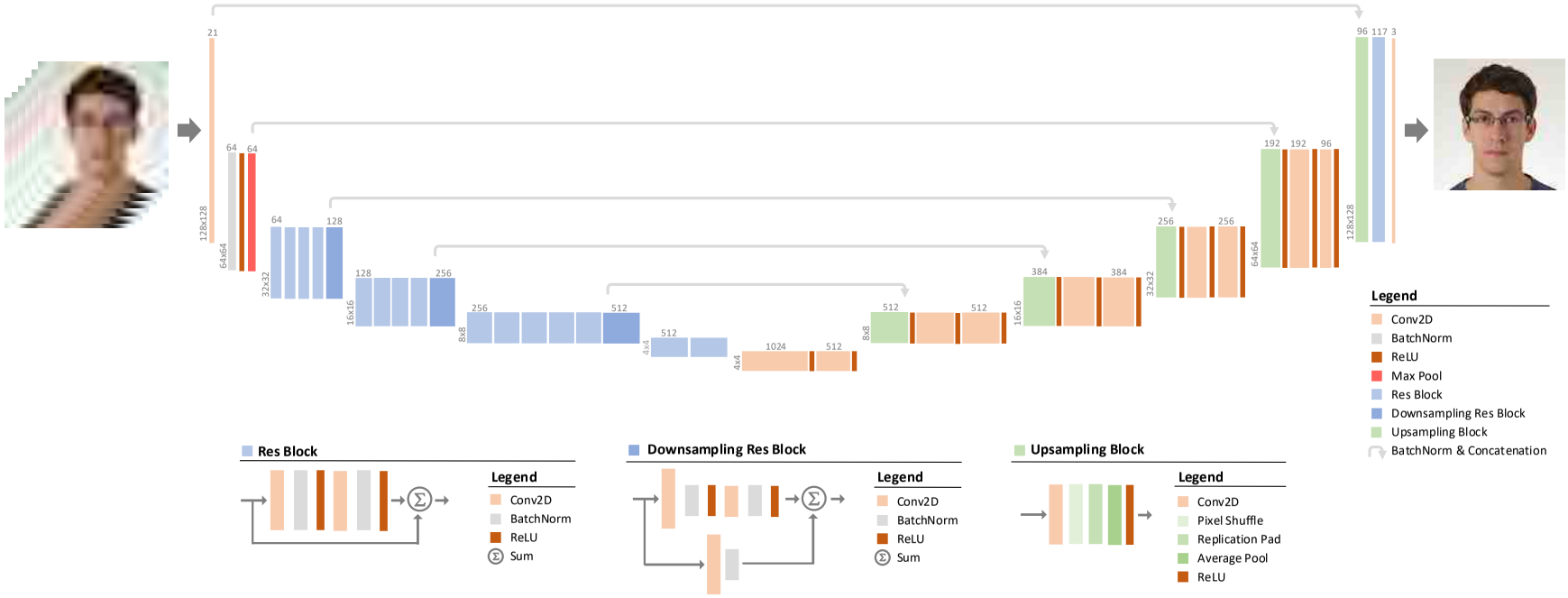

Fig. 2.4: U-Net model with ResNet34 encoder

trix computation and again compared image using MSE in order to obtain Gram loss. All the mentioned loss values are weighted and summed. The mean value of the whole batch is then returned as a final loss value.

Although many related works use a pretrained VGG19 model as a Feature loss network, this thesis will work with VGG16 architecture. After a short comparison it was revealed that VGG19 does help U-Net model train any better. Moreover, it consumes significantly more memory and slows down the training since the batch size needs to be reduced. It does not present limits in case of a simple U-Net architecture, though it causes problems when used in combination with another memory demanding loss function. Simply put, VGG19 is not perspective on limited resources. The last important note is, that the VGG16 model was pretrained on a face recognition problem.

2.1.3 U-Net with SER-FIQ Loss

The architecture described in this section fully matches the one presented in the previous chapter U-Net. It is a pure U-Net model, which also uses SER-FIQ critic, that is why will be referring to it as "U-Net with SER-FIQ loss". A theoretical intro to SER-FIQ was provided in a chapter SER-FIQ Loss. This chapter will discuss the implementation details.

SER-FIQ, as presented in its implementation paper [11], evaluates the quality of a given image. As already explained, thanks to passing the image though stochastic subnetworks pretrained on face recognition problem, SER-FIQ algorithm assigns a high score to high quality images and low score to low-quality images. The returned value is an output of a sigmoid activation and thus, theoretically, value zero stands for a very blurry and damaged image while value one represents an absolutely clear, sharp image.

In order to obtain SER-FIQ quality of an image a face recognition model with at least one Dropout layer is required. After some testing, Arcface IR SE 50 model has proven to evaluate image quality the best. The network architecture is presented in Figure 2.5. It greatly relies on Squeeze-and-Excitation Res blocks, so called SE Res blocks. These blocks work generally as Res block, but they also embed SE module inside them. SE module passes the input through Average Pool and then through two fully connected layers separated by ReLU. The output is forwarded into a sigmoid activation and multiplied with the original input. The Downsampling SE Res block operates in the same manner, except it decreases the resolution (such as Downsampling Res block) and increases the number of output channels. As it can be observed from Figure 2.5, IR SE 50 architecture contains quite a lot SE Res blocks. In return, though, it provides accurate low level visual features at

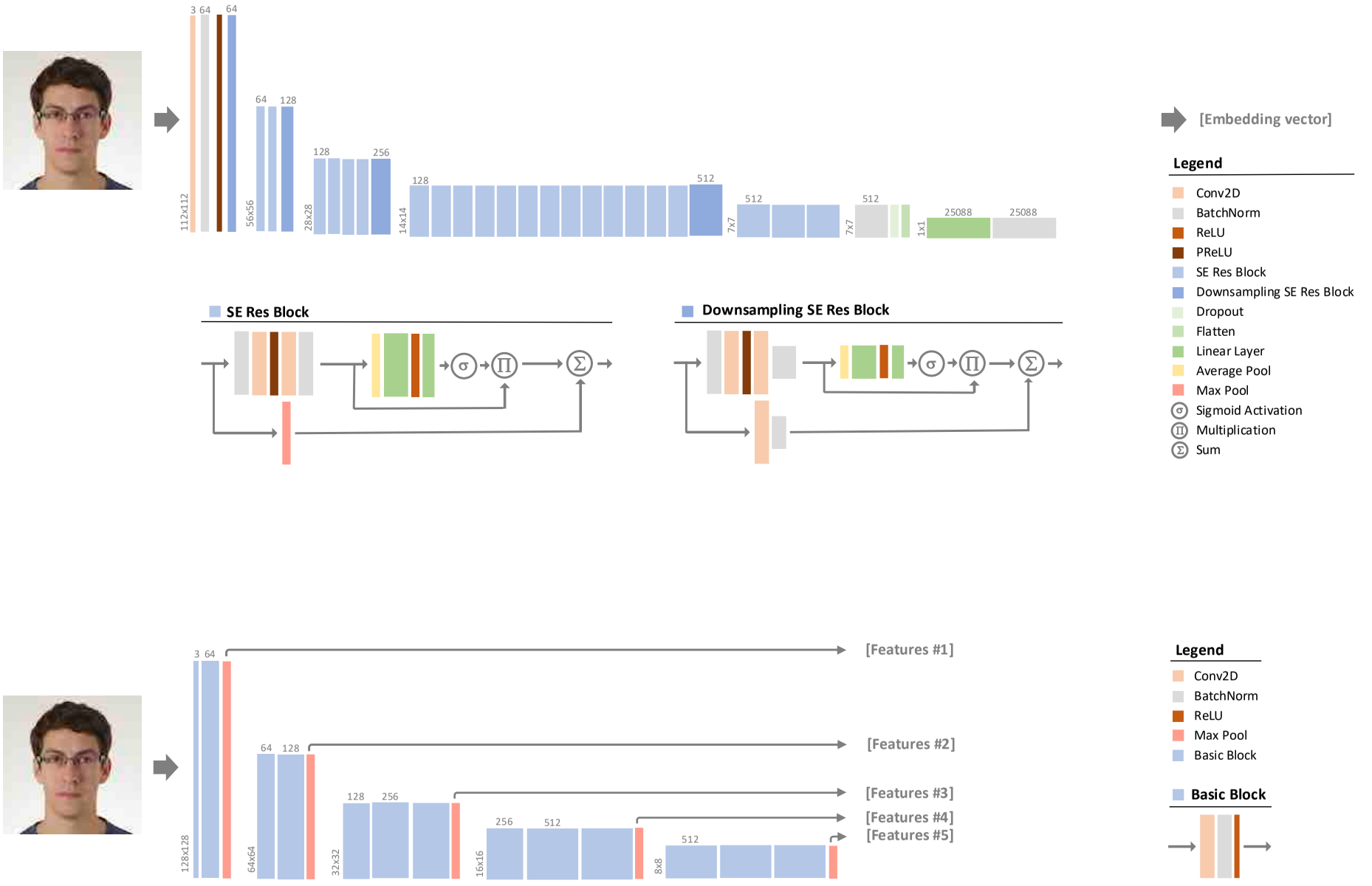

Fig. 2.5: SER-FIQ loss network (top) and Feature loss network (bottom)

the end of the encoder's body. The body is followed by the head, which is also required for the purposes of the whole system. It contains a fixed Dropout layer, which performs dropout operation even during evaluation. The 25088 activations representing low-level visual features of the face are then remapped to a so called emdedding vector representing identity features. Embedding vector could then be remapped to individual identities using another fully connected layer. Though, that is useless for quality evaluation and it is not even shown in the Figure.

Note that, due to use of pretrained IR SE 50 model, the input images need to be in a resolution 112x112. That is why the U-Net model will be producing final images in resolution 112x112. Theoretically, the system could generate images at 128x128 and then downsample them to 112x112 before passing them into SER-FIQ critic, though, as of now, it causes failures during backpropagation.

Ten copies of an image generated by U-Net are passed into SER-FIQ for evaluation. Thus, ten stochastic subnetworks of IR SE 50 model are generated and used for quality estimation. The count of ten was chosen here, as it turned out to be the best compromise between algorithm's time complexity and estimation confidence. Furthermore, memory usage of the SER-FIQ critic is huge and it was necessary to decrease the batch size from 64 to 8 image sequences per batch which training on 15 GB of GPU memory.

Practically speaking, SER-FIQ quality evaluator returns values in a range from 0.78 to 0.84 for low-quality images and values in a range from 0.85 to 0.91 for high quality images. Since this range is impractical, it was created a simple rescaler, which rescales the values into a more practical range from 0.75 to 1.0. Once transformed value is obtained, it is compared with desired value 1.0 using MSE. Final SER-FIQ loss is weighted and summed with a Pixel loss, Feature loss and Gram loss.

2.1.4 BiO-Net

Figure 2.6 shows another architecture implemented and which will be used for training on the presented MF datasets. The figure is simplified since the architecture greatly resembles U-Net already presented in Figure 2.4. As a matter of a fact, Bio-Net architecture is identical to U-Net, but it also introduces back-skip connections and recurrency. In the first iteration, back-skip connections are ignored since the decoder layers do not contain any activation yet. In the following iterations, though, decoder's activations are read and summed with encoder's activations. In order to adjust decoder's data before summing it with the encoder, a small computation is performed. It includes double connection of convolutional layer, batch normalization and ReLU. Multiple subnetworks inside back-skip connections were also tried out, but they did not yield any better results compared to a simple U-Net. It is

important to note that, the same input image sequence is passed to the BiO-Net on each iteration. Passing into BiO-Net its own outputs would be possible only in case of SF approach, but, more importantly, it would imply a different architecture. Such architecture has been tested as well, but it did not prove to work well.

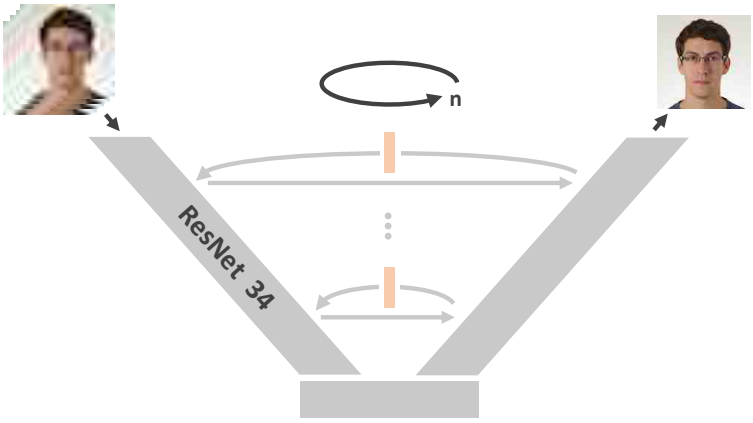

Fig. 2.6: Simplified architecture of BiO-Net.

Finally, the n in the Figure 2.6 symbolizes the number of iterations the network is supposed to perform. It was chosen $n = 2$. Larger values of n increase time complexity of the training over the reasonable amount and the results are almost unnoticeably better.

2.1.5 Feature-Merge U-Net

As a pure experiment, another architecture named Feature-Merge U-Net is presented now. As the name implies, this architecture will be merging features extracted from the input images. Motivation behind this architecture is the temptation to try out different approaches in merging input images. U-Net like architectures rely on stacking the input images in the first layer or preprocessing them in a small subnetwork. Feature-Merge architecture will demonstrate how merging low-level visual features affects the resulting quality of a prediction.

The simplified architecture is presented in the Figure 2.7. First of all, a pretrained SF U-Net model is required ResNet34 encoder will be extracted. In case of this thesis, the model was pretrained using the original SF CelebA dataset. During training the middle image from the input sequence will be passed through this SF U-Net model. Remaining images will be passed through the frozen copy of extracted ResNet34 encoder. With the help of Pytorch hooks, the training will be paused once the middle input image leaves the encoder and merge the activations

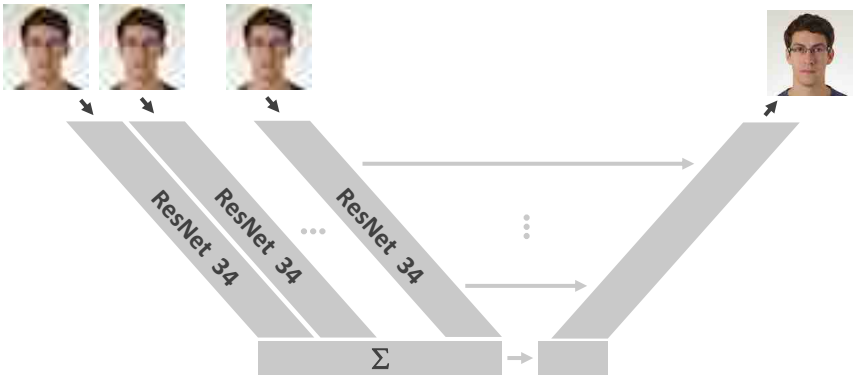

Fig. 2.7: Simplified architecture of Feature-Merge U-Net.

with activations obtained from the copy of ResNet34 encoder. In order to adjust the activations of remaining input images, they are passed through double convolutional block including batch normalization and ReLU nonlinearities. Once the features are merged, they are passed to bottleneck and decoder section of SF U-Net.

2.2 Single-Frame Finetuning

Preliminary results show that multi-frame U-Net based sharpening approaches are not able to generate images at a high resolution. Although the predictions can leave the U-Net model in any desired resolution, their quality always matches images roughly at resolution 80x80. Some of the blurry areas in the predictions could be easily further sharpened. It is just not in the capacity of the multi-frame models to perform such final finetuning, because they act more as aggregators (or at least that is how we want them to operate). The need to further enhance the resolution of the predictions is the main motivation in this chapter. Multiple single-frame models will be presented. Some of them will demonstrate custom implementation and others will be download from online sources.

2.2.1 U-Net

The first SF finetuning model to be implemented is based on U-Net network as presented in Figure 2.4. The architecture will remain the same except for the number of input channels, which will be reduced to three. Moreover, the resolution of predictions will increase to 144x144. CelebA dataset will be used during training. Input images will be first downsampled to resolution 60x60 and fed into the model. Labels will have resolution 144x144 so that they match the resolution of predictions. Although it may sound promising to train this finetuning model on the prediction

from MF U-Net models, it serves no purpose. It would not work at all. Because if it was possible, then the MF U-Net models would have learnt it in the first place.

2.2.2 PULSE

PULSE algorithm presents another possibility as a SF sharpening model. Although, multiple related works in the field of image sharpening imply that PULSE algorithm tends to be extremely creative and produces results far from the ground truth, it will be evaluated anyway. The operation of PULSE algorithm was briefly introduced in the theoretical part of this work. It will not be reimplemented, rather, the source code will be downloaded from an online source.

PULSE experiment does not need to be described in detail. PULSE will be approach as a black box. Inputs will be represented by images sharpened by multi-frame U-Net models and outputs will be presented in final comparison table. As authors recommend, PULSE will perform 100 internal iterations before it produces result. Moreover, even if PULSE fails to converge, its results will be presented.

2.2.3 PSFR-GAN

PSFR-GAN along with the pretrained model provided on GitHub is another SF finetuning model to be evaluated. Preliminary testing shows that the pretrained model works well enough for the majority of use-cases posed by this thesis. That is why it will not be reimplemented nor retrained. Just some custom utility functions will be implemented which will simplify model's usage.

Preliminary testing has also shown that the model does not correctly sharpen heavily damaged input images. The model becomes very creative and predictions no more match the input images. That is the reason why it can only by utilized as a finetuning model. Inputs to this model will be output of multi-frame U-Net models further upsampled to 512x512 using bicubic interpolation.

2.3 Multi-frame Reconstruction

Reconstruction of a single image from a sequence of damaged images is a complex task. Generally, it includes evaluating quality of images, finding the best facial image, finding the image with sharpest eyes, nose, mouth and other facial parts, aligning images and crossing over required visual features. While in the previous chapter a single neural network was used to perform the whole reconstruction task, this chapter will split the Multi-frame Reconstruction process into multiple simpler tasks. Each task will implement a dedicated neural network specifically trained

for given problem. This way, the reconstruction system will not need to rely on a single neural network and hope it will learn to generalize well all the partial tasks. Moreover, system comprised of multiple smaller subsystems is easier to modify and to fix.

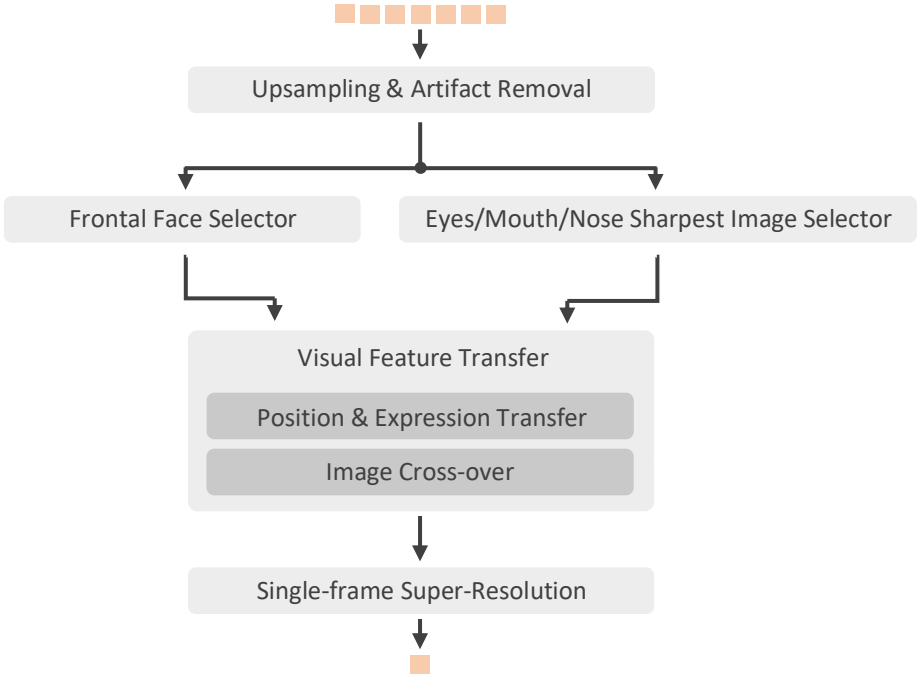

Fig. 2.8: Full pipeline of Multi-frame Reconstruction system.

High level overview of the system implemented in this chapter is presented in Figure 2.8. The first subsystem called "Upsampling & Artifact Removal" is responsible for upsampling of input images from low resolution 32x32 to a higher resolution required by the following subsystems. These systems will generally perform face masking or face alignment task and require resolution at least 512x512. Furthermore, input image sequence will most probably contain some video-compression artifacts. These artifacts need to be removed right in the first stage otherwise the following subsystems will fail. Face alignment gives better results for clear images than images damaged with dark pixels left after compression algorithms.

System "Frontal Face Selector" will be responsible for choosing frontal face image from input sequence. Generally, input sequence may contain images from different angles and it is advantageous to reconstruct the image showing frontal face rather than image showing side or back of the head. This step will not be presented in this chapter since there are no reliable metrics able to evaluate sharpness of parts of the

damaged facial image. It will have to be performed manually.

System "Eyes/Mouth/Nose Sharpest Image Selector" will be designed to select an image with sharpest eyes, mouth, nose or other facial part. Such image will be used to inpaint corresponding blurry area in the frontal face image. Selected image of higher quality will be called as source image in the following text and frontal face image will be referred to as target image. Since input sequence will contain multiple source images, they will need to be inpainted into target image in multiple iterations. Note that, the task of selecting image with sharpest facial parts will also have to be performed manually since it requires working evaluator of image quality.

"Visual Feature Transfer" block in the schematic represents a system able to transfer specific parts of the face from one image to the other. It is a complex system and needs to be split into smaller units. One unit "Position & Expression Transfer" will ensure that the source image is aligned into the position of target image. The other unit "Image Cross-over" will then perform physical transfer of the pixels from one image to the other while respecting given mask.

Although no masking system is presented in the schematic, it will be an important backbone of multiple subsystems. The purpose of masking is to select specific facial parts. Mask will be represented by a separate three-channel PNG image holding values either 0 or 255. 0 will represent pixels to be transferred between images and 255 will represent pixels not to be transferred between images. Masking system will rely on a dedicated neural network performing face segmentation.

The last step in Multi-frame Reconstruction will be "Single-frame Super-Resolution" performing final fine-tuning. It will be based on neural networks already presented in the previous chapters. Ideally, this system should use some custom-trained network not only able to enhance visual quality of the reconstructed image, but also to remove undesired artifacts created in the process. Such artifacts will be introduced by imperfect alignment followed by image cross-over or imperfect position and expression transfer.

To summarize the process of Multi-frame Reconstruction – source images with sharpest eyes, nose, mouth and other facial parts will be manually selected and inpainted into a manually selected image. But the inpainting process itself will be fully automated.

2.3.1 Face Alignment

Face alignment is an important step in almost every deep learning task working with facial images. Its aim is to center facial images to a fixed position. That way, all facial images cropped from generic pictures end up having eyes, mouth, nose and other facial parts at the same position after alignment. It helps increase the

performance of the neural network being trained. To name a few, PULSE, StyleGAN and PSFR-GAN use alignment and even require testing images to be aligned as well. Otherwise the quality of predictions is not guaranteed. In the task of Multi-frame Reconstruction, face alignment will be used by almost every block. Though, it will not be apparent to the outside world, because the aligned predictions will be aligned back during post processing.

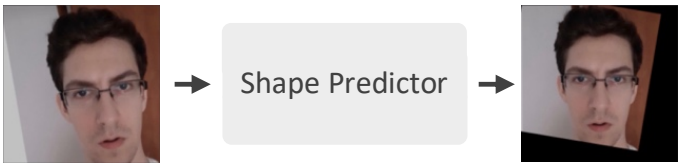

Fig. 2.9: Operation of Shape Predictor.

The majority of works in the field of deep learning uses DLIB shape predictor[2]. There are multiple versions of it differing only in the number of recognizable facial parts. Otherwise, they perform the same alignment to the same fixed position. Most used version of shape predictor recognizes 68 facial points, so called landmarks. After these landmarks are recognized in an input image, affine transformations are performed on each pixel in order to move, rotate and warp the image into standardized position. More sensitive versions of shape predictor define over 100 facial landmarks, but it is useless in case of Multi-frame Reconstruction. This thesis expects to have input images in resolution 32x32 or less and thus, the quality of images is too low for such sensitive detector. An example operation of 68-landmarks shape predictor is presented in Figure 2.9.

2.3.2 Face Masking

The purpose of face masking is to select desired parts of the facial image. This problem has already been solved in numerous related works and this thesis will utilize a face segmentation network trained on CelebA dataset available on GitHub[3]. The operation of face masking system is presented in the Figure 2.10.

As can be seen from the figure, an input image is fed into the neural network performing face segmentation on it. The immediate output is a three-channel PNG image containing values from 0 to 17. Value 0 at any position $[x, y]$ in the mask represents background pixels at the corresponding position $[x, y]$ in an input image. Value 1 represents face skin pixels, value 2 represents right eye pixels and so on.

[2]Source: http://dlib.net/intro.html
[3]Source: https://github.com/switchablenorms/CelebAMask-HQ/tree/master/face-parsing/

Fig. 2.10: Operation of Face Masking system.

Though, other subsystems of the whole Multi-frame Reconstruction pipeline will find it more useful if the masking system could return a mask only stating which pixels in an input image contain given facial parts. That is why, the final output of masking system will be a three-channel PNG image only holding values 0 and 255. Value 0 will state, that input image contains given facial pixel at corresponding position $[x, y]$. Value 255 will state, that given facial pixel is not present in input image.

There is one prerequisite for face masking system to operate reliably – input image needs to be in a good quality. This is not the case of many images which multiple subsystems of Multi-frame Reconstruction system work with. Images are often too blurry for masking system to confidently state which parts of the image belong to eyes, mouth and so forth. That is why, PULSE prediction will be computed from input image first, and passed as input to masking system. This is a bulletproof approach which never fails. The only downside is that it takes long to perform PULSE prediction just to reliably mask input image.

2.3.3 Image Crossover

The purpose of image cross-over system is to perform physical copying of pixels from one image into the other. In order for this operation to be successful, the two images need to be aligned. The process of alignment is not a responsibility of this system and will not be provided here. Since there are multiple approaches available to this problem, they will be examined and compared to one another.

The simplest approach to image cross-over is pure copying of specified pixels from image A to image B. Pixels to copy are specified using mask M. Suppose P stands for final crossed-over image, then cross-over operation can be expressed using formula 2.1:

$$P = AM + B(1 - M) \tag{2.1}$$

The greatest limitation of pure copying is the fact, that it creates clear, sharp border around inpainted region. That is why this approach will be recognized as

"Sharp Cross-over" in the rest of the thesis. Possible fix to this significant downside is application of Gaussian filter on the mask M before the cross-over process. Then, inpainted region will smoothly blend into target image. This approach will be referred to as "Soft Cross-over".

Sharp and Soft Cross-overs do not contain any complex logic. They simply copy pixels from one image to the other. Though, many times it is advantageous if crossing-over system can reason and make some adjustments in the process. Sometimes images are not perfectly aligned or they show slightly different emotions. This is the motivation why cross-over techniques based on neural networks will be evaluated as well. Related fields generally apply the idea presented in Figure 2.11. Neural network accepts nine-channel input. Three channels are allocated for one source image, another three channels are for target image to be inpainted and the last three belong to mask. Alternatively, mask can be passed in in a single channel. Neural network is trained to perform the copy operation from one image to the other. In order for the training to be successful, it is supervised by the critic. Critic evaluates the difference between pixel values of input image and prediction provided that mask value is 255. In case mask value is 0, then critic evaluates the difference between target image and the prediction. Since U-Net architecture will be used for the neural network, this approach will be referred to as "U-Net Cross-over" in the rest of the thesis.

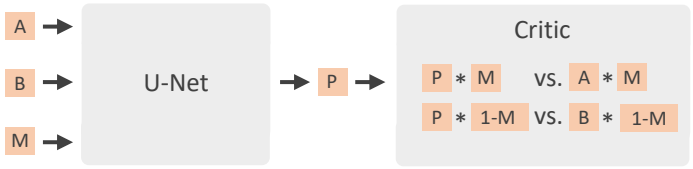

Fig. 2.11: U-Net based cross-over.

Last alternative to image cross-over methods is StyleGAN editor. As presented in the theoretical part of this work, StyleGAN editor allows to turn input images into the latent space of StyleGAN and cross them over by fusing their latent vectors.

More specifically, StyleGAN can synthetize facial images based on an input vector. As of now, the relationship between the input vector and synthetized image is not clear and is a matter of research. It is a challenging task, because StyleGAN's (input) latent space is complex and simple neural networks struggle to find patterns in it. Thus, finding input latent vector is an approximation task. Latent vector is being guessed in multiple iterations. After each iteration, it is passed into StyleGAN and generated image is compared to input image. The process keeps going until the error falls below specific threshold.

The cross-over process operates in a similar manner. Latent vectors are being guessed from both input images. Generated latent vectors are then fed into StyleGAN and synthetized images are compared to input images using provided mask. The process keeps iterating until such latent vector is found which yields StyleGAN image correctly crossed-over. In the rest of the thesis, this approach will be recognized as "StyleGAN Cross-over".

2.3.4 Position and Expression Transfer

Position and Expression Transfer system is responsible for aligning source and target image into the same position so that they can be crossed-over by Cross-over system. Ideally, the source image will be rotated and warped into the same position as target image and its expression will match the expression of face in the target image. Practically, it is challenging to fulfill these objectives and systems presented in this chapter will only respect some of the requirements. As already implied, there are multiple solutions available for this task. They will be explored and compared to each other.

The simplest approach to position transfer is presented in Figure 2.12. It only targets to handle simplest use-cases possible, i.e. frontal face images showing very similar facial expressions. In this approach, both input images are aligned to the center, then the whole face of source image is inpainted into target image and the inpainted image is aligned back to its original position. This way, the output image shows source image in the position of target image and they can be easily crossed over later on. Since both images are first aligned and then blended, this approach will be referred to as "Align and Blend" in the rest of the thesis.

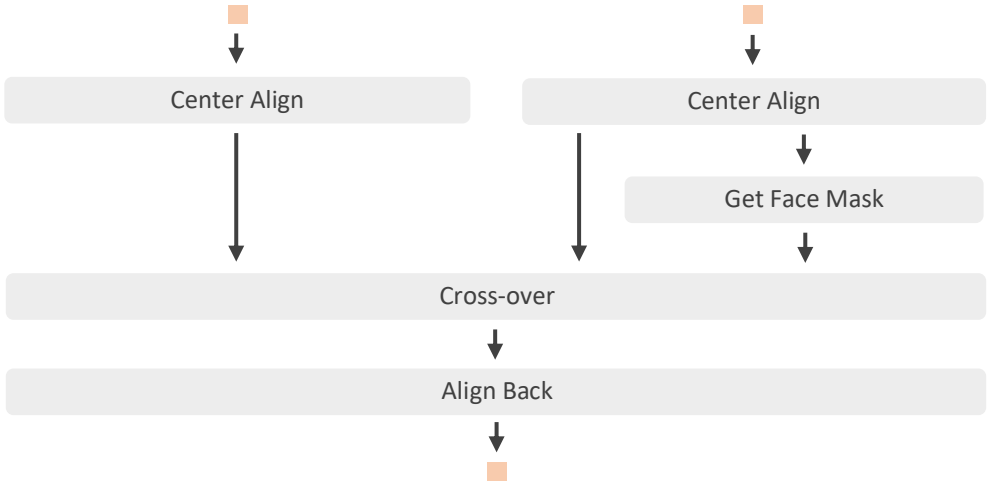

Fig. 2.12: Position and Expression Transfer using Align and Blend approach.

Very similar approach to Align and Blend has already been presented in GitHub project[4]. It uses more sensitive DLIB face landmark detector and after inpainting face from one image into the other, it also performs simple warping and color correction to enhance the outputs. This system can be thought of as a black box shown in Figure 2.13. It has two input images, source and target, and one output image showing face from source image inpainted into the target image. This system will be referred to as "DLIB".

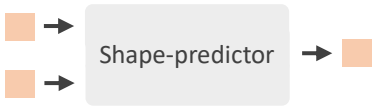

Fig. 2.13: Position and Expression Transfer using DLIB approach.

Multiple works in the field of face reenactment train neural network on multiple images taken from various angles of the same face. Such network is then used to inpaint new image (showing any face) with the face the network was trained on. Although they achieve good results, these approaches will not be explored in this thesis. This thesis expects to have a sequence of few images on its input and such number will not be high enough to train any face reenactment network. In practice such networks get trained on long video sequences so as to learn all details of person's head.

There is one alternative to face reenactment networks which claims to be able to inpaint target image from a single source image. It is called "Few-Shot" and it can be obtained from GitHub[5]. Again, it can be pictured as a black box taking two source images and returning source image in the position of target image. Moreover, it promises to handle basic expression transfer and eye movement. Again, it is shown as a black-box in Figure 2.14.

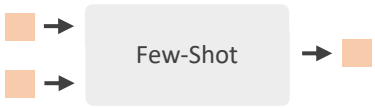

Fig. 2.14: Position and Expression Transfer using Few-Shot approach.

[4]Source: https://github.com/matthewearl/faceswap
[5]Source: https://github.com/shaoanlu/fewshot-face-translation-GAN

Finally, last alternative to position and expression transfer is provided by InterFaceGAN algorithm. InterFaceGAN is also a group of utility functions built on top of StyleGAN. It aims to modify input latent vector so that output latent vector represents a face with some facial attribute aplified or attenuated. InterFaceGAN managed to find patterns in the latent space of StyleGAN using neural networks by examining the effects of changing attributes of facial images on latent vectors. InterFaceGAN takes in a single input image and returns multiple output images with some specific attribute amplified or attenuated. For the purpose of position transfer, attribute pose will be altered. An input frontal face image will be fed into InterFace and the output images will show the same face but from different angles. This may be advantageous since no other presented system is able to fully synthetize new views of the same face. "Best Angle Selector" will choose such output image, whose view angle matches the other input image (target). Slight differences in head position are removed thanks to the alignment block and the area of whole face is swapped using soft cross-over technique. InterFaceGAN approach will be referred to in the rest of the work as "InterFace". The system diagram is presented in Figure 2.15.

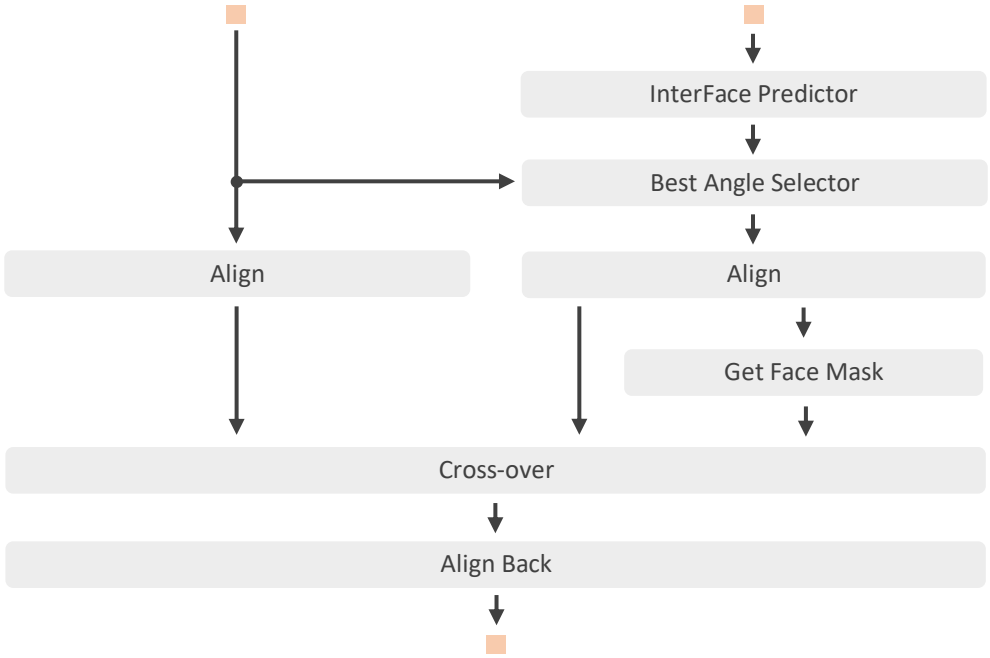

Fig. 2.15: Position and Expression Transfer using InterFaceGAN approach.

3 Results and Discussion

This chapter will be devoted to a discussion of achieved results and a comparison of the models implemented with related works. At the beginning, objective and subjective methods used for model evaluation will be presented. The remaining sections are split into the three groups corresponding to main experiment groups introduced in the previous chapter. The first group aims to evaluate implemented multi-frame U-Net models. All these models accept all input images from input sequence and sharpen the middle image. Second group discusses achievements in the field of single-frame image sharpening. Implemented models will be compared to existing related works. Finally, the third group evaluates novel approach to image reconstruction. It introduces a suite of tools able edit facial images. Thanks to these tools, an arbitrary input image can be chosen and sharpened.

3.1 Objective Evaluation Methods

The following few subchapters will present methods used for objective evaluation of generated images by the model. Peak Signal-to-Noise Ratio will be presented as well as Structural Similarity Index Measure and SER-FIQ image quality estimation.

3.1.1 Peak Signal-to-Noise Ratio

Peak Signal-to-Noise Ratio (PSNR) represents a ratio between the maximum possible signal power and the power of noise degrading the fidelity of original signal. The higher value of PSNR, the higher image quality. Since signals usually have a wide dynamic range, PSNR is expressed in a logarithmic scale. PSNR can be computed by Formula 3.1, where MAX stands for maximum possible value in RGB matrix of a input images (i.e. typically 1.0 or 255) and MSE represents the Mean Squared Error between generated image x by the model and the target image y [27].

$$PSRN(x, y) = 10 \cdot log(\frac{MAX^2}{MSE(x, y)}) \tag{3.1}$$

3.1.2 Structural Similarity

Structural Similarity Index Measure (SSIM) is an objective method for evaluating the perceived quality of a given image. It measures similarity between the generated x and target image y using Formula 3.2:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.2}$$

where

- μ_x and μ_y - the average of x and y respectively,
- μ_x^2 and μ_y^2 - the variance of x and y respectively,
- σ_{xy} - the covariance of x and y,
- C_1 and C_2 - two variables helping stabilize the division.

The implementation of SSIM presented in this thesis will use the recommended values for $C_1 = 0.01^2$ and $C_2 = 0.03^2$. Moreover, it will not compute SSIM from the whole images directly, instead, they will be separated them into smaller chunks using Gaussian windows of size 11x11 and the total SSIM will be computed by averaging obtained partial SSIM values.

SSIM index extracts 3 key features from given image - structure, luminance and contrast. Luminance can be computed by averaging all the pixel values, that is why Formula 3.2 includes parameters μ_x and μ_y. Contrast can be computed by taking the standard deviation of all the pixel values. In the formula, σ_{xy} represents a comparison of the contrasts of the input images. Finally, representation of the structure is hidden in the formula. It is not represented by any variable [28].

3.1.3 Blur Detection

A No Reference Image Blur Detection Using Cumulative Probability Blur Detection (CPBD) represents no reference image blurriness metric [29]. Authors of this metric studied human perception of images and utilized a model to state the probability of detection of blurry areas in the image. Total CPBD is then computed by cumulating the partial CPBD of all blocks of the image as shown in Formula 3.3:

$$CPBD = P(P_{BLUR} \leq P_{JNB}) = \sum_{P_{BLUR}=0}^{P_{BLUR}=P_{JNB}} P(P_{BLUR}), \tag{3.3}$$

$$P_{BLUR} = P(e_i) = 1 - exp(-|\frac{w(e_i)}{w_{JNB}(e_i)}|^{\beta}) \tag{3.4}$$

where

- e_i - edge of i-th block,
- $w(e_i)$ - width of an edge e_i,
- $w_{JNB}(e_i)$ - width of an edge with "Just Noticeable Blur",
- β - output of least squares fitting.

The implementation of CPBD was obtained from online source in accordance with the license terms.

3.1.4 SER-FIQ

SER-FIQ algorithm has already been presented in chapter SER-FIQ Loss as a method for image quality estimation. Later, in chapter U-Net with SER-FIQ Loss, it was also used as a critic for the U-Net model presented in previous chapters. In this short section it will be presented how to utilize SER-FIQ algorithm for quality estimation of generated images.

At first, generated image needs to be denormalized using Imagenet parameters. Unless image's resolution equals 112x112, a bicubic interpolation needs to be applied. SER-FIQ's face recognition network requires that the input images are sampled to resolution 112x112. Face network used matches exactly the one already presented in Figure 2.5. It is a pretrained network with a single fixed Dropout layer in its head. Thus, if the input image is passed through it m times, it will be obtained m different embedding vectors. For quality evaluation this thesis will be using $m = 10$. These m embedding vectors are then evaluated by SER-FIQ algorithm and the returned value represents the quality estimation. Since the values returned fall into a narrow range, they will be automatically rescaled to a range from 0.75 to 1.00.

3.2 Subjective Evaluation Method

Subjective evaluation will be purely based on our opinion. The generated and target images will be compared and judged based on which parts of the image were restored well and where the model failed. The primary focus will be on blurriness in generated images, unwanted artifacts which model left behind and how well the generated image matches the identity in the target image.

3.3 Multi-frame U-Net Based Sharpening

Following sections evaluate U-Net model and its alternatives. These networks aim to reconstruct a middle image from input sequence of damaged images. They should extract all information possible from all input images and use it to reconstruct given middle image. The ultimate goal is to restore damaged images to such extent, that humans will not be able to recognize the difference between original and generated images.

A sample of damaged test images will be loaded and they will be restored using the trained models presented in previous chapters. Images will then be evaluated both by objective metrics and by our subjective opinion. Some example pictures will be presented as well in order to highlight where the models excel and in which tasks they fail.

3.3.1 Formal Comparison of Implemented Models

Table 3.1 and Table 3.2 represent a formal comparison of the implemented models. All the model architectures have been described in the previous chapters. The plus symbol in the name represents the fact that the outputs of the model were also passed through another finetuning model described in chapter U-Net. Words MLFDB and CelebA in the model names state which dataset was used during the training.

Due to the limitations of MLFDB test dataset explained in Visual Comparison of Implemented Models, it was found convenient to use new random 1000 images from CelebA dataset downsampled to 32x32 using bicubic interpolation for testing purposes. MF dataset from the SF images was created by applying random transformations multiple times to the same image.

Although PSNR and SSIM metrics are generally known not to evaluate image quality the way humans perceive it, their evaluations are accurate in this case. Comparing the numbers with actual results presented in Figure 3.1 a simple U-Net architecture trained with CelebA dataset with finetuning performs the best. On the other hand, the same architecture trained on MLFDB performs the worst. More on this matter in the Visual Comparison of Implemented Models section.

SER-FIQ algorithm as a metric fails. Its numerical results do match the visual results. The models performing relatively worse are marked by SER-FIQ as the better models.

	MSE	PSNR	SSIM	SER-FIQ
	[-]	[dB]	[-]	[%]
U-Net (MLFDB)	180.54	24.05	0.73	38.29
U-Net (CelebA)	80.32	26.42	0.82	38.21
U-Net+ (CelebA)	76.47	26.48	0.82	37.63
SER-FIQ+ (CelebA)	90.37	25.40	0.81	38.64
BiO-Net+ (CelebA)	105.20	25.27	0.78	38.27
Feature-Net+ (CelebA)	78.93	26.46	0.82	37.63

Tab. 3.1: Comparison of MSE, PSNR, SSIM and SER-FIQ metrics of implemented models.

3.3.2 Visual Comparison of Implemented Models

The following section is devoted to a visual evaluation of the achieved results. For the testing purposes a few images from the testing directory of MLFDB dataset were chosen. The images were not chosen randomly, but purposely so as to show how

	CPBD before	CPBD after	CPDB diff
	[-]	[-]	[-]
U-Net (MLFDB)	0.09	0.02	-0.07
U-Net (CelebA)	0.09	0.48	0.39
U-Net+ (CelebA)	0.09	0.55	0.46
SER-FIQ+ (CelebA)	0.09	0.52	0.43
BiO-Net+ (CelebA)	0.09	0.38	0.29
Feature-Net+ (CelebA)	0.09	0.54	0.45

Tab. 3.2: Comparison of CPBD Blur Detection metric of implemented models.

models perform on faces of different age, gender and race. Note that, the Figure 3.1 only shows the middle image from the input image sequence. Remaining images were skipped to save up the space. Moreover, all the input images look similar except for the last sequence g. In this sequence, the middle image is damaged the most compared to other input images in the sequence. This image will be used to show how well the models copy features from other input images while restoring the middle image. Finally note that, input image sequence c is damaged using video compression while input images a, b, d, e, f, g are damaged using JPG compression.

Rows in the Figure 3.1 start with a middle input picture in the resolution 32x32 – the same way as the models receive their inputs. Image on the right to the input image is the label image. Remaining columns in the row represent outputs of the models, i.e. their predictions.

Figure 3.2 show the level of details implemented models can generate. For comparison MF U-Net model with SF finetuning and MF Feature-Merge U-Net with SF finetuning were chosen. As it can be seen from the figure, both the models perform the same. Though, it will be shown later, that Feature-Merge U-Net handles MF input poorly and thus it implies, that U-Net model ignores other input images except for the middle one.

3.3.3 U-Net

In the Figure 3.1 "U-Net (MLFDB)" represents architecture described in chapter U-Net and trained on MLFDB dataset. This model performs the worst out of all presented models. The predictions are still blurry and their quality actually matches the quality of input images upsampled to 128x128 using bicubic interpolation. The only task the model is good at is removing video compression artifacts (see prediction c). The reason why MLFDB fails to train the model is most probably because it only contains around 6,000 unique identities. It seems that the model overfit and it

Fig. 3.1: Examples of predictions of implemented models.

did not learn to generalize the sharpening task on an arbitrary input face image.

"U-Net (CelebA)" represents the same architecture but trained on the smaller subset of CelebA dataset. The predictions look sharp and the model operates well on the majority of input images which were tried out. The two obvious flaws of the model can be observed in the figure. Once input images are damaged using other compression techniques than the model was trained for, the quality of generated predictions becomes worse than the input images. In case of this thesis, the model was trained on images damaged using JPG compression, whereas input sequence c was damaged using random video compression. This presents a good lesson, that the training dataset needs to include a wide range of compression artifacts. The second obvious flaw is presented in the prediction g. Although all input images were in a good quality except for the middle one, the prediction's quality is poor. More

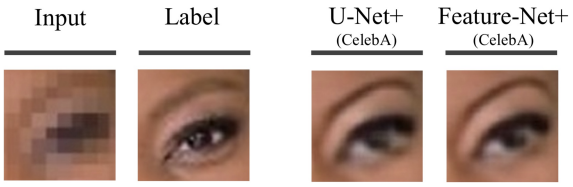

Fig. 3.2: Example of level of details of implemented models.

on t his in section Evaluating Models" MF Performance.

Finally, "U-Net+ (CelebA)" is identical to "U-Net (CelebA)", but predictions were also passed through a finetuning model described in chapter U-Net. Thanks to this approach it can be observed, that the finetuning really works. It slightly enhances the sharpness of predictions which gives them more realistic look. Though, there is one caveat – finetuning also amplifies imperfections and mistakes in predictions. Thus, in order to safely utilize finetuning, it is crucial to train the aggregator model carefully.

3.3.4 U-Net with SER-FIQ Critic

"SER-FIQ+ (CelebA)" in Figure 3.1 stands for architecture presented in chapter U-Net with SER-FIQ Loss and the predictions were also enhanced using finetuning. Compared to U-Net model with finetuning, this model gives slightly worse results. It is possible that it is not caused because of SER-FIQ critic. The culprit in this case will most probably be a small batch size used during training because SER-FIQ critic consumed lots of GPU memory. Though, the summary in this case would be, that SER-FIQ as a critic is absolutely useless. It does not guide the model to produce higher quality predictions.

A practical question may be what would happen if SEF-FIQ critic was used without the help of pixel and perceptual loss? Such experiment was performed and the results are presented in the Figure 3.3. Seven images on the left represent the input to the model followed by model's prediction on the right. This experiments has clearly highlighted the greatest imperfection of SER-FIQ as critic. Since it utilizes unsupervised approach and does not need label image to evaluate the prediction, the model being trained on such critic learns to produce utterly anything that outsmarts SER-FIQ. Thus, SER-FIQ critic cannot operate on its own. It needs to be combined with other critics so as to produce meaningful predictions.

Fig. 3.3: Example prediction of a model trained with SER-FIQ only.

3.3.5 BiO-Net

"BiO-Net+ (CelebA)" in Figure 3.1 represents architecture described in chapter BiO-Net and the predictions were also sharpened using finetuning. All the BiO-Net's predictions imply the same – the model is almost not doing anything and even after finetuning the predictions still require further sharpening. For some reason BiO-Net does not operate well for MF input. On the other hand, when the same model was trained for SF input, it definitely worked better than a simple U-Net.

3.3.6 Feature-Merge U-Net

"Feature-Net+ (CelebA)" in Figure 3.1 refers to architecture described in chapter Feature-Merge U-Net and the predictions were also sharpened using finetuning. Although this model used considerably different approach to merging input images compared to other networks, its predictions do not look any different compared to U-Net predictions. The greatest contribution of this network is the lesson it gives us. It clearly demonstrates, that merging features 'at the end' of the encoder serves no purpose, because the network is extremely insensitive to any input at such stage. In order to explain this idea more visually, the noise was fed instead of input images to the network except for the middle image. Middle image will be a regular image of a face. This input is presented in Figure 3.4 by seven images on the left. The image on the right is a final prediction. The noise did not distort the prediction at all. The quality matches the prediction d in Figure 3.1. This experiment reveals a big concern – the models learn to ignore other input images and perform SF sharpening on the middle image. More on this in the following section Evaluating Models" MF Performance.

3.3.7 Evaluating Models" MF Performance

In the Figure 3.4 it can be seen that feeding noise instead of some input images did not affect the quality of prediction. This gave the motivation to perform few more tests to demonstrate how well MF models handle MF input.

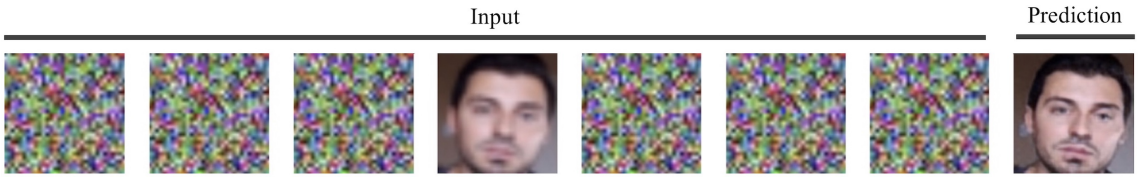

Fig. 3.4: Feeding noise into Feature-Merge U-Net.

The first test performed was an application of U-Net model on an inpainting task. The subset of CelebA dataset was modified to contain white squares and circles at random locations and trained the model on it. An example input and prediction is presented in Figure 3.5. As it can be seen, the model easily learnt to restore the original image.

Fig. 3.5: Training U-Net to inpaint random crops.

Further tests were carried out. This time CelebA dataset was used but with random blurry sections instead of white patches and trained the model. Example input and prediction is presented in Figure 3.6. Though, the performance is not as good as previously. The original image was not fully correctly restored.

Fig. 3.6: Training U-Net to inpaint random blurry sections.

The conclusion is this case it that the models can learn to merge features from multiple input images as long as the task is simple. For example, models can easily learn to inpaint large white square in the input image. On the other hand, blurry squares were not easy to recognize – the models needed to reason which parts of input image to inpaint. And they fail at it. The explanation is simple – the critics are not made for this. By looking more carefully at the distorted prediction in

Figure 3.6, it becomes clear that the pixel loss of such prediction would be small, because the colors of pixels match the label image relatively well. Feature loss would be small too, because the prediction contains nose, mouth and other features which face is supposed to have. Gram loss would be small as well, because the style of the prediction and the label image is the same. Thus, the model believes it is doing a good job. In order to convince it of the opposite, more suitable critic needs to be found.

3.3.8 Summary

The greatest observation after comparing the models is that a simple U-Net architecture can easily outperform its modifications. And even if the modifications would yield better results, the improvements would be subtle. For this reason U-Nets should be preferred to work with even in the future.

The discussion section mostly relied on a subjective evaluation of the implemented models. It was found that objective metrics are often misleading and related works have come to the same conclusion as well. It is not difficult to observe PSNR and SSIM metrics evaluating an image of lower visual quality as superior to an image of considerably higher quality. SER-FIQ metric seemed promising while performing single-image evaluations, but statistically, it gives worse results than PSNR. Due to unsatisfying image quality evaluation of the mentioned objective metrics, subjective opinions were presented instead.

When it comes to the dataset, it was found that it is absolutely essential that the dataset consists of numerous unique identities. For moderate training it is recommended to use around 60,000 unique identities of different age, gender and race. It is also crucial to damage the images using various compression algorithms. A great help would be having dataset of faces from different angles. It would greatly simplify training true MF models.

Finally, in order to build MF models, a better way of feeding input images into the model needs to be found. Stacking the images and passing them into the network directly performs poorly. This point also implies implementing a better critic. Although MSE ensures that predictions resemble the ground-truth images, it has a huge blurring effect. It smooths out the areas of high variance so that the predicted image is, on average, more pixelwise correct. In fact, the ideal solution according to MSE is a pixelwise average of a super-resolved input image which downscales correctly to the LR input. The drawbacks of MSE are slightly compensated by the feature loss. Though, a critic being able to evaluate how well the predictions resemble a human face is still the key point to success.

3.4 Single-frame Finetuning

The following chapter discusses results of single-frame finetuning models. Visual comparison is presented in Figure 3.7 and formal comparison is presented in Table 3.3. LR input image sequences at resolution 32x32 were passed through MF U-Net model and the prediction at resolution 128x128, "Input" image in the figure, was then fed into SF finetuning model. In case of PSFR-GAN, the prediction from MF U-Net model was first upsampled to resolution 512x512 and then passed into the finetuning network. "Ground Truth" image in the figure represents original high-resolution image from CelebA dataset.

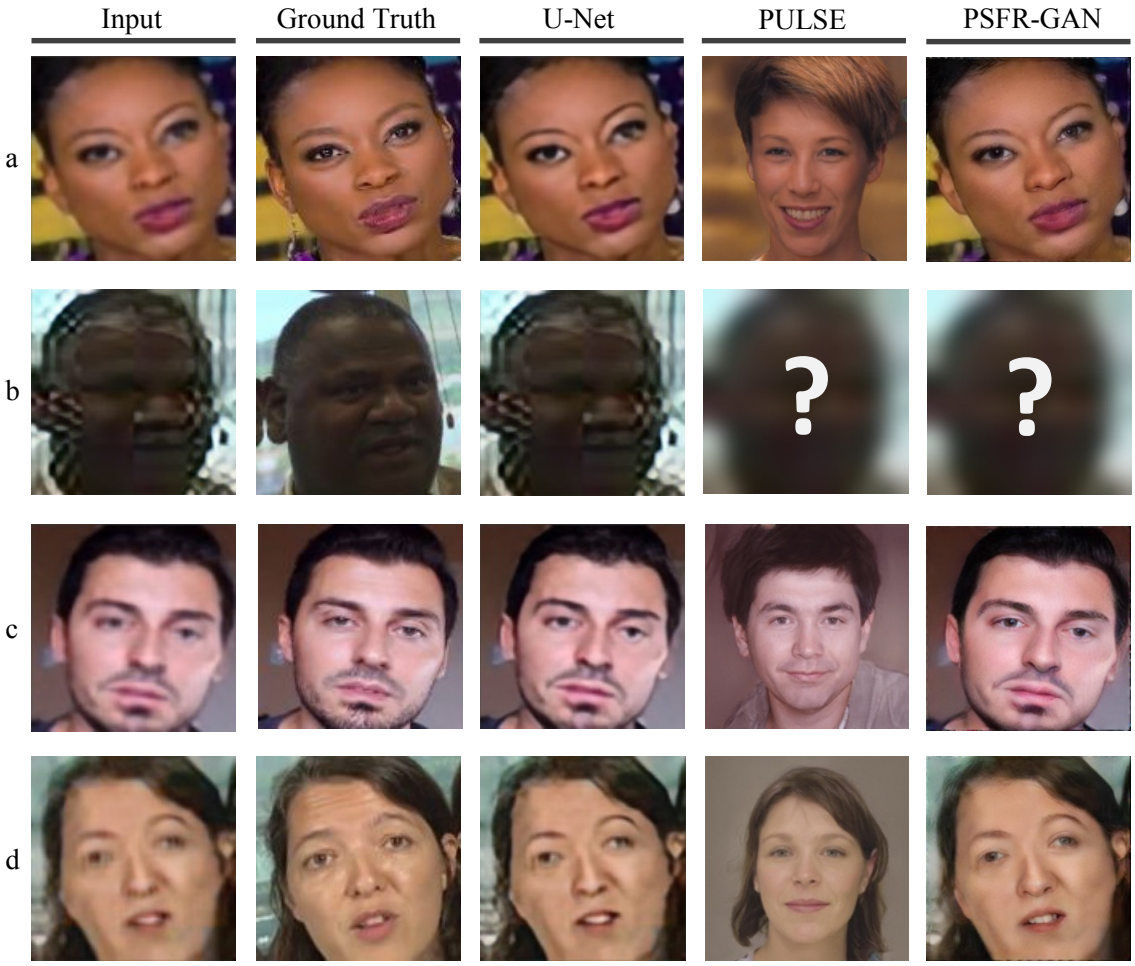

Fig. 3.7: Visual comparison of SF finetuning models.

	PSNR	SSIM	SER-FIQ
	[dB]	[-]	[%]
U-Net	27.23	0.88	37.18
PULSE	14.97	0.66	47.57
PSFR-GAN	24.83	0.84	36.84

Tab. 3.3: Formal comparison of SF finetuning models.

3.4.1 U-Net

U-Net finetuning model slightly improves the visual quality of input images and gives them more realistic look. It is important to note that, this architecture cannot do much better. Trying to train it for higher resolution or using bigger training dataset makes no significant difference. As a matter of the fact, due to increasing resolution of labels, the batch size needs to decrease and it has negative impact on the predictions. There is one caveat when using U-Net finetuning model – finetuning also amplifies imperfections and mistakes in predictions. Thus, in order to safely utilize finetuning, it is essential to train the MF aggregator models carefully so that they do not include undesired artifacts in their predictions.

U-Net finetuning has reached the highest value of PSNR and SSIM metric out of the models tested. The value of PSNR is high because the sharpened prediction is not much different from the original input image. On one hand it is desirable since it implies that U-Net does not change identity, though, on the other hand it underlines the fact that U-Net does poor sharpening job. PSFR-GAN, giving much sharper predictions of slightly different the identity results in worse PSNR rating.

3.4.2 PULSE

As can be seen in Figure 3.7, PULSE does produce images in a high quality with lots of visual details. Though, they have nothing to do with desired ground-truth images. Note that, PULSE first aligns input image and then performs prediction. It explains why all PULSE predictions show different zoom level and view angle compared to *Input* image. Further note, that PULSE would give better results if it performed more iterations. Even after taking these facts into account, PULSE's predictions would almost always show different identity. The interesting fact is, though, if such PULSE predictions were downsampled back to a resolution 32x32, it would exactly match the input image also downsampled to 32x32. PULSE simply demonstrates the ambiguity faced during sharpening process. This ambiguity is even better demonstrated by the fact, that PULSE can generate n unique images at a resolution 1024x1024 for a single input image.

PULSE is a leader in SER-FIQ metric rating. The reason for it is the fact, that SER-FIQ does not take into account ground-truth image. SER-FIQ only evaluates visual image quality regardless of the identity. And, PULSE truly produces images at highest quality with resolution 1024x1024.

3.4.3 PSFR-GAN

"PSFR-GAN" in Figure 3.7 represents architecture described in chapter PSFR-GAN. PSFR-GAN gives predictions in a very high quality. The textures are clear and do not need any extra sharpening. All the salient features of the face hidden in the blurry parts of the image are greatly amplified. Though, there is a significant drawback to this strong sharpening nature of PSFR-GAN. Input images which were not sharpened enough by the MF U-Net model are finetuned by PSFR-GAN incorrectly. All the mistakes introduced in the predictions produced by MF U-Net model are greatly highlighted by PSFR-GAN. Note that, in case of image *b*, the "Input" image was so damaged that PSFR-GAN preprocessor could not recognize any face in it and terminated further sharpening process.

PSFR-GAN is a powerful SF sharpening tool, but it should be only used for final "soft" finetuning. Alternatively, it could be retrained on a dataset with labels in a lower quality. That way the sharpening nature of the whole model could be adjusted.

3.4.4 Summary

Single-frame finetuning plays an important role in final image reconstruction step. It does a good job in sharpening fine details of input image. Overall downside which applies to all the tested methods is that finetuning has a tendency to amplify imperfections and unwanted artifacts left in the images after previous reconstruction steps. It is challenging to build such sharpening model which would reduce this significant drawback. Such model would need to understand human perception of the face perform sharpening accordingly.

3.5 Multi-frame Reconstruction

The following sections evaluate the whole Multi-frame Reconstruction system with its subsystems. Multiple alternatives to image cross-over techniques and position and expression transfer will be compared to each other. Since there are no tools or metrics available for evaluation of image cross-over quality, only subjective opinion will be presented. Note that, for testing purposes images of faces from HeadPose

Annotations dataset [30] will be used as well. The dataset provides pictures of the same identity taken from different angles.

3.5.1 Image Cross-over

The purpose of image cross-over is to transfer visual features from source image the target image according to mask. Four different approaches have been tested and the example results are presented in Figure 3.8 and Table 3.4. Both input images have been aligned into the same position and target image always contains blurry region. In case of target image in the first row it is an area of nose and in case of the second-row image, it is an area of eyes and eye brows. Source image is generally expected to show the same identity as target image, but for this comparison the identities were chosen different. To be more specific, source image is PULSE prediction of target image.

Fig. 3.8: Visual comparison of Image Cross-over methods.

	PSNR	SSIM	SER-FIQ
	[dB]	[-]	[%]
U-Net	21.86	0.82	41.82
StyleGAN	19.07	0.78	40.73
Sharp	21.90	0.81	40.87
Soft	21.98	0.81	42.14

Tab. 3.4: Formal comparison of Image Cross-over methods.

"U-Net" column represents cross-over results when custom U-Net architecture specifically trained for this task was applied. As can be seen, results are not satisfactory. U-Net tends to inpaint areas specified by mask with generic shapes. It does not really copy visual features from source image into the target. After more

thorough testing, when, for example, eye color between source and target image differs, the prediction always ends up having generic eyes. Generic eyes in a sense, that color and the shape match the color and shape of the majority of eyes in the training dataset. U-Net as cross-over system fails. The reason why again lies in critics which cannot effectively evaluate the quality of cross-over image and do not force U-Net to train well enough.

"StyleGAN" column represents cross-over results of a system based on StyleGAN editor. I.e. latent vectors of both input images are created first and then fused in multiple iterations. Finally, resulting latent vector is passed through StyleGAN and new synthetized image is generated. This image is supposed to have visual features of both the input images while respecting the mask. In case of inpainting large regions, such as nose, the cross-over works relatively well. Though, in case of small regions, inpainting fails completely. General conclusion is, that the bigger region to inpaint, the better cross-over results. In case of small regions such as eyes or eye brows, inpainting is fully ignored. More extensive testing has also shown, that StyleGAN editors does not truly replace visual features of target image, it rather inpaints it with some level of intelligence. For example, in case of crossing over eyes (in large images of eyes), they are not fully replaced, just their color is transferred instead. This makes StyleGAN editor good cross-over system for some use-cases.

"Sharp" column stands for simple copying process in order to transfer visual features. It works well in terms of quality of transferred source features into target image. Source pixels are transferred with maximum reliability. On the other hand, very sharp and distinct edges are created around the border of inpainted area, which makes the image look unreal and slightly changes the perception of the same identity. Moreover, color difference between source and target image has a significant effect on final image. The same negative effect has any imperfection in alignment of both images and difference in facial expressions. Though, this cross-over system serves well enough for many use-cases.

"Soft" column shows the results of the same system as sharp cross-over, but Gaussian filter was applied on the mask before blending. It has the same advantages as sharp cross-over system, plus it removes the undesired border around inpainted area. Inpainted areas fit in correctly and do not change the identity. This method serves well enough too provided that alignment and color of images is perfectly matched. The slight downside of this approach is presented in case of crossing over small areas, such as eyes. Gaussian filter should not operate with fixed kernel value. It should be adjusted along with padding parameter according to the size of inpainted area. Simply put, some type of intelligence is required to adjust the Gaussian filter parameters before it is applied on the mask.

"Soft" image cross-over method has been rated by objective metrics as the best

method. It gives the best results not only formally, but subjectively as well. Moreover, it is simple to implement and fast to run.

3.5.2 Position and Expression Transfer

The goal of Position and Expression Transfer system is to align the two source images into the same position and adjust their expressions so that cross-over system can then copy visual features between them. Multiple different approaches have been tested and will be evaluated in this section.

The simplest approach presented under name "Align and Blend" aligns both the images to the center, copies whole facial area from source image into the target and aligns images back. This way face of target image is swapped with source image and aligned to the original position. Images taken from internal stages of this system are presented in Figure 3.9. "Source" image represents face having some specific area sharper (e.g. nose in this case). "Target" image represents image to be inpainted with visual features taken from source image. Both the images were cropped out from different pictures. First step in position transfer is to align them to center, see "S-Aligned" and "T-Aligned". Next step is to generate "Mask" of the whole target face and perform soft cross-over between images. "Cross-over" represents blended image. Since it is still aligned, it needs to be aligned back using inverse transformations to the ones used for alignment of target image. Unaligned image is presented as "C-Unaligned".

Fig. 3.9: Step by step operation of Align and Blend method.

Inner workings of "Few-Shot" and "DLIB" approaches cannot be presented since they operate as black boxes. FSGAN [31], relatively new network for face reenactment, seemed perfectly suited for this task as well. Though, the library contains some bugs and it did not work at all. Moreover, it requires that the target image is a video. It is not a big limitation but serves another reason why not to utilize FSGAN as position and expression transfer system.

The operation of "InterFace" system is based on one of the first methods able to modify StyleGAN's output images in controllable manner. Given an input image, InterFace generates image's latent vector first. Then it modifies the latent vector

in multiple iteration so that StyleGAN generates the same image but from different view angle. Once numerous images of the same identity are generated, image with the most suitable view angle needs to be picked manually (as of now). Chosen image should perfectly match the view angle of target image. Generated and target image are then aligned to the center, the area of whole face is crossed over and the blended image is aligned back. The main advantage of InterFace is that it is the only system able to synthetize new views of the same identity. Though, it is still very limited. It works well enough only for images showing frontal face image, see Figure 3.10.

Fig. 3.10: Operation of InterFace method.

Images showing side of the face cannot be used for InterFace synthesis. Not because image taken from side does not carry enough information to reconstruct frontal face image. This limitation comes from InterFace algorithm directly. The algorithm always tries to find a face in input image even if the image shows side of the face. That is why it always fails in such scenarios. See Figure 3.11 as an example of such failure.

Fig. 3.11: Operation of InterFace method when it fails.

Final comparison of position and expression transfer methods is presented in Figure 3.12 and Table 3.5. First two columns in the figure show source image and target image. The goal was to shape source image so that it is in the same angle as target image and shows similar expression. "A & B" represents the result of Align and Blend system. This system works well for input images which can be easily aligned and show roughly same facial expressions. Typical use-case may be video sequences. Note that, no expression transfer is performed here. Moreover, if any of the input images shows side of the face, the system completely fails such as example in the second row.

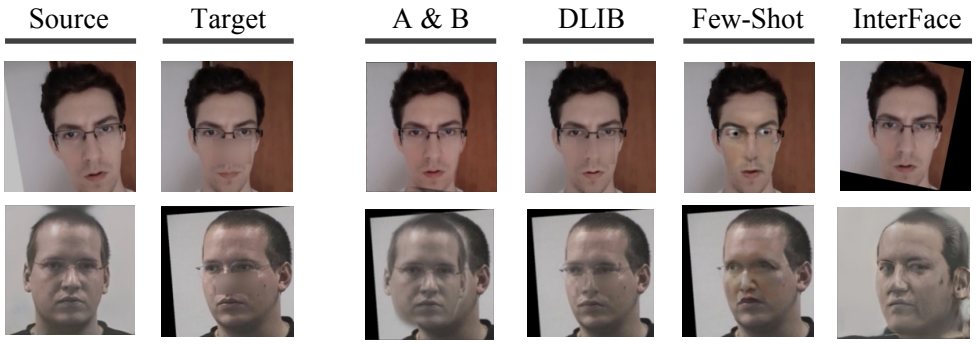

Fig. 3.12: Visual comparison of Position and Expression Transfer methods.

	PSNR	SSIM	SER-FIQ
	[dB]	[-]	[%]
A & B	20.86	0.78	40.11
DLIB	21.16	0.79	42.19
Few-Shot	18.96	0.77	43.08
InterFace	22.49	0.82	45.27

Tab. 3.5: Formal comparison of Position and Expression Transfer methods.

"DLIB" column represents predictions of DLIB system which promises to swap faces using a single pair of images. This system resembles Align and Blend system, but it also performs extra warping and rotating transformation. DLIB prediction perfectly matches Align and Blend prediction if input images do not require any complex rotations. In case they do, the predictions fail. As a clear example of failure second row of DLIB column serves great purpose. Area of nose is not rotated at all, it was simply copied over as for Align and Blend method. More extensive testing has shown that DLIB system is not useful for position and expression transfer.

"Few-Shot" column shows prediction of Few-Shot system. It also promises to perform face reenactment using two input images, but in practice it fails in absolutely all testing images.

Finally, "InterFace" column represents prediction of InterFace system. It performs equally well as Align and Blend system for simple use-cases. In case of more complex use-cases, such as second row, InterFace is the only system able to perform position transfer relatively well. It still has many limitations and faults, such as glasses were removed from the face, birth mark on right cheek became considerably bigger, some discontinuities can be found in the area of hair and identity is not exactly the same as in source image. Finally, InterFace cannot change head pose in

vertical direction, just along horizontal axis. For real use-cases, this may be considerable limitation. Though, it is a matter of time, when authors further enhance it.

InterFace method has reached the highest value of PSNR, SSIM and SER-FIQ metric. Since other methods than InterFace cannot generate new view angles, InterFace beats them during testing. These metrics also show that Align and Blend, DLIB and Few-Shot methods are only useful for use-cases when source and target image show the frontal face taken from roughly the same angle.

3.5.3 Full Pipeline

The following text discusses the full pipeline of Multi-frame Reconstruction system on a simple example. Input will be represented by three images taken from a video sequence in resolution 32x32 showing frontal face. Generally, any number of input images is acceptable. Moreover, each input image will contain some blurry areas.

The whole experiment is captured in Figure 3.13. The goal is to inpaint the middle input image damaged with three blurry areas using visual features taken from first and third input image, each containing a single blurry area. At the same time, blurry areas from other images cannot be transferred into the target image. The first step in this process is application of U-Net sharpening model trained on MLFDB dataset annotated as "U-Net" in the figure. Although it poorly enhances image quality, its main role is to remove video-compression artifacts. These artifacts would have negative impact on the following systems if not removed. The third row in figure shows cross-over process between target image (second input image) and source image (first or third input image) according to mask. The forth row in the figure shows another iteration of cross-over process. Note that, there can be any number of these iterations. At the end of second cross-over iteration, input image is inpainted will all sharper regions of other input images. Finally, single-frame sharpening is performed. For this step PSFR-GAN was utilized. In both the scenarios, it did not work perfectly. There is still lots of room for improvement. Though, PSFR-GAN did not even try sharpening the first input image because it could not find a picture of face.

Note that, masks in case of male image sequence were created manually. In case of female image sequence, masks were created using automated process. As can be seen from the pictures, manual masks capture larger areas than necessary. It has a positive impact on final image quality.

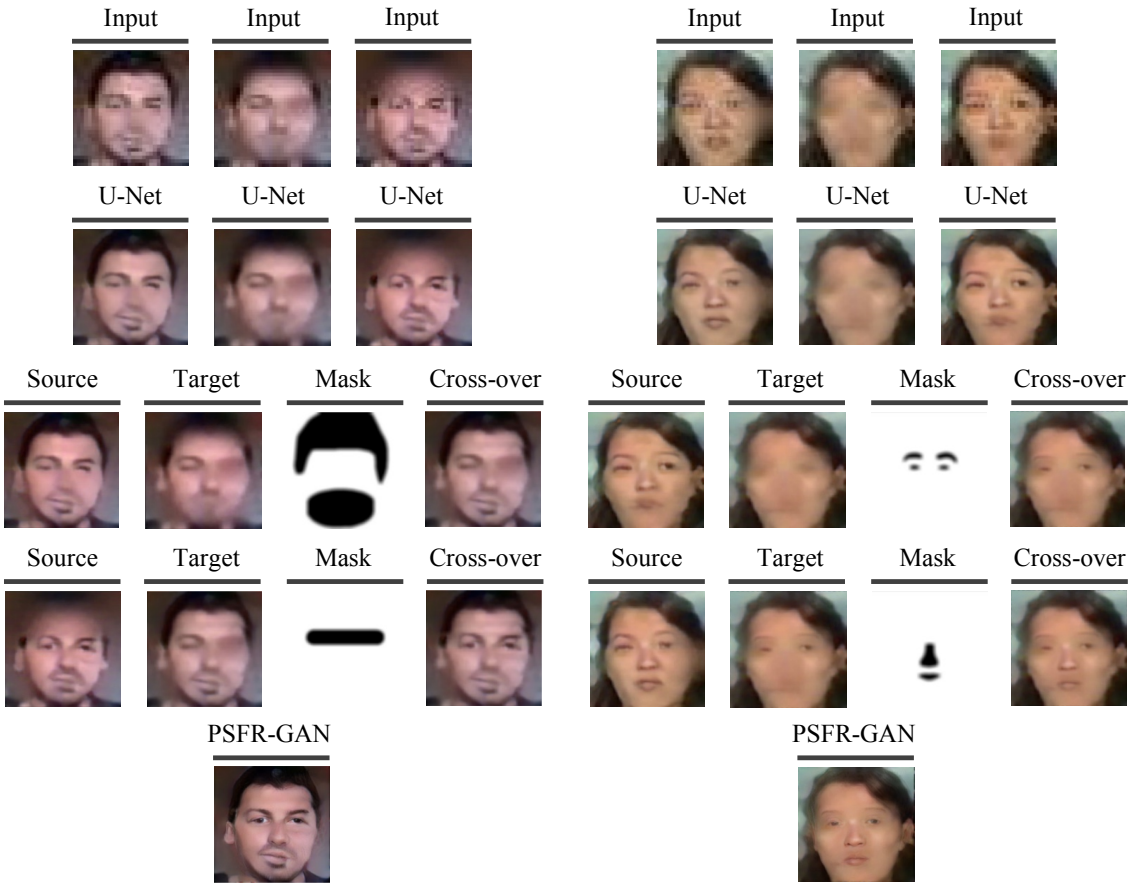

Fig. 3.13: Example operation of the whole Multi-frame Reconstruction system.

3.5.4 Summary

Multi-frame Reconstruction system presents new way of approaching task of image restoration from multiple damaged images. Instead leaving this complex task to a single multi-frame neural network, multi-frame Reconstruction system splits the task into smaller processes and designs a dedicated subsystem for each process. This way, it is easier to have control over each subtask and replace some module with better alternative. Overall, multi-frame reconstruction works well only for simple use-cases, i.e. input images showing frontal face with similar expression. The tools required for this task still have not been developed and it is a challenging task for a single person to develop and finetune each module.

Currently, there are multiple possibilities how to enhance Multi-frame Reconstruction system. First of all, it has again proven advantageous to have a network able to remove video-compression artifacts from input image. U-Net trained on MLDFB works well, but it could be still greatly enhanced by training on much larger dataset.

On the other hand, enhancing single-frame finetuning network used at the end

of Multi-frame Reconstruction process is not needed. Numerous researchers are working on this task and will soon introduce better solutions. When it comes to manual steps in Multi-frame Reconstruction process, they all could be automated. These processes are: choosing the most suitable image from input sequence for reconstruction, choosing images with sharpest eyes, nose, mouth or other facial parts and choosing the most suitable InterFace prediction. These have not been automated yet, because they are simple for user to perform, but challenging to implement in code.

Finally, the most challenging part of Multi-frame Reconstruction process is position and expression transfer. Methods presented in this thesis work poorly on real-world scenarios. This step requires lots of research in the field of single image face reenactment. Although, many papers claim to provide satisfactory results, real-world testing fails. It is very likely that StyleGAN and tools built on top of it will soon offer unique and finally working solution.

3.6 Multi-frame Models vs. Multi-frame Reconstruction

In order to compare multi-frame models and Multi-frame Reconstruction system, new testing dataset will be used. It contains 90 image sequences at resolution 32x32 from MLFDB testing dataset. Each sequence was manually edited to include blurry areas at different places of the face. The middle image from the sequence contains at least three blurry areas while other input images contain just a single blurry area. The idea is to successfully reconstruct the middle image from other input images. Ideally, multi-frame models should extract all the visual features and encode them into a sharp image. On the other hand, Multi-frame Reconstruction system will perform the same operations explicitly. Resolution of reconstructed images will be 128x128.

An example input sequence and visual comparison of multi-frame models and multi-frame reconstruction system is presented in Figure 3.14. As can be seen from the figure, multi-frame U-Net model ignores other input images than the middle one. That is why the prediction is blurry even though sharp visual features were provided in the other input images. Multi-frame Reconstruction system does better job. Though, final output image could be restored even better. The root of imperfection in the male example is final finetuning stage. It negatively affects the original identity. Details can be seen when the image is zoomed in. In case of the female example, the root of imperfection is the masking system. Unlike for male example, where masks were prepared manually, in case of female example an auto-

mated computation of masks was applied. Thus, masks do not specify areas broader than, for example, eye balls, which explains why the prediction looks patched up. Still, Multi-frame Reconstruction system gives better visual results and it is more suitable for image restoration from multiple images.

Fig. 3.14: Visual comparison of MF U-Net and MF Reconstruction system.

Formal comparison of all the systems implemented is presented in Table 3.6. It includes all the objective metrics presented before as well as results of subjective questionnaire. 38 people were asked two questions – Which prediction out of all implemented methods results in images of highest quality? Which prediction out of all implemented methods matches the ground-truth image the most?

PSNR of all the methods implemented stays in range from 17 dB to 18.5 dB which is low for image sharpening systems. It is caused by the fact that the center image in the sequence was damaged to such an extent, that it could not be used for facial reconstruction without other input images. SSIM metric with values around 0.77 is also almost the same for all the tested methods. SER-FIQ assesses reconstructed images with prediction quality of more than 36 % except for Multi-frame Reconstruction system. It reached higher value 46.52 %. MSE measures pixel differences between generated and ground-truth images. In case of Multi-frame Reconstruction system, MSE is high because final reconstructed image is a union of other input images which are not aligned with the ground-truth image. The difference between sharpness (CPBD) before and after reconstructing input image results in values in range from -0.0396 to 0.4166. Negative values were measured in case of interpolation techniques, moderate values were measured in case of U-Net models and the highest value was achieved in case of Multi-frame Reconstruction system.

Subjective survey shows, that Multi-frame Reconstruction system returns the sharpest images and the identity in reconstructed image matches the ground-truth image the most. U-Net model trained on a custom CelebA dataset was ranked as number two in the survey. Finally, U-Net model trained on MLFDB dataset and Bicubic Interpolation were also chosen as valid systems for image reconstruction.

	Objective Metrics							Subjective Metrics	
	PSNR	SSIM	SER-FIQ	MSE	CPBD before	CPDB after	CPBD diff	Highest visual quality	Best matches original identity
	[dB]	[-]	[%]	[-]	[-]	[-]	[-]	[%]	[%]
Bilinear Interpolation	18.12	0.78	36.09	185.69	0.0742	0.0346	-0.0396	0	0
Bicubic Interpolation	18.27	0.78	36.02	178.38	0.0742	0.0591	-0.0151	0	10.52
U-Net (MLFDB)	16.05	0.70	37.04	236.84	0.0742	0.0624	-0.0118	0	0
U-Net (CelebA)	17.15	0.77	36.93	190.63	0.0742	0.2690	0.1948	15.79	0
U-Net+ (CelebA)	17.26	0.77	36.83	196.71	0.0742	0.3173	0.2431	26.32	36.84
SER-FIQ+ (CelebA)	16.88	0.72	37.02	201.67	0.0742	0.2764	0.2022	0	0
BiO-Net+ (CelebA)	16.53	0.71	36.94	203.43	0.0742	0.2281	0.1539	0	0
Feature-Net+ (CelebA)	17.28	0.75	36.79	194.28	0.0742	0.3201	0.2459	0	0
MF Reconstruction (A & B, Soft cross-over, PSFR-GAN)	17.11	0.77	46.52	221.46	0.0742	0.4908	0.4166	57.89	52.63

Tab. 3.6: Formal comparison of MF U-Net and MF Reconstruction system.

Remaining methods do not outperform already mentioned ones according to the people who participated in the survey.

Results of subjective and objective metrics do not match. Multi-frame Reconstruction system was chosen as the best approach when considering subjective opinion of a group of people. Though, objective metrics except for SER-FIQ and CPBD do not rate it any better compared to other implemented systems. Generally speaking, based on the results presented in the table, objective metrics can only help roughly categorize the quality of predictions. When the difference between two values is high, then the 'better' number correctly identifies superior approach. Though, when the difference is small, it cannot be stated which approach is better based on given metric.

3.7 Summary and Future Work

The following section will summarize the strengths and drawbacks of all presented methods for image reconstruction from a sequence of damaged images. It will also discuss how these methods could be improved in future works.

Implemented U-Net model and its alternatives generally perform the same and the visual quality of the results is not surprising. The quality is slightly better compared to simple upsampling using bicubic interpolation. Although, the models could be trained to increase the quality considerably more, they would become creative and start changing identity of person in the image. What these models truly lack is the ability to extract information from other input images. Ideally, new architecture should be introduced which would be able to perform semantic analysis of the face in the input images and reconstruct final facial image. Though, no works have focused on this domain yet and it requires lots of experimenting.

MLFDB dataset's main strength is the variety of video-compression algorithms it included in the images. It really helps the model train for real-world images. Though, it could be greatly enhanced if the size of dataset raised to at least 60,000 sequences. Renowned datasets use typically around 200,000 to train models for real-world application. Moreover, since testing dataset contains identities also included in the training set, trained model's performance cannot be objectively measured. Unique identities should be used across the whole dataset. Another significant improvement would be including images from different angles. As of now, the dataset shows faces taken from a short video sequence. Thus, models tend to align images and merge them. In real world, the input to any multi-frame reconstruction system will be images of some identity taken under different lighting conditions and considerably different angles. MLFDB does not take this fact into account. Finally,

MLFDB lacks picture-compression algorithms in image sequences. They should be included as well.

Multi-frame Reconstruction system represents novel approach to image reconstruction. Although it is suited for simplest use-cases, some of its subsystems such as image cross-over and image masking operate flawlessly. The greatest bottleneck is posed by position and expression transfer. It is absolutely essential to be able to merge images taken from different angles under different lighting conditions. Currently presented implementations operate well only with facial images which are roughly the same. Alignment block works well too, but it is only able to precisely align images to the center. It would be advantageous to have a system able to align one image into the position of the other. Systems presented here are able to do it, but they need to center image first, which poses significant limitation. Finally, Multi-frame Reconstruction system will never be fully automated unless a system able to evaluate image quality is found. Such system would need to give the same results as human would. Moreover, it would need to be able to evaluate parts of the image, such as quality of eyes or hair.

Single-frame sharpening works well. Moreover, many more related woks are released on regular basis and bring in considerable improvements. Still, there are multiple improvements possible as well. For example, sharpening should be enforced as much as possible as long as the identity in the prediction is not altered. This also implies having a system able to sharpen some parts of the face more than the others. Hair or eye brows affect the identity a little. Whereas eyes or mouth have a significant effect on human perception of the face. Generally, when single-frame sharpening model is being trained, large dataset should be used. Training on small dataset always results in a model not able to work with real-world images. Finally, the most important point. Better critics are needed. Pixel loss or perceptual loss work well, but they are not able to finetune the model. Often, these critics return small loss for images of poor quality. They are simply not able to recognize the true quality of prediction well enough.

Conclusion

The aim of this thesis was to increase the quality of facial images using image sequences. The ideal goal was to restore given damaged facial image using multiple different damaged facial images taken from different angles at a different point of time. The presumption was that each input frame contained different facial parts at different visual quality.

Although, few related works have presented novel approaches to single-frame facial image sharpening, almost no works have focused on multi-frame techniques. This missing part is provided by this thesis. The first contribution of this work lies in the implementation of multiple multi-frame alternative architectures to U-Net. These models accept all input images at once and sharpen the middle image. It was found out, that all the models perform almost the same. Although they can learn to slightly increase quality of the middle image, they all absolutely fail at extracting information from other input images. No matter which critic was used and what change in U-Net architecture was made, they always ended up ignoring the other input images. Moreover, these models tend to inpaint blurry areas of middle image instead of copying visual features from other sharp input images.

It is also important to note that, novel approach to image quality assessment called SER-FIQ has been reimplemented and used as a critic and a metric. It failed in both the roles. MLFDB dataset, on the other hand, has proven useful. There is still room for improvement, though. It ought to considerably increase in size and identities in testing and evaluation sets should not resemble the identities in the training set.

Another important contribution of this thesis is a tool suite for multi-frame image reconstruction. Multi-frame Reconstruction is a novel approach to sharpening images. Instead of feeding them to some pretrained network at once, it splits this complex sharpening logic into multiple smaller tasks. For example: video-compression artifact removal, position and expression transfer from image to image, image cross-over based on the mask or final stage single-frame finetuning. Having a dedicated tool for each subtask brings in more control over what is happening and how it is performed. This is definitely the way the problem of multi-frame sharpening should be approached even in future. Though, as of now, the tools implemented operate well only for the simplest use-cases. There is still considerable room for improvement.

Multi-frame Reconstruction system is superior to multi-frame U-Net models and has a huge potential. Image sharpening is an ill-posed problem, i.e. from single input, multiple outputs can be produced. Though, just one output is correct. While U-Net models try to guess how to produce such desired output, Multi-frame recon-

struction system approaches it systematically. It also performs sharpening, but at the very end after all information from other images was successfully included in the target image.

There are two main domains of research which could push this work further in future. Having a tool evaluating facial image quality from the human perspective is absolutely essential. It is necessary as a critic, metric and a helper for automated tool selecting the sharpest image in the sequence. Though, very few related works have focused on this problem. The second domain is position and expression transfer. It is the greatest challenge in Multi-frame Reconstruction process. It aims to align images to such angle that visual features can be crossed-over. Once this problem is solved, there is no other bottleneck in the whole system.

Bibliography

[1] Tiange Xiang, Chaoyi Zhang, Dongnan Liu, Yang Song, Heng Huang, and Weidong Cai. Bio-net: Learning recurrent bi-directional connections for encoder-decoder architecture, 2020.

[2] Yuanfeng Ji, Ruimao Zhang, Zhen Li, Jiamin Ren, Shaoting Zhang, and Ping Luo. Uxnet: Searching multi-level feature aggregation for 3d medical image segmentation, 2020.

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[4] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space?, 2019.

[5] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images?, 2020.

[6] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models, 2020.

[7] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing, 2020.

[8] Chaofeng Chen, Xiaoming Li, Xianhui Lin, Yang Lingbo, Lei Zhang, and KKY Wong. Progressive semantic-aware style transformation for blind face restoration. 2020.

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[10] Anish Athalye. Neural style. `https://github.com/anishathalye/neural-style`, 2015.

[11] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness, 2020.

[12] G Mastroianni. *Interpolation processes : basic theory and applications*. Springer, Berlin, 2008.

[13] Andriy Burkov. *The hundred-page machine learning book*. Andriy Burkov, Quebec City, Canada, 2019.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[15] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[16] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017.

[17] Mohammad Sadegh Ebrahimi and Hossein Karkeh Abadi. Study of residual networks for image recognition, 2018.

[18] Chaoyue Wang, Chang Xu, Xin Yao, and Dacheng Tao. Evolutionary generative adversarial networks. *CoRR*, abs/1803.00657, 2018.

[19] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan, 2018.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[21] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[22] Roger Horn. *Matrix analysis*. Cambridge University Press, Cambridge New York, 2012.

[23] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[24] Carl Edward Rasmussen. Gaussian processes for machine learning. MIT Press, 2006.

[25] Muhammad Waleed Gondal, Bernhard Schölkopf, and Michael Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution, 2018.

[26] Jeremy Howard et al. fastai. `https://github.com/fastai/fastai`, 2018.

[27] D Salomon. *Data compression : the complete reference*. Springer, London, 2007.

[28] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[29] N. D. Narvekar and L. J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011.

[30] Yunus Emre Kara, Gaye Genc, Oya Aran, and Lale Akarun. Actively estimating crowd annotation consensus. *Journal of Artificial Intelligence Research*, 61:363–405, 2018.

[31] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019.

List of Symbols, Quantities and Abbreviations

ANN	Artifician Neural Network
BiO-Net	Bi-directional O-shape Network
CNN	Convolutional Neural Network
CPBD	A No Reference Image Blur Detection Using Cumulative Probability Blur Detection
ESRGAN	Enhanced Super-Resolution Generative Adversarial Network
GAN	Generative Adversarial Network
HR	High-resolution
JS	Jensen-Shannon
LR	Low-resolution
MF	Multi-frame
MLFDB	Multi-frame Labeled Faces Database
MSE	Mean Squared Error
NAS	Neural Architecture Search
PSRN	Peak Signal-to-Noise Ratio
PULSE	Photo Upsampling via Latent Space Exploration
SF	Single-frame
SR	Super-resolution
SSIM	Structural Similarity Index Measure
RGB	Red, Green, Blue
SE	Squeeze-and-Excitation
SER-FIQ	Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness
SRGAN	Super-Resolution Generative Adversarial Network
WGAN	Wasserstein Generative Adversarial Network

Content of the Attachment

- /
  - sharpening
    - data
      - mframe_dl.py ..... multi-frame data loader
    - metrics
      - github
      - metrics.py ..... objective metrics (PSNR, SSIM, SER-FIQ, others)
    - models
      - basic_blocks
        - blocks.py ..... single and double convolutional layers
      - bionet
        - model.py ..... BiO-Net architecture
      - face_align
        - utils.py ..... tools for face alignment
      - face_seg
        - github
        - utils.py ..... tools for masking facial parts
      - feature_net
        - model.py ..... Feature-Merge U-Net architecture
      - inpaint_net
        - model.py ..... Inpainting U-Net architecture
      - psfr_gan
        - github
        - utils.py ..... tools simplifying use of PSFR-GAN
      - pulse
        - github
        - pulse_transfer.py ..... tools simplifying use of PULSE
        - utils.py ..... raw PULSE predictor
      - serfiq_net
        - github
        - ir_se_model.py ..... IR-SE architecture
        - model.py ..... SER-FIQ U-Net architecture
      - unet
        - model.py ..... MF U-Net architecture and loss functions
    - utils
      - crossover_utils.py ..... sharp and soft cross-over
      - face_seg_utils.py ..... helper functions for face masking
      - feature_transfer_utils.py ... transfer of visual features between images
      - image_utils.py ..... basic image editing methods
      - smoothing_utils.py ..... Gaussian smoothing of images
      - test_utils.py ..... functions for testing models
      - train_utils.py ..... functions for training models
      - unet_utils.py ..... helper functions for U-Net
  - README.md