ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

*На правах рукописи*

**Аланов Айбек**

# РАЗРАБОТКА ЭФФЕКТИВНЫХ ПАРАМЕТРИЗАЦИЙ ДЛЯ ГЕНЕРАТИВНЫХ СОСТЯЗАТЕЛЬНЫХ СЕТЕЙ В ЗАДАЧАХ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ И РЕЧИ

РЕЗЮМЕ

диссертации на соискание учёной степени

кандидата компьютерных наук

Научный руководитель:

кандидат физ.-мат. наук

Ветров Дмитрий Петрович

Москва — 2024

## Тема диссертации

В последние годы GANs [1, 2, 3, 4, 5] добились впечатляющих результатов в генерации данных, которые по качеству неотличимы от реальных данных. Они позволяют обучать генератор, который преобразует латентное пространство с простой распределением в пространство реальных объектов с очень сложным распределением. Благодаря своей способности генерировать высококачественные данные, GANs широко используются в различных задачах и областях, включая компьютерное зрение [6, 7, 8, 9, 10, 11, 12] и обработку сигналов [13, 14]. Однако для достижения такого высокого качества генерации во время обучения GANs требуется доступ к крупномасштабным наборам данных, сбор которых занимает много времени и требует значительных затрат. Например, для обучения передовой модели StyleGAN для генерации фотореалистичных человеческих лиц потребовался сбор набора данных FFHQ [3], включающего 70 тысяч изображений лиц очень высокого разрешения (1024x1024).

Проблема обучения GANs на небольших наборах данных остается значительным вызовом. Один из основных подходов к решению этой проблемы заключается в адаптации к домену, когда GAN обучается на новом домене с ограниченным количеством примеров путем дообучения модели, предобученной на другом домене с доступом к крупномасштабному набору данных. Например, для генерации лиц в стиле определенных художников, когда сбор большого набора данных затруднителен, можно дообучить GAN, предобученный на большом наборе данных фотореалистичных лиц (например, FFHQ), используя несколько примеров картин конкретного художника. В адаптации к домену важно, какой поднабор параметров базовой модели оптимизируется. Эта оптимизация определяет, насколько эффективно знания базовой модели могут быть перенесены на новый домен и помогает избежать коллапса мод, к которому GANs сильно склонны.

Данная диссертация предложит новые эффективные параметризации StyleGAN для задачи адаптации к домену и новые компактные архитектуры для задачи улучшения записи речи, которые также эффективно используют тренировочные данные. В частности, в этой работе предлагается техника модуляции домена, позволяющая обучать тысячи раз меньше параметров для модели StyleGAN по сравнению с полной параметризацией для адаптации к домену. Эта инновация позволила предложить модель HyperDomainNet [15], которая решает задачу адаптации к многим доменам.

Дальнейшее развитие этих идей привело к обнаружению более эффективных параметризаций, таких как StyleSpace и Affine+ [16]. Кроме того, был проведен более глубокий анализ того, какие части модели StyleGAN критичны для адаптации к домену, и были раскрыты интересные свойства направлений из StyleSpace. В области улучшения записи речи были предложены модели HiFi++ [17] и FFC-SE [18], демонстрирующие превосходное качество в этой задаче по сравнению с существующими подходами, при этом имея значительно меньшее количество параметров.

Анализируя проблему эффективного обучения GANs, работа стремится ответить на фундаментальные вопросы, такие как: как можно дообучить GANs для новых доменов с ограниченными тренировочными данными? Какие факторы наиболее важны при адаптации генератора для контента, специфичного для домена? Можно ли уменьшить вычислительные затраты, сохранив или даже улучшив производительность в задачах генерации и улучшения аудио? Эти вопросы составляют ядро нашего исследования, и последующие главы этой диссертации направлены на предоставление всестороннего понимания этих важных тем.

В этом введении закладывается основа для детального исследования каждого из четырех статей, подчеркивая их конкретный вклад, идеи и значимость в области генеративных моделей на основе GAN. К концу этого анализа предлагается более глубокое понимание того, как эффективные параметризации могут способствовать большей адаптируемости, устойчивости и ресурсной эффективности GANs, тем самым способствуя дальнейшему развитию технологий генерации изображений и речи.

### Актуальность работы

Данная работа представляет собой ценный вклад в решение критических задач при обучении GANs с ограниченными данными и вычислительными ресурсами, что оказывает значительное влияние на множество приложений в генерации изображений и улучшении записи речи. Здесь подчеркивается значимость и важность данного исследования:

1. **Развитие доменной адаптации GANs:** Первые две статьи, *HyperDomainNet* и *StyleDomain*, вносят значительный вклад в область адаптации к домену для GANs. С учетом растущей необходимости адаптации моделей GAN к конкретным доменам с ограниченными данными, эти статьи предлагают эффективные и

легковесные параметризации. Это исследование позволяет практическое использование GANs в ситуациях, где нехватка данных является критической проблемой, расширяя их применимость в реальных условиях.

2. **Снижение вычислительных ресурсов:** Статьи *HyperDomainNet* и *HiFi++* подчеркивают важность снижения вычислительных ресурсов при сохранении или даже улучшении качества генерируемого контента. Учитывая, что вычислительная эффективность является критическим фактором при развертывании GANs в условиях ограниченных ресурсов, это исследование способствует тому, чтобы модели на основе GANs стали более доступными и экономически эффективными. Это соответствует текущей тенденции в области ИИ к оптимизации моделей глубокого обучения для практического применения.

3. **Универсальная применимость:** Разработка *HyperDomainNet*, способной адаптироваться к нескольким доменам с одной моделью, особенно актуальна в эпоху данных, управляемых ИИ. Во многих практических сценариях поддержание отдельных моделей для различных доменов представляет собой сложную задачу, что делает идею универсальной адаптации весьма привлекательной. Способность одной модели обобщаться и адаптироваться к множеству доменов является ключевой для эффективных, гибких и масштабируемых систем ИИ.

4. **Эффективное улучшение записи речи:** В области улучшения записи речи статьи *HiFi++* и *FFC-SE* вводят новые и эффективные архитектуры. В статье *HiFi++* показано, что GANs могут превосходить традиционные методы расширения полосы пропускания и улучшения записи речи, при этом имея значительно меньше параметров и уменьшенную вычислительную сложность. В то же время, статья *FFC-SE* применяет новые техники для улучшения записи речи с помощью быстрой свертки Фурье, делая архитектуру еще более легковесной и достигая лучшей производительности на практике.

**Целью** данной работы является разработка новых эффективных параметризаций для GAN-моделей, позволяющих значительно сократить количество оптимизируемых параметров и объем необходимых обучающих данных.

# 1 Основные результаты и выводы

Основные вклады данного исследования можно описать следующим образом:

1. В статье *HyperDomainNet* предложена новая параметризация StyleGAN на основе техник модуляции домена и новую модель HyperDomainNet. Наша параметризация сократила количество обучаемых параметров StyleGAN в несколько тысяч раз для адаптации к домену, сохраняя при этом качество, сравнимое с существующими подходами, которые обучают почти все параметры генератора StyleGAN. Также была представлена новая модель HyperDomainNet, которая позволяет решать проблему многодоменной адаптации, то есть когда StyleGAN адаптируется к нескольким доменам одновременно. Это открывает новые возможности для случаев, когда у нас много разных доменов, на которых нужно обучаться, и не хочется обучать отдельную модель для каждого. Наш подход значительно улучшает эффективность и применимость модели для таких случаев.

2. В работе *StyleDomain* была более глубоко проанализирована задача адаптации домена для StyleGAN. Было исследовано, какие части этой модели важны для адаптации к различным доменам в зависимости от схожести целевого домена с исходными доменами. В результате этого анализа были предложены новые эффективные параметризации StyleSpace и Affine+. StyleSpace является наиболее простой параметризацией для решения проблемы адаптации домена для близких доменов и достигает такого же качества, как и другие подходы, которые обучают значительно больше параметров. Параметризация Affine+ предназначена для более отдаленных доменов и показывает наилучшие результаты в задаче обучения на малом числе примеров, при этом имея меньше обучаемых параметров, чем базовые модели. Также были обнаружены удивительные свойства этих параметризаций, которые могут быть использованы для еще большего числа приложений.

3. В статьях *HiFi++* и *FFC-SE* были предложены новые эффективные модели для задачи улучшения записи речи. В *HiFi++* были представлены новые модули в архитектуре генератора GAN, которые значительно улучшают конечное качество модели при очень малом числе параметров. Было показано, что с этой

архитектурой модель работает на уровне или даже лучше существующих подходов, имея значительно меньше параметров. В *FFC-SE* архитектура генератора была дополнительно улучшена с помощью свертки Фурье, что позволило учитывать и использовать больше информации. Это сократило размер модели и улучшило конечное качество.

**Теоретическая и практическая значимость.** Теоретическая значимость данной работы заключается в новых подходах к параметризации и адаптации архитектуры StyleGAN, а также в усовершенствовании моделей улучшения речи. Благодаря внедрению фреймворков HyperDomainNet и StyleDomain в работе представлены методы сокращения числа обучаемых параметров StyleGAN для адаптации к домену, что позволяет достичь качества, сравнимых с существующими подходами, при этом имея существенно меньше параметров. Сюда входят новые параметризации, такие как StyleSpace и Affine+, которые оптимизируют адаптацию как для близких, так и для далеких доменов, обнаруживая неожиданные свойства, расширяющие потенциальные области применения. На практике эти усовершенствования приводят к созданию более эффективных, адаптируемых к нескольким доменам моделей, что повышает их практическую полезность в сценариях с множеством различных доменов. Кроме того, модели HiFi++ и FFC-SE предлагают новые архитектуры для улучшения записи речи, используя модули GAN и свертку Фурье для значительного повышения производительности модели при меньшем количестве параметров, что способствует созданию более эффективных и высококачественных решений для обработки речи.

**Методология и методы исследования.** В данной работе применяются глубокое обучение, генеративные модели, генеративные состязательные сети, методы доменной адаптации, методы улучшения записи речи, а также стандартные методы оптимизации.

**Воспроизводимость.** Подробно описаны предложенные методы и эксперименты, а код всех работ опубликован в открытом доступе.

**Результаты, выносимые на защиту.**

1. Техника *доменной модуляции* для эффективной адаптации доменов и *HyperDomainNet* для обучения мультидоменной адаптации.

2. Эффективные параметризации, *StyleSpace* и *Affine+*, для доменной адаптации StyleGAN в задачах близких и далеких доменов.

3. Эффективные модели улучшения записи речи: *HiFi++*, улучшающая качество с небольшим числом параметров, и *FFC-SE*, повышающая производительность модели с помощью свертки Фурье.

**Вклад автора.** Исследование, представленное в данной диссертации, является результатом нескольких лет упорной работы и совместных усилий. В этом разделе описываются конкретные вклады автора в каждую из четырех статей, составляющих данную диссертацию. В первой статье *HyperDomainNet* автор предложил технику доменной модуляции для эффективной доменной адаптации StyleGAN. Также автор отвечал за реализацию экспериментов по one-shot доменной адаптации и подготовил основную часть текста для всех секций статьи. Во второй статье *StyleDomain* автор предложил параметризации *StyleSpace* и *Affine+* и подготовил текст всех секций статьи, кроме секции экспериментов. В статье *HiFi++* автор предложил идею использования нескольких простых и легковескных дискриминаторов, подобрал оптимальные размеры каждой части архитектуры и отвечал за эксперименты по нахождению лучшей конфигурации дискриминаторов. Кроме того, автор сыграл значительную роль в написании текста введения и основных секций статьи. В четвертой статье *FFC-SE* автор участвовал в написании кодовой базы и в разработке дизайна экспериментов. Также автор занимался редактированием текста статьи и участвовал в обсуждениях относительно анализа полученных результатов.

## Публикации и апробация работы

* обозначает равный вклад соавторов

**Публикации повышенного уровня.**

1. ***Айбек Алanов****, *Вадим Титов* и Дмитрий Ветров.* HyperDomainNet: универсальная адаптация домена для генеративных состязательных сетей (HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks). // В материалах конференции Neural Information Processing Systems, 2022 (NeurIPS 2022). Том 35, страницы 29414–29426. Конференция уровня CORE A*.

2. ***Айбек Аланов****, Вадим Титов*, Максим Находнов* и Дмитрий Ветров.* StyleDomain: эффективные и легковесные параметризации StyleGAN для адаптации домена с одного и нескольких примеров (StyleDomain: Efficient and Lightweight Parameterizations of StyleGAN for One-shot and Few-shot Domain Adaptation). // В материалах конференции International Conference on Computer Vision, 2023 (ICCV 2023). Страницы 2184-2194. Конференция уровня CORE A*.

3. *Иван Щекотов*, Павел Андреев*, Олег Иванов, **Айбек Аланов** и Дмитрий Ветров.* FFC-SE: Быстрая свертка Фурье для улучшения записи речи (FFC-SE: Fast Fourier Convolution for Speech Enhancement). // В материалах конференции InterSpeech Conference, 2022. Страницы 1188-1192. Конференция уровня CORE A.

**Публикации стандартного уровня**

1. *Павел Андреев*, **Айбек Аланов***, *Олег Иванов* и Дмитрий Ветров.* HiFi++: единая платформа для расширения полосы пропускания и улучшения записи речи (HiFi++: a Unified Framework for Bandwidth Extension and Speech Enhancement). // В материалах конференции International Conference on Acoustics, Speech, and Signal Processing, 2023 (ICASSP 2023). Страницы 1-5. Конференция уровня CORE B (согласно CORE2018).

**Доклады на конференциях**

1. Доклад на тему “Синтез аудио и расширение полосы пропускания”, Семинар исследовательской группы по байесовским методам, Москва, апрель 2021.

2. Презентация постера на тему “FFC-SE: Быстрая свертка Фурье для улучшения записи речи”, конференция InterSpeech, Сеул, Республика Корея, сентябрь 2022.

3. Презентация постера на тему “HyperDomainNet: универсальная адаптация домена для генеративных состязательных сетей”, конференция Neural Information Processing Systems, Новый Орлеан, США, декабрь 2022.

4. Доклад на тему “Доменная адаптация GANs”, семинар исследовательской группы по байесовским методам, Москва, декабрь 2022.

5. Доклад на тему “HyperDomainNet: универсальная адаптация домена для генеративных состязательных сетей”, конференция Fall into ML, Москва, ноябрь 2022.

6. Доклад на тему “HyperDomainNet: универсальная адаптация домена для генеративных состязательных сетей”, семинар AIRI AIschnitsa, Москва, декабрь 2022.

7. Доклад на тему “HyperDomainNet: универсальная адаптация домена для генеративных состязательных сетей”, конференция факультета компьютерных наук, Вороново, июнь 2022.

**Объем и структура работы**. Диссертация содержит введение, содержание публикаций и заключение. Полный объем диссертации 142 страницы.

# 2 Содержание работы

## 2.1 HyperDomainNet: Универсальная адаптация домена для генеративных состязательных сетей

В области компьютерного зрения генеративные состязательные сети (GAN) [1, 2, 3, 4, 5] показали выдающиеся результаты в различных задачах, таких как улучшение изображений [6, 7], редактирование [8, 9] и перевод изображений [10, 11, 12]. Однако обучение современных GAN требует большого количества образцов, что ограничивает их применение в областях с недостаточным количеством данных. Для преодоления этого ограничения часто используется трансферное обучение (TL), когда предварительно обученная модель дообучается для нового домена с ограниченными данными.

Текущие методы TL для GAN обычно дообучают почти все веса предварительно обученной модели [19, 20, 21, 22, 4, 23, 24, 25, 26]. Хотя это подходит для далеких целевых доменов, для доменов, схожих с исходным, это часто излишне. В таких случаях дообучение всех весов может быть избыточным. Данное исследование предлагает более эффективный подход, называемый *модуляцией домена*, который оптимизирует только один 6000-мерный вектор для каждого целевого домена, что значительно сокращает пространство параметров по сравнению с традиционным дообучением всех 30 миллионов весов.

Техника модуляции домена применяется к двум передовым методам доменной адаптации, StyleGAN-NADA [25] и MindTheGAP [26], демонстрируя сравнимую производительность с полной параметризацией, но будучи значительно более лёгкой. Кроме того, предлагается новая регуляризация функции потерь для улучшения разнообразия дообученного генератора.

Исследование также затрагивает многодоменную адаптацию, когда одна модель адаптируется к нескольким доменам на основе текстовых описаний или примеров изображений. Вместо дообучения отдельных генераторов для каждого целевого домена, вводится гиперсеть под названием *HyperDomainNet*. Эта гиперсеть предсказывает вектор для StyleGAN2 на основе целевого домена, значительно сокращая время обучения и количество обучаемых параметров. Наблюдается, что этот подход может обобщаться на новые домены при условии достаточного количества доменов, используемых для обучения.

Исследование представляет обширные эксперименты для проверки предлагаемых методов в различных доменах. Результаты показывают, что модуляция домена достигает качества, сопоставимого с полной параметризацией, а регуляризация потерь улучшает разнообразие дообученного генератора. Кроме того, HyperDomainNet демонстрирует перспективную обобщаемость на различные новые домены.

В заключение, данное исследование предлагает три ключевых вклада:

1. Техника модуляции домена, которая сокращает пространство параметров для адаптации домена в StyleGAN2 на несколько порядков.

2. Новый регуляризационный подход для улучшения разнообразия дообученных генераторов.

3. Введение HyperDomainNet для многодоменной адаптации, демонстрирующей обобщаемость на новые домены.

#### Основные понятия

StyleGAN2 [3] генерирует изображения через нейросеть отображения $M(z)$, которая преобразует начальные случайные векторы $z \in \mathcal{Z}$ в промежуточное латентное пространство $\mathcal{W}$, которое затем проходит через аффинные преобразования $A(w)$ для создания параметров стиля $s = A(w) \in \mathcal{S}$. Эти параметры влияют на окончательные карты признаков, создаваемые синтезирующей сетью $G_{sys}$. Слои ToRGB $G_{tRGB}$ используются для генерации выходного изображения.

Задача доменной адаптации: адаптация обученного генератора StyleGAN2 из одного домена (исходного) в другой (целевой), руководствуясь либо изображением, либо текстовым описанием из целевого домена.

Модель CLIP [27]: CLIP — это модель, которая выравнивает текстовые и визуальные эмбеддинги в общем пространстве, измеряя семантическое сходство объектов на основе косинусного расстояния.

StyleGAN-NADA [25]: Этот метод использует CLIP для выравнивания исходных и целевых доменов в пространстве CLIP. Он оптимизирует синтезирующую сеть целевого домена с использованием функции потерь направления между изображениями и текстовыми описаниями.

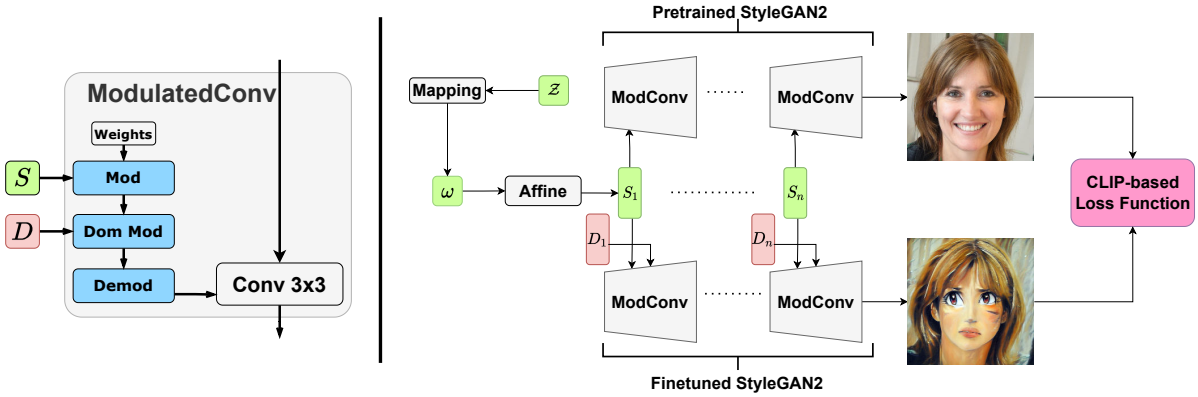
MindTheGap [26]: Разработанный для доменной адаптации на основе картинки, MindTheGap стремится предотвратить потерю разнообразия в целевых изображениях. Он вводит регуляризаторы, использующие эмбеддинг целевого изображения в исходном домене, улучшая качество адаптации.

В заключение, эти методы адаптируют StyleGAN2 к новым доменам с использованием техник выравнивания на основе CLIP, улучшая качество синтезированных изображений.

**Метод**

Цель данного исследования заключается в улучшении адаптации StyleGAN к новому домену посредством оптимизации сети синтеза $G_{sys}(\cdot, \cdot)$ с использованием компактного параметрического пространства. Этот компонент сети в основном изменяется во время тонкой настройки для нового домена. Предлагаемый подход вводит технику *доменной модуляции*, операцию, которая корректирует веса свертки признаков внутри сети синтеза. Модуляция настраивает веса на основе параметров стиля, что приводит к более эффективной форме адаптивной нормализации экземпляров (AdaIN) [28, 29]. Этот метод вдохновлен методами переноса стиля, использующими AdaIN для стилизации изображений.

Техника доменной модуляции уменьшает параметрическое пространство для тонкой настройки StyleGAN2, оптимизируя только вектор $d$ с той же размерностью, что и параметры стиля. Этот вектор интегрируется в архитектуру StyleGAN через дополнительную операцию модуляции (см. диаграмму 1a). Вместо оптимизации всех весов $\theta$ компонента $G_{sys}$, обучается только вектор $d$. Размерность вектора $d$ составляет 6 тысяч, что в 4 тысячи раз меньше исходного пространства весов $\theta$ размером 30 миллионов в компоненте $G_{sys}(\cdot, \cdot)$.




(a) Domain-modulation technique (b) Fine-tuning StyleGAN2 by optimizing the domain vector D

Рис. 1: Подробная схема предложенного метода. (a) Пересмотренный блок ModulatedConv с введенной операцией доменной модуляции. (b) Полностью детализированный процесс обучения доменной адаптации с предложенной техникой доменной модуляции.

**Улучшение разнообразия адаптации области на основе CLIP**

Существующие методы адаптации области на основе CLIP, такие как StyleGAN-NADA и MindTheGap, используют регуляризатор $\mathcal{L}_{direction}$ (или $\mathcal{L}_{clip\_across}$) для решения проблем с коллапсом мод. Однако этот регуляризатор частично сохраняет разнообразие и приводит к коллапсу после нескольких итераций, что особенно проблематично для доменов, требующих значительной тонкой настройки. Проблема с $\mathcal{L}_{direction}$ заключается в том, что она вычисляет косинусные расстояния между эмбеддингами, которые больше не находятся на сфере CLIP, способствуя коллапсу мод.

Для решения этой проблемы вводится новый регуляризатор, называемый регуляризатором *согласованности углов в области*. Этот регуляризатор вычисляет косинусные расстояния CLIP исключительно между эмбеддингами CLIP. Он направлен на поддержание попарных косинусных расстояний между изображениями до и после доменной адаптации, что эффективно увеличивает разнообразие генератора по сравнению с исходными функциями потерь:

$$\mathcal{L}_{indomain-angle}(\{G_d^B(w_i)\}_{i=1}^n, \{G^A(w_i)\}_{i=1}^n, B, A) = \tag{1}$$

$$= \sum_{i,j}^{n} (\langle E_I(G^A(w_i)), E_I(G^A(w_j))\rangle - \langle E_I(G_d^B(w_i)), E_I(G_d^B(w_j))\rangle)^2, \tag{2}$$

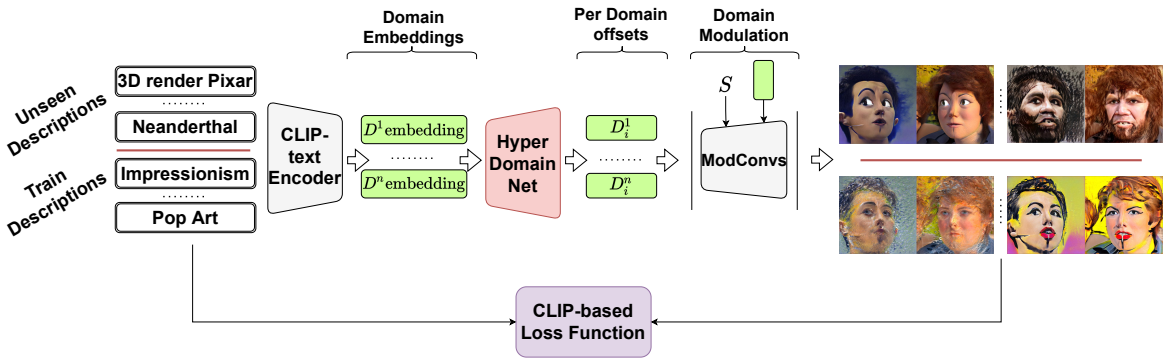



Рис. 2: Подробный процесс обучения HyperDomainNet.

**Проектирование HyperDomainNet для универсальной доменной адаптации**

Предлагается техника доменной модуляции для эффективной многодоменной адаптации StyleGAN2. Наша цель состоит в обучении *HyperDomainNet*, который предсказывает параметры домены для тонкой настройки генераторов. В частности, эта работа фокусируется на сценарии, когда целевые домены представлены текстовыми описаниями.

HyperDomainNet принимает текстовые эмбеддинги в качестве входных данных и выдает параметры домена. Модель использует комбинацию функций потерь, включая $\mathcal{L}_{direction}$, $\mathcal{L}_{tt-direction}$ и $\mathcal{L}_{domain-norm}$, для обучения сети. Эти члены обеспечивают, что предсказанные параметры области эффективно направляют доменную адаптацию и предотвращают смешивание доменов. Процесс обучения описан на диаграмме 2.

В заключение, в этой работе представляется подход доменной модуляции для многодоменной адаптации в StyleGAN2, фокусируясь на целевых доменах на основе текста. HyperDomainNet обучается с набором функций потерь для обеспечения эффективной тонкой настройки, специфичной для каждого домена. Для подробных описаний функций потерь и процесса оптимизации, пожалуйста, обратитесь к оригинальной статье.

**Результаты**

В этом разделе представлены результаты для адаптации на основе текста, изображения и многодоменной адаптации с использованием предложенного подхода.


Рис. 3: Сравнение с оригинальным методом StyleGAN-NADA [25] (слева) и его версией с нашей параметризацией.


Рис. 4: Сравнение с методами доменной адаптации на основе изображения. Левый блок — это MindTheGap+indomain, правый блок — StyleGAN-NADA [26]. Средний блок — MindTheGap+indomain с нашей параметризацией.

**Текстовая доменная адаптация** В этой части сравнивается наша параметризация со StyleGAN-NADA [25] на различных доменах. Наша параметризация соответствует выразительности StyleGAN-NADA, что позволяет адаптироваться к изменениям стиля и текстуры. Качественные результаты представлены на Рис. 3, демонстрируя сопоставимую производительность.

**Доменная адаптация на основе изображения** В этой части наша параметризация и индоменная угловова согласованность применяется к методу MindTheGap [26]. Результаты в Таблице 1 и на Рис. 4 показывают, что наш подход достигает аналогичной производительности с оригинальным, используя значительно меньше параметров. TargetCLIP [30] и другие методы демонстрируют низкое качество адаптации, в основном пригодные для редактирования в пределах домена. Идоменная угловая согласованность значительно улучшает метрики FID и точности.

**Мультидоменная адаптация** В этой части HyperDomainNet используем в двух сценариях: (i) фиксированное количество доменов и (ii) произвольное количество доменов. Результаты на Рис. 5 показывают эффективность нашего метода в обоих сценариях, с многообещающей адаптацией к новым доменам. Исследование влияния исключений подтверждает важность предлагаемых функций потерь в обучении HyperDomainNet для мультидоменной адаптации.

Таблица 1: Оценка методов доменной адаптации на основе изображения. Результаты для методов TargetCLIP, Cross-correspondence и StyleGAN-NADA взяты из [26].

| | **Model quality** | | | **Model complexity** |
|---|---|---|---|---|
| Model | FID | Precision | Recall | # trainable parameters |
| TargetCLIP [30] | 199.33 | 0.000 | 0.293 | 9K |
| Cross-correspondence [24] | 158.86 | 0.001 | 0 | 30M |
| StyleGAN-NADA [25] | 124.55 | 0.118 | 0 | 24M |
| MindTheGap [26] | 78.35 | 0.326 | 0.017 | 24M |
| MindTheGap (our param.) | 79.83 | 0.452 | 0.017 | 6k |
| MindTheGap+indomain | 71.46 | 0.503 | 0.014 | 24M |
| MindTheGap+indomain (our param.) | 72.71 | 0.472 | 0.028 | 6k |


Рис. 5: Сравнение тренировочных доменов. Верхний ряд представляет реальные изображения, встроенные в латентное пространство StyleGAN2, которые затем используются для вывода HyperDomainNet. Левый блок представляет результаты, полученные из текстовых описаний, представленных в списке тренировки. Правый блок представляет результаты вывода HyperDomainNet на новые текстовые описания.

## 2.2 StyleDomain: Эффективные и легковесные параметризации StyleGAN для доменной адаптации по одному и нескольким экземплярам

Недавние достижения в области генеративно-состязательных сетей (GAN) [1, 2, 3, 31, 5], особенно моделей StyleGAN, оказались чрезвычайно эффективными в различных приложениях синтеза изображений, включая улучшение, редактирование и перевод изображений. Однако обучение моделей StyleGAN требует больших и высококачественных наборов данных, что ограничивает их полезность в доменах с небольшим количеством изображений. Переносное обучение, дообучение предварительно обученной модели из одного домена в другой, является распространенным подходом к решению этой проблемы.

Существуют несколько методов доменной адаптации для StyleGAN [4, 32, 23, 33, 34, 35, 24, 36, 15, 25, 26, 37], но большинство из них предполагают, что адаптация к новому домену требует дообучения большинства весов модели, даже для схожих доменов. Это предположение не имеет эмпирического подтверждения, и мало исследований посвящено анализу, какие части StyleGAN важны для различных сценариев данных и схожести доменов.

В данном исследовании был проведен систематический анализ для решения этой проблемы. Наше исследование состоит из двух основных частей. Во-первых, определяются, какие части StyleGAN нуждаются в адаптации в зависимости от сходства между исходным и целевым доменами. Было обнаружено, что для схожих доменов часто достаточно дообучения только аффинных слоёв. Для более несхожих доменов необходимо оптимизировать дополнительные параметры, но не обязательно всю сеть. Это указывает на возможность более эффективных и лёгких параметризаций StyleGAN для доменной адаптации.

Во второй части нашего анализа предлагаются две новые параметризации StyleGAN. Для схожих доменов вводится концепция *StyleSpace*, где можно оптимизировать направления для адаптации к схожим целевым доменам без дообучения всех весов StyleGAN. Для более далеких доменов представляется параметризация *Affine+*, которая значительно уменьшает количество обучаемых параметров при сохранении качества. Дальнейшие улучшения достигаются с параметризацией *AffineLight+*, которая использует разложение низкого ранга для весов аффинных

слоёв. Эти параметризации превосходят сложные базовые методы в адаптации по нескольким экземплярам для несхожих доменов.

Более того, исследуются свойства направлений *StyleDomain*, обнаруживая их смешиваемость и переносимость. Эти направления могут быть объединены для создания совершенно новых стилей или применены к моделям StyleGAN, дообученным для других доменов. Эти выводы используются в различных задачах компьютерного зрения, включая перевод изображений и междоменное морфингование.

#### Важность каждой части StyleGAN

В этом разделе оценивается важность различных компонентов StyleGAN, в частности StyleGAN2, для доменной адаптации. Исходный домен - FFHQ, и исследуются различные целевые домены. StyleGAN2 состоит из трех основных компонентов:

- Сеть отображения: она преобразует входной шум в промежуточный латентный вектор.
- Аффинные слои: эти слои отображают латентный вектор в векторы стиля, которые формируют StyleSpace.
- Сеть синтеза: состоящая из модулированных сверток, она генерирует выходное изображение из входного шума.

Описание диаграммы архитектуры StyleGAN2 предоставляется на Рис. 6.

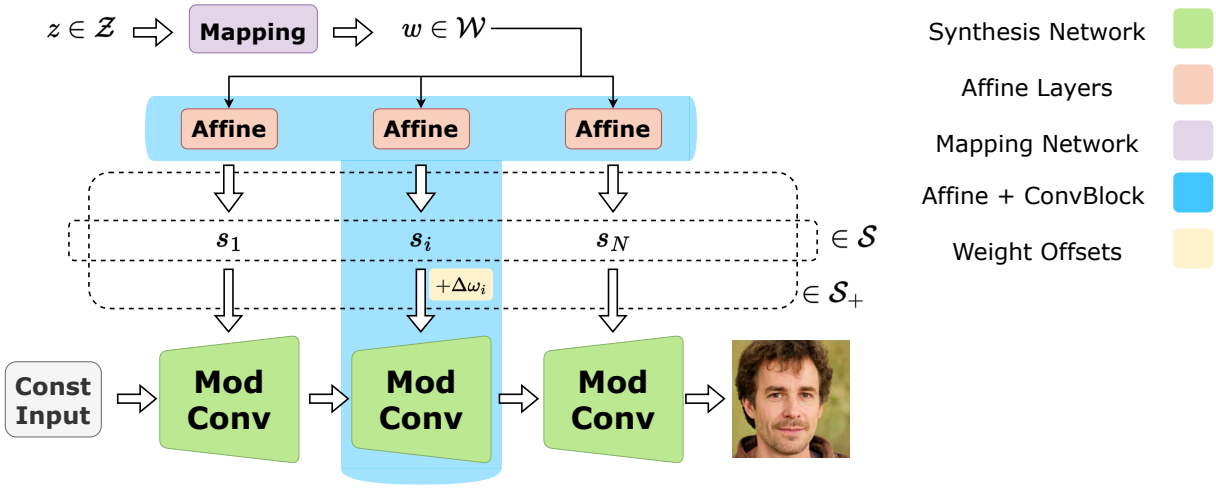



Рис. 6: Архитектура StyleGAN2. Вводится новое латентное пространство $S+$ для доменной адаптации, которое сочетает StyleSpace и смещения весов для одного блока из сети синтеза.

Сеть синтеза традиционно считалась наиболее важной для адаптации, тогда как сеть отображения и аффинные слои в основном обрабатывают семантические манипуляции в пределах исходного домена. Эта работа стремится подтвердить это предположение.

В наших экспериментах также рассматривается комбинированное влияние аффинных слоёв и свёрточного блока из сети синтеза на доменную адаптацию, предлагая промежуточный анализ.

В этой работе предлагается метод для анализа влияния каждого компонента. В то время как предыдущие работы сбрасывали веса компонентов дообученного генератора к их предварительно обученным значениям, эта работа предлагает дообучение только одного компонента для определения того, какой из них достаточен для доменной адаптации.

Цель оптимизации для доменной адаптации - минимизация потерь доменной адаптации, $\mathcal{L}_D$, с использованием сгенерированных образцов от генератора $G_\theta(s(z))$. Обычно генератор оптимизируется по всем компонентам:

$$\mathcal{L}_D\left(\{G_\theta(s(z_i))\}_{i=1}^K\right) \to \min_{\theta, f^A, f_M}. \tag{3}$$

В этой работе исследуются настройки, где оптимизируем только один компонент: *SyntConv* для сети синтеза, *Affine* для аффинных слоёв и *Mapping* для сети отображения. Полная оптимизация всех компонентов называется параметризацией *Full*.

Эта работа рассматривает две настройки доменной адаптации: на основе одного или нескольких экземпляров. Для каждой настройки используются разные домены, которые различаются по схожести с исходным доменом (FFHQ). Домены с одним экземпляром сохраняют геометрию и идентичность лица, изменяя стиль. Домены на основе нескольких экземпляров, с другой стороны, значительно изменяют форму, геометрию и идентичность лица. Разные функции потерь доменной адаптации применяются в зависимости от режима данных.

В случае адаптации по одному экземпляру используются метрики качества и разнообразия. Для адаптации по нескольким экземплярам вычисляется метрика FID. Дополнительные детали о функциях потерь доменной адаптации можно найти в приложении.

**Анализ для доменов на основе одного экземпляра.**

В нашем анализе исследуются текстовые домены и домены с изображениями на основе одного экземпляра.

В наших экспериментах рассматриваются четыре параметризации: Full, SyntConv, Affine и Mapping. Наши качественные результаты представлены на Рис. 7.

В этой работе было обнаружено, что параметризации Full, SyntConv и Affine работают аналогично с точки зрения визуального качества и объективных метрик. Это согласуется с предыдущими исследованиями [37]. Удивительно, но параметризация Affine также оказывается эффективной, позволяя нам изменять домены изображений без повторного обучения сети синтеза. Однако сеть Mapping демонстрирует низкое визуальное качество и ограниченное разнообразие генерируемых изображений, подчеркивая важность обновления вектора стиля из $\mathcal{S}$ для успешной адаптации, а не промежуточного латентного вектора из $\mathcal{W}$.

**Анализ для доменов на основе нескольких экземпляров.**

В этом исследовании анализируются два набора данных, AFHQ Dogs и Cats [38]. Результаты представлены на Рис. 8 и в Таблице 2. Были обнаружены различия в результатах для Dogs и Cats по сравнению с похожими доменами. В частности, параметризация Affine дает более низкое качество, что проявляется в ухудшении визуального результата и увеличении метрики FID. Удивительно, но даже без тонкой настройки адаптированные изображения демонстрируют приемлемое визуальное ка-

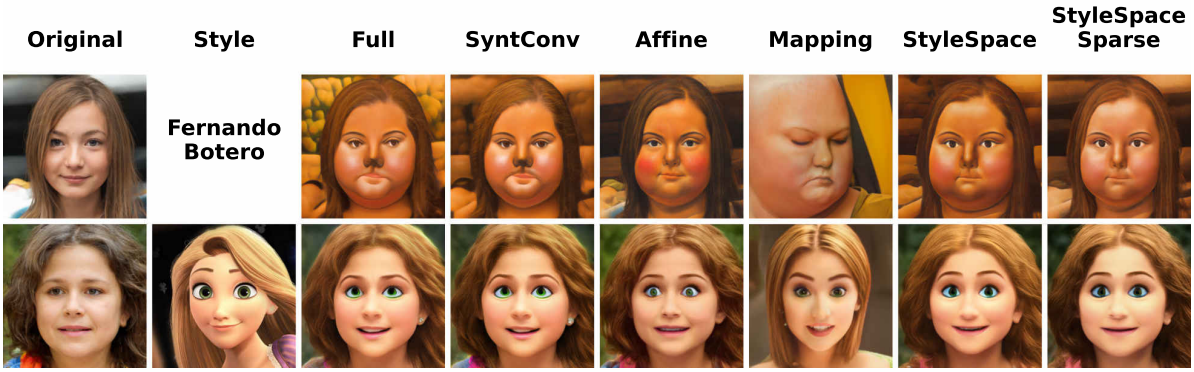



Рис. 7: Текстовая и образная адаптация для различных параметризаций. Параметризации Affine, StyleSpace и StyleSpaceSparse демонстрируют производительность, сопоставимую с Full. Это стильное изображение называется "Disney".

чество. SyntConv соответствует параметризации Full в результатах, в то время как Mapping стабильно показывает низкое качество на всех наборах данных.

**StyleSpace и StyleSpaceSparse.**

В исследовании рассматривается модификация стилевого вектора в StyleSpace $\mathcal{S}$ для изменения сгенерированного домена изображения. Авторы оптимизируют направление $\Delta s^D$ во время тонкой настройки StyleGAN2 для достижения этого изменения. Эти оптимизированные направления они называют направлениями "StyleDomain":

$$\mathcal{L}_D\left(\{G_\theta(s(z_i)+\Delta s)\}_{i=1}^K\right)\to\min_{\Delta s}, \tag{4}$$

где $\Delta s = (\Delta s_1, \ldots, \Delta s_N) \in \mathcal{S}$ является оптимизированным направлением в $\mathcal{S}$ для адаптации генератора $G_\theta$ к домену $D$.

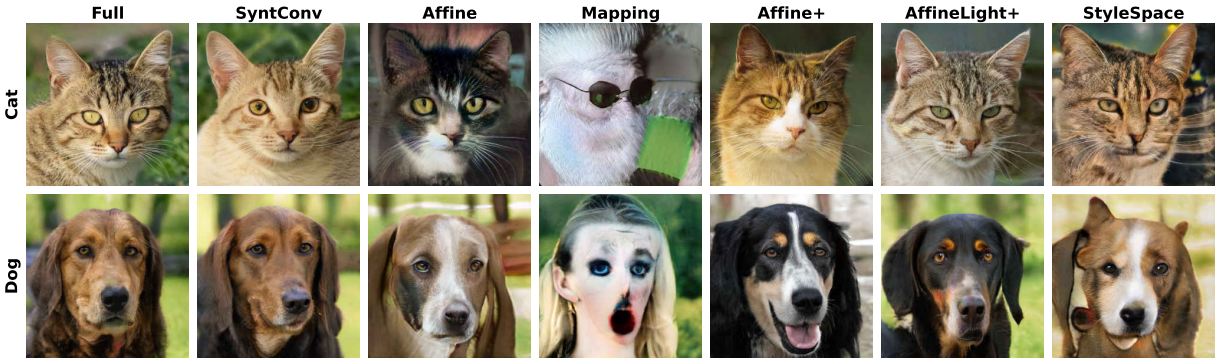



Рис. 8: Доменная адаптация для непохожих доменов. Параметризация Affine+ дает результаты, сопоставимые с Full.

Таблица 2: Оценки FID для доменной адаптации с различными параметризациями. Наблюдается значительный разрыв между параметризациями Affine и Full, который, однако, может быть значительно уменьшен при введении параметризации Affine+.

| | | **Домены** | |
|---|---|---|---|
| Параметрическое пространство | Размер | Собака | Кошка |
| Full | 30.3M | 20.3 | 7.1 |
| SyntConv | 23.6M | 19.7 | 7.2 |
| Affine | 4.6M | 70.1 | 27.6 |
| Mapping | 2.1M | 208.2 | 226.1 |
| **Affine+** | 5.1M | **18.6** | **7.0** |
| **AffineLight+** | 0.6M | 20.6 | 8.9 |
| **StyleSpace** | **6.0K** | 75.8 | 22.0 |

Они обнаружили, что можно удалить большинство координат направлений StyleDomain без ухудшения качества. Они используют стандартную технику обрезки, сохраняя 20% наибольших абсолютных значений и устанавливая остальные равными нулю, что они называют "StyleSpaceSparse".

Авторы применяют эти параметризации к доменам на основе одного экземпляра и нескольких экземпляров и делают следующие наблюдения:

Для доменной адаптации на основе одного экземпляра оптимизация направления StyleDomain достигает тех же результатов, что и полная параметризация, как визуально, так и количественно. Это позволяет генерировать образцы из вне-доменных областей реалистичных человеческих лиц.

Для доменов на основе нескольких экземпляров StyleSpace недостаточен, что приводит к значительному ухудшению качества. Они вводят новую параметризацию, подходящую для более отдаленных доменов.

**Affine+ и AffineLight+.**

В этой работе предлагается улучшение параметризации Affine для доменной адаптации в синтезе изображений, особенно для собак и кошек. Вводится компактная параметризация для некоторых слоев, используя смещения вместо тонкой настройки всех весов. Цель оптимизации — минимизировать функцию потерь для этой параметризации. Это пространство называется Affine+, которое выбирается для конкретного блока в сети синтеза с разрешением 64x64, так как оно показывает наилучшие результаты. Таким образом, для этой параметризации процедура оптимизации имеет следующий вид:

$$\mathcal{L}_D\left(\{G_{\theta,\Delta\theta_1,\Delta\theta_2}(s(z_i))\}_{i=1}^K\right) \to \min_{\Delta\theta_1,\Delta\theta_2,f^A}, \tag{5}$$

где $G_{\Delta\theta_1,\Delta\theta_2}$ является генератором со смещениями весов $\Delta\theta_1, \Delta\theta_2$ для одного блока из сети синтеза.

Affine+ уже имеет значительно меньше параметров, чем полная параметризация. Его размер еще больше уменьшается с помощью разложения с низким рангом и называем его AffineLight+. Он имеет гораздо меньше параметров при сохранении хорошего качества, особенно в условиях малого объема данных.

Эти две параметризации применяются к доменам на основе нескольких экземпляров и достигаем многообещающих результатов. Для получения дополнительных де-

талей и результатов см. приведенные рисунки и таблицы. Affine+ сокращает разрыв в производительности с полной параметризацией, указывая на то, что стилевые векторы помогают адаптировать генератор даже к отдаленным доменам. AffineLight+ показывает хорошие результаты с гораздо меньшим количеством параметров, что делает его подходящим для ситуаций с малым объемом данных.

**Свойства направлений StyleDomain.**

В этой работе исследуются две примечательные особенности направлений StyleDomain. Во-первых, они демонстрируют возможность смешивания, позволяя комбинировать направления, соответствующие различным схожим доменам, что приводит к семантически смешанной адаптации (см. Рис. 10 для примеров).

Во-вторых, направления StyleDomain передаваемы между различными моделями StyleGAN2. Это демонстрируется применением направлений, оптимизированных для базового генератора $G_\theta$, для адаптации тонко настроенных генераторов в различных доменах (например, собаки, кошки) (см. Рис. 9 для результатов).

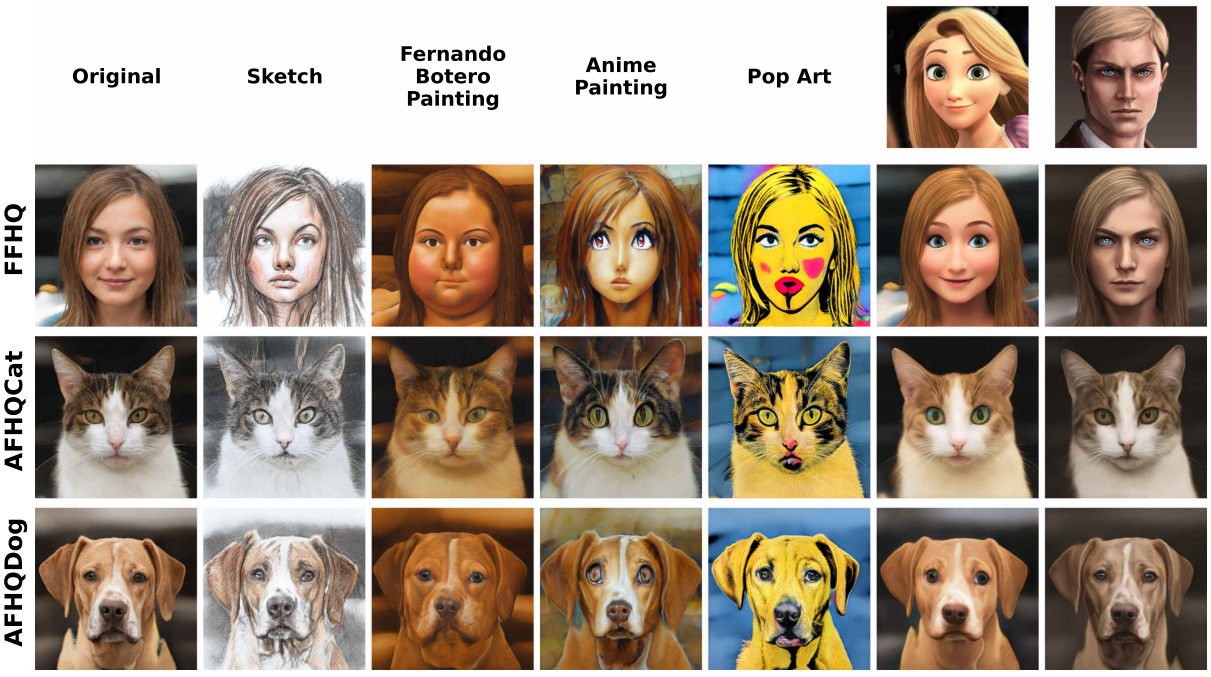



Рис. 9: Передача направлений StyleSpace от текстовой и изображенческой доменной адаптации к другим тонко настроенным моделям. Можно успешно передавать стиль, сохраняя содержание изображения.

**Результаты.**

**Доменная адаптация на основе одного экземпляра.** В данном исследовании рассматривается доменная адаптация на основе одного экземпляра для задач на основе изображений, используя различные базовые модели, такие как TargetCLIP, JoJoGAN, MTG, GOSA, DiFa и DomMod. Параметризации StyleSpace и StyleSpaceSparse применяются к модели DiFa, что приводит к улучшению производительности. Эксперименты используют StyleGAN2 с исходным доменом FFHQ, сохраняя базовые конфигурации для справедливого сравнения. Различные стилевые изображения служат целевыми доменами. Количественные и качественные результаты представлены в Таблице 3 и на Рис. 11, указывая на то, что DiFa достигает наилучшей метрики качества, но уступает в разнообразии. Предложенные параметризации улучшают производительность по этим метрикам, превосходя другие базовые модели. DomMod также показывает хорошие результаты, но сопоставим с StyleSpaceSparse, который более параметроэффективен. Особенно стоит отметить, что StyleSpaceSparse требует значительно меньше памяти, что важно для масштабирования на множество целевых доменов. TargetCLIP, несмотря на ограниченное количество обучаемых параметров, дает плохие визуальные и качественные результаты. Представлены визуальные результаты для всесторонней оценки.

**Доменная адаптация на основе нескольких экземпляров.** В контексте доменной адаптации на основе нескольких экземпляров исследование сравнивает пара-


Рис. 10: Пример смешивания направлений StyleDomain. Можно комбинировать различные направления для выполнения адаптации в семантически смешанный домен.

Таблица 3: Метрики качества и разнообразия [15] для доменных адаптаций на основе одного экземпляра с использованием различных методов. Память обозначает память, необходимую для хранения адаптированных генераторов для всех 12 доменов для каждого метода. Параметризации StyleSpace и StyleSpaceSparse достигают результатов, сопоставимых с другими базовыми методами, при этом имея значительно меньшее количество обучаемых параметров.

| | | | **Titan Erwin** | | **Disney** | | **Across 12 domains** | |
|---|---|---|---|---|---|---|---|---|
| Method | Size | Memory | Quality | Diversity | Quality | Diversity | Quality | Diversity |
| JoJoGAN [39] | 30M | 1.80GB | 0.572 | 0.292 | 0.591 | 0.260 | $0.590 \pm 0.048$ | $0.257 \pm 0.025$ |
| MTG [26] | 30M | 1.80GB | 0.607 | 0.269 | 0.509 | 0.234 | $0.586 \pm 0.054$ | $0.263 \pm 0.028$ |
| GOSA [40] | 30M | 1.80GB | 0.547 | 0.283 | 0.617 | 0.216 | $0.584 \pm 0.034$ | $0.252 \pm 0.030$ |
| DiFa [41] | 30M | 1.80GB | 0.719 | 0.226 | 0.699 | 0.263 | $0.734 \pm 0.047$ | $0.215 \pm 0.038$ |
| TargetCLIP [30] | 9.0K | 420KB | 0.474 | 0.306 | 0.502 | 0.333 | $0.491 \pm 0.043$ | $0.322 \pm 0.015$ |
| DomMod (DiFa) [15] | 6.0K | 280KB | 0.705 | 0.250 | 0.625 | 0.294 | $0.679 \pm 0.049$ | $0.253 \pm 0.037$ |
| **StyleSpace (DiFa)** | 6.0K | 280KB | 0.672 | 0.296 | 0.627 | 0.308 | $0.644 \pm 0.041$ | $0.298 \pm 0.025$ |
| **StyleSpaceSparse (DiFa)** | **1.2K** | **56.4KB** | 0.659 | 0.303 | 0.617 | 0.304 | $0.638 \pm 0.046$ | $0.299 \pm 0.026$ |

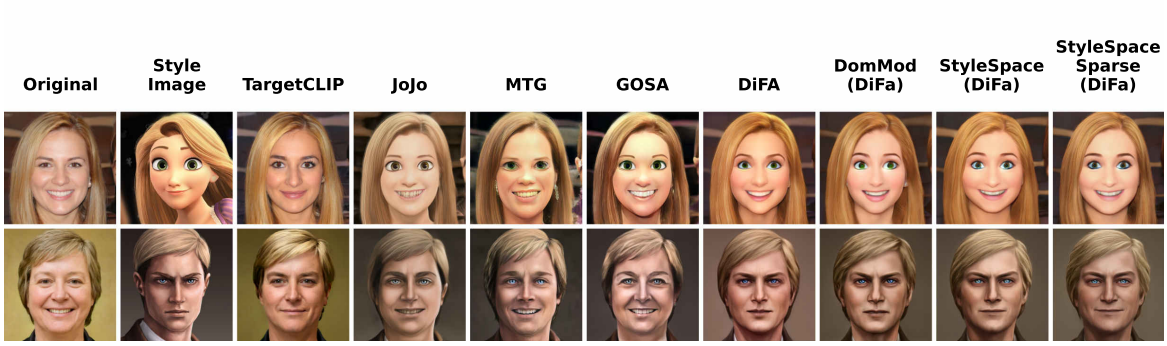



Рис. 11: Сравнение с базовыми моделями для доменной адаптации на основе одного экземпляра. Параметризации StyleSpace и StyleSpaceSparse достигают сопоставимого качества с другими методами, имея значительно меньшее количество обучаемых параметров.

метризации (Affine+ и AffineLight+), примененные к стандартному StyleGAN-ADA с базовыми моделями ADA, CDC и AdAM, используя датасет Dogs и Cats. Эффективность этих методов оценивается с использованием различного количества целевых образцов, и строго следуются тренировочным установкам. Результаты представлены на Рис. 12 и в Таблице 4. Стоит отметить, что количество тренировочных итераций увеличено до 50K для всех методов, чтобы избежать недообучения. Результаты показывают, что производительность AdAM не превосходит стандартный ADA при достаточном количестве тренировок. ADA (Affine+) и ADA (AffineLight+), улучшен-

ные предложенными параметризациями, постоянно превосходят другие методы при различном количестве экземпляров, особенно в условиях малого объема данных.

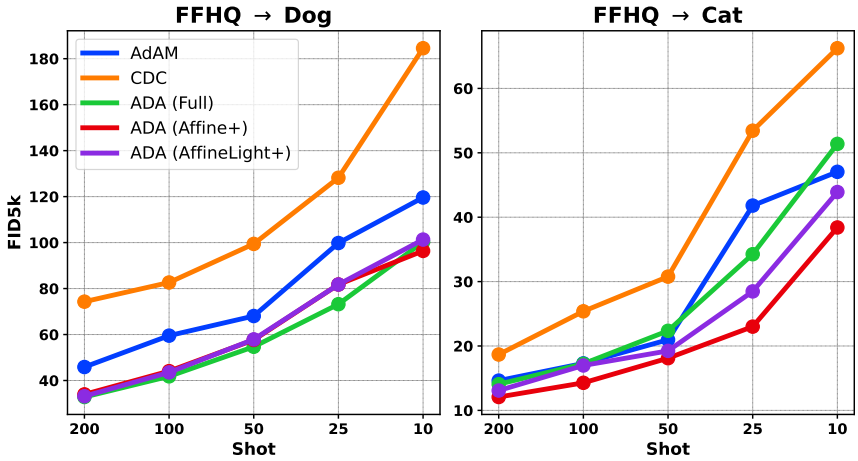



Рис. 12: Результаты тренировок на основе нескольких экземпляров для различного количества экземпляров. Предложенные ADA (Affine+) и ADA (AffineLight+) показывают равномерно лучшую производительность по сравнению с базовыми моделями.

Таблица 4: Результаты тренировок на основе нескольких экземпляров с 10 экземплярами. Предложенные ADA (Affine+) и ADA (AffineLight+) достигают лучшей производительности.

| | | **Domains (10-shots)** | |
|---|---|---|---|
| Method | Size | Cat | Dog |
| CDC [24] | 30M | 66.24 | 184.56 |
| AdAM [42] | 19M | 47.05 | 119.61 |
| ADA (Full) [4] | 30M | 51.38 | 100.25 |
| **ADA (Affine+)** | 5.1M | **38.40** | **96.38** |
| **ADA (AffineLight+)** | **0.6M** | 43.91 | 101.31 |

## 2.3 HiFi++: единая структура для расширения полосы пропускания и улучшения записи речи

Проблема условной генерации речи имеет значительную практическую важность, охватывая такие приложения, как нейронный вокодинг, расширение полосы пропускания (BWE), улучшение записи речи (SE) и другие. Недавний прорыв в этой области использует генеративные состязательные сети (GAN) [13, 14]. В частности, было показано, что вокодеры на основе GAN превосходят общедоступные нейронные вокодеры по качеству записи речи и скорости.

В этом исследовании адаптируется модель HiFi [14] для задач расширения полосы пропускания и улучшения записи речи, вводя новую архитектуру генератора HiFi++. Эта архитектура включает новые компоненты, такие как спектральная предобработка (SpectralUnet), сверточная кодировочно-декодировочная сеть (WaveUNet) и обучаемое спектральное маскирование (SpectralMaskNet). Эти улучшения позволяют нашему генератору эффективно решать задачи расширения полосы пропускания и улучшения записи речи.

Наши обширные эксперименты показывают, что наша модель демонстрирует конкурентоспособные результаты по сравнению с передовыми решениями в области расширения полосы пропускания и улучшения записи речи, при этом будучи значительно более легкой и сохраняя при этом превосходное или эквивалентное качество.

**Адаптация генератора HiFi-GAN для расширения полосы пропускания и улучшения записи речи**

В данной статье представлена архитектура HiFi++, которая расширяет генератор HiFi для решения проблем SE и BWE путем включения трех новых модулей: SpectralUNet, WaveUNet и SpectralMaskNet (см. рисунок 13). Генератор HiFi++ основан на версии V2 генератора HiFi-GAN, принимая в качестве входных данных обогащенную мел-спектрограмму от SpectralUNet и проходя постобработку через WaveUNet и SpectralMaskNet. Изменение порядка этих модулей постобработки не привело к значительным улучшениям.

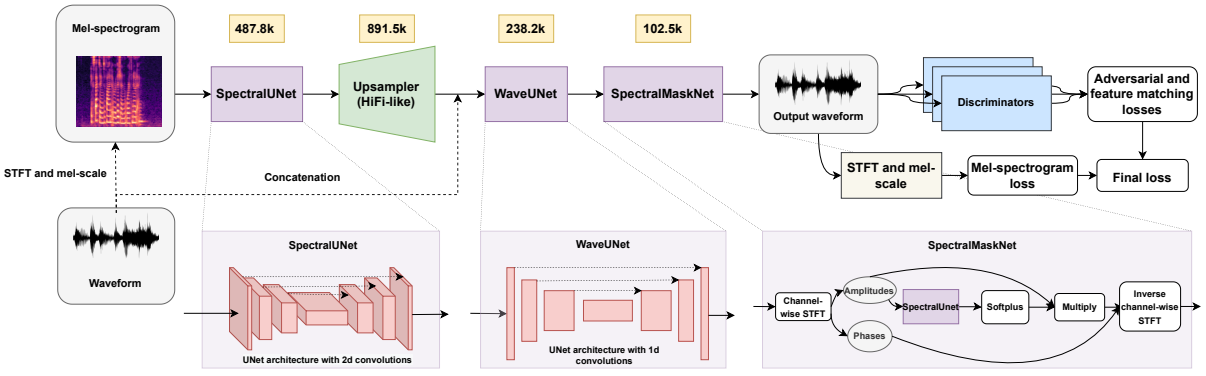



Рис. 13: Архитектура HiFi++ и алгоритм обучения. Генератор HiFi++ состоит из Upsampler, аналогичного HiFi, и трех введенных модулей SpectralUNet, WaveUNet и SpectralMaskNet (их размеры указаны в желтых блоках). Архитектура генератора идентична для BWE и SE.

SpectralUNet: Модуль SpectralUNet служит начальным этапом генератора HiFi++. Он улучшает разрешение мел-спектрограммы, двухмерного представления исходного сигнала, чтобы упростить последующее преобразование в одномерную последовательность. Эта архитектура, подобная UNet, использует двумерные свертки и выполняет функцию предобработки, извлекая важную информацию для целевой задачи, что особенно полезно для расширения полосы пропускания и улучшения записи речи.

WaveUNet: Размещенный после Upsampler HiFi, модуль WaveUNet принимает несколько одномерных последовательностей, объединенных с исходным сигналом. Он работает в временной области, улучшая выход Upsampler и объединяя предсказанный сигнал с исходным. WaveUNet использует архитектуру Wave-U-Net, полностью сверточную одномерную сеть, создавая двумерный тензор из $m$ одномерных последовательностей, которые затем обрабатываются SpectralMaskNet.

SpectralMaskNet: В качестве завершающего этапа генератора, SpectralMaskNet вводит обучаемое спектральное маскирование. Он принимает двумерный тензор из $m$ одномерных последовательностей, применяет поканальное коротковременное преобразование Фурье (STFT) и предсказывает мультипликативные коэффициенты для амплитуд, затем выполняет обратное STFT для изменения спектра. Этот механизм постобработки в частотной области эффективно удаляет артефакты и шум из выходного сигнала в обучаемом режиме.

Цель обучения: В статье используется многодискриминаторная состязательная структура обучения, вдохновленная работой [14]. Вместо использования многопериодных и многомасштабных дискриминаторов используется несколько идентичных дискриминаторов, работающих на тех же разрешениях с меньшим количеством весов. Потери, используемые в обучении, включают LS-GAN потери, потери сопоставления признаков и потери мел-спектрограммы:

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta) \tag{6}$$

$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \ldots, k. \tag{7}$$

Общая потеря для генератора включает эти потери с соответствующими весами, в то время как каждый дискриминатор оптимизируется индивидуально. В экспериментах веса были установлены следующим образом: $\lambda_{fm} = 2$, $\lambda_{mel} = 45$, и количество дискриминаторов $k = 3$.

#### Результаты

**Расширение полосы пропускания:** Для экспериментов по расширению полосы пропускания использованы набор данных VCTK, состоящий из 44 200 записей речи от 110 говорящих. Шесть говорящих были исключены из учебного набора данных, чтобы предотвратить утечку данных. Для оценки использовались 48 высказываний от этих исключенных говорящих. Важно отметить, что текст в оценочных высказываниях не присутствовал в учебных данных.

**Удаление шума из записи речи:** Наши эксперименты по удалению шума проводились с использованием набора данных VCTK-DEMAND, содержащего 11 572 учебных высказывания с различными отношениями сигнал/шум (SNR) и 824 тестовых высказывания. Дополнительные подробности можно найти в оригинальной статье.

#### Оценка

**Объективная оценка:** Улучшение записи речи оценивалось, используя стандартные метрики, такие как WB-PESQ, STOI, SI-SDR и DNSMOS. Кроме того, была введена метрика WV-MOS, прямое предсказание MOS на основе донастроенной модели wave2vec2.0, которая показала лучшую корреляцию с субъективными оценками качества.

**Субъективная оценка:** Субъективная оценка качества проводилась с использованием тестов MOS по шкале 5 баллов, при этом аудиоклипы нормализовались для учета различий в громкости. Участие принимали судьи, говорящие на английском языке и обладающие подходящим оборудованием для прослушивания.

#### Расширение полосы пропускания

В экспериментах по расширению полосы пропускания были обучены модели независимо для трех входных полос пропускания (1 кГц, 2 кГц и 4 кГц). Как видно из таблицы 5, HiFi++ превосходил другие техники по размеру модели и качеству расширения полосы пропускания, будучи в пять раз меньше ближайшего аналога, SEANet. Парные сравнения подтвердили статистическое превосходство HiFi++ над SEANet.

Эти результаты подчеркивают важность состязательных целей в моделях расширения полосы пропускания записи речи. Примечательно, что SEANet, который также использует состязательные цели, оказался сильнейшим аналогом среди изученных моделей, оставив позади модели с супервизорным восстановлением, такие как TFilm и 2S-BWE, особенно для низких входных полос пропускания.

Таблица 5: Результаты расширения полосы пропускания на наборе данных VCTK. * указывает на перевоспроизведение.

| | BWE (1kHz) | | | | BWE (2kHz) | | | | BWE (4kHz) | | | | Размер |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Модель** | **WB-PESQ ↑** | **SI-SDR ↑** | **STOI ↑** | **WV-MOS ↑** | **WB-PESQ ↑** | **SI-SDR ↑** | **STOI ↑** | **WV-MOS ↑** | **WB-PESQ ↑** | **SI-SDR ↑** | **STOI ↑** | **WV-MOS ↑** | **Параметры (M) ↓** |
| **WaveUNet** | 2.38 | 8.40 | 0.76 | 3.38 | 2.80 | 9.40 | 0.84 | 3.57 | 3.30 | 13.60 | 0.90 | 3.86 | 8.67 |
| **TFilm** | 2.64 | 9.10 | 0.81 | 3.45 | 3.01 | 9.80 | 0.87 | 3.61 | 3.40 | 13.80 | 0.91 | 3.89 | 27.00 |
| **2S-BWE** | 2.82 | 9.20 | 0.80 | 3.53 | 3.22 | 10.60 | 0.88 | 3.74 | 3.60 | 14.10 | 0.92 | 3.94 | 9.00 |
| **SEANet*** | 2.91 | 9.40 | 0.83 | 3.60 | 3.35 | 11.20 | 0.89 | 3.82 | 3.71 | 14.80 | 0.93 | 4.02 | 6.30 |
| **HiFi++ (предлагаемая)** | **2.99** | **9.80** | **0.84** | **3.68** | **3.41** | **11.70** | **0.90** | **3.91** | **3.80** | **15.10** | **0.94** | **4.05** | **1.20** |

## 2.4 FFC-SE: Быстрое сверточное преобразование Фурье для улучшения записи речи

Улучшение записи речи играет важную роль в телекоммуникации и привлекает значительное внимание в аудиообработке. Традиционные методы обработки сигналов решают эту задачу, но часто зависят от конкретных моделей шума. В последние годы методы, основанные на данных и использующие глубокое обучение, стали доминирующими в современном улучшении записи речи.

Широко распространенный подход в глубоком обучении для улучшения записи речи включает извлечение сигнала во временной области с использованием структуры сверточного кодера-декодера (CED). Замечательные работы, такие как [43] и [44], используют адверсариальное обучение и сети CED. Некоторые также включают нейронные компоненты, такие как длинные краткосрочные ячейки памяти [45] и трансформеры [46]. Эти методы напрямую отображают шумные волны в чистые сигналы, но часто игнорируют информацию о спектре сигнала, что может приводить к потенциальным неэффективностям. Недавнее исследование стремится явно включить спектральную информацию в процесс генерации, что приводит к достижению передовых результатов.

Другая линия исследований сосредотачивается на представлениях короткосрочного преобразования Фурье (STFT). Подходы в этой категории нацелены на прогнозирование чистых коэффициентов STFT непосредственно или на коррекцию шумных

спектров с использованием масок для изменения амплитуды или амплитуды и фазы [47]. Статьи, такие как MetricGAN и MetricGAN+ [48], используют двунаправленную LSTM для прогнозирования бинарных масок и сообщают о передовых результатах в метриках качества записи речи. Прямое оценивание фаз представляет собой вызов, что приводит к различным стратегиям, включая разделение оценки амплитуды и фазы [49] и использование отдельных вокодерных сетей для синтеза волновой формы. Эти методы часто требуют больших нейронных сетей и значительных вычислительных ресурсов. Для улучшения прогнозирования фаз было предложено ввести нелокальные нейронные операторы, которые уменьшают размер модели, улучшая при этом качество.

В этой работе предлагаются новые нейронные архитектуры, основанные на операторе быстрого сверточного преобразования Фурье (FFC), изначально предназначенном для задач компьютерного зрения. Глобальное рецептивное поле FFC представляет собой преимущество для сложного прогнозирования спектра, особенно для обработки периодических структур в спектрограммах. Наши эксперименты показывают, что большое рецептивное поле FFC помогает в производстве согласованных фаз. Используя эти знания, были разработаны новые нейронные архитектуры для прямого оценивания комплексных спектрограмм в задаче улучшения записи речи. Эти модели достигают передовых результатов на наборах данных VoiceBank-DEMAND [50] и Deep Noise Suppression с гораздо меньшим количеством параметров по сравнению с базовыми методами.

#### Предложенный подход

В этой работе рассматривается проблема удаления шума из одноканальной записи речи с целью отобразить зашумленную волну $y = x + n$, где $n$ - добавленный шум, на чистый сигнал $x$. Наша стратегия включает в себя нейронные архитектуры, улучшенные с помощью нелокального нейронного оператора, называемого быстрой сверткой Фурье (FFC) [51], который адаптируется для обработки сложных спектров. Были представлены две нейронные архитектуры, которые включают этот оператор как основной компонент.

### Быстрая свертка Фурье (FFC)

Быстрая свертка Фурье (FFC) - это нейронный оператор, позволяющий осуществлять нелокальное рассуждение внутри нейронной сети. FFC применяет канальное

быстрое преобразование Фурье, за которым следует поканальная свертка и обратное преобразование Фурье, глобально воздействуя на входной тензор по всем измерениям, вовлеченным в преобразование Фурье. FFC разделяет каналы на локальные и глобальные ветви:

1. Локальная ветвь использует обычные свертки для локального обновления карт признаков.

2. Глобальная ветвь выполняет преобразование Фурье карт признаков в частотной области, воздействуя на глобальный контекст.

В нашей работе выполняется преобразование Фурье только по частотным измерениям карт признаков, соответствующим представлениям короткого временного преобразования Фурье (STFT), в отличие от задач компьютерного зрения, где преобразование Фурье распространяется на обе измерения изображения [51, 52]. Глобальная ветвь слоя FFC состоит из трех шагов:

1. Реальное быстрое преобразование Фурье по частотной размерности входной карты признаков, за которым следует конкатенация реальных и мнимых частей спектра по размерности каналов:

$$\mathbb{R}^{C\times F\times T} \xrightarrow{\text{fft1d}} \mathbb{C}^{C\times F/2\times T} \xrightarrow{\text{concat}} \mathbb{R}^{2C\times F/2\times T}. \tag{8}$$

2. Применение сверточного блока с ядром $1 \times 1$ в частотной области:

$$\mathbb{R}^{2C\times F/2\times T} \xrightarrow{\text{conv}-\text{bn}-\text{relu}} \mathbb{R}^{2C\times F/2\times T}. \tag{9}$$

3. Обратное преобразование Фурье:

$$\mathbb{R}^{2C\times F/2\times T} \xrightarrow{\text{concat}} \mathbb{C}^{C\times F/2\times T} \xrightarrow{\text{ifft1d}} \mathbb{R}^{C\times F\times T}. \tag{10}$$

Здесь $C$, $F$ и $T$ представляют собой количество каналов, размерность, соответствующую частоте, и размерность, соответствующую времени, соответственно. Глобальная и локальная ветви взаимодействуют через суммирование активаций. Общая схема этого модуля приведена на рисунке 14.

В этой работе используется вариация FFC из [52] для восстановления изображений, используя одномерное преобразование Фурье по частотной размерности.

**FFC-AE**

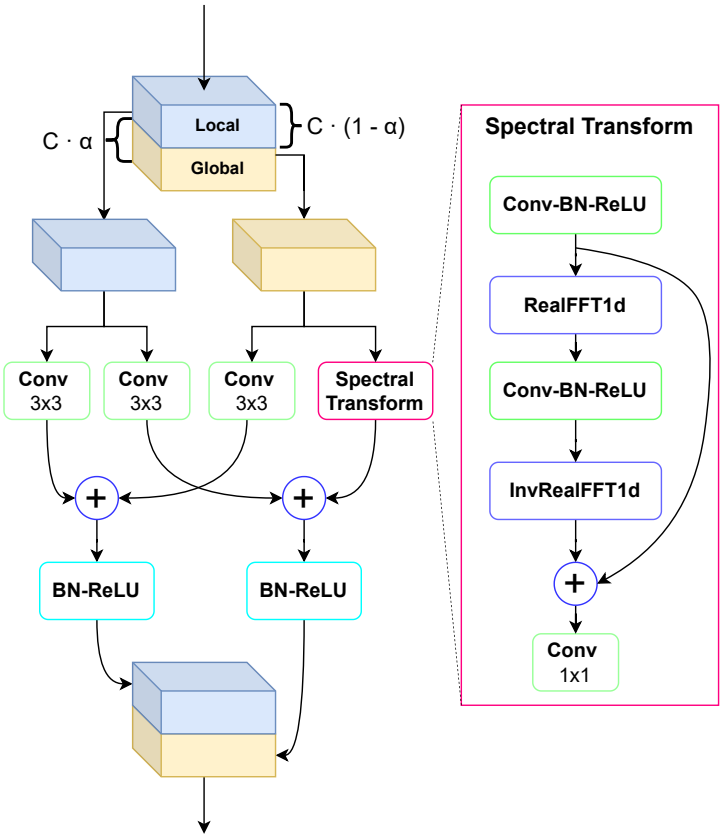



Рис. 14: Нейронный модуль быстрой свертки Фурье для улучшения записи речи. Параметр $\alpha \in [0, 1]$ контролирует соотношение каналов, используемых в глобальной ветви модуля.

Для улучшения записи речи были предложены две архитектуры нейронных сетей. Первая, FFC-AE, вдохновлена работой [52]. FFC-AE включает в себя сверточный энкодер, который уменьшает размер представления входного короткопериодического преобразования Фурье (STFT) по временным и частотным измерениям в два раза. За энкодером следуют остаточные блоки, каждый из которых состоит из двух последовательных модулей быстрого сверточного преобразования Фурье. Выход этих блоков увеличивается с помощью транспонированной свертки и используется для предсказания действительной и мнимой частей денормализованной комплекснозначной спектрограммы. Архитектура изображена на рисунке 15 (слева), и обозначается как быстрый сверточный автокодировщик преобразования Фурье (FFC-AE).

Было обнаружено, что коэффициент уменьшения в 2 обеспечивает подходящий баланс между производительностью и вычислительной сложностью для STFT с окном размером 1024 и шагом 256.

**FFC-UNet**

Вторая архитектура черпает вдохновение из классической модели U-Net [53]. Слои FFC интегрируются в архитектуру U-Net, как показано на рисунке 15 (справа). На каждом уровне структуры U-Net несколько остаточных блоков FFC интегрируются с сверточным уменьшением или увеличением размера. Параметр $\alpha$, представляющий собой соотношение каналов, направляемых в глобальную ветвь быстрого сверточного преобразования Фурье, адаптируется в зависимости от уровня U-Net. Более высокие уровни U-Net работают с данными более высокого разрешения, богатыми периодическими структурами, в то время как более низкие уровни оперируют на более грубой шкале, лишенной таких периодических структур. $\alpha$ уменьшается от 0.75 на верхнем уровне до 0 на нижнем слое с шагом 0.25.

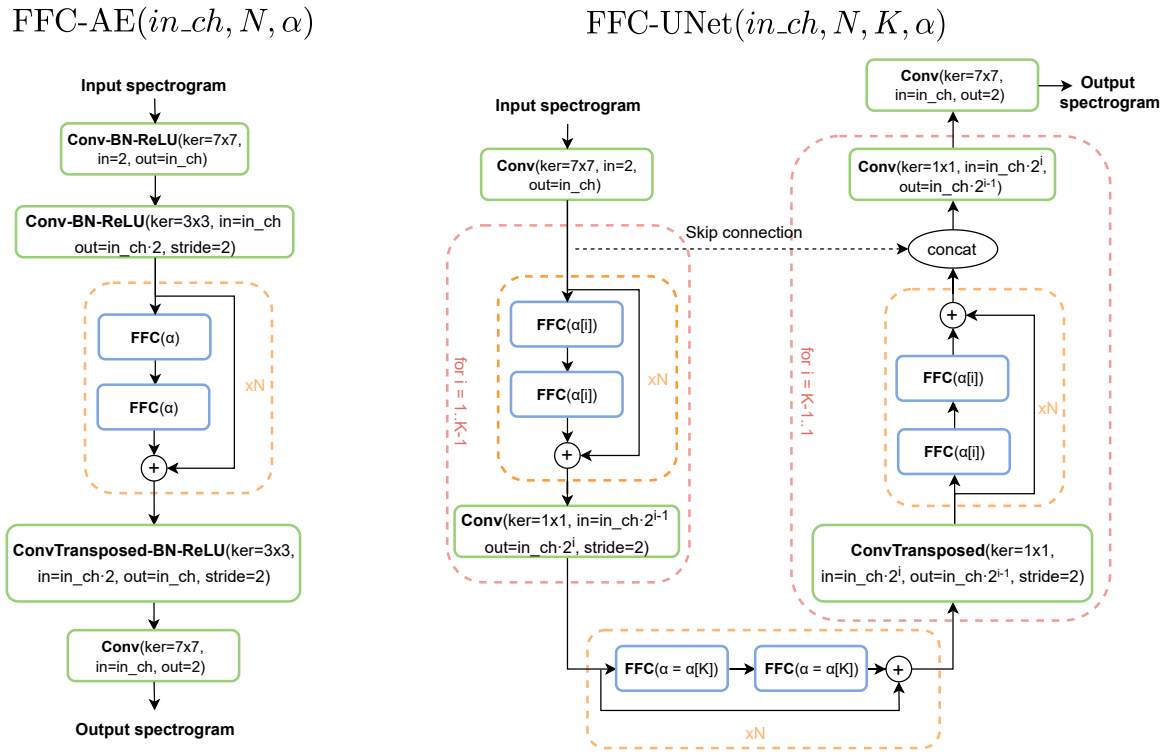



Рис. 15: Предложенные архитектуры для улучшения записи речи. *Слева:* быстрый сверточный автокодировщик преобразования Фурье, который принимает архитектуру, представленную в [52], для задачи улучшения записи речи. *Справа:* быстрый сверточный U-Net преобразования Фурье. Параметр in_ch контролирует общую ширину сетей, N определяет количество остаточных блоков FFC, $K$ - глубина архитектуры FFC-UNet, $\alpha$ (вещественное число $\in [0, 1]$ в случае FFC-AE, K чисел $\in [0, 1]$ в случае FFC-Unet) управляет долей каналов, идущих в глобальную ветвь.

### Обучение

Чтобы преобразовать предсказанное представление STFT в волновую форму, используется обратное короткопериодическое преобразование Фурье. Обучение следу-

ет за мульти-дискриминаторным адверсариалом. Оно включает три потери: потерю LS-GAN $\mathcal{L}_{GAN}$, потерю сопоставления признаков $\mathcal{L}_{FM}$ и потерю мел-спектрограммы $\mathcal{L}_{Mel}$. Потери определены следующим образом:

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta)$$
$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \ldots, k.$$

Здесь $\mathcal{L}(\theta)$ обозначает потери для генератора с параметрами $\theta$, а $\mathcal{L}(\varphi_i)$ обозначает потери для i-го дискриминатора с параметрами $\varphi_i$. Все дискриминаторы идентичны, но инициализируются по-разному. Во всех экспериментах мы устанавливаем $\lambda_{fm} = 2$, $\lambda_{mel} = 45$ и $k = 3$.

#### Результаты

**Наборы данных** В этой работе эффективность моделей удаления шума из записи речи оценивается, используя два бенчмарка. Аудиозаписи были отобраны с частотой дискретизации 16 кГц.

Первый бенчмарк - это набор данных VoiceBank-DEMAND [50], состоящий из обучающего набора с 28 дикторами и 11572 фразами при 4 уровнях отношения сигнал-шум (SNR) (15, 10, 5 и 0 дБ). Тестовый набор (824 фразы) содержит записи 2 дикторов, не участвовавших в обучении, с 4 уровнями SNR (17.5, 12.5, 7.5 и 2.5 дБ).

Второй бенчмарк - это соревнование Deep Noise Suppression (DNS), в рамках которого мы синтезировали 100 часов обучающих данных без искусственной реверберации. Модели были протестированы на двух тестовых наборах: DNS-INDOMAIN (отложенные данные из обучающего набора продолжительностью 100 часов) и DNS-BLIND (реальные шумные записи).

**Метрики** Для объективной оценки мы использовали стандартные метрики, включая WB-PESQ [54], расширенный STOI, SI-SDR [55], COVL, CBAK и CSIG. Кроме того, мы рассматривали объективную меру качества записи речи (WV-MOS), основанную на прямом прогнозировании оценочного MOS с использованием настроенной модели wav2vec2.0, которая показала сильную корреляцию с субъективными мерами качества.

Для субъективной оценки качества мы проводили тесты MOS на 5-балльной шкале. Аудиофрагменты были нормализованы, а референты были англоязычными носителями с правильным акустическим оборудованием.

**Характеристики эксперимента** В наших экспериментах сигналы были преобразованы в спектральное представление с использованием короткопериодического преобразования Фурье (STFT) с окном Ханна размером 1024 и шагом 256. Были использованы конкретные значения параметров для различных версий моделей. FFC-AE использовал $\alpha = 0.75$, $N = 9$, $in\_ch = 32$ для версии V0 и $in\_ch = 64$ для версии V1. FFC-UNet использовал $K = 4$, $N = 4$, $in\_ch = 32$ и постепенное уменьшение $\alpha$. Все модели обучались в течение 800,000 итераций с размером батча 8 и оптимизатором Adam с коэффициентом обучения 0.0002. Модель ResUNet-Decouple+ использовала те же параметры обучения, что и указано в оригинальной статье, с 800,000 итерациями и коэффициентом обучения 0.0002.

**Экспериментальные результаты**

Наши предложенные модели были сравнены с различными базовыми моделями из литературы, включая FullSubNet и DEMUCS, а также модели, такие как обычный U-Net и FFC-AE (abl.). Сравнение проводилось на обоих бенчмарках.

Для VoiceBank-DEMAND (см. Таблицу 6), наши модели достигли значительно более высоких оценок MOS, чем все базовые модели, и продемонстрировали конкурентоспособные результаты по объективным метрикам. На бенчмарке DNS (см. Таблицу 7), наши модели проявили высокое качество по сравнению с конкурентами на тестовом наборе DNS-INDOMAIN и конкурентоспособную производительность с FullSubNet на тестовом наборе DNS-BLIND, что является значимым достижением, учитывая высокое рейтинговое положение FullSubNet в соревновании DNS Challenge 2021.

Замечательно, что наши модели достигли этих результатов без использования динамического синтеза данных, моделирования реверберации или методов аугментации, на которые полагались некоторые ближайшие базовые модели. Дальнейшие улучшения в обобщении к слепому тестовому набору могут быть исследованы с помощью продвинутых пайплайнов генерации данных.

Таблица 6: Результаты удаления шума из записи речи на наборе данных Voicebank-DEMAND. Лучшие три результаты выделены жирным шрифтом.

| Model | MOS | WV-MOS | SI-SDR | STOI | PESQ | CSIG | CBAK | COVL | # Params (M) | # GMAC on 16k |
|---|---|---|---|---|---|---|---|---|---|---|
| Ground Truth | $4.46 \pm 0.06$ | 4.50 | - | 1.00 | 4.64 | 5.0 | 5.0 | 5.0 | - | - |
| Input | $3.44 \pm 0.06$ | 2.99 | 8.4 | 0.79 | 1.97 | 3.34 | 2.82 | 2.74 | - | - |
| MetricGAN+ [48] | $3.82 \pm 0.06$ | 3.90 | 8.5 | 0.83 | **3.13** | 4.12 | 3.16 | 3.62 | 2.7 | - |
| ResUNet-Decouple+ [49] | $3.94 \pm 0.04$ | 4.13 | **18.4** | 0.84 | 2.45 | 3.38 | 3.15 | 2.89 | 102.6 | - |
| DEMUCS (non-caus.) [45] | $4.06 \pm 0.03$ | **4.37** | **18.5** | **0.87** | **3.03** | **4.36** | **3.51** | **3.72** | 60.8 | - |
| VoiceFixer [56] | $4.10 \pm 0.03$ | 4.14 | -18.5 | 0.75 | 2.38 | 3.6 | 2.37 | 2.96 | 122.1 | 34.4 (x2) |
| HiFi++ [57] | $4.15 \pm 0.07$ | 4.27 | **18.4** | 0.86 | 2.76 | 4.09 | 3.35 | 3.43 | **1.7** | 1.5(x2) |
| FFC-AE-V0 (ours) | $\mathbf{4.24 \pm 0.09}$ | 4.34 | 17.9 | 0.86 | 2.88 | 4.25 | 3.40 | 3.57 | **0.42** | 4.39 |
| FFC-AE-V1 (ours) | $\mathbf{4.33 \pm 0.03}$ | **4.37** | 17.5 | **0.87** | 2.96 | **4.34** | **3.42** | **3.66** | **1.7** | 16.33 |
| FFC-UNet (ours) | $\mathbf{4.28 \pm 0.03}$ | **4.38** | 18.1 | **0.87** | **2.99** | **4.35** | **3.47** | **3.69** | 7.7 | 19.81 |
| FFC-AE-V1 (abl.) | $3.98 \pm 0.07$ | 4.05 | 16.7 | 0.84 | 2.68 | 3.94 | 3.23 | 3.31 | 2.9 | 2.25 |
| vanilla UNet | $4.10 \pm 0.07$ | 4.11 | 17.2 | 0.85 | 2.73 | 3.94 | 3.28 | 3.34 | 20.7 | 11.2(x2) |

Таблица 7: Результаты удаления шума из записи речи на наборе данных DNS. * указывает на результаты на тестовом наборе DNS-BLIND. Лучшие три результаты выделены жирным шрифтом.

| Model | MOS | MOS* | WV-MOS | WV-MOS* | SI-SDR | STOI | PESQ | CSIG | CBAK | COVL | # Params (M) | # GMAC on 16k |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ground Truth | $4.40 \pm 0.08$ | - | 3.845 | - | - | 1.00 | 4.64 | 5.0 | 5.0 | 5.0 | - | - |
| Input | $2.75 \pm 0.07$ | $2.43 \pm 0.08$ | 1.195 | 0.80 | - | 0.69 | 1.49 | 2.59 | 2.32 | 1.99 | - | - |
| DEMUCS [45] | $3.52 \pm 0.15$ | $2.94 \pm 0.08$ | **3.32** | **2.83** | **15.56** | 0.82 | 2.20 | 3.44 | 3.21 | 2.81 | 33.5 | 7.84 |
| HiFi++ [57] | $3.54 \pm 0.08$ | $2.75 \pm 0.06$ | 2.91 | 2.32 | 11.69 | 0.82 | 2.20 | 3.65 | 3.00 | 2.92 | **1.7** | - |
| ResUNet-Dec+ [49] | $3.63 \pm 0.04$ | $2.51 \pm 0.08$ | 2.94 | 1.86 | 14.78 | 0.81 | 2.09 | 2.82 | 3.06 | 2.43 | 102.6 | - |
| FullSubNet [**?** ] | $3.73 \pm 0.02$ | $\mathbf{3.08 \pm 0.09}$ | 2.90 | 2.41 | **14.96** | 0.82 | 2.43 | 3.59 | **3.27** | 3.0 | 5.6 | - |
| FFC-AE-V0 (ours) | $\mathbf{3.92 \pm 0.09}$ | $2.88 \pm 0.09$ | 3.20 | 2.58 | 12.86 | **0.83** | **2.44** | **3.84** | 3.17 | **3.15** | **0.42** | 4.39 |
| FFC-AE-V1 (ours) | $\mathbf{4.02 \pm 0.05}$ | $\mathbf{3.10 \pm 0.07}$ | **3.33** | **2.76** | 14.12 | **0.85** | **2.61** | **3.98** | **3.31** | **3.31** | **1.7** | 16.33 |
| FFC-UNet (ours) | $\mathbf{4.00 \pm 0.06}$ | $\mathbf{3.11 \pm 0.08}$ | **3.35** | **2.70** | **15.48** | **0.86** | **2.69** | **4.08** | **3.44** | **3.41** | 7.7 | 19.81 |

# 3 Заключение

Этот раздел представляет краткое изложение основных вкладов нашей работы. Основные результаты работы включают в себя эффективные параметризации и архитектурные модули для генераторов GAN, предназначенных для решения проблемы доменной адаптации в компьютерном зрении и улучшения записи речи в обработке сигналов.

1. В *HyperDomainNet* была предложена новая параметризацию StyleGAN для доменной адаптации, которая содержит всего 6 тысяч обучаемых параметров по сравнению с 30 миллионами весов в обычной полной параметризации. Эта параметризация основана на технике модуляции домена, которая позволяет эффективно изменять веса генератора с помощью небольшого вектора обучения. В серии обширных экспериментов по адаптации текста и изображений было показано, что эта параметризация достигает такого же качества, как и текущие методы, использующие полную параметризацию генератора StyleGAN. Также был предложен новый HyperDomainNet, который решает проблему мультидоменной адаптации. Идея заключается в том, что из текстового описания домена или примера изображения домена гиперсеть предсказывает вектор домена, который генератор адаптирует с помощью техники модуляции домена. Это позволяет адаптироваться к сотням или тысячам новых доменов сразу, без необходимости повторного обучения генератора для каждого домена индивидуально. В экспериментах было показано, что HyperDomainNet позволяет адаптировать генератор к новым доменам так же, как и обычные методы, работающие в одиночной доменной адаптации. Кроме того, эта модель показала многообещающие результаты обобщения для новых доменов.

2. В *StyleDomain* был проводен систематический анализ для решения проблемы адаптации StyleGAN между доменами. Наше исследование разворачивается в двух частях: сначала определяется, какие части StyleGAN требуют адаптации на основе сходства между исходными и целевыми доменами. Для схожих доменов часто достаточно тонкой настройки только аффинных слоев, в то время как более различные домены требуют оптимизации дополнительных параметров, хотя и не всей сети, что указывает на потенциал более эффективных параметриза-

ций. Во второй части представляются две новые параметризации: для схожих доменов предлагается *StyleSpace*, который оптимизирует направления адаптации без тонкой настройки всех весов, и для более удаленных доменов представляется *Affine+*, значительно сокращающий количество обучаемых параметров при сохранении качества. Дальнейшая доработка с *AffineLight+* использует низкоранговое разложение для весов аффинных слоев, превосходя сложные базовые подходы в адаптации на основе нескольких экземпляров. Кроме того, исследуются свойства направлений *StyleDomain*, раскрывая их смешиваемость и переносимость, которые могут создавать новые стили или применяться к другим тонко настроенным моделям StyleGAN. Эти результаты используются в различных задачах компьютерного зрения, таких как перевод изображения в изображение и морфинг между доменами.

3. В статье *HiFi++* была представлена новая архитектура генератора HiFi++, разработанная для расширения полосы пропускания и задач улучшения записи речи. Эта архитектура включает новые компоненты, включая спектральную предобработку (SpectralUnet), сверточную кодировщик-декодировщик сеть (WaveUNet) и обучаемую спектральную маску (SpectralMaskNet), что позволяет нашему генератору эффективно решать эти задачи. Обширные эксперименты показывают, что наша модель выступает конкурентоспособно по сравнению с передовыми решениями в BWE и SE, при этом она значительно более легкая и поддерживает превосходное или эквивалентное качество. Кроме того, в работе *FFC-SE* предлагаются новые нейронные архитектуры на основе оператора быстрой свертки Фурье (FFC), изначально разработанного для задач компьютерного зрения. Глобальное рецептивное поле FFC имеет преимущества для сложного прогнозирования спектра, особенно для обработки периодических структур в спектрограммах, что помогает производить согласованные фазы. Используя эти исследования, были разработаны новые нейронные архитектуры для прямой оценки комплексных значений спектрограммы в улучшении записи речи, достигая передового качества на наборах данных VoiceBank-DEMAND и Deep Noise Suppression с значительно меньшим количеством параметров по сравнению с базовыми методами.

## Список литературы


[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[4] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[7] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021.

[8] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

[9] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[12] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

[13] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019.

[14] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020.

[15] Aibek Alanov, Vadim Titov, and Dmitry Vetrov. Hyperdomainnet: Universal domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2210.08884*, 2022.

[16] Aibek Alanov, Vadim Titov, Maksim Nakhodnov, and Dmitry Vetrov. Styledomain: Efficient and lightweight parameterizations of stylegan for one-shot and few-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2184–2194, 2023.

[17] Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Dmitry Vetrov. Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement. *arXiv preprint arXiv:2203.13086*, 1(2), 2022.

[18] Ivan Shchekotov, Pavel Andreev, Oleg Ivanov, Aibek Alanov, and Dmitry Vetrov. Ffc-se: Fast fourier convolution for speech enhancement. *arXiv preprint arXiv:2204.03042*, 2022.

[19] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.

[20] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.

[21] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018.

[22] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020.

[23] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

[24] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021.

[25] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[26] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398*, 2021.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[28] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[29] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[30] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. *arXiv preprint arXiv:2110.12427*, 2021.

[31] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[32] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021.

[33] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020.

[34] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2020.

[35] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020.

[37] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021.

[38] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[39] Min Jin Chong and David Forsyth. Jojogan: One shot face stylization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 128–152. Springer, 2022.

[40] Zicheng Zhang, Yinglu Liu, Congying Han, Tiande Guo, Ting Yao, and Tao Mei. Generalized one-shot domain adaption of generative adversarial networks. *arXiv preprint arXiv:2209.03665*, 2022.

[41] Yabo Zhang, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Wangmeng Zuo, et al. Towards diverse and faithful one-shot adaption of generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2022.

[42] Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Few-shot image generation via adaptation-aware kernel modulation. *arXiv preprint arXiv:2210.16559*, 2022.

[43] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. Seanet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095*, 2020.

[44] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[45] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.

[46] Eesung Kim and Hyeji Seo. SE-Conformer: Time-Domain Speech Enhancement Using Conformer. In *Proc. Interspeech 2021*, pages 2736–2740, 2021. doi: 10.21437/Interspeech.2021-2207.

[47] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations*, 2018.

[48] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*, 2021.

[49] Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. *arXiv preprint arXiv:2109.05418*, 2021.

[50] Cassia Valentini-Botinhao et al. Noisy speech database for training speech enhancement algorithms and tts models. 2017.

[51] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020.

[52] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[54] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001.

[55] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[56] Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731*, 2021.

[57] Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Dmitry Vetrov. Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement. *arXiv preprint arXiv:2203.13086*, 2022.