Обзор методов глубокого обучения в задаче слепого восстановления лица

С.Р. Шарипов, Б.М. Нутфуллин, Н.Г. Малоян

***Аннотация*—Актуальность исследования методов для решения задачи слепого восстановления лица (англ. Blind Face Restoration, BFR) обусловлена их возможными практическими применениями в разнообразных областях. Примерами таких областей являются диджитал-искусство и компьютерная графика для воссоздания и анимации лиц персонажей, а также социальные сети и мобильные приложения, где они способствуют улучшению качества изображений и видео.**

В данной статье мы проводим обзор современных методов и подходов, используемых для решения задачи BFR. Мы рассматриваем различные виды моделей, основанные на генеративно-состязательных сетях, автокодировщиках, дифузионных моделях, которые продемонстрировали значительный прогресс в данной области. В частности, мы анализируем ключевые аспекты, такие как архитектура сети, функции потерь, метрики качества и датасеты.

Кроме того, мы обсуждаем проблемы и ограничения существующих методов, а также возможные направления для будущих исследований. В частности, мы акцентируем внимание на необходимости разработки алгоритмов, устойчивых к разнообразным деградациям и способных адаптироваться к различным условиям освещения, позам и выражениям лица. В заключение, мы предоставляем систематическое сравнение существующих методов и подводим итоги об их достоинствах и недостатках.

***Ключевые слова*—слепое восстановление лица, низкое разрешение, шум, артефакты сжатия, размытие, глубокое обучение, диффузионная модель, генеративно-состязательная сеть**

I. Введение

Человечество создаёт, обрабатывает и хранит колоссальное количество изображений лиц, поскольку они:

- Являются основой коммуникации в эпоху социальных сетей и цифровизации.
- Выступают средством идентификации и верификации личности для обеспечения безопасности и контроля доступа.
- Служат средством сохранения воспоминаний, что порождает множество фото- и видеоматериалов.
- Широко используются в маркетинге для привлечения внимания и создания эмоциональной связи с продуктом или услугой.
- Занимают центральное место в развлечениях и искусстве - кинематограф, телевидение, живопись.

Однако изображения лиц, полученные в реальном мире могут страдать от различных деградаций (Рис. 1), таких, как *низкое разрешение* (англ. low resolution) [1], [2], [3], [4], [5], [6], *шум* (англ. noise) [7], [8], [9], *размытие* (англ. blur) [10], [11], [12], [13], *артефакты сжатия* (англ. compression artifacts) [14], [15] и др., а также от их комбинации (англ. blind). Кроме того изображения лиц, искусственно синтезированные с помощью стремительно развивающихся генеративных моделей также могут быть подвержены к искажениям различного рода.

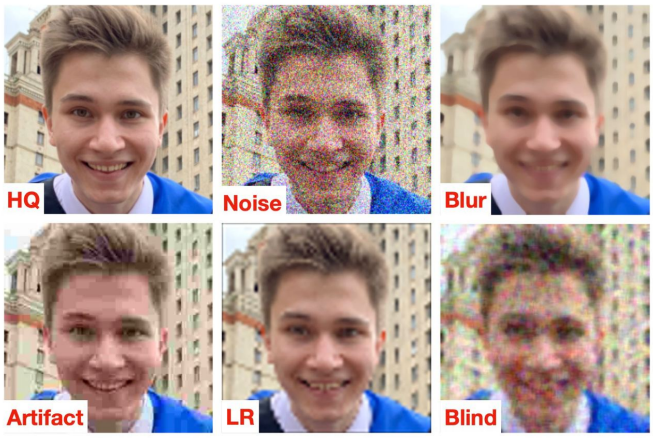


Рис. 1: *Примеры низкокачественных (LQ) изображений лиц, полученных из высококачественного (HQ) изображения: добавление шума (англ. noise), размытие (англ. blur), артефакты сжатия (англ. compression artifacts), низкое разрешение (англ. low resolution) и комбинация вышеперечисленных деградаций (англ. blind).*

Таким образом, в связи с широким распространением изображений лиц и их подверженности к различным искажениям, актуальной является задача *слепого восстановления лица* (англ. blind face restoration, BFR), суть которой заключается в получении высококачественного (англ. high-quality) изображения лица I_{HQ} из соответствующего ему низкокачественного (англ. low-quality) аналога I_{LQ}, страдающего от неизвестных заранее деградаций.

В реальных условиях задача BFR осложняется более сложными деградациями, а также разнообразием выражений и уникальных черт лиц людей. На процесс восстановления также могут повлиять такие факторы как освещение, окружающая среда и фон изображения, качество камеры, возраст изображений, тип генеративной модели для синтеза лиц и т.п.

В последние годы наблюдается стремительное развитие в области глубокого обучения и увеличение доступности крупномасштабных наборов данных. Благодаря этому искусственные нейронные сети (ИНС) демонстрируют превосходные результаты в различных задачах обработки изображений, опережая традиционные методы

Статья получена 2 мая 2023.

Саит Равильевич Шарипов, МГУ им. М.В. Ломоносова, (email: ssharipovr@gmail.com).

Булат Маратович Нутфуллин, МГУ им. М.В. Ломоносова, (email: Bulat15g@gmail.com).

Нарек Гагикович Малоян (email: maloyan.narek@gmail.com).

компьютерного зрения [16]. На сегодняшний день наилучшие результаты в задаче BFR также демонстрируют методы основанные на глубоком обучении. Основная идея большинства из них заключается в изучении отображения из I_{LQ} в I_{HQ}, параметризованного с помощью глубоких нейронных сетей с использованием большого набора предварительно собранных пар изображений I_{LQ} и I_{HQ}. Разные подходы к решению задачи BFR имеют свои преимущества, недостатки и ограничения.

В данном обзоре рассматриваются современные методы для слепого восстановления лиц с использованием нейросетевых подходов. В разделе II представлена классификация задач восстановления лица в зависимости от используемой модели деградации, а также приводится формальная постановка задачи слепого восстановления лица. Раздел III посвящен обзору используемых метрик оценки качества восстановления изображения. В разделе IV приведена классификация методов восстановления лица на основе использования априорных знаний. В разделе V подробно рассмотрены передовые методы, основанные на глубоком обучении, для решения задачи BFR. Раздел VI посвящен наборам данных, используемых исследователями для обучения и тестирования методов BFR. Наконец, в разделе VII подводятся итоги обзора.

II. Постановка задачи слепого восстановления изображения лица

A. Общий вид (Face Restoration)

В процессе создания, обработки, передачи и хранения изображений возникают искажения, которые могут быть представлены различными формами, включая аддитивный шум, размытие, снижение разрешения и артефакты сжатия. Общую модель деградации изображения лица можно сформулировать в следующем виде:

$$I_{LQ} = \mathcal{D}(I_{HQ}) \tag{1}$$

где I_{LQ} и I_{HQ} – это низкокачественное и высококачественное изображения соответственно, а $\mathcal{D}$ – функция деградации. Тогда общая задача восстановления изображения лица заключается в поиске такой модели $\mathcal{D}^{-1}$, что:

$$I_{HQ} = \mathcal{D}^{-1}(I_{LQ}) \tag{2}$$

Таким образом, определив вид деградации $\mathcal{D}$ можно уточнить модель деградации и тем самым определить подзадачу общей задачи восстановления изображения лица.

B. Удаление шума (Face Denoising)

$$I_{LQ} = \mathcal{D}(I_{HQ}) = I_{HQ} + n_\delta \tag{3}$$

где n_δ – аддитивный белый Гауссов шум с уровнем δ.

C. Устранение размытия (Face Deblurring)

Обычно причинами размытия на изображении являются движение объекта съемки относительно камеры и ошибки в фокусировке. Размытие может быть задано следующим образом:

$$I_{LQ} = \mathcal{D}(I_{HQ}) = I_{HQ} * k_\sigma \tag{4}$$

где k_σ – ядро размытия, $*$ – операция свёртки.

D. Увеличение разрешения (Face Super-Resolution)

$$I_{LQ} = \mathcal{D}(I_{HQ}) = (I_{HQ}) \downarrow_s \tag{5}$$

где $\downarrow_s$ – это операция уменьшения разрешения изображения (англ. downsampling) с коэффициентом масштабирования s.

E. Удаление артефактов (Face Artifact Removal)

Методы сжатия с потерями (например, $JPEG$, $Webp$ и др.) широко применяются для уменьшения размеров изображений, что ведёт к возникновению артефактов сжатия.

$$I_{LQ} = \mathcal{D}(I_{HQ}) = JPEG_q(I_{HQ}) \tag{6}$$

где $JPEG_q$ соответствует распространенному способу сжатия JPEG с коэффициентом качества q.

F. Слепое восстановление (Blind Face Restoration)

Как правило, методы восстановления изображения лица, разработанные под конкретный тип деградации плохо справляются с искажениями, встречающихся в реальных сценариях. Поэтому наиболее актуальной является задача слепого восстановления лица (BFR). Модель деградации в BFR является случайной комбинацией всех вышеперечисленных искажений (шум, размытие, низкое разрешение, артефакты сжатия), и потому намного лучше имитирует повреждения, наблюдаемые в реальном мире.

$$I_{LQ} = \mathcal{D}(I_{HQ}) = \{JPEG_q\big((I_{HQ} * k_\sigma) \downarrow_s + n_\delta\big)\} \uparrow_s \tag{7}$$

где $\uparrow_s$ – это операция увеличения разрешения изображения (англ. upsampling) с коэффициентом масштабирования s.

III. Метрики оценки качества восстановления изображений

Эффективность реконструкции методов слепого восстановления лица может быть оценена различными способами. Обычно для оценки качества изображений используются два основных метода: *субъективная* и *объективная* оценка.

Определение оптимального метода оценки качества изображений требует отдельного внимания, поскольку субъективные и объективные методы имеют свои преимущества и ограничения, связанные с доступностью ресурсов, времени соответствию предварительно поставленным целям. Для того чтобы обеспечить оптимальные результаты оценки качества слепого восстановления лица при выборе метода оценки необходимо учитывать конкретные цели и требования, а также учитывать различия между математическими моделями и визуальным восприятием человека.

A. Субъективная оценка

Субъективная оценка качества изображений базируется на восприятии людей и требует их участия для оценки качества сгенерированных изображений. Хотя этот метод и предоставляет результаты, согласующиеся с человеческим восприятием, он требует значительных временных и финансовых затрат.

1) Mean Opinion Score (MOS): Это широко используемая субъективная метрика оценки качества изображений, основанная на мнениях людей. Она используется для получения общей оценки качества изображения, которая может быть использована для сравнения с другими изображениями или для оценки производительности алгоритмов обработки изображений. Для получения MOS эксперты оценивают качество восприятия тестируемых изображений, после чего вычисляется среднее арифметическое значение присвоенных оценок. Количество оценщиков может сильно повлиять на предвзятость MOS – чем меньше экспертов, тем более смещённой и неправдоподобной может оказаться результирующая метрика.

B. Объективная оценка

Объективная оценка качества изображений в основном опирается на статистические данные и математические модели, которые могут давать результаты, отличающиеся от субъективной оценки, основанной на визуальном восприятии человека. Это объясняется тем, что методы объективной оценки не учитывают все аспекты качества изображения, и могут быть нечувствительны к некоторым визуальным артефактам, которые влияют на восприятие изображения человеком.

1) Peak Signal-to-Noise Ratio, PSNR: Это широко используемая метрика объективной оценку в задаче BFR. Пусть имеется эталонное высококачественное изображение I_{HQ} и восстановленное $\hat{I}_{HQ}$. Сначала вычисляется сумма квадратов разностей между соответствующими пикселями I_{HQ} и $\hat{I}_{HQ}$:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\Big(I_{HQ}(i) - \hat{I}_{HQ}(i)\Big)^2 \tag{8}$$

где N – количество пикселей в I_{HQ}. Затем вычисляется PSNR:

$$PSNR = 10 \cdot \log_{10}\Big(\frac{L^2}{MSE}\Big) \tag{9}$$

где L — максимальное возможное значение пикселя (например, для 8-bit RGB изображений $L = 255$). Чем меньше разница в соответствующих пикселях между двумя изображениями, тем выше PSNR. Таким образом, PSNR сосредотачивается на разности пикселей, что приводит к плохой интерпретации при представлении качества реконструкции в реальных условиях, когда важно соответствие с человеческим восприятием.

2) Structural Similarity Index Measure, SSIM: Предложенная в [17] метрика структурного сходства между двумя изображениями I и $\hat{I}$ основана на вычислении трёх аспектов: яркости, контрастности и структуры. Для изображения I из N пикселей яркость μ_I и контрастность σ_I определяется следующим образом:

$$\mu_I = \frac{1}{N}\sum_{i=1}^{N} I(i), \quad \sigma_I = \Big(\frac{1}{N-1}\sum_{i=1}^{N}(I(i)-\mu_I)^2\Big)^{\frac{1}{2}} \tag{10}$$

где $I(i)$ — интенсивность (значение) i-ого пикселя изображения. Тогда может быть вычислена схожесть по яркости (англ. luminance) и контрастности (англ. contrast):

$$C_l(I,\hat{I}) = \frac{2\mu_I\mu_{\hat{I}} + C_1}{\mu_I^2 + \mu_{\hat{I}}^2 + C_1},\ C_c(I,\hat{I}) = \frac{2\sigma_I\sigma_{\hat{I}} + C_2}{\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2} \tag{11}$$

где $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$ – константы для избежания нестабильности вычислений ($k_1 << 1$, $k_2 << 1$), L – максимальное возможное значение пикселя (например, для 8-bit RGB изображений $L = 255$).

Структура изображения может быть представлена нормированными значениями пикселей, т.е. $\frac{I-\mu_I}{\sigma_I}$. Тогда с помощью их корреляции (т.е. скалярного произведения) может быть вычислено структурное сходство:

$$\sigma_{I,\hat{I}} = \frac{1}{N}\sum_{i=1}^{N}((I(i)-\mu_I)(\hat{I}(i)-\mu_{\hat{I}})) \tag{12}$$

$$C_s(I,\hat{I}) = \frac{\sigma_{I,\hat{I}} + C_3}{\sigma_I\sigma_{\hat{I}} + C_3} \tag{13}$$

где C_3 – константа для стабильности вычислений. В результате SSIM может быть вычислена следующим образом:

$$SSIM(I,\hat{I}) = [C_l(I,\hat{I})]^{\alpha}[C_c(I,\hat{I})]^{\beta}[C_s(I,\hat{I})]^{\gamma} \tag{14}$$

где α, β, γ — гиперпараметры для настройки относительной важности сходства яркости, контрастности и структуры. Область значений SSIM – отрезок $[-1, 1]$, где -1 соответствует полной антикорреляции, 0 – отсутствие сходства, 1 соответствует полному сходству. SSIM часто применяют к патчам изображений, проходя по ним скользящим окном. На практике SSIM неплохо отражает человеческое восприятие.

3) Learned Perceptual Image Patch Similarity, LPIPS: Авторы работы [18] показали, что признаки извлеченные из предварительно обученной глубокой нейронной сети для классификации могут быть использованы для измерения сходства между двумя изображениями. На рисунке 2 и в формуле 15 отражено вычисление расстояния между эталонным патчем x и искаженным патчем x_0 с помощью нейронной сети $\mathcal{F}$. Сначала извлекаются и нормализуются вдоль канального измерения признаки $\hat{y}^l, \hat{y}_0^l \in \mathbb{R}^{H_l \times W_l \times C_l}$, полученные после применения каждого слоя l сети $\mathcal{F}$. Затем их разность масштабируются вдоль измерения каналов путём умножения на вектор $w^l \in \mathbb{R}_l^C$ и вычисляется L_2 норма. Наконец, производится усреднение вдоль пространственных измерений и суммирование по каналам.

$$d_0 = d(x, x_0) = \sum_l \frac{1}{H_l W_l}\sum_{h,w}\|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l)\|_2^2 \tag{15}$$

Чем больше похожи два изображения, тем меньше метрика LPSIS. Авторы работы продемонстрировали, что

метрика LPSIS больше приближена к человеческому восприятию, чем другие распространенные метрики, такие как PSNR и SSIM.

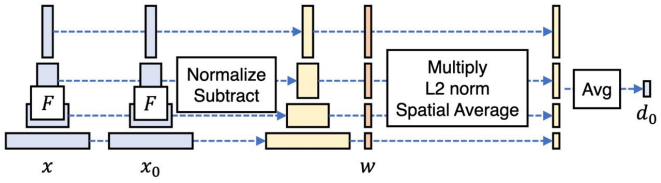


Рис. 2: *Схема вычисления метрики LPIPS между эталонным патчем x и искаженным патчек x_0 с помощью нейронной сети $\mathcal{F}$.*

4) Fréchet Inception Distance, FID: В [19] была предложена метрика, которая позволяет оценить близость двух вероятностных распределений. В задаче слепого восстановления лица это распределение высококачественных изображений и распределение восстановленных низкокачественных изображений. С помощью предобученной нейронной сети для классификации изображений (например, VGG, Inception) вычисляются эмбеддинги множества эталонных высококачественных изображений и множества восстановленных изображений. В предположении, что полученные векторные представления из распределения Гаусса можно вычислить расстояние Фреше между двумя многомерными нормальными распределениями X и Y:

$$FID = \|\mu_X - \mu_Y\|^2 + Tr(\Sigma_X + \Sigma_X - 2\sqrt{\Sigma_X \Sigma_Y}) \quad (16)$$

где μ – математическое ожидание, Σ – матрица ковариации, Tr – след матрицы. Недостатком FID является зависимость от изображений, на которых обучалась нейронная сеть для получения векторов представлений. Однако на практике FID больше соответствует человеческому восприятию, чем PSNR и SSIM. Чем меньше FID, тем лучше восстановление.

5) Natural Image Quality Evaluator, NIQE: Метод позволяет оценить качество визуального восприятия отдельного изображения без заданного эталона. В NIQE [20] с помощью определённой последовательности действий конструируются набор признаков, отражающих качество изображений(я). По извлечённым признакам с помощью метода наибольшего правдоподобия строится многомерная Гауссова модель (англ. Multivariate Gaussian Model, MVG). В результате качество тестируемого изображения вычисляется на основе расстояния между MVG моделью, построенной на основе признаков, извлечённых из тестируемого изображения и MVG моделью, обученной на признаках, излеченных из набора данных естественных изображений:

$$NIQE = \sqrt{(\nu_1 - \nu_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\nu_1 - \nu_2)} \quad (17)$$

где ν_1, ν_2 и Σ_1, Σ_2 – векторы средних и матрицы ковариации MVG модели естественных изображений и MVG модели тестируемого изображения. Так как метрика отражает расхождение между моделями, то чем меньше NIQE, тем считается выше визуальное качество тестируемого изображения.

IV. Классификация методов восстановления лица на основе использования априорной информации

С точки зрения использования априорной информации методы слепого восстановления лица могут быть разделены следующим образом:

- Методы, не использующие априорную информацию.
- Методы, использующие априорную информацию, которые могут разделены на три типа:
  1) Методы, использующие в качестве априорной информации геометрию лица (англ. geometric prior)
  2) Методы, использующие в качестве априорной информации эталонные данные (англ. reference)
  3) Методы, использующие в качестве априорной информации знания обученных генеративных нейронных сетей (англ. generative prior)

Хотя есть некоторые работы [21], [22], [23], стремящиеся восстановить высококачественное изображение I_{HQ} только на основе информации из низкокачественного изображения I_{LQ}, большинство существующих работ продемонстрировали, что априорная информация играет решающую роль в задаче BFR, ведь человеческое лицо имеет сложную структуру и специфические характеристики, которые следует учитывать. Поэтому далее мы рассмотрим подробнее особенности методов восстановления на основе предварительных знаний. Методы восстановления, использующие априорную информацию могут быть разделены на три группы:

A. Методы, использующие в качестве априорной информации геометрию лица (англ. geometric prior).

В этих методах как правило используется информация об уникальной геометрии и пространственном распределении лиц на изображении, чтобы помочь модели постепенно восстанавливать высококачественные изображения лиц. В качестве априорной, например, могут выступать: ключевые точки лица (англ. facial landmark) [24], тепловая карта лица (англ. facial heatmaps) [25], карта разбиения лица на атрибуты (facial parsing map) [26], [27], [28], 3D форма лица [29], [30] и др. Однако такая априорная информация не может быть точно получена от изображений, подверженных деградациям. Более того, геометрическая априорная информация не может полностью обеспечить богатую детализацию для качественного восстановления лица.

B. Методы, использующие в качестве априорной информации эталонные данные (англ. reference)

Методы этой группы, как правило, используют в качестве априорной информации структуру лица или словари компонентов лица, полученных из дополнительных высококачественных изображений лица. Подходы этой группы менее подвержены к ограничениям методов, основанных на геометрии лица, но имеют свои недостатки. Так, в [31] используются эталонные данные в виде дополнительного высококачественного изображения той же идентичности, не доступные в общем случае. А в DFDNet [32] предварительно конструируются словари,

состоящие из компонент (глаз, рта и т.п.) высококачественных изображений лиц, однако ограниченный набор заранее заданных компонент не позволяет качественно восстановить изображение лица в реальных условиях. Чтобы решить эту проблему, недавние методы [33], [34], [35] используют идею векторного квантования, представленную в VQVAE [36] и VQGAN [37], обучая словарь признаков высококачественных изображений, который содержит более обобщенные и подробные детали для восстановления лица.

C. Методы, использующие в качестве априорной информации знания обученных генеративных нейронных сетей (англ. generative prior)

Предварительно обученные генеративно-состязательные сети (англ. Generative Adversarial Networks, GANs), такие, как StyleGAN2 демонстрируют поразительную возможность синтеза высококачественных изображений и могут быть использованы для предоставления богатой и разнообразной информации о лице в задаче BFR. Некоторых методы этого типа основаны на инверсии GAN, например, авторы PULSE [38] производят градиентный спуск в скрытом пространстве предобученного StyleGAN [39], чтобы найти такое изображение $\hat{I}_{HQ}$, что $DS(\hat{I}_{HQ}) \approx I_{LQ}$, где DS – функция уменьшения масштаба изображения (англ. downscale). Недостаток этих методов заключается в недостаточной точности или ”правильности”(англ. fidelity) восстановленного изображения лица – оно может сильно отличаться от I_{LQ}. Другие методы, такие как [40], [41] используют архитектуру кодировщик-декодировкщик: сначала для достижения большей точности (англ. fidelity) из I_{LQ} извлекается структурна информация, а затем для достижения наилучшего визуального качества, в качестве декодировщика используется предварительно обученный GAN. Чтобы достичь большей точности (англ. fidelity), эти методы в значительной степени полагаются на входные данные через пропускные соединения (англ. skip connections), что может привести к артефактам в результатах, когда входные данные сильно повреждены. Кроме того, сложности также может вызывать сам процесс обучения сети, из-за состязательной природы GAN. Недавний успех диффузионных моделей в генерации изображений [42] вдохновил исследователей к использованию генеративных возможностей диффузионных моделей для восстановления лица. Недавние исследования [43], [44] показали, что диффузионные модели могут успешно использоваться для восстановления изображений лиц, в том числе и в слепой постановке, когда деградация изображения неизвестна. Таким образом, использование диффузионных моделей для восстановления лиц является перспективным направлением исследований.

V. Обзор методов слепого восстановления лица

В таблице 1 представлены краткие описания методов слепого восстановления лица. В дальнейшем мы рассмотрим каждый метод более подробно.

A. Методы, не использующие априорных знаний

1) ***STUNet****:* Вдохновившись успехом Swin Transformer [51], достигающего state-of-the-art результатов в различных задачах компьютерного зрения, Zhang et al. для решения задачи BFR разработали Swin Transformer U-net (STUNet) [48](Рис. 3). Сначала, чтобы извлечь низкоуровневые признаки из изображения I_{LQ}, к нему применяется свёрточный слой с ядром 3×3. Затем, полученные признаки проходят через симметричную 4-х уровневую архитектуру кодировщик-декодировщик, состоящую из блоков Swin Transformer. В кодировщике на каждом из уровней применяется блок Swin Transformer и уменьшается размерность выходных признаков с помощью (pixel-unshuffle operation). Таким образом, кодировщих преобразует входящие признаки с небольшим количеством каналов в скрытое представление низкого разрешения, но с большим количеством каналов. Далее декодировщик постепенно, симметрично кодировщику восстанавливает из скрытого представления исходные признаки. Для объединения информации из признаков одного уровня используется skip-connections. Наконец, к агрегации выхода декодеровщика и признаков, полученных из I_{LQ}, применяется свёрточный слой с ядром 3×3, результатом которого является высококачественное изображение I_{HQ}.

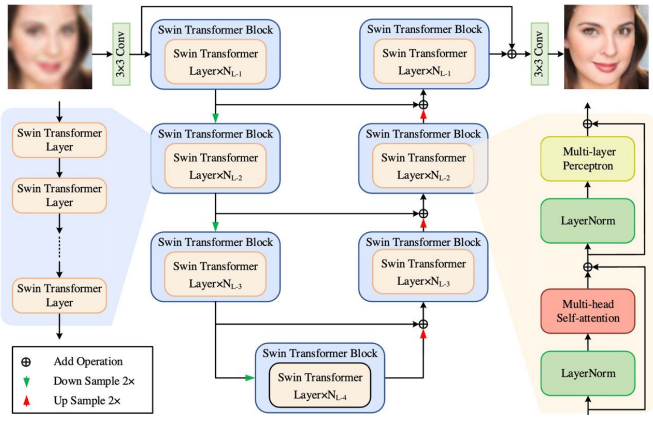


Рис. 3: *Архитектура Swin Transformer U-net (STUNet).*

2) ***HiFaceGAN****:* В работе [28] предложено использование U-net подобной архитектуры, содержащей иерархические CSR блоки (англ. collaborative suppression and replenishment).

Кодировщик, состоящий из CSR блоков, извлекает семантические признаки для последующего восстановления в декодировщике. Улучшение достигается за счет использования адаптивных сверток LIP [52] и PAC [53] вместо классических свёрточных слоёв. LIP адаптация вычисления свертки на основе локальной важности признаков в изображении. Этот подход позволяет модели учитывать различные контексты и динамически приспосабливать свертку, чтобы извлечь более информативные признаки из изображения. PAC – это другой подход к адаптивным фильтрам, который объединяет признаки на основе их позиции и соседства. PAC слои позволяют модели адаптивно выбирать и агрегировать признаки в зависимости от их расположения на изображении.

Таблица 1: *Методы слепого восстановления лица с помощью глубоких нейронных сетей.*

Метод	Априорная информация	Архитектура	Ключевая идея	Публикация
DFDNet	Reference prior Словарь лицевых компонентов	VGG, DFT block	Предложено использование алгоритма K-средних для создания словарей, содержащих компоненты лица эталонных изображений I_{HQ}. Полученные словари применяются для передачи компонент лица высокого качества на деградированное изображение.	Li *et al.* 2020 [32]
PULSE	Generative prior Инверсия GAN	StyleGAN	Инверсия предобученного GAN. Поиск в скрытом пространстве StyleGAN такого вектора z, что уменьшение масштаба изображения $\hat{I}_{HQ}$, синтезированного с помощью z, приведет к получению изображения низкого разрешения I_{LQ}.	Menon *et al.* 2020 [38]
HiFaceGAN	Не используется	U-Net подобная	В работе предложено использование адаптивных сверток в декодировщике вместо классических сверточных слоев. Также предложено использование SPADE нормализации в декодировщике.	Yang *et al.* 2020 [28]
SPARNetHD	Generative prior	FAU block (residual block + spatial attention)	В работе предложено использование механизма пространственного внимания в классических residual блоках.	Chen *et al.* 2020 [45]
GFP-GAN	Generative prior	StyleGan2, U-Net	Использование U-Net подобной нейросетевой модели для восстановления изображения с помощью априорных знаний StyleGAN.	Wang *et al.* 2021 [41]
GPEN	Generative prior	U-Net подобная, декодировщик – GAN	Предварительное обучение GAN (подобного StyleGAN) для синтеза высококачественных изображений. Встраивание GAN в качестве декодировщика в U-Net подобную архитектуру и её дообучение для задачи восстановления.	Yang *et al.* 2021 [40]
PSFRGAN	Geometric prior	VGG19, GauGAN and StyleGAN inspired architecture	Идея метода заключается в использовании прогрессивном увеличении разрешения восстановленных изображений, с использованием семантической информации. В работе предложена функция потерь semantic-aware style loss, использующая активации слоев VGG19.	Chen *et al.* 2021 [26]
GLEAN	Generative prior	Encoder - latent bank(StyleGAN) - decoder	В работе реализована идея использования предобученного StyleGan для извлечения признаков, используемых декодировщиком для восстановления изображения.	Chan *et al.* 2021 [46]
FaceFormer	Generative prior	StyleGan2,SWIN	Предложено использование архитектуры Трансформер для извлечения признаков деградированного лица с последующим восстановлением с помощью GAN. Также в работе предложена замена интерполяции для увеличения разрешения изображения: использование upsampling слоев с дополнительной информацией об изменении масштаба изображения.	Li *et al.* 2022 [47]
STUNet	Не используется	U-Net подобная с Swin Transformer блоками	Baseline для задачи BFR. Применение Swin Transformer блоков в U-Net подобной архитектуре без использования априорной информации.	Zhang *et al.* 2022 [48]
RestoreFormer	Reference prior Словарь лицевых компонентов	VQVAE, Multi-Head Cross-Attention (MHCA)	Предварительное обучение словаря HQ высококачественных лицевых признаков. Использование трансформеров с multi-head cross-attention для слияния признаков искажённого изображения и их высококачественных аналогов из словаря HQ.	Wang *et al.* 2022 [35]
CodeFormer	Reference prior Словарь лицевых компонентов	VQVAE, ViT	Использование двухэтапной архитектуры, основанной на комбинации квантованного кодировщика с кодовой книгой, обученной для реконструкции высококачественного лица и Трансформера, необходимого, чтобы распутать несоответствия между кодовой книгой и выходом кодировщика для низкокачественного изображения.	Zhou *et al.* 2022 [33]
VQFR	Reference prior Словарь лицевых компонентов	VQ-GAN, U-net подобная архитектура с использованием кодовой книги	VQFR состоит из кодировщика, предобученной кодовой книги и параллельного декодировщика, который восстанавливает отдельно структурные особенности и текстуру лица и потом совмещает для получения результата.	Gu *et al.* 2022 [34]
DDRM	Generative prior	Diffusion model	Используется предобученная модель DDPM. Задача восстановления изображения рассматривается как линейная обратная задача. Марковская цепь диффузионного процесса строится в спектральном пространстве оператора деградации и обуславливается на искажённое изображение. Подход не требует дополнительного обучения.	Kawar *et al.* 2022 [49]
DDNM	Generative prior	Diffusion model	Используется предобученная модель DDPM. Задача восстановления изображения рассматривается как линейная обратная задача. Используется Range-Null space Decomposition (RND) для изменения обратного диффузионного процесса. Подход не требует дополнительного обучения.	Wang *et al.* 2022 [43]

Метод	Априорная информация	Архитектура	Ключевая идея	Публикация
DifFace	Generative prior	Diffusion model	Сначала восстановить изображение с помощью стандартной модели для восстановления. Затем добавить шум согласно прямому диффузионному процессу. Удалить шум с помощью обратного процесса диффузии.	Yue *et al.* 2022 [44]
DR2	Generative prior	Diffusion model	Используется предварительно обученная диффузионная модель для синтеза лиц. Изображение с деградациями подвергается прямому диффузионному процессу до некоторого шага ω. При обратном диффузионном процессе в качестве условия используется изображение с деградациями на соответствующем шаге прямой диффузии. Обратный диффузионный процесс прерывается на шаге τ. Для достижения наилучшего результата к выходу диффузионной модели применяется передовой нейросетевой метод BFR.	Wang *et al.* 2023 [50]

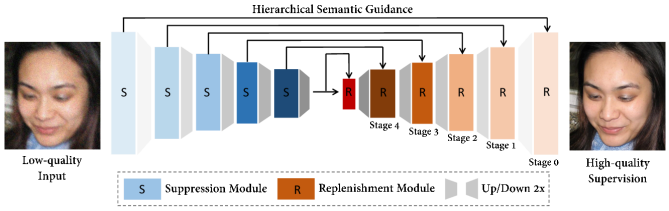


Рис. 4: *Архитектура HiFaceGAN.*

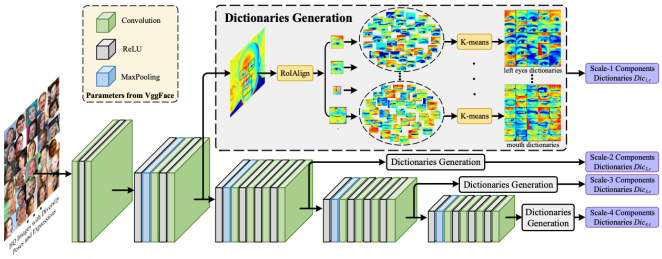


Рис. 5: *Архитектура DFDNet. Автономное создание словарей компонентов разного масштаба из большого количества изображений высокого качества с разнообразными позами и выражениями лица. K-means используется для создания K кластеров для каждого компонента (то есть левого/правого глаза, носа и рта) на разных масштабах признаков.*

Авторы отмечают, что использование LIP [52] и PAC [53] помогает лучше отфильтровывать деградации на изображении.

В архитектуре декодировщика предложено использование SPADE (англ. Spatially-Adaptive Denormalization) блоков [54] в CSR блоках. Основная идея SPADE заключается в модуляции нормализации признаков в генераторе изображений с помощью семантической карты, что позволяет управлять детализацией и структурой сгенерированных изображений.

SPADE-блок состоит из следующих компонентов:

- Семантическая карта используется для модуляции аффинных параметров, которые применяются к нормализованным признакам.
- Обучаемый свёрточный слой преобразует семантическую карту в аффинные параметры для каждого канала признаков.
- Аффинные параметры применяются к нормализованным признакам поканально, что позволяет генератору учитывать семантическую информацию.

B. Методы, использующие словари лицевых компонент в качестве априорных знаний

1) ***DFDNet****:* Идея Deep Face Dictionary Network (DFDNet) [32] заключается в использовании словарей компонентов лица для извлечения признаков и восстановления деталей лица.

Авторы используют набор данных FFHQ для создания компонентных словарей, покрывающих разные типы лиц. Из 70 000 изображений выбирают 10 000, учитывая разнообразие атрибутов (возраст, этническая принадлежность, позы, выражения лица и т.д). Для выделения признаков используют предварительно обученную модель VggFace. Четыре компонента (левый и правый глаз, нос, рот) обрезаются и семплируются с использованием RoIAlign на разных масштабах. Затем методом K-средних генерируются кластеры для каждого компонента, формируя компонентные словари для каждого из маштабов. В частности, для обработки изображений размером 256×256 пикселей, размеры признаков левого/правого глаза, носа и рта на масштабе-1 устанавливаются равными 40/40, 25, 55 соответственно. Размеры уменьшаются вдвое для следующих масштабов-2, 3, 4. Эти признаки словаря могут быть сформулированы следующим образом:

$$Dic_{s,c} = F_{Dic}(I^h | L^h; \Theta_{Vgg}),$$

где $s \in \{1, 2, 3, 4\}$ - масштаб словаря, $c \in \{\text{левый глаз}, \text{правый глаз}, \text{нос}, \text{рот}\}$ - тип компонентов, и Θ_{Vgg} - фиксированные параметры от VggFace.

На второй стадии DFDNet переносит лицевые признаки из словаря компонентов на входное изображение. В качестве кодировщика входного изображения используется VggFace, чтобы гарантировать, что признаки входного изображения и словаря компонентов находятся в одном пространстве признаков.

Для переноса признаков из словаря компонентов на изображение авторы предложили блок DFT, состоящий из пяти частей: RoIAlign [55], CAdaIN, сопоставление признаков, оценка уверенности и обратный RoIAlign.

В блоке DFT Сначала используется RoIAlign для создания четырех компонентных областей: левый/правый глаз, нос, рот. Затем, так как входные компоненты могут иметь разное распределение или стиль, такие как освещение, цвет кожи, предлагается CAdaIN для нормализации каждого кластера в словарях. Нормированные

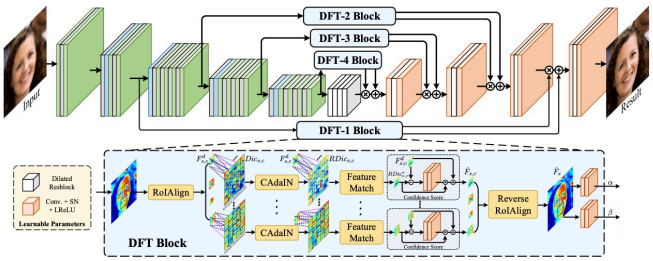


Рис. 6: *Архитектура DFDNet. Процесс восстановления и блок передачи признаков словаря (DFT), который используется для предоставления ссылочных деталей прогрессивным образом. Здесь блок DFT-i использует словари компонентов масштаба-i в качестве ссылки на одном и том же уровне признаков.*

словари $RDic^k_{s,c}$ с использованием CAdaIN получаются следующим образом:

$$RDic^k_{s,c} = \sigma(F^d_{s,c}) \frac{Dic^k_{s,c} - \mu(Dic^k_{s,c})}{\sigma(Dic^k_{s,c})} + \mu(F^d_{s,c})$$

Здесь $F^d_{s,c}$ и $Dic^k_{s,c}$ обозначены как c-й компонент признаков входного изображения I^d и k-й кластер из словаря компонентов масштаба s. При этом возможные значения $c \in \{\text{left eye}, \text{right eye}, \text{nose}, \text{mouth}\}$, $s \in 1, 2, 3, 4$.

После шага с нормализацией словаря компонентов применяется сопоставление признаков для выбора кластера с похожей текстурой. Для этого используется скалярное произведение между признаками $F^d_{s,c}$ и всеми нормированными кластерами в $RDic^k_{s,c}$. Таким образом для k-го кластера в компонентном словаре сходство определяется следующим образом:

$$S^{s,c}_k = \langle F^{s,c}_d, RDick^{s,c} \rangle, (4)$$

Среди всех оценок $S_{s,c}$ выбирается нормированный кластер с наибольшим сходством $RDic^*_{s,c}$.

Для регуляции действия признаков из словаря компонентов вводится оценка уверенности. Заметим, что небольшое ухудшение входного изображения (например, увеличение разрешения в 2 раза) слабо влияет на словарь компонентов. Для адаптации DFDNet к различным изменениям входного изображения, мы рассчитываем разность между $F^d_{s,c}$ и $RDic^*_{s,c}$ и используем ее для подсчета оценки уверенности, которая воздействует на выбранный словарный признак $RDic^*_{s,c}$. Результат должен содержать отсутствующие детали высокого качества, которые могут быть восстановлены в $F^d_{s,c}$. Формула для оценки уверенности выглядит следующим образом:

$$\hat{F}^{s,c} = F^d_{s,c} + RDic^*_{s,c} * F^{Conf}(RDic^*_{s,c} - F^d_{s,c}; \Theta_C), (5)$$

где Θ_C - обучаемые параметры блока коэффициента достоверности F_{Conf}.

После того как все компоненты изображения прошли обработку в предыдущем разделе, мы применяем обратную операцию RoIAlign, вернув $\hat{F}_{s,c}$ и ($c \in$ левый/правый глаз, нос, рот) на их исходные позиции $F^d_{s,c}$.

*2) **VQFR**:* В данной работе предложено использование словаря векторов (codebook), впервые предложенного в VQ-VAE и VQ-GAN.

Архитектура VQFR [34] состоит из трех компонент - кодировщика, codebook и параллельного декодировщика.(Рис. 7).

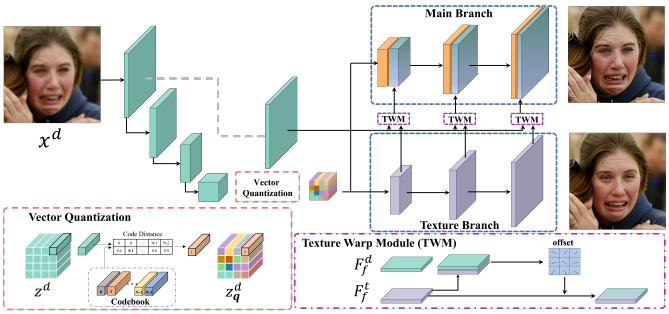


Рис. 7: *Архитектура VQFR.*

Кодировщик отвечает за сжатие признаков из искаженных изображений лиц перед квантованием с помощью codebook и передачей декодировщику для восстановления. В методе используется codebook из VQ-GAN, обученный на лицах высокого разрешения. Параллельный декодировщик состоит из двух декодировщиков, которые работают параллельно для восстановления изображения. Один декодировщик восстанавливает структурные особенности лица, а другой - текстуру и детали. Восстановленные изображения затем совмещаются для получения окончательного результата.

Совмещение осуществляется с помощью texture warping module, основной идеей которого является применение слоев деформируемой свертки [56] для сохранения высококачественных деталей лица и текстурных особенностей.

*3) **CodeFormer**:* Модель предложенная в [33] обучается в два этапа:

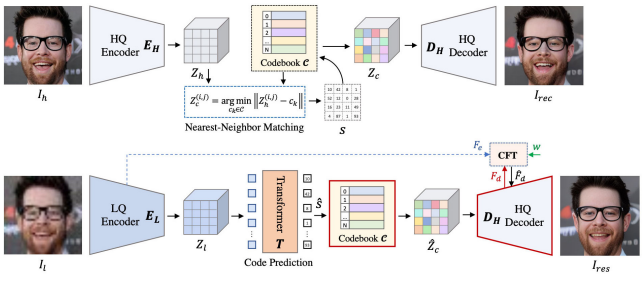
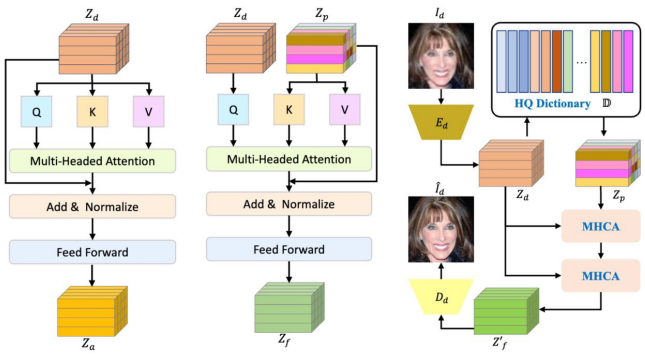


Рис. 8: *Двухэтапное обучение архитектуры CodeFormer.*

1) Сначала авторы используют идею векторного квантования, представленную в VQVAE [36] и VQGAN [37]. Для уменьшения неопределённости при отображении из I_{LQ} в I_{HQ} и дополнения высококачественных деталей для восстановления сначала обучается квантованный автоэнкодер посредством реконструкции входного высококачественного изображения I_{HQ}, чтобы получить кодовую книгу (англ. codebook) (Рис. 8(a)).
2) На втором этапе фиксируются кодовая книга и декодеровщик, архитектура дополняется Трансформером, а сеть обучается по изображению I_{LQ} восстанавливать I_{HQ}. Предсказания кодировщика для

низкокачественного изображения и его высококачественного аналога могут быть различными, поэтому Трансформеру необходимо распутать это несоответствие (Рис. 8(b)).

4) ***RestoreFormer****:* Архитектура RestoreFormer [35] изображена на Рис. 9(c). Сначала кодировщик E_d извлекает скрытое представление Z_d ухудшенного лица I_d, и ближайшие высококачественные априорные представления Z_p извлекаются из HQ словаря $\mathbb{D}$, полученного при предварительном обучении квантованного автокодировщика посредством реконструкции высококачественного изображения. Затем два последовательных трансформера, реализованных с помощью multi-head cross-attention (MHCA, Рис. 9(b)), используются для слияния признаков ухудшенных изображений и априорных данных в Z'_f. Стоит отметить, что нельзя было прямо применить multi-head self-attention (MHSA, Рис. 9(a)), поскольку при восстановлении лица следует комбинировать информацию из повреждённого изображения и априорных представлений. Наконец, декодеровщик D_d применяется к объединенному представлению Z'_f для восстановления высококачественного лица $\hat{I}_d$.



(a) **MHSA** (b) **MHCA** (c) **RestoreFormer**

Рис. 9: *Архитектура RestoreFormer. (a) MHSA – трансформер с multi-head self-attention, используемый в большинстве предыдущих ViT. В нём запросы (Q), ключи (K) и значения (V) взяты из ухудшенной информации Z_d. (b) MHCA – трансформер с multi-head cross-attention, предложенный в RestoreFormer. Он предназначен для пространственного слияния как ухудшенной информации Z_d, так и соответствующих ей высококачественных априорных данных Z_p, принимая Z_d в качестве запросов (Q) и Z_p в качестве пар ключ (K) и значение (V). (c) Архитектура RestoreFormer.*

C. Методы, использующие знания предобученных генеративно-состязательных сетей (GAN)

1) ***GFP-GAN****:* Архитектура GFP-GAN [41] (Рис. 10(a)) состоит из модуля U-Net для удаления деградации и заранее обученной генеративно-состязательной сети (StyleGAN2) для синтеза лиц.

Модуль удаления деградации U-Net разработан для удаления сложных деградаций и извлечения двух видов признаков:

- скрытые признаки F_{latent} для отображения входного изображения на ближайшее скрытое представление W в StyleGAN2;
- многоуровневые пространственные признаки $F_{spatial}$ для нормализации признаков StyleGAN2.

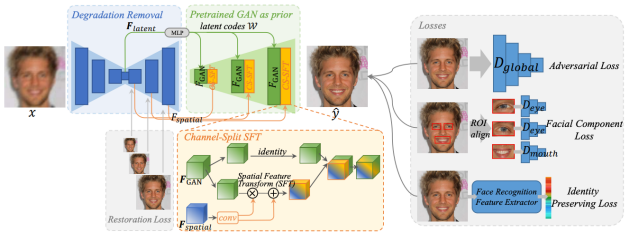


Рис. 10: *Архитектура GFP-GAN.*

U-Net и GAN связаны с помощью отображения в скрытое представления W и нескольких слоев Channel-Split Spatial Feature Transform (CS-SFT).

Channel-Split Spatial Feature Transform необходим для разделения пространственного разделения признаков на две ветки обработки (Рис. 10(a)).

F_{latent} отображается на скрытое представление W с помощью нескольких полносвязных слоев. (MLP на Рис. 10(a)) Используя ближайшее скрытое представление W к входному изображению, StyleGAN2 генерирует промежуточные сверточные признаки F_{GAN}, которые предоставляют богатые детали лица, захваченные весами предварительно обученного GAN.

Многоуровневые пространственные признаки $F_{spatial}$ используются для пространственной модуляции признаков FGAN с помощью предложенных слоев CS-SFT в порядке от крупного к мелкому, обеспечивая генерацию реалистичных результатов.

2) ***GPEN****:* Модель, представленная в [40] объединяет в себе преимущества GAN и DNN (англ. deep neural network). Сначала авторы обучают GAN (Рис. 11(a)), имеющего mapping network и блоки (Рис. 11(b)) подобно StyleGAN [39], для создания высококачественных изображений лиц, а затем встраивают его в U-образную кодировщик-декодировщик архитектуру в качестве декодировщика (Рис. 11(c)), дообучая всю архитектуру на парах изображений I_{LQ} и I_{HQ}.

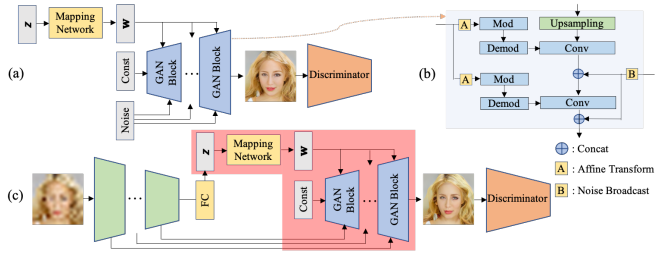


Рис. 11: *Архитектура GPEN. (a) Архитектура GAN. (b) Подробная структура блока GAN. (c) Полная архитектура GPEN.*

3) ***SPARNetHD****:* Архитектура SPARNetHD [45] (Рис. 12) состоит из трех модулей - модуля уменьшения размерности, модуля извлечения признаков и модуля увеличения размерности, каждый из которых состоит из последовательности блоков Face Attention Unit(FAU).

В FAU используется механизм, который сфокусирован на ключевых частях лица, таких как глаза, брови, нос и рот, и придает им большее значение при увеличении разрешения, чем другим частям лица.

Для достижения этого в классические residual блоки вводится механизм пространственного внимания(англ.

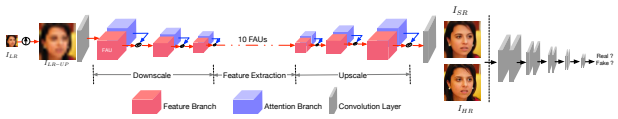


Рис. 12: *Архитектура сети SPARNet*

spatial attention).Spatial Attention - это метод внимания, применяемый в свёрточных нейронных сетях, который позволяет сети сосредоточиться на важных частях изображения, учитывая их пространственное расположение и отношения между ними. Это достигается путем генерации карты пространственного внимания, которая указывает, на каких областях следует усиливать или подавлять признаки.

Авторами отмечено, что последовательность блоков FAU улучшает метрики качества решения задачи увеличения разрешения лица.

4) ***PULSE****:* В традиционных подходах задачу увеличения разрешения (Super-Resolution) часто сводят к обучению функции SR для минимизации среднего расстояния по пикселям между эталонным высококачественным изображением I_{HQ} и восстановленным $SR(I_{LQ})$:

$$L_{SR} = \|I_{HQ} - SR(I_{LQ})\|_p^p. \tag{18}$$

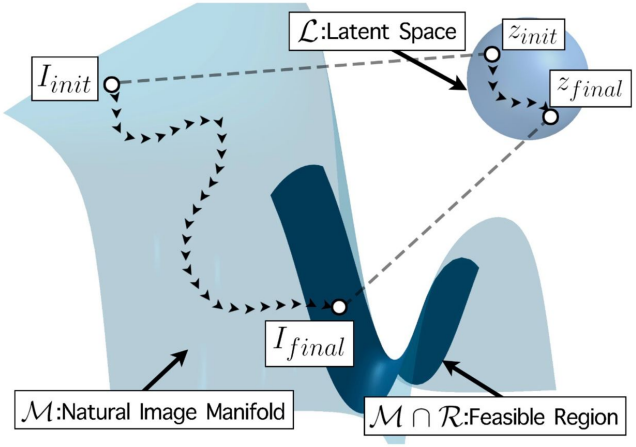


Рис. 13: *Иллюстрация идеи PULSE. $\mathcal{M}$ – естественное многообразие изображений. $\mathcal{R}$ – множество изображений, для которых уменьшение масштаба происходит корректно, т.е. $\mathcal{R} = \{I \in \mathbb{R}^{N \times M} : DS(I) = I_{LQ}\}$. Во время градиентного спуска от z_{init} к z_{final} в скрытом пространстве $\mathcal{L}$ изображение проходит от $I_{init} \in \mathcal{M}$ к $I_{final} \in \mathcal{M} \cap \mathcal{R}$.*

Такая оптимизация зачастую приводит к размытию на восстановленном изображении, особенно в детализированных областях с высокой дисперсией. Поэтому авторы PULSE [38] предлагают искать изображение $\hat{I}_{HQ}$ в естественном многообразии изображений $\mathcal{M}$. Имея дифференцируемую параметризацию многообразия $\mathcal{M}$, можно с помощью градиентного спуска найти $\hat{I}_{HQ}$ путём минимизации функции потерь:

$$L_{DS} = \|DS(\hat{I}_{HQ}) - I_{LQ}\|_p^p. \tag{19}$$

где DS – функция уменьшения масштаба (англ. downscale). Такой подход не требует эталонного I_{HQ} и может быть использован в моделях без учителя. Для аппроксимации многообразия $\mathcal{M}$, авторы предлагают использовать предобученную генеративную модель G со скрытым пространством $\mathcal{L}$, например, StyleGAN [39] и минимизировать:

$$L'_{DS} = \|DS(G(z)) - I_{LQ}\|_p^p. \tag{20}$$

где $z \in \mathcal{L}$. Для того чтобы в градиентном спуске не выходить за пределы $\mathcal{L}$, в [57] было предложено добавить в функцию потерь 20 компоненту l_2 регуляризации для z. Однако такой штраф заставляет векторы стремиться к $\vec{0}$, что не согласуется с тем, что d-мерное стандартное нормальное распределение (которому соответствует $\mathcal{L}$) в пространстве большой размерности очень близко к равномерному распределению на сфере радиуса $\sqrt{d}$ [58]. Поэтому авторы предложили искать вектор z, удовлетворяющий 20 в $\mathcal{L}' = \sqrt{d}\mathbb{S}^{d-1}$, где $\mathbb{S}^{d-1}$ – это единичная сфера в d-мерном Евклидовом пространстве.

5) ***PSFRGAN****:* В данной статье [26] авторы предлагают новый метод для восстановления поврежденных изображений лиц, который называется ”Progressive Semantic-Aware Style Transformation”(PSAST). Идея метода заключается в использовании прогрессивном увеличении разрешения восстановленных изображений, с использованием семантической информации, для эффективного восстановления лиц, даже когда нет информации о повреждениях.

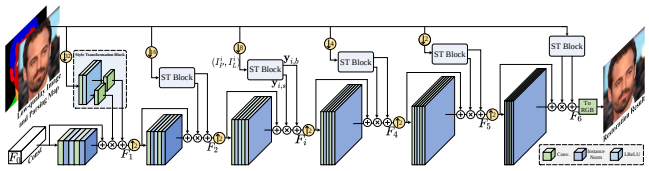


Рис. 14: *Архитектура сети PSFRGAN*

Архитектура модели (Рис. 14) является комбинацией двух подходов: GauGAN [54] и StyleGAN [39], [59].

- Пусть I_d - поврежденное изображение, а I_r - целевое восстановленное изображение;
- Используется ST block после операций понижения размерности для извлечения семантической информации из поврежденного изображения I_d;
- Извлеченная семантическая информация используется для нормирования признаков в прогрессивной сети увеличения разрешения восстановленного изображения.

Модель GauGAN имеет несколько особенностей - в ней нет операций уменьшения размерности и используется слой SPADE Normalization Layer[54] в residual block.

В методе предложена функция потерь semantic-aware style loss(L_{ss}), которая использует значения активации определенных слоев VGG19. Авторами отмечено, что использоване этой функции потерь помогает улучшить восстановление текстур.

6) ***GLEAN****:* Архитектура GLEAN [46] состоит из трех основных компонентов: кодировщика, latent bank и декодировщика. (Рис. 15)

Кодировщик принимает на вход уменьшенное изображение низкого разрешения и извлекает из него многоуровневые свёрточные признаки(на Рис. 15 стрелки, идущие вверх из E_i) и скрытое представление(на Рис. 15

стрелка идущая вниз из E_0), который получается с помощью алгоритма RRDBNet [3].

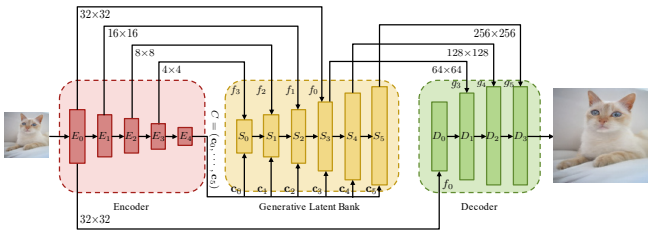


Рис. 15: *Архитектура GLEAN.*

Признаки после сверток содержат в себе информацию о структуре изображения, а скрытое представление представляет собой компактное представление контента и стиля изображения.

Латентный банк представляет собой генеративную модель, использующуюся для генерации новых признаков, которые затем используются декодером для генерации изображения высокого разрешения. В данной работе метод GLEAN использует предобученный StyleGAN в качестве латентного банка.

Декодировщик принимает на вход признаки от кодировщика и латентного банка и генерирует изображение высокого качества. Декодировщик включает в себя многоуровневые свёрточные слои, которые могут повышать разрешение изображения путем интерполяции признаков, полученных от кодировщика и латентного банка. За каждым свёрточным слоём следует слой pixelshuffle [60], за исключением последнего выходного слоя.

Благодаря skip connection между кодировщиком и декодировщиком, информация, извлеченная кодировщиком, может быть использована и, следовательно, латентный набор векторов (на Рис. 15 стрелки, идущие из S_3,S_4,S_5 в D_1,D_2,D_3) может больше сосредоточиться на текстуре и генерации деталей.

Авторами отмечено, что возможность использования высокоуровневых признаков и высокоуровневой структуры изображения позволяет улучшить результаты решения задачи увеличения разрешения.

7) ***FaceFormer****:* В работе [47] предлагается использование архитектуры Трансформер для извлечения признаков деградированного лица с последующим восстановлением с помощью GAN. Также в работе предложена замена интерполяции для увеличения разрешения изображения: использование upsampling слоев с дополнительной информацией об изменении масштаба изображения (Рис. 16 FUP).

Архитектура модели FaceFormer состоит из следующих компонент (Рис. 16):

- Facial Feature Up-sampling (FFUP) - модуль увеличения разрешения изображения лица с учетом коэффициентов масштабирования (s_h, s_v) и относительного расстояния ($R(x)$, $R(y)$) (они помогают извлекать признаки лица с учетом разницы размеров изображений I_{hq} и I_{lq}).
- Facial Feature Embedding (FFE) - модуль, предназначенный для извлечения семантических признаков лица $F_{semantic}$. Он необходим для устранения деградации FFUP. Модуль FFE состоит из Swin Transformer blocks(STB).

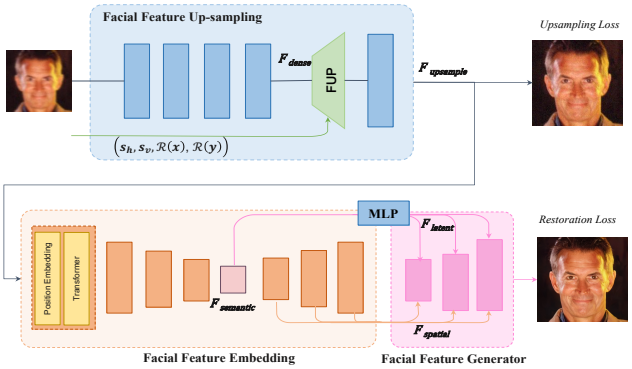


Рис. 16: *Архитектура FaceFormer. Состоит из Facial Feature Up-sampling Module, блока с вниманием для конструирования векторов представлений и генератора черт лица по полученным эмбедингам.*

- Facial Feature Generator (FFG) - необходим для восстановления изображения. В работе используется предобученный StyleGan2. Скрытые представления содержат два вида информативных признаков: скрытые признаки F_{latent} и пространственные признаки $F_{spatial}$. F_{latent} строится из $F_{semantic}$ при помощи нескольких полносвязных слоев $F_{latent} = MLP(F_{semantic})$. $F_{spatial}$ используется для нормирования признаков в GAN.

D. Методы использующие знания предобученных диффузионных моделей

1) ***Диффузионная модель****:* В последнее время наблюдается стремительное развитие диффузионных моделей. Их генеративные способности показывают конкурентоспособные результаты по сравнению с генеративно-состязательными сетями (GAN). Применение диффузионных моделей в задаче слепого восстановления лиц также демонстрирует впечатляющие результаты. Для облегчения последующего изложения методов BFR, использующих в качестве априорной информации знания предобученных диффузионных моделей мы представим краткое введение в диффузионные модели (англ. Denoising Diffusion Probabilistic models, DDPM) [61], [62].

Диффузионная модель имеет прямой процесс из T шагов и обратный процесс из T шагов. Прямой процесс итеративно добавляет случайный шум к данным, а обратный процесс итеративно восстанавливает данные из полученного шума.

Итерация прямого диффузионного процесса получения $\mathbf{x}_t$ из предыдущего состояния $\mathbf{x}_{t-1}$ может быть представлена следующей формулой:

$$\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (21)$$

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (22)$$

где $\mathbf{x}_t$ – зашумленный экземпляр данных (в нашем случае изображение) на шаге t, $\beta_1, \ldots, \beta_t$ – фиксированные, а не обучаемые, константы (заданы в планировщике (англ. scheduler)), $\mathcal{N}$ – нормальное распределение.

Проведя репараметризацию (англ. reparametrization trick), можно выразить зашумленный экземпляр $\mathbf{x}_t$ через исходный, незашумленный $\mathbf{x}_0$:

$$\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (23)$$

$$\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i \quad (24)$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (25)$$

С помощью теоремы Байеса можно представить итерацию обратного диффузионного процесса получения предыдущего состояния $\mathbf{x}_{t-1}$ из текущего $\mathbf{x}_t$ и известного начального $\mathbf{x}_0$ по следующей формуле:

$$\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2\mathbf{I}) \quad (26)$$

$$\mu_(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad (27)$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \quad (28)$$

Но, как правило, $\mathbf{x}_0$ неизвестен, выразив его из уравнения 25 и подставив в формулу 27 получим:

$$\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}) \quad (29)$$

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) \quad (30)$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t\varepsilon \quad (31)$$

где $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, что предполагает, что каждая генерация случайная.

Таким образом, при прямом диффузионном процессе обучаемых параметров нет, а при обратном обучается искусственная нейронная сеть $\epsilon_\theta(\mathbf{x}_t, t)$ с размерностью входа, совпадающей с размерностью выхода (например, U-Net), которая предсказывает шум необходимый для итеративного восстановления данных в уравнении 31.

*2) **Denoising Diffusion Restoration Models, DDRM** :* Авторы DDRM [49] рассматривают задачу восстановления изображений как линейную обратную задачу:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \quad (32)$$

где $\mathbf{x}$ – это оригинальное изображение, которое необходимо восстановить из искаженного $\mathbf{y}$, $\mathbf{H}$ – оператор деградации и $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2\mathbf{I})$ – дополнительный гауссов шум с известной дисперсией.

Чтобы восстановить оригинальное изображение $\mathbf{x}$ авторы определяют DDRM как цепь Маркова $\mathbf{x}_T \to \mathbf{x}_{T-1} \to \cdots \to \mathbf{x}_1 \to \mathbf{x}_0$, обусловленную на $\mathbf{y}$. Результирующий $\mathbf{x}_0$ будет являться восстановленным изображением.

Марковская цепь задаётся в спектральном пространстве оператора $\mathbf{H}$, с помощью элементов из его сингулярного разложения $\mathbf{H} = U\Sigma V^T$ (англ. Singular Value Decomposition, SVD). С помощью вариационного вывода авторы строят целевую функцию для оптимизации DDRM, а также показывают её связь с целевой функцией DDPM/DDIM. Так авторы мотивируют использование предварительно обученной модели DDPM в DDRM.

Таким образом, модифицировать потребуется лишь итерационный процесс, согласно предложенным авторами формулам, а также вид $\mathbf{H}$ и его SVD разложения для различных линейных обратных задач (Рис. 17) (устранение размытия (англ. deblurring), увеличение разрешения (англ. super-resolution), раскрашивание (англ. colorization) и другие). Получается, что DDRM является методом без учителя (англ. unsupervised). Узким местом метода является требование к вычислению сингулярного разложения (SVD), что требует значительных затрат по времени и памяти, если матрица $\mathbf{H}$ имеет большую размерность.

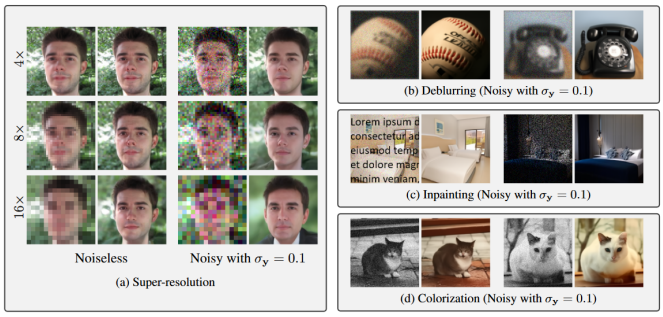


Рис. 17: *Результаты работы DDRM для различных задач: (a) увеличение разрешения (англ. super-resolution) (b) устранение размытия (англ. deblurring) (c) восстановление, дорисовка (англ. inpaiting) (d) раскрашивание (англ. colorization)*

*3) **Denoising Diffusion Null-space Model, DDNM**:* Авторы DDNM [43] пытаются восстановить высококачественное изображение $\hat{\mathbf{x}}$ из низкокачественного изображения $\mathbf{y}$ рассматривая следующую модель деградации:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \quad (33)$$

где $\mathbf{x}$ – эталонное высококачественное изображение, $\mathbf{n}$ – нелинейный шум, $\mathbf{A}$ – линейный оператор деградации.

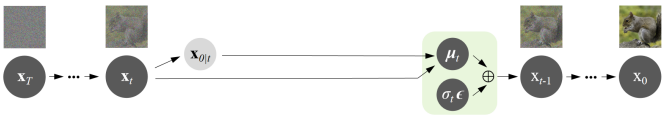


Рис. 18: *Оригинальный обратный диффузионный процесс DDPM.*

Авторы пытаются найти изображение $\hat{\mathbf{x}}$, для которого выполняются два свойства:

$$\textit{Согласованность}: \quad \mathbf{A}\hat{\mathbf{x}} = \mathbf{y} \quad (34)$$

$$\textit{Реалистичность}: \quad \hat{\mathbf{x}} \sim q(\mathbf{x}) \quad (35)$$

где $q(\mathbf{x})$ – распределение эталонных изображений.

Рассмотрим задачу восстановления при отсутствии шума при деградации:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (36)$$

Пусть $\mathbf{A} \in \mathbb{R}^{D\times d}$, $\mathbf{A}^\dagger \in \mathbb{R}^{d\times D}$ – псевдообратная матрица, удовлетворяющая равенству $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$. Заметим, что $\hat{\mathbf{x}}$ вида:

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\bar{\mathbf{x}} \quad (37)$$

удовлетворяет свойству *согласованности* при любом $\bar{\mathbf{x}}$. Действительно:

$$\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\mathbf{A}^\dagger\mathbf{y} + (\mathbf{A} - \mathbf{A}\mathbf{A}^\dagger\mathbf{A})\bar{\mathbf{x}} = \mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{y} \quad (38)$$

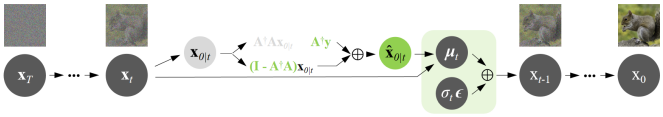


Рис. 19: *Обратный диффузионный процесс DDNM.*

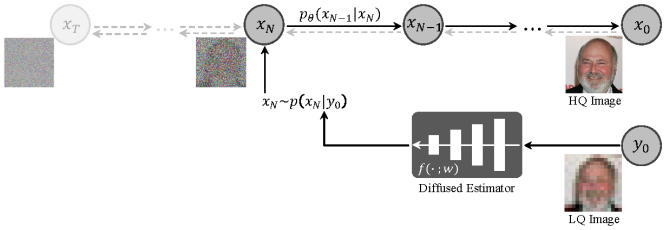


Рис. 20: *Иллюстрация работы DifFace. Сплошными линями обозначены этапы восстановления изображения в DifFace. Для большей наглядности идеи метода, пунктирными линиями обозначен прямые и обратные шаги диффузионной модели.*

От определения $\bar{\mathbf{x}}$ зависит выполнение свойство *реалистичности*, для достижения которой авторы прибегают к диффузионным моделям. Выразив $\mathbf{x}_0$ из уравнения 25:

$$\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)) \tag{39}$$

в DDNM не сразу подставляют его в формулу 27 как в диффузионных моделях, а сначала используют его в качестве $\bar{\mathbf{x}}$ в формуле 37:

$$\hat{\mathbf{x}}_{0|t} = \mathbf{A}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}_{0|t} \tag{40}$$

и уже затем подставляют его в формулу 27 и переходят к $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. Рис. 18 и рис. 19 демонстрируют отличие оригинального обратного диффузионного процесса от обратного процесса, используемого в DDNM. Так как в DDNM используется предварительно обученная диффузионная модель, то благодаря такому подходу будет выполняться свойство *реалистичности*, т.е. финальный результат $\mathbf{x}_0 \sim q(\mathbf{x})$.

Кроме того, авторы представляют улучшенную версию DDDM$^+$, которая решает задачу восстановления 33 с ненулевым шумом $\mathbf{n}$. В отличие от DDRM [49] DDNM не требует трудоемкого вычисления сингулярного разложения матрицы деградации – авторы предварительно строят матрицы $\mathbf{A}$ и $\mathbf{A}^\dagger$ для задач раскрашивания (англ. colorization), дорисовки (англ. inpainting) и увеличения разрешения (англ. super-resolution).

4) ***DifFace*** *[44]:* Авторы вдохновились недавними успехами диффузионных моделей в задаче генерации изображений и представили метод способный справляться со сложными деградациями, используя сильные генеративные возможности предварительно обученной диффузионной модели без её переобучения на каких-либо предполагаемых вручную ухудшениях с помощью заданной модели деградации. Сначала исходное низкокачественное изображение I_{LQ} поступает на вход в так называемый *"diffused estimator"* $f(\cdot; w)$, представляющего собой некоторую стандартную модель для восстановления изображения, например SRCNN [1] или SwinIR [63]. Далее к выходу diffused estimator добавляется шум по формуле 25, переводя его в промежуточное состояние x_N на шаге N прямого диффузионного процесса. Наконец, это промежуточное состояние x_N подвергается обратному диффузионному процессу, результатом которого является I_{HQ}.

5) ***DR2****:* Вдохновившись идеей управляемой генерации изображений с помощью обусловливания диффузионного процесса DDPM, представленной в ILVR [64], авторы DR2 [50] предложили метод на основе диффузионных моделей для слепого восстановления лиц.

Общая схема работы DR2 проиллюстрирована на Рис 21. Сначала в изображение $\mathbf{y}$ с неизвестными деградациями добавляется шум с помощью прямого диффузионного процесса (формула 22), в результате получается

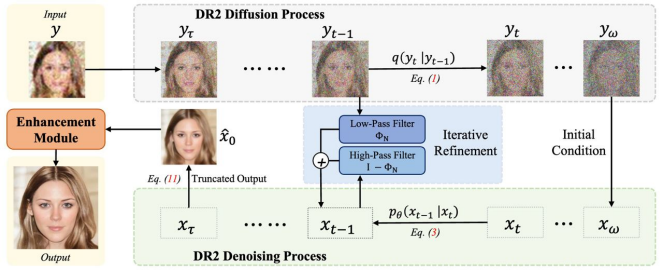


Рис. 21: *Схема работы DR2.*

$\mathbf{y}_\omega$, которое используется в качестве начального $\mathbf{x}_\omega$ при обратном диффузионном процессе. Затем при обратном диффузионном процессе к $\mathbf{x}_{t-1}$, полученному из $\mathbf{x}_t$ на итерации t по формуле 31 применяется *итеративное улучшение* по формуле:

$$\mathbf{x}_{t-1} = \Phi_N(y_{t-1}) + (\mathbf{I} - \Phi_N)(\mathbf{x}_{t-1}) \tag{41}$$

где $\Phi_N(\cdot)$ – это фильтр низких частот (англ. low-pass filter), реализованный путём понижения разрешения изображения (англ. downsampling) и последующим повышением разрешения (англ. upsampling) с общим коэффициентом N. Тогда $(\mathbf{I} - \Phi_N)$ можно рассматривать как фильтр верхних частот (англ. high-pass filter). Таким образом, отбрасывается высокочастотная часть $\mathbf{y}$, так как она содержит мало информации из-за деградации. Так обратный диффузионный процесс обуславливается на $\mathbf{y}$, гарантируя, что результат будет содержать базовую семантику исходного изображения с деградациями. Чем меньше t, то есть на поздних стадиях обратного диффузионного процесса расстояние между распределениями $q(\mathbf{x}_t|\mathbf{x}_w)$ и $q(\mathbf{y}_t|\mathbf{y})$ будет становиться больше, поэтому на некотором шаге τ $(0 < \tau < \omega)$ обратный процесс прекращается и авторы предсказывают $\mathbf{x}_{0|\tau}$ выразив его из формулы 25:

$$\mathbf{x}_{0|\tau} = \frac{1}{\sqrt{\bar{\alpha}_\tau}}(\mathbf{x}_\tau - \sqrt{1-\bar{\alpha}_\tau}\epsilon_\theta(\mathbf{x}_\tau, \tau)) \tag{42}$$

Такой подход позволит справиться со сложными деградациями. После дальнейшего применения модуля улучшения (англ. enhancement module) к $\mathbf{x}_{0|\tau}$, в качестве которого может служить другая передовая искусственная нейронная сеть для слепого восстановления лица (например, VQFR [34], CodeFormer [33]), будет получено результирующее высококачественное восстановленное изображение. Преимущество подхода заключается в том, что достаточно взять предварительно обученную диффузионную модель для генерации человеческих лиц без необходимости в её дообучении.

VI. Наборы данных

Большинство исследователей задачи BFR самостоятельно адаптируют существующие наборы данных изображений лиц, вводя деградации по формуле 7. В настоящий момент доступно только два набора данных [48], подготовленных специально для задачи BFR. Далее предоставлено краткое описание наборов данных, используемых исследователями при решении задачи BFR:

CelebA [65] - это набор данных с атрибутами лица, в котором использовались изображения лиц из набора данных CelebFaces [66]. Он содержит $202,599$ изображений лиц с $10,177$ уникальными идентификаторами человека. Каждое изображение в CelebA аннотировано 40 атрибутами лица и 5 ключевыми точками. На основе набора данных CelebA существующие методы [35], [32] создают набор данных CelebA-Test для валидации модели. CelebA-Test - это синтетический набор данных с 3000 изображениями CelebA-HQ из тестового набора данных CelebA.

FFHQ содержит $70,000$ высококачественных изображений размера 1024×1024 извлеченных из Интернета.

CASIA-WebFace [67] была выпущена в 2014 году. Она состоит из $494,414$ изображений лиц из $10,575$ разных субъектов. Каждое изображение имеет разрешение $250{\times}250$.

VGGFace2 [68] - это большой набор данных лиц, который включает в себя $3,31$ миллиона изображений от $9,131$ людей. По каждому человеку в этом наборе данных находится в среднем $362,6$ изображений. Изображения в VGGFace2 собраны из Google Image Search и разнообразны по положению, возрасту и фонам. Кроме того, каждое изображение лица в этом наборе данных имеет ограничивающую рамку, проверенную человеком вокруг лица и пять эталонных ключевых точек, оцененных моделью [69].

IMDB-WIKI [70] состоит из $524,230$ изображений лиц, собранных с сайтов IMDB и Wikipedia. Среди них $461,871$ изображение лица получено с IMDB, а $62,359$ - с веб-сайтов Википедии.

Helen [71] - это сложный набор данных по локализации атрибутов лица, содержащий $2,330$ изображений лиц в высоком разрешении, сделанных в разных местах (в домашних условиях, на улице, в фотостудии и др.). Этот набор данных содержит 194 метки для каждого изображения лица.

WIDER-face dataset [72] это набор данных для распознавания лиц, изображения которого выбирались из общедоступного набора данных WIDER [73]. Он состоит из $32,203$ изображений и $393,703$ лиц.

LWF [74] содержит взятые из Интернета изображения низкого качества со средними по уровню деградациями.

BioID [75] был создан в 2001 году и включает в себя $1,521$ изображение лиц 23 людей, сделанных в оттенках серого.

AFLW [76] - это крупный набор данных для выравнивания лиц, собранный с Flickr. Он включает в себя $25,993$ изображения лиц, аннотированных до 21 метки для каждого изображения. В данном наборе данных присутствуют лица с разнообразными эмоциональными выражениями.

EDFace-Celeb-1M [77] - это набор данных для задачи увеличения разрешения фотографий. По сравнению с существующими наборами данных для лиц, EDFaceCeleb-1M полностью учитывает расовое распределение между людьми в процессе создания набора. Он содержит $1,7$ миллиона изображений лиц, охватывающих людей из разных стран. Набор предоставляет собой пары изображений лиц низкого и высокого разрешения. Для обучения и тестирования моделей было использовано $1,5$ миллиона пар изображений лиц, а также имеется $140,000$ изображений реальных лиц маленького размера для проведения визуальных сравнений.

EDFace-Celeb-1M (BFR128) [48] - это набор данных для оценки производительности алгоритмов по восстановлению изображений лиц, созданный для слепого восстановления изображений (BFR). Высококачественные изображения в этом наборе данных выбраны из EDFace-Celeb-1M [77]. Авторы используют модели деградации (размытие, шум, низкое разрешение, искусственные артефакты JPEG-сжатия и полная деградация) для синтеза низкокачественных изображений на основе высококачественных изображений. С помощью этих разных моделей этот набор данных может использоваться для задач размытия, шумоподавления, удаления искусственных артефактов, увеличения разрешения и слепого восстановления лиц. Для каждой деградации в наборе содержится $1,5$ миллиона изображений с разрешением $128{\times}128$. $1,36$ миллиона изображений лиц используются для обучения и $145,000$ - для тестирования.

EDFace-Celeb-150K (BFR512) [48] - еще один набор данных для восстановления лица вслепую. Деградация этого набора данных такая же, как и у **EDFace-Celeb-1M (BFR128)** [48]. Он также имеет пять моделей деградаций, включая размытие, шум, низкое разрешение, артефакты сжатия JPEG и их комбинацию. Набор содержит 149 тысяч изображений с разрешением $512{\times}512$. Количество обучающих и тестовых изображений составляет около $132,000$ и $17,000$ соответственно.

VII. Заключение

В данной обзорной статье мы подробно рассмотрели современные методы и подходы, используемые в области восстановления изображений лиц с применением глубокого обучения. Основные аспекты нашего исследования включают анализ моделей деградаций, особенности изображений лиц, вызовы, связанные с реконструкцией лиц. Проведен анализ разнообразных методов восстановления лиц, включая геометрические подходы, которые учитывают структуру и особенности лица, эталонные подходы, опирающиеся на сравнение с образцами, и генеративные подходы, использующие генеративные сети для синтеза реалистичных изображений лиц.

Кроме того, изучены подходы к реконструкции лиц с точки зрения архитектуры нейронной сети, базовых компонентов, функций потерь и наборов данных.

Несмотря на достигнутые успехи в области восстановления изображений лиц, существуют все еще нерешенные проблемы и вызовы, такие как реконструкция лиц при сильной деградации, улучшение реконструкции текстур и деталей лица.

Список литературы

[1] Image super-resolution using deep convolutional networks / Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang // IEEE

transactions on pattern analysis and machine intelligence. — 2015. — Vol. 38, no. 2. — P. 295–307.
[2] Enhanced deep residual networks for single image super-resolution / Bee Lim, Sanghyun Son, Heewon Kim et al. // Proceedings of the IEEE conference on computer vision and pattern recognition workshops. — 2017. — P. 136–144.
[3] Esrgan: Enhanced super-resolution generative adversarial networks / Xintao Wang, Ke Yu, Shixiang Wu et al. // Proceedings of the European conference on computer vision (ECCV) workshops. — 2018. — P. 0–0.
[4] Image super-resolution using very deep residual channel attention networks / Yulun Zhang, Kunpeng Li, Kai Li et al. // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 286–301.
[5] Photo-realistic single image super-resolution using a generative adversarial network / Christian Ledig, Lucas Theis, Ferenc Huszár et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 4681–4690.
[6] Sajjadi Mehdi SM, Scholkopf Bernhard, Hirsch Michael. Enhancenet: Single image super-resolution through automated texture synthesis // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 4491–4500.
[7] Variational denoising network: Toward blind noise modeling and removal / Zongsheng Yue, Hongwei Yong, Qian Zhao et al. // Advances in neural information processing systems. — 2019. — Vol. 32.
[8] Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising / Kai Zhang, Wangmeng Zuo, Yunjin Chen et al. // IEEE transactions on image processing. — 2017. — Vol. 26, no. 7. — P. 3142–3155.
[9] Zhang Kai, Zuo Wangmeng, Zhang Lei. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising // IEEE Transactions on Image Processing. — 2018. — Vol. 27, no. 9. — P. 4608–4622.
[10] Deblurgan: Blind motion deblurring using conditional adversarial networks / Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 8183–8192.
[11] Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better / Orest Kupyn, Tetiana Martyniuk, Junru Wu, Zhangyang Wang // Proceedings of the IEEE/CVF international conference on computer vision. — 2019. — P. 8878–8887.
[12] Deblurring by realistic blurring / Kaihao Zhang, Wenhan Luo, Yiran Zhong et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2020. — P. 2737–2746.
[13] Adversarial spatio-temporal learning for video deblurring / Kaihao Zhang, Wenhan Luo, Yiran Zhong et al. // IEEE Transactions on Image Processing. — 2018. — Vol. 28, no. 1. — P. 291–301.
[14] Compression artifacts reduction by a deep convolutional network / Chao Dong, Yubin Deng, Chen Change Loy, Xiaoou Tang // Proceedings of the IEEE international conference on computer vision. — 2015. — P. 576–584.
[15] Jpeg artifacts reduction via deep convolutional sparse coding / Xueyang Fu, Zheng-Jun Zha, Feng Wu et al. // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2019. — P. 2501–2510.
[16] Deep learning vs. traditional computer vision / Niall O'Mahony, Sean Campbell, Anderson Carvalho et al. // Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1 / Springer. — 2020. — P. 128–144.
[17] Wang Zhou, Bovik Alan C. A universal image quality index // IEEE signal processing letters. — 2002. — Vol. 9, no. 3. — P. 81–84.
[18] The unreasonable effectiveness of deep features as a perceptual metric / Richard Zhang, Phillip Isola, Alexei A Efros et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 586–595.
[19] Gans trained by a two time-scale update rule converge to a local nash equilibrium / Martin Heusel, Hubert Ramsauer, Thomas Unterthiner et al. // Advances in neural information processing systems. — 2017. — Vol. 30.
[20] Mittal Anish, Soundararajan Rajiv, Bovik Alan C. Making a "completely blind" image quality analyzer // IEEE Signal processing letters. — 2012. — Vol. 20, no. 3. — P. 209–212.
[21] Attention-aware face hallucination via deep reinforcement learning / Qingxing Cao, Liang Lin, Yukai Shi et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 690–698.
[22] Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution / Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 1689–1697.
[23] Learning to super-resolve blurry face and text images / Xiangyu Xu, Deqing Sun, Jinshan Pan et al. // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 251–260.
[24] Fsrnet: End-to-end learning face super-resolution with facial priors / Yu Chen, Ying Tai, Xiaoming Liu et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 2492–2501.
[25] Face super-resolution guided by facial component heatmaps / Xin Yu, Basura Fernando, Bernard Ghanem et al. // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 217–233.
[26] Progressive semantic-aware style transformation for blind face restoration / Chaofeng Chen, Xiaoming Li, Lingbo Yang et al. // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — P. 11896–11905.
[27] Deep semantic face deblurring / Ziyi Shen, Wei-Sheng Lai, Tingfa Xu et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 8260–8269.
[28] Hifacegan: Face renovation via collaborative suppression and replenishment / Lingbo Yang, Shanshe Wang, Siwei Ma et al. // Proceedings of the 28th ACM international conference on multimedia. — 2020. — P. 1551–1560.
[29] Face super-resolution guided by 3d facial priors / Xiaobin Hu, Wenqi Ren, John LaMaster et al. // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16 / Springer. — 2020. — P. 763–780.
[30] Face video deblurring using 3d facial priors / Wenqi Ren, Jiaolong Yang, Senyou Deng et al. // Proceedings of the IEEE/CVF international conference on computer vision. — 2019. — P. 9388–9397.
[31] Learning warped guidance for blind face restoration / Xiaoming Li, Ming Liu, Yuting Ye et al. // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 272–289.
[32] Blind face restoration via deep multi-scale component dictionaries / Xiaoming Li, Chaofeng Chen, Shangchen Zhou et al. // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16 / Springer. — 2020. — P. 399–415.
[33] Zhou Shangchen, Chan Kelvin C. K., Li Chongyi, Loy Chen Change. Towards robust blind face restoration with codebook lookup transformer. — 2022. — 2206.11253.
[34] Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder / Yuchao Gu, Xintao Wang, Liangbin Xie et al. // Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII / Springer. — 2022. — P. 126–143.
[35] Restoreformer: High-quality blind face restoration from undegraded key-value pairs / Zhouxia Wang, Jiawei Zhang, Runjian Chen et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — P. 17512–17521.
[36] Van Den Oord Aaron, Vinyals Oriol et al. Neural discrete representation learning // Advances in neural information processing systems. — 2017. — Vol. 30.
[37] Esser Patrick, Rombach Robin, Ommer Bjorn. Taming transformers for high-resolution image synthesis // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — P. 12873–12883.
[38] Pulse: Self-supervised photo upsampling via latent space exploration of generative models / Sachit Menon, Alexandru Damian, Shijia Hu et al. // Proceedings of the ieee/cvf conference on computer vision and pattern recognition. — 2020. — P. 2437–2445.
[39] Karras Tero, Laine Samuli, Aila Timo. A style-based generator architecture for generative adversarial networks // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2019. — P. 4401–4410.
[40] Gan prior embedded network for blind face restoration in the wild / Tao Yang, Peiran Ren, Xuansong Xie, Lei Zhang // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2021. — P. 672–681.
[41] Towards real-world blind face restoration with generative facial prior / Xintao Wang, Yu Li, Honglun Zhang, Ying Shan // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2021. — P. 9168–9178.
[42] Dhariwal Prafulla, Nichol Alexander. Diffusion models beat gans on image synthesis // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — P. 8780–8794.
[43] Wang Yinhuai, Yu Jiwen, Zhang Jian. Zero-shot image restoration using denoising diffusion null-space model // arXiv preprint arXiv:2212.00490. — 2022.
[44] Yue Zongsheng, Loy Chen Change. Difface: Blind face restoration with diffused error contraction // arXiv preprint arXiv:2212.06512. — 2022.
[45] Learning spatial attention for face super-resolution / Chaofeng Chen, Dihong Gong, Hao Wang et al. // IEEE Transactions on Image Processing. — 2020. — Vol. 30. — P. 1219–1231.

[46] Glean: Generative latent bank for large-factor image super-resolution / Kelvin CK Chan, Xintao Wang, Xiangyu Xu et al. // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — P. 14245–14254.
[47] Faceformer: Scale-aware blind face restoration with transformers / Aijin Li, Gen Li, Lei Sun, Xintao Wang // arXiv preprint arXiv:2207.09790. — 2022.
[48] Blind face restoration: Benchmark datasets and a baseline model / Puyang Zhang, Kaihao Zhang, Wenhan Luo et al. // arXiv preprint arXiv:2206.03697. — 2022.
[49] Denoising diffusion restoration models / Bahjat Kawar, Michael Elad, Stefano Ermon, Jiaming Song // arXiv preprint arXiv:2201.11793. — 2022.
[50] Dr2: Diffusion-based robust degradation remover for blind face restoration / Zhixin Wang, Xiaoyun Zhang, Ziying Zhang et al. // arXiv preprint arXiv:2303.06885. — 2023.
[51] Swin transformer: Hierarchical vision transformer using shifted windows / Ze Liu, Yutong Lin, Yue Cao et al. // Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — P. 10012–10022.
[52] A new class of efficient adaptive filters for online nonlinear modeling / Danilo Comminiello, Alireza Nezamdoust, Simone Scardapane et al. // IEEE Transactions on Systems, Man, and Cybernetics: Systems. — 2023. — mar. — Vol. 53, no. 3. — P. 1384–1396.
[53] Cai Changjiang, Mordohai Philippos. Do end-to-end stereo algorithms under-utilize information? — 2020. — 10.
[54] Park Taesung, Liu Ming-Yu, Wang Ting-Chun, Zhu Jun-Yan. Semantic image synthesis with spatially-adaptive normalization. — 2019. — 1903.07291.
[55] He Kaiming, Gkioxari Georgia, Dollár Piotr, Girshick Ross. Mask r-cnn. — 2018. — 1703.06870.
[56] Dai Jifeng, Qi Haozhi, Xiong Yuwen et al. Deformable convolutional networks. — 2017. — 1703.06211.
[57] Compressed sensing using generative models / Ashish Bora, Ajil Jalal, Eric Price, Alexandros G Dimakis // International Conference on Machine Learning / PMLR. — 2017. — P. 537–546.
[58] Vershynin Roman. Random vectors in high dimensions // Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. — 2018. — Vol. 3. — P. 38–69.
[59] Karras Tero, Laine Samuli, Aittala Miika et al. Analyzing and improving the image quality of stylegan. — 2020. — 1912.04958.
[60] Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network / Wenzhe Shi, Jose Caballero, Ferenc Huszár et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — P. 1874–1883.
[61] Deep unsupervised learning using nonequilibrium thermodynamics / Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli // International Conference on Machine Learning / PMLR. — 2015. — P. 2256–2265.
[62] Ho Jonathan, Jain Ajay, Abbeel Pieter. Denoising diffusion probabilistic models // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 6840–6851.
[63] Swinir: Image restoration using swin transformer / Jingyun Liang, Jiezhang Cao, Guolei Sun et al. // Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — P. 1833–1844.
[64] Ilvr: Conditioning method for denoising diffusion probabilistic models / Jooyoung Choi, Sungwon Kim, Yonghyun Jeong et al. // arXiv preprint arXiv:2108.02938. — 2021.
[65] Deep learning face attributes in the wild / Ziwei Liu, Ping Luo, Xiaogang Wang, Xiaoou Tang // Proceedings of the IEEE international conference on computer vision. — 2015. — P. 3730–3738.
[66] Deep learning face representation by joint identification-verification / Yi Sun, Yuheng Chen, Xiaogang Wang, Xiaoou Tang // Advances in neural information processing systems. — 2014. — Vol. 27.
[67] Learning face representation from scratch / Dong Yi, Zhen Lei, Shengcai Liao, Stan Z Li // arXiv preprint arXiv:1411.7923. — 2014.
[68] Vggface2: A dataset for recognising faces across pose and age / Qiong Cao, Li Shen, Weidi Xie et al. // 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018) / IEEE. — 2018. — P. 67–74.
[69] Joint face detection and alignment using multitask cascaded convolutional networks / Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao // IEEE signal processing letters. — 2016. — Vol. 23, no. 10. — P. 1499–1503.
[70] Rothe Rasmus, Timofte Radu, Van Gool Luc. Dex: Deep expectation of apparent age from a single image // Proceedings of the IEEE international conference on computer vision workshops. — 2015. — P. 10–15.
[71] Interactive facial feature localization / Vuong Le, Jonathan Brandt, Zhe Lin et al. // Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12 / Springer. — 2012. — P. 679–692.
[72] Yang Shuo, Luo Ping, Loy Chen Change, Tang Xiaoou. Wider face: A face detection benchmark. — 2015. — 1511.06523.
[73] Recognize complex events from static images by fusing deep channels / Yuanjun Xiong, Kai Zhu, Dahua Lin, Xiaoou Tang // Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on / IEEE. — 2015.
[74] Labeled faces in the wild: A database for studying face recognition in unconstrained environments : Rep. : 07-49 / University of Massachusetts, Amherst ; Executor: Gary B. Huang, Manu Ramesh, Tamara Berg, Erik Learned-Miller : 2007. — October.
[75] Jesorsky Oliver, Kirchberg Klaus J., Frischholz Robert W. Robust face detection using the hausdorff distance // Audio- and Video-Based Biometric Person Authentication / Ed. by Josef Bigun, Fabrizio Smeraldi. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2001. — P. 90–95.
[76] Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization / Martin Köstinger, Paul Wohlhart, Peter M. Roth, Horst Bischof // 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). — 2011. — P. 2144–2151.
[77] Zhang Kaihao, Li Dongxu, Luo Wenhan et al. Edface-celeb-1m: Benchmarking face hallucination with a million-scale dataset. — 2022. — 2110.05031.

Blind Face Restoration Survey

Sait Sharipov, Bulat Nutfullin, Narek Maloyan

***Abstract*—The importance of researching methods for blind face restoration (BFR) arises from their potential practical applications in various domains. Examples of such areas include digital art and computer graphics for character face reconstruction and animation, as well as social networks and mobile applications, where they contribute to improving the quality of images and videos.**

In this paper, we conduct a review of contemporary methods and approaches used for solving the BFR problem. We examine various types of models based on generative adversarial networks, autoencoders, and diffusion models, which have demonstrated significant progress in this field. Specifically, we analyze key aspects such as network architecture, loss functions, quality metrics, and datasets.

Furthermore, we discuss the issues and limitations of existing methods, as well as possible directions for future research. In particular, we emphasize the need for developing algorithms that are robust to various degradations and capable of adapting to different lighting conditions, poses, and facial expressions. In conclusion, we provide a systematic comparison of existing methods and summarize their merits and drawbacks.

***Keywords*—blind face restoration, low resolution, noise, compression artifacts, blur, deep learning, diffusion model, generative adversarial network, GAN, image restoration**

References

[1] Image super-resolution using deep convolutional networks / Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang // IEEE transactions on pattern analysis and machine intelligence. — 2015. — Vol. 38, no. 2. — P. 295–307.

[2] Enhanced deep residual networks for single image super-resolution / Bee Lim, Sanghyun Son, Heewon Kim et al. // Proceedings of the IEEE conference on computer vision and pattern recognition workshops. — 2017. — P. 136–144.

[3] Esrgan: Enhanced super-resolution generative adversarial networks / Xintao Wang, Ke Yu, Shixiang Wu et al. // Proceedings of the European conference on computer vision (ECCV) workshops. — 2018. — P. 0–0.

[4] Image super-resolution using very deep residual channel attention networks / Yulun Zhang, Kunpeng Li, Kai Li et al. // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 286–301.

[5] Photo-realistic single image super-resolution using a generative adversarial network / Christian Ledig, Lucas Theis, Ferenc Huszár et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 4681–4690.

[6] Sajjadi Mehdi SM, Scholkopf Bernhard, Hirsch Michael. Enhancenet: Single image super-resolution through automated texture synthesis // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 4491–4500.

[7] Variational denoising network: Toward blind noise modeling and removal / Zongsheng Yue, Hongwei Yong, Qian Zhao et al. // Advances in neural information processing systems. — 2019. — Vol. 32.

[8] Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising / Kai Zhang, Wangmeng Zuo, Yunjin Chen et al. // IEEE transactions on image processing. — 2017. — Vol. 26, no. 7. — P. 3142–3155.

[9] Zhang Kai, Zuo Wangmeng, Zhang Lei. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising // IEEE Transactions on Image Processing. — 2018. — Vol. 27, no. 9. — P. 4608–4622.

[10] Deblurgan: Blind motion deblurring using conditional adversarial networks / Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 8183–8192.

[11] Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better / Orest Kupyn, Tetiana Martyniuk, Junru Wu, Zhangyang Wang // Proceedings of the IEEE/CVF international conference on computer vision. — 2019. — P. 8878–8887.

[12] Deblurring by realistic blurring / Kaihao Zhang, Wenhan Luo, Yiran Zhong et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2020. — P. 2737–2746.

[13] Adversarial spatio-temporal learning for video deblurring / Kaihao Zhang, Wenhan Luo, Yiran Zhong et al. // IEEE Transactions on Image Processing. — 2018. — Vol. 28, no. 1. — P. 291–301.

[14] Compression artifacts reduction by a deep convolutional network / Chao Dong, Yubin Deng, Chen Change Loy, Xiaoou Tang // Proceedings of the IEEE international conference on computer vision. — 2015. — P. 576–584.

[15] Jpeg artifacts reduction via deep convolutional sparse coding / Xueyang Fu, Zheng-Jun Zha, Feng Wu et al. // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2019. — P. 2501–2510.

[16] Deep learning vs. traditional computer vision / Niall O'Mahony, Sean Campbell, Anderson Carvalho et al. // Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1 / Springer. — 2020. — P. 128–144.

[17] Wang Zhou, Bovik Alan C. A universal image quality index // IEEE signal processing letters. — 2002. — Vol. 9, no. 3. — P. 81–84.

[18] The unreasonable effectiveness of deep features as a perceptual metric / Richard Zhang, Phillip Isola, Alexei A Efros et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 586–595.

[19] Gans trained by a two time-scale update rule converge to a local nash equilibrium / Martin Heusel, Hubert Ramsauer, Thomas Unterthiner et al. // Advances in neural information processing systems. — 2017. — Vol. 30.

[20] Mittal Anish, Soundararajan Rajiv, Bovik Alan C. Making a "completely blind" image quality analyzer // IEEE Signal processing letters. — 2012. — Vol. 20, no. 3. — P. 209–212.

[21] Attention-aware face hallucination via deep reinforcement learning / Qingxing Cao, Liang Lin, Yukai Shi et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 690–698.

[22] Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution / Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 1689–1697.

[23] Learning to super-resolve blurry face and text images / Xiangyu Xu, Deqing Sun, Jinshan Pan et al. // Proceedings of the IEEE international conference on computer vision. — 2017. — P. 251–260.

[24] Fsrnet: End-to-end learning face super-resolution with facial priors / Yu Chen, Ying Tai, Xiaoming Liu et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 2492–2501.

[25] Face super-resolution guided by facial component heatmaps / Xin Yu, Basura Fernando, Bernard Ghanem et al. // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 217–233.

[26] Progressive semantic-aware style transformation for blind face restoration / Chaofeng Chen, Xiaoming Li, Lingbo Yang et al. // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — P. 11896–11905.

[27] Deep semantic face deblurring / Ziyi Shen, Wei-Sheng Lai, Tingfa Xu et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 8260–8269.

[28] Hifacegan: Face renovation via collaborative suppression and replenishment / Lingbo Yang, Shanshe Wang, Siwei Ma et al. // Proceedings of the 28th ACM international conference on multimedia. — 2020. — P. 1551–1560.

[29] Face super-resolution guided by 3d facial priors / Xiaobin Hu, Wenqi Ren, John LaMaster et al. // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16 / Springer. — 2020. — P. 763–780.

[30] Face video deblurring using 3d facial priors / Wenqi Ren, Jiaolong Yang, Senyou Deng et al. // Proceedings of the IEEE/CVF

international conference on computer vision. — 2019. — P. 9388–9397.
[31] Learning warped guidance for blind face restoration / Xiaoming Li, Ming Liu, Yuting Ye et al. // Proceedings of the European conference on computer vision (ECCV). — 2018. — P. 272–289.
[32] Blind face restoration via deep multi-scale component dictionaries / Xiaoming Li, Chaofeng Chen, Shangchen Zhou et al. // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16 / Springer. — 2020. — P. 399–415.
[33] Zhou Shangchen, Chan Kelvin C. K., Li Chongyi, Loy Chen Change. Towards robust blind face restoration with codebook lookup transformer. — 2022. — 2206.11253.
[34] Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder / Yuchao Gu, Xintao Wang, Liangbin Xie et al. // Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII / Springer. — 2022. — P. 126–143.
[35] Restoreformer: High-quality blind face restoration from undegraded key-value pairs / Zhouxia Wang, Jiawei Zhang, Runjian Chen et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — P. 17512–17521.
[36] Van Den Oord Aaron, Vinyals Oriol et al. Neural discrete representation learning // Advances in neural information processing systems. — 2017. — Vol. 30.
[37] Esser Patrick, Rombach Robin, Ommer Bjorn. Taming transformers for high-resolution image synthesis // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — P. 12873–12883.
[38] Pulse: Self-supervised photo upsampling via latent space exploration of generative models / Sachit Menon, Alexandru Damian, Shijia Hu et al. // Proceedings of the ieee/cvf conference on computer vision and pattern recognition. — 2020. — P. 2437–2445.
[39] Karras Tero, Laine Samuli, Aila Timo. A style-based generator architecture for generative adversarial networks // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2019. — P. 4401–4410.
[40] Gan prior embedded network for blind face restoration in the wild / Tao Yang, Peiran Ren, Xuansong Xie, Lei Zhang // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2021. — P. 672–681.
[41] Towards real-world blind face restoration with generative facial prior / Xintao Wang, Yu Li, Honglun Zhang, Ying Shan // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2021. — P. 9168–9178.
[42] Dhariwal Prafulla, Nichol Alexander. Diffusion models beat gans on image synthesis // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — P. 8780–8794.
[43] Wang Yinhuai, Yu Jiwen, Zhang Jian. Zero-shot image restoration using denoising diffusion null-space model // arXiv preprint arXiv:2212.00490. — 2022.
[44] Yue Zongsheng, Loy Chen Change. Diffface: Blind face restoration with diffused error contraction // arXiv preprint arXiv:2212.06512. — 2022.
[45] Learning spatial attention for face super-resolution / Chaofeng Chen, Dihong Gong, Hao Wang et al. // IEEE Transactions on Image Processing. — 2020. — Vol. 30. — P. 1219–1231.
[46] Glean: Generative latent bank for large-factor image super-resolution / Kelvin CK Chan, Xintao Wang, Xiangyu Xu et al. // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — P. 14245–14254.
[47] Faceformer: Scale-aware blind face restoration with transformers / Aijin Li, Gen Li, Lei Sun, Xintao Wang // arXiv preprint arXiv:2207.09790. — 2022.
[48] Blind face restoration: Benchmark datasets and a baseline model / Puyang Zhang, Kaihao Zhang, Wenhan Luo et al. // arXiv preprint arXiv:2206.03697. — 2022.
[49] Denoising diffusion restoration models / Bahjat Kawar, Michael Elad, Stefano Ermon, Jiaming Song // arXiv preprint arXiv:2201.11793. — 2022.
[50] Dr2: Diffusion-based robust degradation remover for blind face restoration / Zhixin Wang, Xiaoyun Zhang, Ziying Zhang et al. // arXiv preprint arXiv:2303.06885. — 2023.
[51] Swin transformer: Hierarchical vision transformer using shifted windows / Ze Liu, Yutong Lin, Yue Cao et al. // Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — P. 10012–10022.
[52] A new class of efficient adaptive filters for online nonlinear modeling / Danilo Comminiello, Alireza Nezamdoust, Simone Scardapane et al. // IEEE Transactions on Systems, Man, and Cybernetics: Systems. — 2023. — mar. — Vol. 53, no. 3. — P. 1384–1396.
[53] Cai Changjiang, Mordohai Philippos. Do end-to-end stereo algorithms under-utilize information? — 2020. — 10.
[54] Park Taesung, Liu Ming-Yu, Wang Ting-Chun, Zhu Jun-Yan. Semantic image synthesis with spatially-adaptive normalization. — 2019. — 1903.07291.
[55] He Kaiming, Gkioxari Georgia, Dollár Piotr, Girshick Ross. Mask r-cnn. — 2018. — 1703.06870.
[56] Dai Jifeng, Qi Haozhi, Xiong Yuwen et al. Deformable convolutional networks. — 2017. — 1703.06211.
[57] Compressed sensing using generative models / Ashish Bora, Ajil Jalal, Eric Price, Alexandros G Dimakis // International Conference on Machine Learning / PMLR. — 2017. — P. 537–546.
[58] Vershynin Roman. Random vectors in high dimensions // Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. — 2018. — Vol. 3. — P. 38–69.
[59] Karras Tero, Laine Samuli, Aittala Miika et al. Analyzing and improving the image quality of stylegan. — 2020. — 1912.04958.
[60] Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network / Wenzhe Shi, Jose Caballero, Ferenc Huszár et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — P. 1874–1883.
[61] Deep unsupervised learning using nonequilibrium thermodynamics / Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli // International Conference on Machine Learning / PMLR. — 2015. — P. 2256–2265.
[62] Ho Jonathan, Jain Ajay, Abbeel Pieter. Denoising diffusion probabilistic models // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 6840–6851.
[63] Swinir: Image restoration using swin transformer / Jingyun Liang, Jiezhang Cao, Guolei Sun et al. // Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — P. 1833–1844.
[64] Ilvr: Conditioning method for denoising diffusion probabilistic models / Jooyoung Choi, Sungwon Kim, Yonghyun Jeong et al. // arXiv preprint arXiv:2108.02938. — 2021.
[65] Deep learning face attributes in the wild / Ziwei Liu, Ping Luo, Xiaogang Wang, Xiaoou Tang // Proceedings of the IEEE international conference on computer vision. — 2015. — P. 3730–3738.
[66] Deep learning face representation by joint identification-verification / Yi Sun, Yuheng Chen, Xiaogang Wang, Xiaoou Tang // Advances in neural information processing systems. — 2014. — Vol. 27.
[67] Learning face representation from scratch / Dong Yi, Zhen Lei, Shengcai Liao, Stan Z Li // arXiv preprint arXiv:1411.7923. — 2014.
[68] Vggface2: A dataset for recognising faces across pose and age / Qiong Cao, Li Shen, Weidi Xie et al. // 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018) / IEEE. — 2018. — P. 67–74.
[69] Joint face detection and alignment using multitask cascaded convolutional networks / Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao // IEEE signal processing letters. — 2016. — Vol. 23, no. 10. — P. 1499–1503.
[70] Rothe Rasmus, Timofte Radu, Van Gool Luc. Dex: Deep expectation of apparent age from a single image // Proceedings of the IEEE international conference on computer vision workshops. — 2015. — P. 10–15.
[71] Interactive facial feature localization / Vuong Le, Jonathan Brandt, Zhe Lin et al. // Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12 / Springer. — 2012. — P. 679–692.
[72] Yang Shuo, Luo Ping, Loy Chen Change, Tang Xiaoou. Wider face: A face detection benchmark. — 2015. — 1511.06523.
[73] Recognize complex events from static images by fusing deep channels / Yuanjun Xiong, Kai Zhu, Dahua Lin, Xiaoou Tang // Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on / IEEE. — 2015.
[74] Labeled faces in the wild: A database for studying face recognition in unconstrained environments : Rep. : 07-49 / University of Massachusetts, Amherst ; Executor: Gary B. Huang, Manu Ramesh, Tamara Berg, Erik Learned-Miller : 2007. — October.
[75] Jesorsky Oliver, Kirchberg Klaus J., Frischholz Robert W. Robust face detection using the hausdorff distance // Audio- and Video-Based Biometric Person Authentication / Ed. by Josef Bigun, Fabrizio Smeraldi. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2001. — P. 90–95.
[76] Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization / Martin Köstinger, Paul Wohlhart, Peter M. Roth, Horst Bischof // 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). — 2011. — P. 2144–2151.
[77] Zhang Kaihao, Li Dongxu, Luo Wenhan et al. Edface-celeb-1m: Benchmarking face hallucination with a million-scale dataset. — 2022. — 2110.05031.