NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Aibek Alanov

EXPLORING EFFICIENT PARAMETERIZATIONS FOR GANS IN IMAGE AND SPEECH GENERATION

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Academic Supervisor:
Candidate of Sciences
Dmitry P. Vetrov

Moscow — 2024

1 Introduction

Topic of the thesis

GANs [1, 2, 3, 4, 5] have, in recent years, achieved impressive results in generating data that is indistinguishable in quality from real data. They enable the learning of a generator that transforms a latent space with a simple distribution into a space of real objects with a very complex distribution. Due to their ability to generate high-quality data, GANs have been widely utilized in various tasks and fields, including computer vision [6, 7, 8, 9, 10, 11, 12] and signal processing [13, 14]. However, achieving such high-quality generation during GAN training requires access to large-scale datasets, which are time-consuming and expensive to collect. For instance, training the state-of-the-art StyleGAN model to generate photorealistic human faces necessitated the collection of the FFHQ dataset [3], comprising 70 thousand very high-resolution (1024x1024) images of faces.

The issue of training GANs on small datasets remains a significant challenge. One primary approach to addressing this problem is domain adaptation, wherein a GAN is trained on a new domain with a limited number of examples by fine-tuning a model pretrained on another domain with access to a large-scale dataset. For example, to generate faces in the style of certain artists, where assembling a large dataset is impractical, a GAN pretrained on a large dataset of photorealistic faces (e.g., FFHQ) can be fine-tuned using a few example pictures of a particular artist. In domain adaptation, it is crucial which subset of the underlying model parameters is optimized. This optimization determines how effectively the underlying model's knowledge can be transferred to the new domain and helps avoid mode collapse, to which GANs are highly prone.

This thesis will propose new efficient StyleGAN parametrizations for the domain adaptation problem and new compact architectures for the speech enhancement problem, which also make efficient use of training data. Specifically, this work proposes a domain modulation technique that allows training thousands of times fewer parameters for the StyleGAN model than the full parametrization for domain adaptation. This innovation enabled the proposal of the HyperDomainNet [15] model, which addresses the multi-domain adaptation problem. Further development of these ideas led to the discovery of more efficient parametrizations, such as StyleSpace and Affine+ [16]. Additionally, there has been a deeper analysis of which parts of the StyleGAN model are crucial in domain adaptation, and interesting properties of directions from StyleSpace have been uncovered. In the realm of speech enhancement, the HiFi++ [17] and FFC-SE [18] models were proposed, demonstrating superior quality in this task compared to existing approaches, while having significantly fewer parameters.

As we analyze the problem of efficient GAN training, we aim to answer fundamental questions, such as: How can we fine-tune GANs for novel domains with limited training data? What are the most important factors in adapting the generator for domain-specific content? Can we reduce computational overhead while maintaining or even improving performance in audio generation and enhancement tasks? These questions form the core of our investigation, and the subsequent chapters of this thesis aim to provide comprehensive insights into these essential topics.

In this introduction, we set the stage for a detailed exploration of each of the four papers, highlighting their specific contributions, insights, and significance in the realm of GAN-based generative models. By the end of this analysis, we hope to offer a deeper understanding of how efficient parameterizations can propel GANs towards greater adaptability, robustness, and resource efficiency, thereby contributing to the continued advancement of image and speech generation technologies.

Relevance

This work offers valuable contributions that address critical challenges in training GANs with limited data and computational resources, which have a significant impact on many applications in image generation and speech enhancement. Here, we highlight the relevance and importance of this research:

1. **Advancing Domain Adaptation in GANs:** The first two papers, *HyperDomainNet* and *StyleDomain*, make substantial contributions to the field of domain adaptation for GANs. With an increasing need to adapt GAN models to specific domains with limited data, these papers propose efficient and lightweight parameterizations. This research enables the practical use of GANs in scenarios where data scarcity is a critical concern, extending their applicability in real-world settings.

2. **Reduction in Computational Resources:** The *HyperDomainNet* and *HiFi++* papers emphasize the importance of reducing computational resources while maintaining or improving the quality of generated content. Given that computational efficiency is a critical factor in deploying GANs in resource-constrained environ-

ments, this research contributes to making GAN-based models more accessible and cost-effective. It aligns with the current trend in AI research toward optimizing deep learning models for practical deployment.

3. **Universal Applicability:** The development of *HyperDomainNet*, which can adapt to multiple domains with a single model, is particularly relevant in the era of data-driven AI. In many practical scenarios, maintaining distinct models for various domains is challenging, making the idea of universal adaptation highly appealing. The capability of a single model to generalize and adapt to multiple domains is crucial for efficient, flexible, and scalable AI systems.

4. **Efficient Speech Enhancement:** In the domain of speech enhancement, the *HiFi++* and *FFC-SE* papers introduce novel and effective architectures. The *HiFi++* paper demonstrates that GANs can outperform traditional methods for bandwidth extension and speech enhancement, while having considerably fewer parameters and reduced computational complexity. Meanwhile, the *FFC-SE* paper applies novel techniques to improve speech enhancement through Fast Fourier convolution, making the architecture even more lightweight and achieving better performance in practice.

The goal of this work is to propose novel efficient parameterizations for GAN models that allows us to significantly reduce the number of optimized parameters and the volume of required training data.

2 Key results and conclusions

The main contributions of this study can be described as:

1. In *HyperDomainNet* paper, we proposed a new parametrization of StyleGAN based on domain modulation techniques and a new HyperDomainNet model. Our parametrization reduced the number of trained StyleGAN parameters by several thousand times for domain adaptation, while achieving comparable quality as existing approaches that train almost all StyleGAN generator parameters. We also introduced a new HyperDomainNet model that allows us to address the problem of multi-domain adaptation, i.e., when we want to adapt StyleGAN to multiple domains simultaneously. This gives new possibilities for cases where we have a lot of different

domains that we want to train on and we don't want to train a separate model for each one. Our approach dramatically improves the efficiency and applicability of the model for such cases.

2. In *StyleDomain* work, we analysed the StyleGAN domain adaptation task in more depth. We investigated which parts of this model are important for adapting to different domains depending on the similarity of the target domain to the source domains. As a result of this analysis, we proposed new efficient parametrisations of StyleSpace and Affine+. StyleSpace is the easiest parametrization to solve the domain adaptation problem for close domains and achieves the same quality as other approaches that train significantly more parameters. The Affine+ parametrization is designed for more distant domains and performs the best in the few-shot learning task, while having fewer trained parameters than baselines. We also discovered surprising properties of these parameterisations that can be used for even more applications.

3. In *HiFi++* and *FFC-SE* papers, we proposed new efficient models for the speech enchancement problem. In *HiFi++*, we presented new modules in the GAN generator architecture that significantly improve the final quality of the model with very few parameters. We have shown that with this architecture, the model performs on par or even better than existing approaches with significantly fewer parameters. In *FFC-SE*, the generator architecture was further improved by Fourier convolution, which allowed more information to be considered and utilised. This reduced the size of the model and improved the final quality.

Theoretical and practical significance. The theoretical significance of this work lies in its novel approaches to parametrizing and adapting the StyleGAN architecture, as well as advancing speech enhancement models. By introducing the HyperDomainNet and StyleDomain frameworks, the study presents methods for reducing the number of trained parameters in StyleGAN for domain adaptation, achieving quality comparable to existing approaches while having significantly less parameters. This includes new parametrizations like StyleSpace and Affine+, which optimize adaptation for both close and distant domains, revealing unexpected properties that broaden potential applications. Practically, these advancements result in more efficient, multi-domain adaptable models, enhancing their practical utility in scenarios with numerous diverse domains. Additionally, the HiFi++ and FFC-SE models propose new architectures for speech enhancement, utilizing GAN modules and Fourier convolution to significantly improve model performance with fewer parameters, thus contributing to more efficient and high-quality speech processing solutions.

Methodology and research methods. In this work, we apply deep learning, generative models, generative adversarial networks, domain adaptation approaches, speech enhancement, as well as standard optimization methods.

Reliability. Detailed descriptions of the proposed methods and experiments are provided, with the code for all papers released publicly.

Key aspects/ideas to be defended.

1. The *domain modulation* technique for efficient domain adaptation and *HyperDomainNet* for multi-domain adaptation training.

2. The efficient parametrizations, *StyleSpace* and *Affine+*, for StyleGAN domain adaptation in close and distant domain tasks.

3. The efficient speech enhancement models: *HiFi++* that improves quality with minimal parameters and *FFC-SE* that enhances model performance using Fourier convolution.

Author contribution. The research presented in this thesis is the result of several years of dedicated work and collaborative effort. This section describes the author's specific contributions to each of the four papers that make up this thesis. In the first paper *HyperDomainNet*, the author proposed a domain modulation technique for efficient domain adaptation of StyleGAN. The author was also responsible for implementing the one-shot domain adaptation experiments and prepared the main body of text for all sections of the paper. In the second *StyleDomain* paper, the author proposed the *StyleSpace* and *Affine+* parameterisations and prepared the text for all sections of the paper except the experiments section. In the *HiFi++* paper, the author proposed the idea of using several simple and lightweight discriminators, found the optimal size for each part of the architecture, and was responsible for the experiments to find the best discriminator configuration. In addition, the author played a significant role in writing the text of the introduction and the main sections of the paper. In the fourth paper *FFC-SE*, the author was involved in writing the code base and the design of the experiments. The author was also involved in editing the text of the paper and participated in discussions regarding the analysis of the results obtained.

Publications and probation of the work

First-tier publications

* denotes equal contribution of coauthors

1. ***Aibek Alanov****, *Vadim Titov*, and Dmitry Vetrov.* HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks. // In Advances in Neural Information Processing Systems, 2022 (NeurIPS 2022). Vol. 35, pages 29414–29426. CORE A* conference.

2. ***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 International Conference on Computer Vision, 2023 (ICCV 2023). Pages 2184-2194. CORE A* conference.

3. *Ivan Shchekotov*, Pavel Andreev*, Oleg Ivanov,* ***Aibek Alanov****, and Dmitry Vetrov.* FFC-SE: Fast Fourier Convolution for Speech Enhancement. // In InterSpeech Conference, 2022. Pages 1188-1192. CORE A conference.

Second-Tier Publications

1. *Pavel Andreev*,* ***Aibek Alanov****, *Oleg Ivanov*, and Dmitry Vetrov.* HiFi++: a Unified Framework for Bandwidth Extension and Speech Enhancement. // In International Conference on Acoustics, Speech, and Signal Processing, 2023 (ICASSP 2023). Pages 1-5. CORE B conference (according to CORE2018).

Reports at Conferences and Seminars

1. Talk on “Audio Synthesis and Bandwidth Extension”, Seminar of Bayesian methods research group, Moscow, April 2021.

2. Poster presentation on “FFC-SE: Fast Fourier Convolution for Speech Enhancement.”, InterSpeech Conference, Seoul, Republic of Korea, September 2022.

3. Poster presentation on “HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks”, Conference on Neural Information Processing Systems, New Orleans, USA, December 2022.

4. Talk on "Domain Adaptation of GANs", Seminar of Bayesian methods research group, Moscow, December 2022.

5. Talk on "HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks", Conference Fall into ML, Moscow, November 2022.

6. Talk on "HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks", Seminar AIRI AIschnitsa, Moscow, December 2022.

7. Talk on "HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks", Conference of the Faculty of Computer Science, Voronovo, June 2022.

Volume and structure of the work. The thesis contains an introduction, contents of publications and a conclusion. The full volume of the thesis is 142 pages.

3 Content of the work

3.1 HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks

In the field of computer vision, Generative Adversarial Networks (GANs) [1, 2, 3, 4, 5] have shown remarkable performance in various tasks like image enhancement [6, 7], editing [8, 9], and image-to-image translation [10, 11, 12]. However, training modern GANs requires a large number of samples, limiting their application to domains with abundant image data. To overcome this limitation, transfer learning (TL) is commonly used, where a pre-trained model is fine-tuned for a new domain with limited data.

Current GAN TL methods typically fine-tune almost all weights of the pre-trained model [19, 20, 21, 22, 4, 23, 24, 25, 26]. While this is suitable for distant target domains, it's often unnecessary for domains similar to the source. In such cases, fine-tuning all weights seems redundant. This study introduces a more efficient approach called *domain-modulation*, which optimizes only a single 6,000-dimensional vector for each target domain, significantly reducing the parameter space compared to traditional fine-tuning of all 30 million weights.

The domain-modulation technique is applied to two state-of-the-art domain adaptation methods, StyleGAN-NADA [25] and MindTheGAP [26], demonstrating comparable performance to full parameterization while being much more lightweight. Additionally, a new regularization loss is proposed to enhance the diversity of the fine-tuned generator.

The study also addresses multi-domain adaptation, where a single model adapts to multiple domains based on input text descriptions or image examples. Instead of fine-tuning separate generators for each target domain, a hyper-network named *HyperDomainNet* is introduced. This hyper-network predicts the vector for StyleGAN2 based on the target domain, significantly reducing training time and the number of trainable parameters. It is observed that this approach can generalize to unseen domains if a sufficient number of domains are used for training.

The research presents extensive experiments to validate the proposed techniques across various domains. The results show that domain-modulation achieves quality comparable to full parameterization, and the regularization loss improves the fine-tuned generator's diversity. Additionally, the HyperDomainNet demonstrates promising generalization to diverse unseen domains.

In summary, this work offers three key contributions:

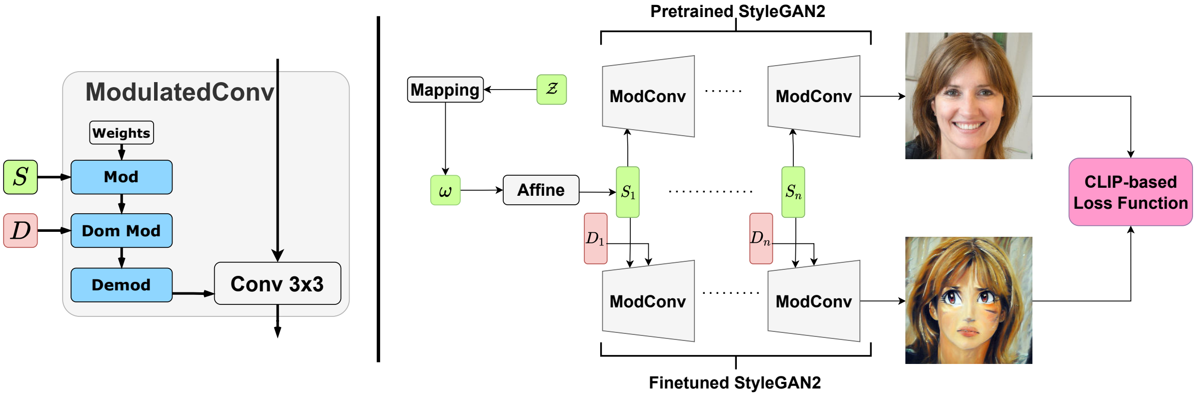
1. A domain-modulation technique that reduces the parameter space for domain adaptation in StyleGAN2 by several orders of magnitude.
2. A novel regularization loss to enhance the diversity of fine-tuned generators.
3. The introduction of a HyperDomainNet for multi-domain adaptation, showcasing generalization to unseen domains.

Background

StyleGAN2 [3] generates images through a mapping network $M(z)$ that transforms initial random vectors $z \in \mathcal{Z}$ into an intermediate latent space $\mathcal{W}$, which is then passed through affine transformations $A(w)$ to create style parameters $s = A(w) \in \mathcal{S}$. These parameters influence the final feature maps produced by a synthesis network G_{sys}. ToRGB layers G_{tRGB} are used to generate the output image.

Domain Adaptation Problem: Adapting a trained StyleGAN2 generator from one domain (source) to another (target), guided by either an image or text description from the target domain.

CLIP Model [27]: CLIP is a model that aligns text and image embeddings in a shared space, measuring the semantic similarity of objects based on cosine distance.



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

Figure 1: Detailed diagram of proposed method. (a) Revised ModulatedConv block with introduced domain-modulation operation. (b) Fully detailed training process of the domain adaptation with the proposed domain-modulation technique.

StyleGAN-NADA [25]: This method uses CLIP to align source and target domains in the CLIP space. It optimizes the synthesis network of the target domain using a direction loss between images and text descriptions.

MindTheGap [26]: Designed for one-shot domain adaptation, MindTheGap aims to prevent the loss of diversity in target images. It introduces a direction loss that uses the embedding of the target image in the source domain, improving adaptation quality.

In summary, these methods adapt StyleGAN2 to new domains using CLIP-based alignment techniques, improving the quality of synthesized images.

Approach

The study aims to enhance StyleGAN domain adaptation by optimizing the synthesis network $G_{sys}(\cdot, \cdot)$ using a compact parameter space. This network component is primarily altered during domain fine-tuning. The approach introduces *domain modulation*, an operation that refines feature convolution weights within the synthesis network. Modulation adjusts weights based on style parameters, leading to a more efficient form of adaptive instance normalization (AdaIN) [28, 29]. This technique is inspired by style transfer methods using AdaIN for image translation.

The domain modulation method reduces the parameter space for fine-tuning StyleGAN2 by optimizing only a vector d with the same dimension as the style parameters. This vector is incorporated into the StyleGAN architecture through an additional modulation operation (see Figure 1a). Instead of optimizing all weights θ of the G_{sys} component, only the d vector is trained. This dimension of the vector d equals 6 thousand that is 4 thousand times less than the original 30 million weights space θ of $G_{sys}(\cdot, \cdot)$ part.

Improving Diversity of CLIP-Guided Domain Adaptation

Existing CLIP-based domain adaptation methods, StyleGAN-NADA and MindTheGap, employ $\mathcal{L}_{direction}$ (or $\mathcal{L}_{clip\_across}$) loss to address mode collapsing issues. However, this loss partially preserves diversity and collapses after some iterations, particularly problematic for domains requiring extensive fine-tuning. The issue with $\mathcal{L}_{direction}$ is that it calculates cosine distances between embeddings that no longer lie on the CLIP sphere, contributing to mode collapse.

To tackle this, a new regularizer called *indomain angle consistency* loss is introduced. This loss computes CLIP cosine distances exclusively between CLIP embeddings. It aims to maintain pairwise cosine distances between images before and after domain adaptation, effectively enhancing generator diversity compared to the original loss functions:

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

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

Designing the HyperDomainNet for Universal Domain Adaptation

We propose a domain-modulation technique for efficient multi-domain adaptation of StyleGAN2. Our goal is to train the *HyperDomainNet*, which predicts domain parameters for fine-tuning generators. Specifically, we focus on the scenario where target domains are represented by text descriptions.

The HyperDomainNet takes text embeddings as input and outputs domain parameters. We use a combination of loss functions, including $\mathcal{L}_{direction}$, $\mathcal{L}_{tt-direction}$ and $\mathcal{L}_{domain-norm}$, to train the network. These losses ensure that the predicted domain parameters effectively guide domain adaptation and prevent domain mixing. The training process is described in the Figure 2.

In summary, we introduce a domain-modulation approach for multi-domain adaptation in StyleGAN2, focusing on text-based target domains. The HyperDomainNet is trained with a set of loss functions to enable effective domain-specific fine-tuning. For detailed descriptions of the losses and optimization process, please see the original paper.

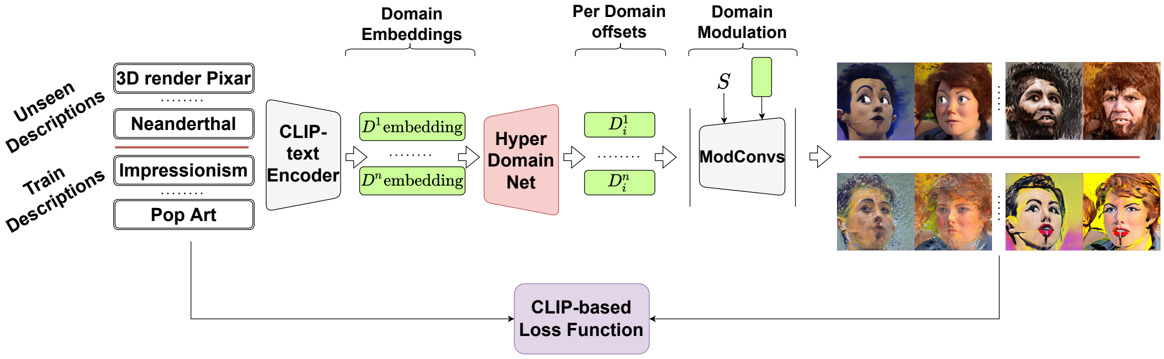


Figure 2: Detailed training process of the HyperDomainNet. On the training phase only reference descriptions are included into CLIP-guided training.

Results

This section presents results for text-based, one-shot, and multi-domain adaptation using our proposed approach.

Text-Based Domain Adaptation We compare our parameterization with StyleGAN-NADA [25] on diverse domains. Our parameterization matches the expressiveness of StyleGAN-NADA, enabling adaptation to style and texture changes. Qualitative results are in Figure 3, demonstrating comparable performance.

One-Shot Domain Adaptation We apply our parameterization and the indomain angle consistency loss to the MindTheGap [26] method. Results in Table 1 and in Figure 4 show our approach achieves similar performance to the original, with significantly fewer parameters. TargetCLIP [30] and other methods exhibit poor adaptation quality, mainly suitable for in-domain editing. Indomain angle consistency significantly improves FID and precision metrics.

Multi-Domain Adaptation We use the HyperDomainNet in two scenarios: (i) fixed number of domains and (ii) arbitrary number of domains. Results in Figure 5 reveal the effectiveness of our method in both scenarios, with promising adaptation to unseen domains. Ablation study supports the importance of the proposed losses in training the HyperDomainNet for multi-domain adaptation.

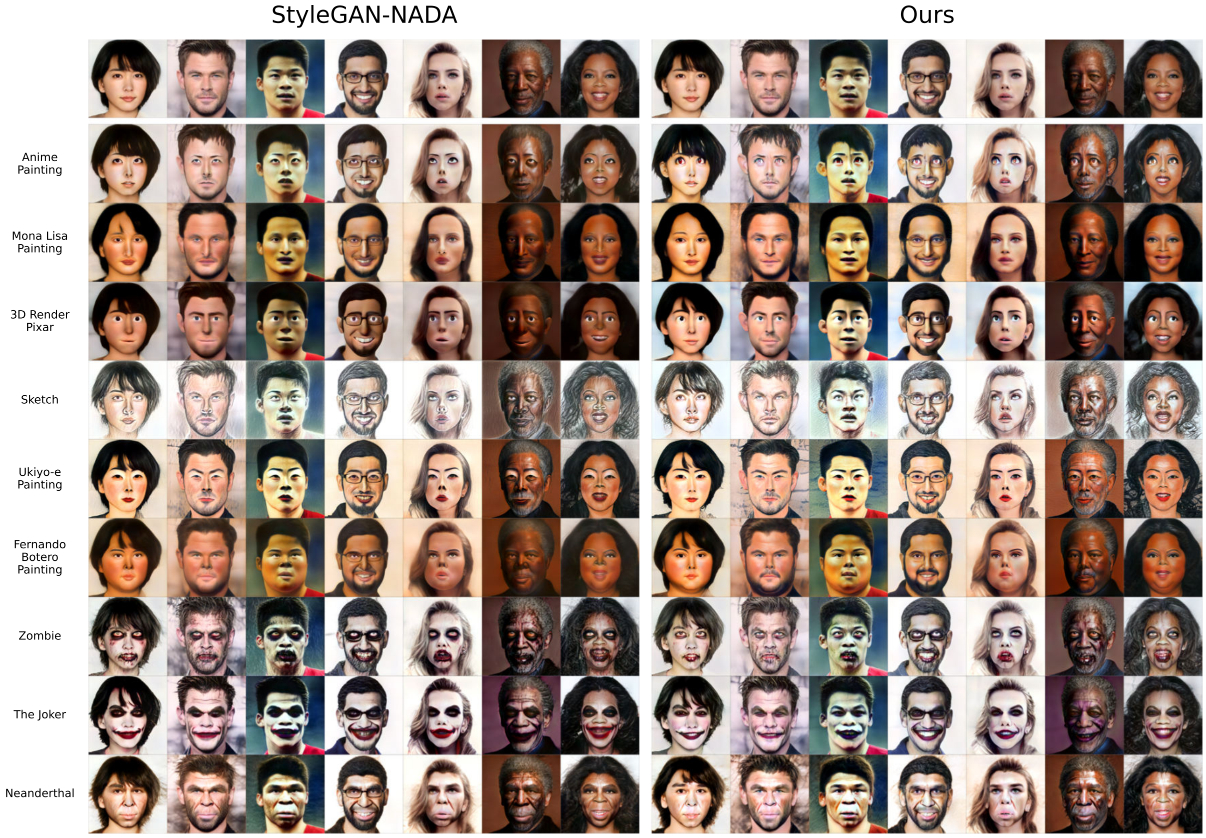


Figure 3: Comparison with the original StyleGAN-NADA [25] method (left) and its version with our parameterization.

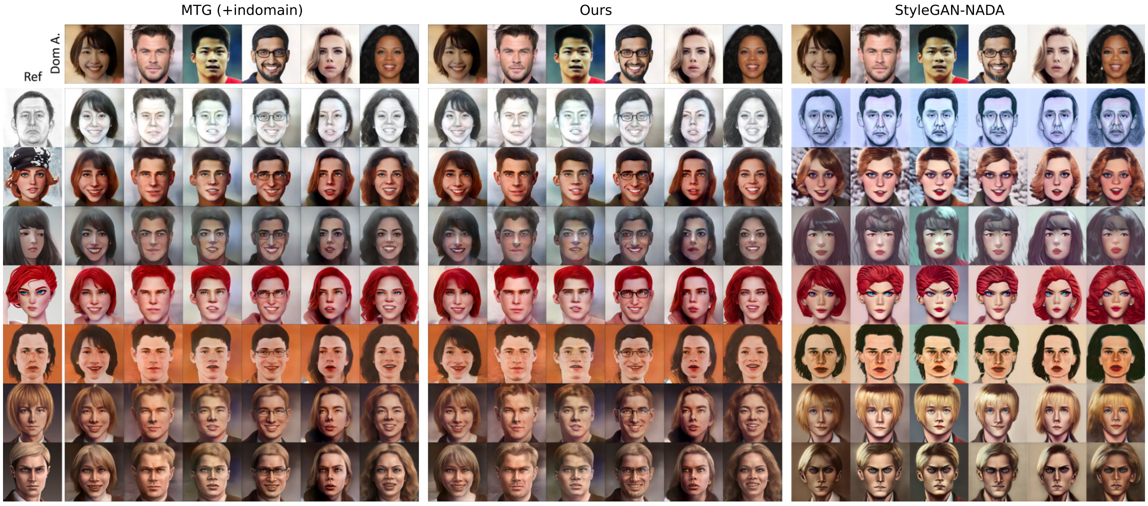


Figure 4: Comparison with one-shot domain adaptation methods. Left block is MindTheGap+indomain and right block is StyleGAN-NADA [26]. The middle block is the MindTheGap+indomain with our parameterization.

3.2 StyleDomain: Efficient and Lightweight Parameterizations of StyleGAN for One-shot and Few-shot Domain Adaptation

Recent advancements in Generative Adversarial Networks (GANs) [1, 2, 3, 31, 5], particularly StyleGAN models, have proven highly effective in various image synthesis applications, including image enhancement, editing, and translation. However, training StyleGAN models demands large, high-quality datasets, limiting their usefulness in domains with few images. Transfer learning, fine-tuning a pretrained model from one domain to another, is a common approach to tackle this issue.

Several domain adaptation methods for StyleGAN exist [4, 32, 23, 33, 34, 35, 24, 36, 15, 25, 26, 37], but most assume that adapting to a new domain requires fine-tuning most model weights, even for similar domains. This assumption lacks empirical validation, and little analysis has been conducted on which parts of StyleGAN are crucial for different data scenarios and domain similarities.

In this study, we conduct a systematic analysis to address this issue. Our investigation has two main parts. First, we identify what parts of StyleGAN need adaptation depending on the similarity between the source and target domains. We find that, for similar domains, fine-tuning only the affine layers is often sufficient. For more dissimilar domains, we need to optimize additional parameters, but not necessarily the entire network. This suggests the potential for more efficient and lightweight parameterizations of StyleGAN for domain adaptation.

Table 1: Evaluation of one-shot adaptation methods. Results for TargetCLIP, Cross-correspondence and StyleGAN-NADA methods are taken from [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

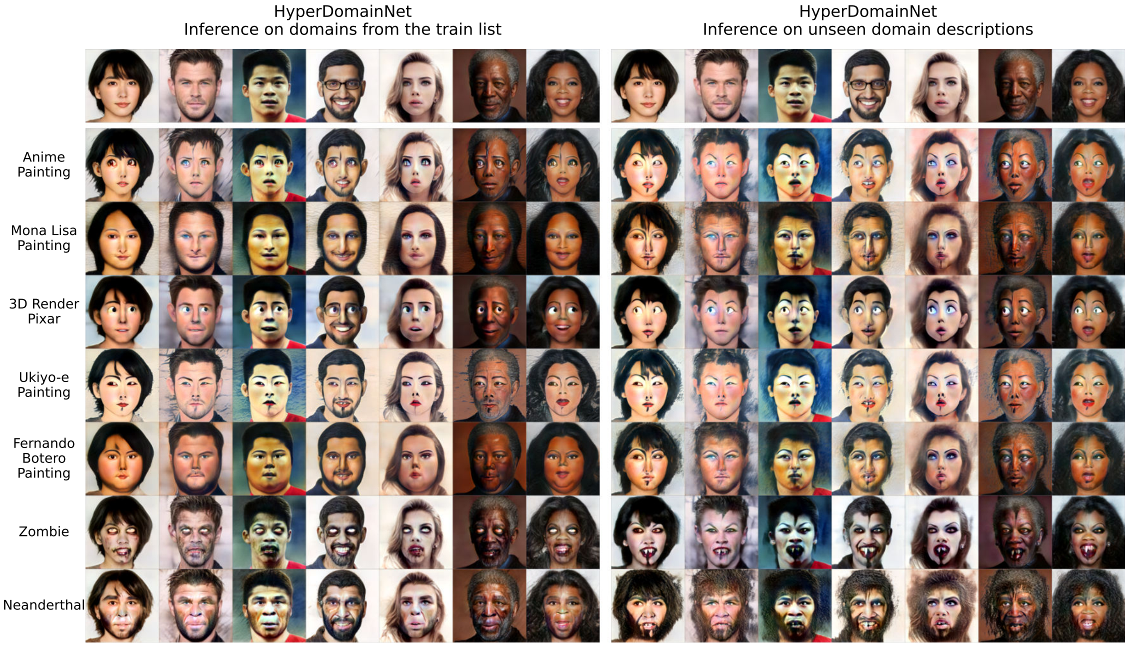


Figure 5: Comparison of training setups. The top row represents the real images embedded into StyleGAN2 latent space which latents are then used for HyperDomainNet inference. The left block represents results obtained from text-descriptions presented in the train list. The right block represents results of HyperDomainNet inference on unseen text-descriptions.

In the second part of our analysis, we propose two new parameterizations of StyleGAN. For similar domains, we introduce the concept of *StyleSpace*, where we can optimize directions to adapt to similar target domains without fine-tuning all StyleGAN weights. For more distant domains, we present the *Affine+* parameterization, which significantly reduces the number of trainable parameters while maintaining quality. Further improvements are made with the *AffineLight+* parameterization, which utilizes low-rank decomposition for affine layer weights. These parameterizations outperform complex baselines in few-shot adaptation for dissimilar domains.

Moreover, we explore the properties of *StyleDomain* directions, discovering their mixability and transferability. These directions can be combined to create entirely new styles or applied to StyleGAN models fine-tuned for other domains. We leverage these findings in various computer vision tasks, including image-to-image translation and cross-domain morphing.

Importance of Each Part of the StyleGAN

In this section, we assess the importance of various components of StyleGAN, particularly StyleGAN2, for domain adaptation. The source domain is FFHQ, and we explore different target domains. StyleGAN2 consists of three primary components:

- Mapping Network: It transforms input noise into an intermediate latent vector.
- Affine Layers: These layers map the latent vector to style vectors, which form the StyleSpace.
- Synthesis Network: Composed of modulated convolutions, it generates the output image from the input noise.

We provide a diagram description of the StyleGAN2 architecture in Figure 6.

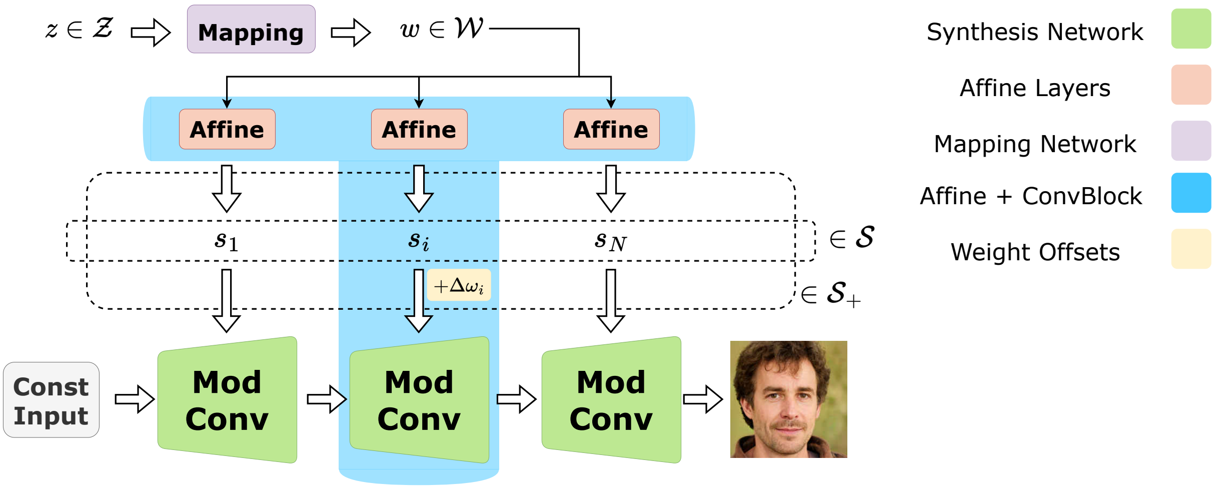


Figure 6: StyleGAN2 architecture. We introduce new latent space $S+$ for the for domain adaptation that combines StyleSpace and weight offsets for one block from the synthesis network.

The synthesis network has been traditionally considered the most critical for adaptation, while the mapping network and affine layers primarily handle semantic manipulations within the source domain. We aim to validate this assumption.

In our experiments, we also examine the combined impact of affine layers and a convolutional block from the synthesis network on domain adaptation, offering an intermediate analysis.

We propose a method to analyze each component's impact. While previous work reset the fine-tuned generator's component weights to their pretrained values, we suggest fine-tuning only one component to determine which is sufficient for domain adaptation.

The optimization objective for domain adaptation is to minimize the domain adaptation loss, $\mathcal{L}_D$, using generated samples from the generator $G_\theta(s(z))$. Typically, the generator is optimized with respect to all components:

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

We explore settings where we optimize with respect to only one component: *SyntConv* for synthesis network, *Affine* for affine layers, and *Mapping* for the mapping network. The full optimization of all components is termed *Full* parameterization.

We consider two domain adaptation settings: one-shot and few-shot. For each setting, we use different domains that vary in similarity to the source domain (FFHQ). One-shot domains maintain face geometry and identity while altering the style. Few-shot domains, on the other hand, change the face form, geometry, and identity more drastically. Different domain loss functions are applied depending on the data regime.

In the case of one-shot adaptation, we utilize Quality and Diversity metrics. For few-shot adaptation, we compute the FID metric. Further details about the domain adaptation loss functions can be found in the appendix.

Analysis for one-shot domains.

In our analysis, we explore text-based and one-shot image-based domains.

In our experiments, we examine four parameterizations: Full, SyntConv, Affine, and Mapping. Our qualitative results are presented in Figure Figure 7.

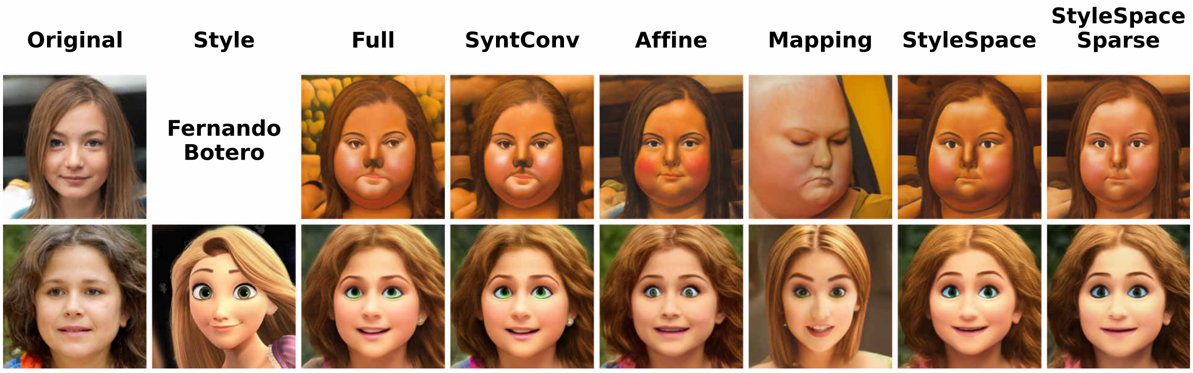


Figure 7: Text-based and image-based adaptation for different parameterizations. Affine, StyleSpace and StyleSpaceSparse parameterizations yield performance comparable with Full one. This style image is called "Disney".

We find that Full, SyntConv, and Affine parameterizations perform similarly in terms of visual quality and objective metrics. This aligns with prior research [37]. Surprisingly, the Affine parameterization alone is also effective, allowing us to change image domains without retraining the synthesis network. However, the Mapping network exhibits poor visual quality and limited diversity in generated images, emphasizing the importance of updating the style vector from $\mathcal{S}$ for successful adaptation rather than the intermediate latent vector from $\mathcal{W}$.

Analysis for few-shot domains.

In this study, we analyze two datasets, AFHQ Dogs and Cats [38]. Results are presented in Figure Figure 8 and Table Table 2. We find differences in results for Dogs and Cats compared to similar domains. Specifically, the Affine parameterization yields lower quality, evident in degraded visual output and increased FID metric. Surprisingly, even without fine-tuning, the adapted images exhibit reasonable visual quality. SyntConv matches Full parameterization in results, while Mapping yields consistently poor quality across all datasets.

StyleSpace and StyleSpaceSparse.

The study explores modifying the style vector in StyleSpace $\mathcal{S}$ to alter the generated image domain. The authors optimize the direction Δs^D during fine-tuning of StyleGAN2 to achieve this change. They call these optimized directions "StyleDomain" directions:

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

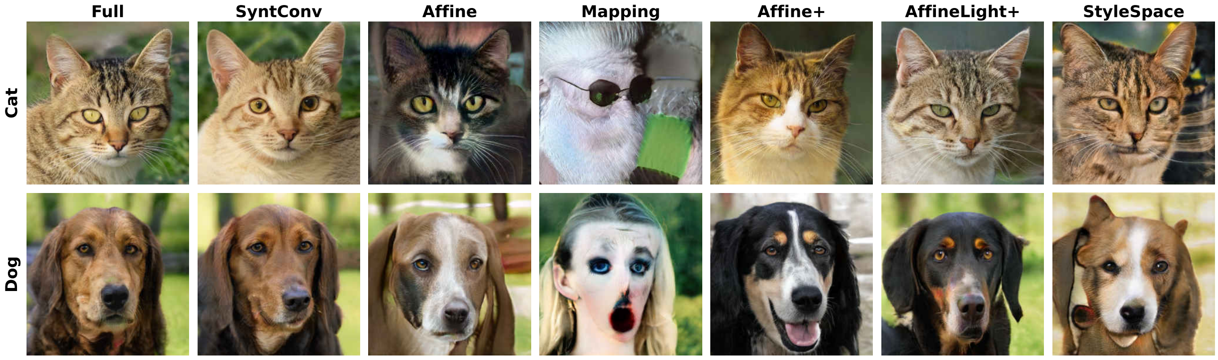


Figure 8: Domain adaptation for dissimilar domains. Affine+ parameterization produces results on par with the Full one.

where $\Delta s = (\Delta s_1, \ldots, \Delta s_N) \in \mathcal{S}$ is the optimized direction in the $\mathcal{S}$ for adapting the generator G_θ to the domain D.

They find that it's possible to prune most coordinates of StyleDomain directions without degrading quality. They use a standard pruning technique, retaining the top 20% of largest absolute values and setting the rest to zero, which they refer to as "StyleSpaceSparse."

The authors apply these parameterizations to one-shot and few-shot domains and make the following observations:

For one-shot adaptation, optimizing the StyleDomain direction achieves the same results as the Full parameterization, both visually and quantitatively. This allows generating samples from out-of-domain regions of realistic human faces.

For few-shot domains, StyleSpace is insufficient, resulting in significant quality degradation. They introduce a new parameterization suitable for more distant domains.

Affine+ and AffineLight+.

We propose improving Affine parameterization for domain adaptation in image synthesis, specifically for Dogs and Cats. We introduce a compact parameterization for certain layers by using offsets instead of fine-tuning all weights. The optimization objective is to minimize a loss function for this parameterization. We call this space "Affine+," which is chosen for a specific block in the synthesis network with a 64x64 resolution, as it performs

Table 2: FID scores for domain adaptation with different parameterizations. We observe a significant gap between Affine and Full parametrizations that, however, can be drastically reduced by introducing Affine+ parameterization.

		Domains	
Parameter Space	Size	Dog	Cat
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

the best. So, for this parameterization the optimization procedure has the following form:

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

where $G_{\Delta\theta_1, \Delta\theta_2}$ is the generator with weight offsets $\Delta\theta_1, \Delta\theta_2$ for the one block from the synthesis network.

Affine+ already has significantly fewer parameters than Full parameterization. We further reduce its size using low-rank decomposition and name it "AffineLight+." It has far fewer parameters while maintaining good quality, especially in low-data scenarios.

We apply these two parameterizations to few-shot domains and achieve promising results. For more details and results, refer to the provided figures and tables. Affine+ narrows the performance gap with the Full parameterization, indicating that style vectors help adapt the generator even to distant domains. AffineLight+ performs well with a much smaller parameter count, making it suitable for low-data situations.

Properties of the StyleDomain directions.

We explore two notable features of StyleDomain directions. First, they exhibit mixability, enabling the combination of directions corresponding to different similar domains, resulting in semantically mixed adaptation (see Figure 10 for examples).

Second, StyleDomain directions are transferable between different StyleGAN2 models. This is demonstrated by applying directions optimized for a base generator G_θ to adapt fine-tuned generators in various domains (e.g., Dogs, Cats) (see Figure 9 for results).

Results.

One-shot domain adaptation. This study explores one-shot domain adaptation for image-based tasks, using various baselines, such as TargetCLIP, JoJoGAN, MTG, GOSA, DiFa, and DomMod. StyleSpace and StyleSpaceSparse parameterizations are applied to the DiFa model, yielding improved performance. The experiments utilize StyleGAN2 with the source domain FFHQ, maintaining baseline configurations for fair comparison. A variety of style images serve as target domains. Quantitative and qualitative results are provided in Table 3 and in Figure 11, indicating that DiFa achieves the best Quality metric but lacks Diversity. The proposed parameterizations enhance performance across these metrics, outperforming other baselines. DomMod also performs well but is comparable to StyleSpaceSparse, which is more parameter-efficient. Notably, StyleSpaceSparse demands significantly less memory, which is crucial for scaling to numerous target domains. TargetCLIP, despite its limited trainable parameters, delivers poor visual and quality results. User studies are presented for a comprehensive evaluation.

Few-shot domain adaptation. In the context of few-shot domain adaptation, the study compares parameterizations (Affine+ and AffineLight+) applied to the vanilla StyleGAN-ADA with ADA, CDC, and AdAM baselines using the Dogs and Cats dataset.

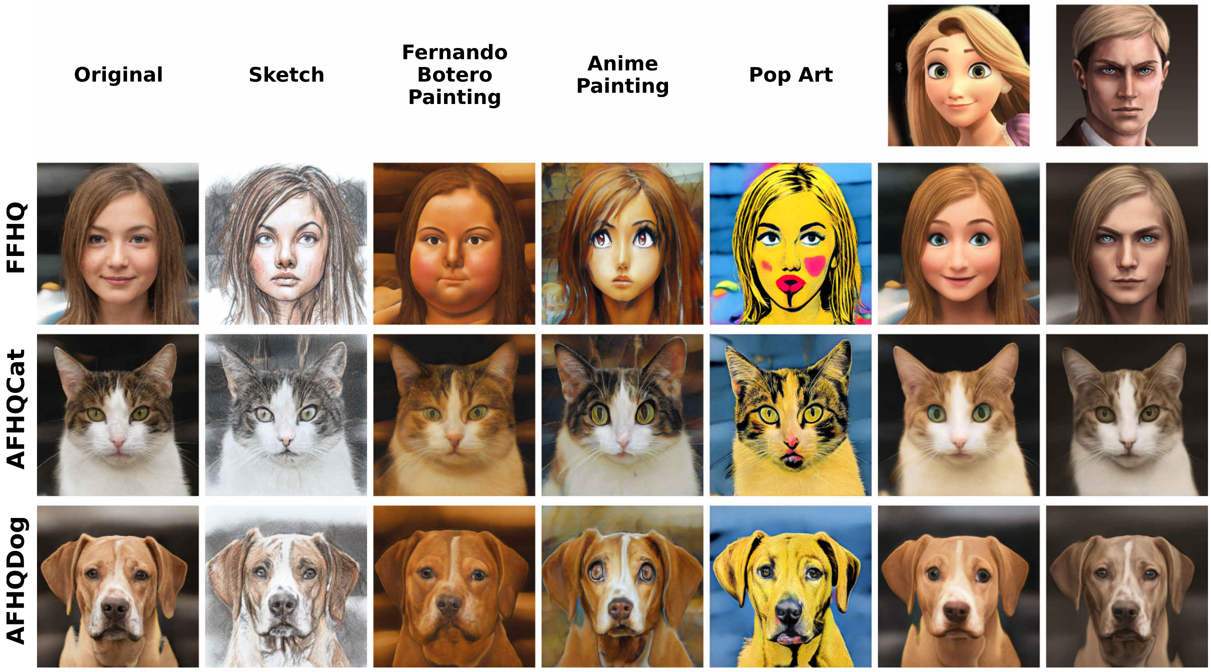


Figure 9: StyleSpace directions transfer from text-based and image-based domain adaptation to other fine-tuned models. We can successfully transfer style while preserving image content.

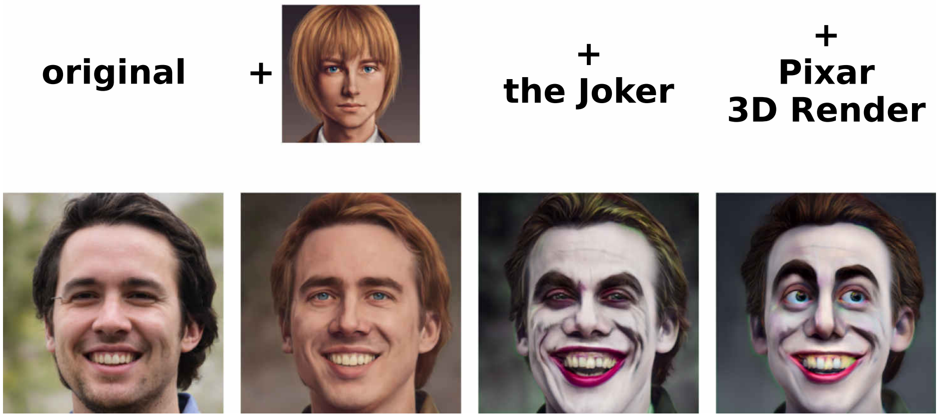


Figure 10: Example of mixing StyleDomain directions. We can combine different directions in order to perform adaptation into a semantically mixed domain.

Table 3: Quality and Diversity metrics [15] for one-shot image-based domain adaptations with different methods. Memory denotes the memory needed for keeping adapted generators for all 12 domains for each method. StyleSpace and StyleSpaceSparse parameterizations achieve results comparable to other baselines while having much less trainable parameters.

			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 ± 0.048	0.257 ± 0.025
MTG [26]	30M	1.80GB	0.607	0.269	0.509	0.234	0.586 ± 0.054	0.263 ± 0.028
GOSA [40]	30M	1.80GB	0.547	0.283	0.617	0.216	0.584 ± 0.034	0.252 ± 0.030
DiFa [41]	30M	1.80GB	0.719	0.226	0.699	0.263	0.734 ± 0.047	0.215 ± 0.038
TargetCLIP [30]	9.0K	420KB	0.474	0.306	0.502	0.333	0.491 ± 0.043	0.322 ± 0.015
DomMod (DiFa) [15]	6.0K	280KB	0.705	0.250	0.625	0.294	0.679 ± 0.049	0.253 ± 0.037
StyleSpace (DiFa)	6.0K	280KB	0.672	0.296	0.627	0.308	0.644 ± 0.041	0.298 ± 0.025
StyleSpaceSparse (DiFa)	**1.2K**	**56.4KB**	0.659	0.303	0.617	0.304	0.638 ± 0.046	0.299 ± 0.026

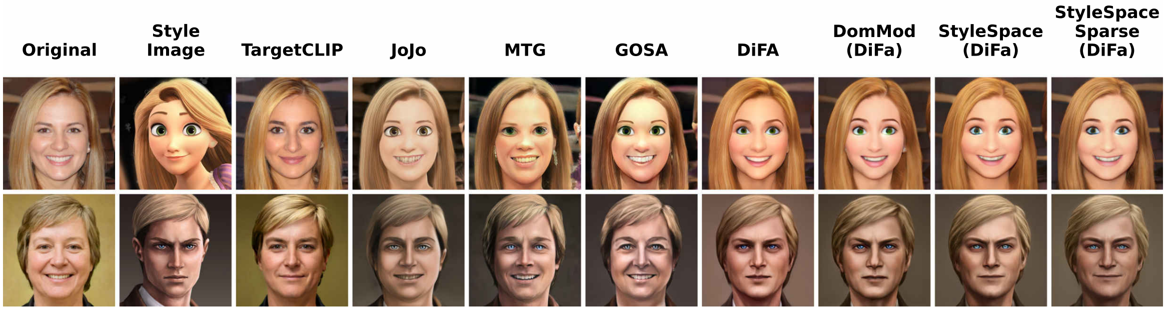


Figure 11: Comparison with baselines for one-shot image-based domain adaptation. StyleSpace and StyleSpaceSparse parameterizations achieve comparable quality as other methods while having much less trainable parameters.

The efficiency of these methods is assessed with varying numbers of target samples, and rigorous training setups are followed. The results are provided in Figure 12 and in Table 4. Notably, training iterations are increased to 50K for all methods to prevent underfitting. Results show that AdAM's performance is not superior to the vanilla ADA when sufficiently trained. ADA (Affine+) and ADA (AffineLight+), enhanced with the proposed parameterizations, consistently outperform other methods across different numbers of shots, especially in low-data scenarios.

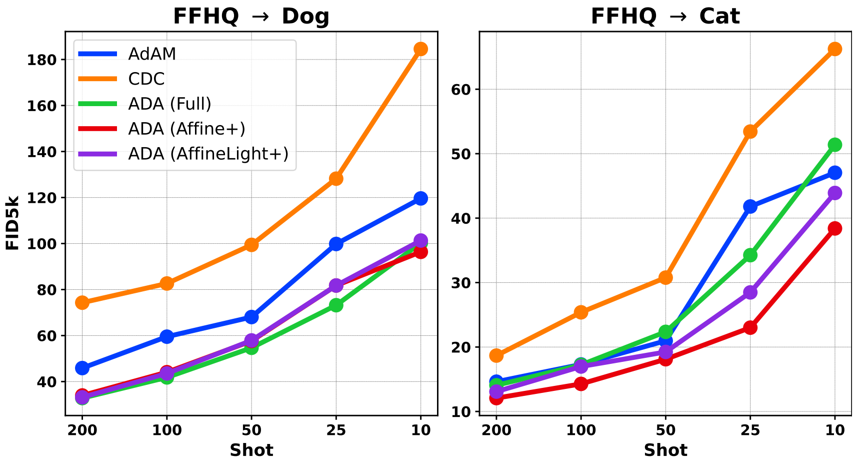


Figure 12: Few-shot training results for different number of shots. Proposed ADA (Affine+) and ADA (AffineLight+) show uniformly better performance than baselines.

Table 4: Results for few-shot training with 10-shots. Proposed ADA (Affine+) and ADA (AffineLight+) achieve better performance.

		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

3.3 HiFi++: a Unified Framework for Bandwidth Extension and Speech Enhancement

The issue of conditional speech generation holds significant practical importance, encompassing applications such as neural vocoding, bandwidth extension (BWE), speech enhancement (SE), and more. A recent breakthrough in this field leverages generative adversarial networks (GANs) [13, 14]. Specifically, it has been shown that GAN-based vocoders outperform publicly available neural vocoders in speech quality and speed.

In this study, we adapt the HiFi model [14] for bandwidth extension and speech enhancement tasks by introducing a novel HiFi++ generator architecture. This architecture incorporates new components, including spectral preprocessing (SpectralUnet), a convolutional encoder-decoder network (WaveUNet), and learnable spectral masking (Spectral-

MaskNet). These enhancements enable our generator to effectively address bandwidth extension and speech enhancement challenges.

Our extensive experiments reveal that our model performs competitively with state-of-the-art solutions in bandwidth extension and speech enhancement, all while being notably more lightweight and maintaining superior or equivalent quality.

Adapting HiFi-GAN Generator For Bandwidth Extension and Speech Enhancement.

This paper introduces the HiFi++ architecture, which extends the HiFi generator to address the SE and BWE problems by incorporating three novel modules: SpectralUNet, WaveUNet, and SpectralMaskNet (see Figure 13). The HiFi++ generator is based on the HiFi-GAN generator's V2 version, taking an enriched mel-spectrogram as input from SpectralUNet and undergoing post-processing via WaveUNet and SpectralMaskNet. Reordering these post-processing modules did not yield significant improvements.

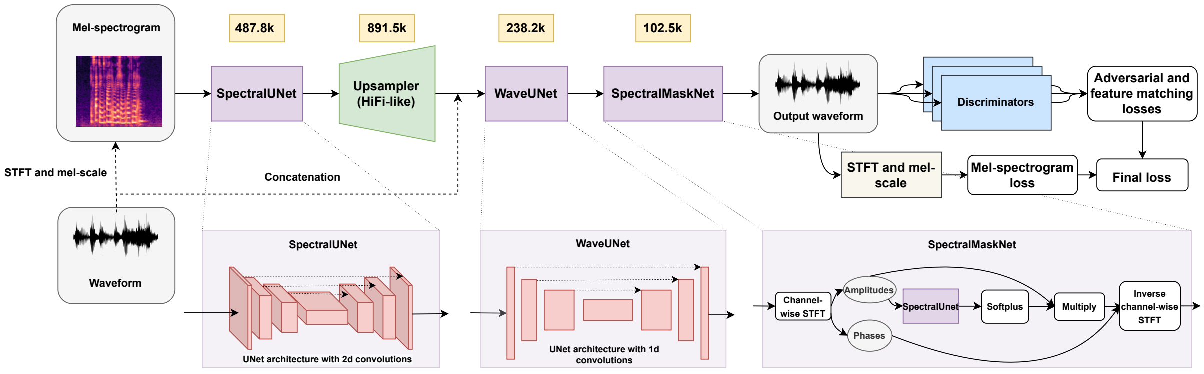


Figure 13: HiFi++ architecture and training pipeline. The HiFi++ generator consists of the HiFi-like Upsampler and three introduced modules SpectralUNet, WaveUNet and SpectralMaskNet (their sizes are in yellow boxes). The generator's architecture is identical for BWE and SE.

SpectralUNet: The SpectralUNet module serves as the initial stage of the HiFi++ generator. It enhances the resolution of the mel-spectrogram, a 2D representation of the raw waveform, to simplify the subsequent transformation into a 1D sequence. This UNet-like architecture employs 2D convolutions and acts as a preprocessing step, extracting essential information for the target task, particularly beneficial for bandwidth extension and speech enhancement.

WaveUNet: Positioned after the HiFi's Upsampler, the WaveUNet module takes multiple 1D sequences concatenated with the input waveform. It operates in the time domain, enhancing the Upsampler's output and merging the predicted waveform with the source. WaveUNet adopts the Wave-U-Net architecture, a fully convolutional 1D-UNet-like network, resulting in a 2D tensor of m 1D sequences, which are further processed by SpectralMaskNet.

SpectralMaskNet: As the final part of the generator, SpectralMaskNet introduces learnable spectral masking. It takes the 2D tensor of m 1D sequences, applies channel-wise short-time Fourier transform (STFT), and predicts multiplicative factors for the amplitudes, followed by inverse STFT to modify the spectrum. This frequency-domain post-processing mechanism effectively removes artifacts and noise from the output waveform in a learnable manner.

Training Objective: The paper employs a multi-discriminator adversarial training framework, inspired by [14] work. Instead of using multi-period and multi-scale discriminators, we utilize several identical discriminators operating on the same resolutions with fewer weights. The losses used in training include LS-GAN loss, feature matching loss, and mel-spectrogram loss:

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

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

The total loss for the generator incorporates these losses with associated weights, while each discriminator is optimized individually. Experiments set the weights as follows: $\lambda_{fm} = 2$, $\lambda_{mel} = 45$, and the number of discriminators $k = 3$.

Results.

Bandwidth Extension: We utilized the VCTK dataset, comprising 44,200 speech recordings from 110 speakers, for bandwidth extension experiments. Six speakers were excluded from the training set to prevent data leakage. The evaluation used 48 utterances from these excluded speakers. Importantly, the text in the evaluation utterances was not present in the training data.

Speech Denoising: Our denoising experiments employed the VCTK-DEMAND dataset, featuring 11,572 training utterances with various signal-to-noise ratios (SNR) and 824 test utterances. Further details can be found in the original paper.

Evaluation

Objective Evaluation: We assessed speech enhancement using conventional metrics such as WB-PESQ, STOI, SI-SDR, and DNSMOS. Additionally, we introduced WV-MOS, a direct MOS score prediction based on fine-tuned wave2vec2.0, which exhibited better correlation with subjective quality measures.

Subjective Evaluation: Subjective quality assessment was conducted using 5-scale MOS tests, with audio clips normalized to account for volume differences. English-speaking referees with suitable listening equipment participated.

Bandwidth Extension

In bandwidth extension experiments, we trained models independently for three input frequency bandwidths (1 kHz, 2 kHz, and 4 kHz). As we observe in Table 5, HiFi++ outperformed other techniques in terms of model size and quality of bandwidth extension, being five times smaller than the closest baseline, SEANet. Pair-wise comparisons confirmed HiFi++'s statistical dominance over SEANet.

These results underscore the importance of adversarial objectives in speech frequency bandwidth extension models. Notably, SEANet, which also employs adversarial objectives, emerged as the strongest baseline among the examined models, leaving supervised reconstruction models like TFilm and 2S-BWE far behind, particularly for low input frequency bandwidths.

Table 5: Bandwidth extension results on VCTK dataset. * indicates re-implementation.

	BWE (1kHz)				**BWE (2kHz)**				**BWE (4kHz)**					
Model	MOS	WV-MOS	STOI	PESQ	MOS	WV-MOS	STOI	PESQ	MOS	WV-MOS	STOI	PESQ	# Param (M)	# MACs (G)
Ground truth	4.62 ± 0.06	4.17	1.00	4.64	4.63 ± 0.03	4.17	1.00	4.64	4.50 ± 0.04	4.17	1.00	4.64	-	-
HiFi++ (ours)	$\mathbf{4.10 \pm 0.05}$	**3.71**	0.86	1.74	$\mathbf{4.44 \pm 0.02}$	**3.95**	0.94	2.54	$\mathbf{4.51 \pm 0.02}$	4.16	1.00	3.74	**1.7**	**2.8**
SEANet	3.94 ± 0.09	3.66	0.82	1.54	$\mathbf{4.43 \pm 0.05}$	**3.95**	0.93	2.43	4.45 ± 0.04	**4.17**	0.99	3.65	9.2	4.5
VoiceFixer	3.04 ± 0.08	3.21	0.73	1.44	3.82 ± 0.06	3.50	0.78	1.73	4.34 ± 0.03	3.77	0.83	2.38	122.1	34.4
TFiLM	1.98 ± 0.02	1.65	0.81	2.11	2.67 ± 0.04	2.27	0.91	2.63	3.54 ± 0.04	3.49	1.00	3.52	68.2	-
input	1.87 ± 0.08	0.39	0.78	2.60	2.46 ± 0.04	1.74	0.88	3.04	3.36 ± 0.06	3.17	0.99	3.65	-	-

Speech Enhancement

We see in Table 6 that comparing HiFi++ with baselines in speech enhancement, our model achieved similar performance to state-of-the-art models like VoiceFixer and DB-

AIAT while being significantly more computationally efficient. VoiceFixer excelled in subjective quality despite lagging behind in objective metrics, primarily due to its utilization of mel-spectrograms rather than raw signal waveforms. HiFi++, which uses the signal spectrum, outperformed SEANet, a model with a similar adversarial approach but lacking spectral information.

An intriguing observation was the performance of the MetriGAN+ model, explicitly trained to optimize PESQ, but not translating this success to other objective and subjective metrics.

Table 6: Speech denoising results on Voicebank-DEMAND dataset. * indicates re-implementation.

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	DNSMOS	# Par (M)	# MACs (G)
Ground truth	4.46 ± 0.05	4.50	-	1.00	4.64	3.15	-	-
DB-AIAT	$\mathbf{4.40 \pm 0.05}$	**4.38**	**19.4**	**0.96**	**3.27**	**3.18**	2.8	41.8
HiFi++ (ours)	4.31 ± 0.05	4.36	17.9	0.95	2.90	3.10	**1.7**	**2.8**
VoiceFixer	4.21 ± 0.06	4.14	-18.5	0.89	2.38	3.13	122.1	34.4
DEMUCS	4.17 ± 0.06	4.37	18.5	0.95	3.03	3.14	60.8	38.1
MetricGAN+	3.98 ± 0.06	3.90	8.5	0.93	3.13	2.95	2.7	28.5
Input	3.45 ± 0.07	2.99	8.4	0.92	1.97	2.53	-	-

3.4 FFC-SE: Fast Fourier Convolution for Speech Enhancement

Speech enhancement plays a crucial role in telecommunication and has garnered significant attention within the audio processing community. Traditional signal processing methods have addressed this challenge but often rely on specific noise models. In recent years, data-driven approaches, leveraging deep learning, have emerged as dominant in modern speech enhancement.

A prevalent approach in deep learning for speech enhancement involves time domain signal retrieval, employing a convolutional encoder-decoder (CED) structure. Notable works, such as [43] and [44], utilize adversarial training and CED networks. Some also incorporate neural components, such as long short-term memory cells [45] and transformers [46]. These methods directly map noisy waveforms to clean signals, but often neglect information about the signal spectrum, leading to potential inefficiencies. A recent endeavor seeks to explicitly incorporate spectral information during generation, yielding state-of-the-art results.

Another research strand focuses on short-time Fourier transform (STFT) representations. Approaches in this category aim to predict clean signal STFT coefficients directly or correct noisy signal spectra using masks for magnitude or both magnitude and phase modification [47]. Papers like MetricGAN and MetricGAN+ [48] employ Bidirectional LSTM for predicting binary masks and report state-of-the-art results in speech quality metrics. Estimating phases directly poses a challenge, leading to various strategies, including decoupling magnitude and phase estimation [49] and employing separate vocoder networks for waveform synthesis. These methods often necessitate large neural networks and significant computational resources. To enhance phase prediction, we introduce non-local neural operators, which reduce model size while improving quality.

We propose novel neural architectures based on the fast Fourier convolution (FFC) operator, originally designed for computer vision tasks. FFC's global receptive field is advantageous for complex spectrum prediction, particularly for handling periodic structures in spectrograms. Our experiments reveal that FFC's large receptive field aids in producing coherent phases. Leveraging these insights, we design new neural architectures for direct complex-valued spectrogram estimation in speech enhancement. These models achieve state-of-the-art performance on VoiceBank-DEMAND [50] and Deep Noise Suppression datasets with significantly fewer parameters than baseline methods.

Proposed Approach

We address the single-channel speech denoising problem, aiming to map noisy waveform $y = x + n$ with additive noise n to the clean signal x. Our strategy involves neural architectures enhanced with a non-local neural operator called fast Fourier convolution (FFC) [51], which we adapt for complex spectrum processing. We present two neural architectures that incorporate this operator as a fundamental component.

Fast Fourier Convolution (FFC)

Fast Fourier convolution (FFC) is a neural operator enabling non-local reasoning within a neural network. FFC applies channel-wise fast Fourier transform, followed by point-wise convolution and inverse Fourier transform, globally influencing the input tensor across dimensions involved in the Fourier transform. FFC divides channels into local and global branches:

1. The local branch employs conventional convolutions for local feature map updates.
2. The global branch conducts a Fourier transform of the feature map in the spectral domain, affecting the global context.

In our work, we perform the Fourier transform solely across the frequency dimensions of feature maps, corresponding to Short-Time Fourier Transform (STFT) representations, in contrast to computer vision tasks where the Fourier transform spans both image dimensions [51, 52]. The global branch of the FFC layer consists of three steps:

1. Real fast Fourier transform across the frequency dimension of the input feature map, followed by the concatenation of real and imaginary parts of the spectrum across the channel dimension:
$$\mathbb{R}^{C\times F\times T} \overset{\text{fft1d}}{\longrightarrow} \mathbb{C}^{C\times F/2\times T} \overset{\text{concat}}{\longrightarrow} \mathbb{R}^{2C\times F/2\times T}. \tag{8}$$
2. Application of a convolutional block with a 1×1 kernel in the frequency domain:
$$\mathbb{R}^{2C\times F/2\times T} \overset{\text{conv}-\text{bn}-\text{relu}}{\longrightarrow} \mathbb{R}^{2C\times F/2\times T}. \tag{9}$$
3. Inverse Fourier transform:
$$\mathbb{R}^{2C\times F/2\times T} \overset{\text{concat}}{\longrightarrow} \mathbb{C}^{C\times F/2\times T} \overset{\text{ifft1d}}{\longrightarrow} \mathbb{R}^{C\times F\times T}. \tag{10}$$

Here, C, F, and T represent the number of channels, the dimension corresponding to frequency, and the dimension corresponding to time, respectively. The global and local branches interact through activation summation. You can see the overall diagram of this module in Figure 14.

We employ a variation of FFC from [52] for image inpainting, utilizing one-dimensional Fourier transform across the frequency dimension.

FFC-AE

For speech enhancement, we implement two neural network architectures. The first, FFC-AE, is inspired by [52]. FFC-AE comprises a convolutional encoder that downsamples the input STFT representation across time and frequency dimensions by a factor of two. Residual blocks follow the encoder, each composed of two sequential fast Fourier convolution modules. The output of these blocks is upsampled by transposed convolution and used to predict the real and imaginary parts of the denoised complex-valued spectrogram. The architecture is depicted in Figure 15 (left), and we refer to it as the fast Fourier convolutional autoencoder (FFC-AE).

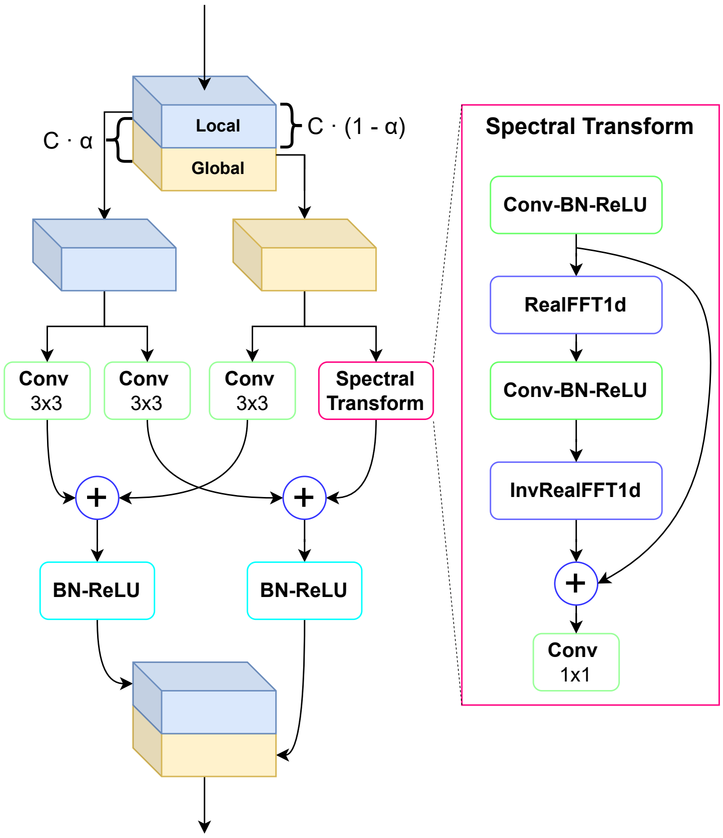


Figure 14: Fast Fourier Convolution neural module for speech enhancement. Parameter $\alpha \in [0, 1]$ controls the ratio of channels used in the global branch of the module.

We found that a downsampling factor of 2 strikes a suitable balance between performance and computational complexity for STFT with a window size of 1024 and a hop length of 256.

FFC-UNet

The second architecture draws inspiration from the classic U-Net model [53]. We incorporate FFC layers into the U-Net architecture, as shown in Figure 15 (right). At each level of the U-Net structure, we integrate several residual FFC blocks with convolutional upsampling or downsampling. We adapt the parameter α, representing the ratio of channels going to the global branch of fast Fourier convolution, based on the U-Net level. Higher U-Net levels work with higher-resolution data, rich in periodic structures, while lower levels operate at a coarser scale lacking such periodic structures. We decrease α from 0.75 at the topmost level to 0 at the bottom layer in steps of 0.25.

Training

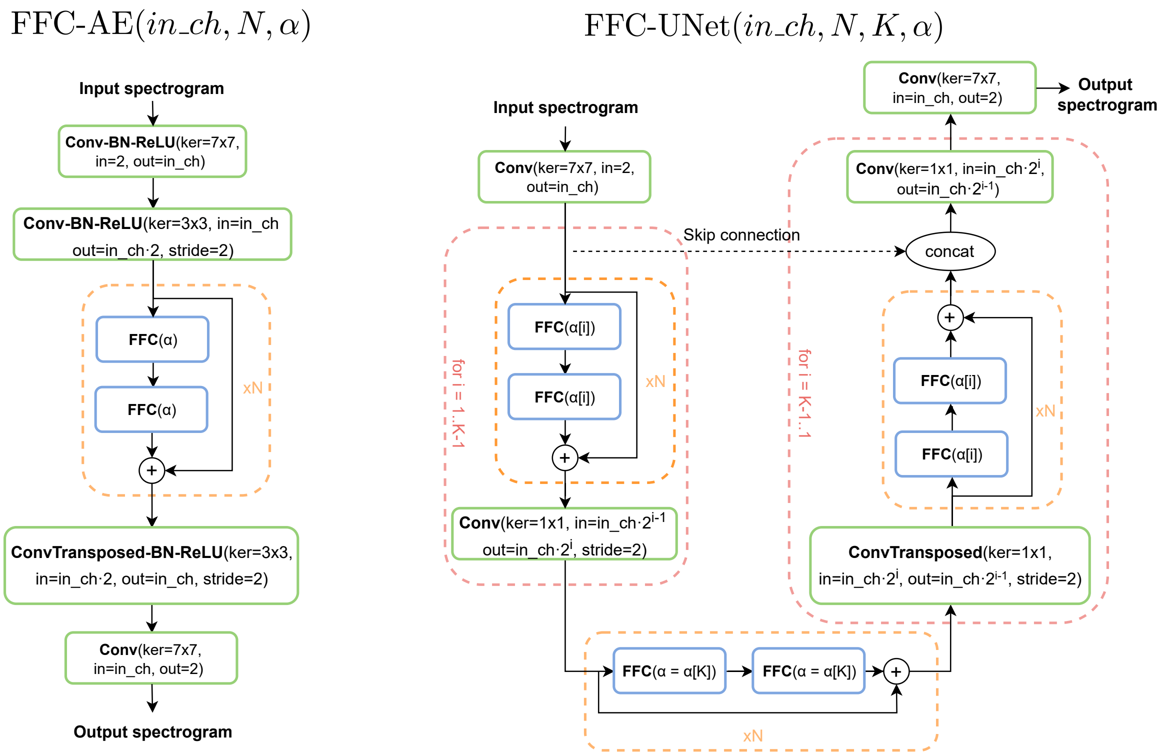

FFC-AE(in_ch, N, α)
Input spectrogram
Conv-BN-ReLU(ker=7x7, in=2, out=in_ch)
Conv-BN-ReLU(ker=3x3, in=in_ch out=in_ch·2, stride=2)
FFC(α)
FFC(α)
xN
+
ConvTransposed-BN-ReLU(ker=3x3, in=in_ch·2, out=in_ch, stride=2)
Conv(ker=7x7, in=in_ch, out=2)
Output spectrogram
FFC-UNet(in_ch, N, K, α)
Input spectrogram
Conv(ker=7x7, in=2, out=in_ch)
for i = 1..K-1
FFC(α[i])
FFC(α[i])
xN
+
Conv(ker=1x1, in=in_ch·2^(i-1) out=in_ch·2^i, stride=2)
Skip connection
FFC(α = α[K])
FFC(α = α[K])
+
xN
ConvTransposed(ker=1x1, in=in_ch·2^i, out=in_ch·2^(i-1), stride=2)
FFC(α[i])
FFC(α[i])
+
xN
for i = K-1..1
concat
Conv(ker=1x1, in=in_ch·2^i, out=in_ch·2^(i-1))
Conv(ker=7x7, in=in_ch, out=2)
Output spectrogram

Figure 15: Proposed architectures for speech enhancement. *Left:* fast Fourier convolutional autoencoder which adopts architecture introduced in [52] for speech enhancement task. *Right:* fast Fourier convolutional U-Net. Parameter in_ch controls the overall width of the networks, N defines the number of FFC residual blocks, K is the depth of the FFC-UNet architecture, α (real number $\in [0, 1]$ in case of FFC-AE, K numbers $\in [0, 1]$ in case of FFC-Unet) controls the proportion of channels going to the global branch.

To convert the predicted STFT representation into a waveform, we use inverse short-time Fourier transform. Training follows a multi-discriminator adversarial. It includes three losses: LS-GAN loss $\mathcal{L}_{GAN}$, feature matching loss $\mathcal{L}_{FM}$, and mel-spectrogram loss $\mathcal{L}_{Mel}$. The losses are as follows:

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

Here, $\mathcal{L}(\theta)$ denotes the loss for the generator with parameters θ, and $\mathcal{L}(\varphi_i)$ denotes the loss for the i-th discriminator with parameters φ_i. All discriminators are identical but initialized differently. In all experiments, we set $\lambda_{fm} = 2$, $\lambda_{mel} = 45$, and $k = 3$.

Results

Datasets We evaluated the effectiveness of speech denoising models using two benchmarks. The audio recordings were sampled at 16 kHz.

The first benchmark is the VoiceBank-DEMAND dataset [50], consisting of a train set with 28 speakers and 11572 utterances at 4 signal-to-noise ratios (SNR) (15, 10, 5, and 0 dB). The test set (824 utterances) features 2 speakers not seen during training, with 4 SNR levels (17.5, 12.5, 7.5, and 2.5 dB).

The second benchmark is the Deep Noise Suppression (DNS) challenge, where we synthesized 100 hours of training data without artificial reverberation. The models were tested on two test sets: DNS-INDOMAIN (hold-out data from the 100-hour training set) and DNS-BLIND (real-world noisy recordings).

Metrics For objective evaluation, we used standard metrics, including WB-PESQ [54], extended STOI, SI-SDR [55], COVL, CBAK, and CSIG. Additionally, we considered an objective speech quality measure (WV-MOS) based on direct MOS score prediction using a fine-tuned wav2vec2.0 model, which showed strong correlation with subjective quality measures.

For subjective quality evaluation, we conducted 5-scale MOS tests. Audio clips were normalized, and the referees were English speakers with proper listening equipment.

Experimental Setup In our experiments, signals were transformed to the spectral domain using STFT with a Hann window of size 1024 and a hop size of 256. We used specific parameter values for different model versions. FFC-AE employed $\alpha = 0.75$, $N = 9$, $in\_ch = 32$ for V0, and $in\_ch = 64$ for V1. FFC-UNet used $K = 4$, $N = 4$, $in\_ch = 32$, and a gradual decrease of α. All models were trained for 800,000 iterations with a batch size of 8 and an Adam optimizer with a learning rate of 0.0002. ResUNet-Decouple+ followed the same training setup as reported in the original paper, with 800,000 iterations and a learning rate of 0.0002.

Experimental Results We compared our proposed models with various baselines from the literature, including FullSubNet and DEMUCS, as well as models like vanilla U-Net and FFC-AE (abl.). The comparison was conducted on both benchmarks.

For VoiceBank-DEMAND (see Table 7), our models achieved significantly better MOS scores than all the baselines and performed competitively in terms of objective metrics. On the DNS benchmark (see Table 8), our models exhibited superior quality compared to competitors on the DNS-INDOMAIN test set and competitive performance with FullSubNet on the DNS-BLIND test set, which is a notable achievement considering FullSubNet's high ranking in the DNS Challenge 2021.

Notably, our models achieved these results without employing dynamic data synthesis, reverberation simulation, or augmentation techniques, which some of the closest baselines relied on. Further improvements in generalization to the blind test set can be explored with advanced data generation pipelines.

Table 7: Speech denoising results on Voicebank-DEMAND dataset. Best three results are highlighted in bold.

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	CSIG	CBAK	COVL	# Params (M)	# GMAC on 16k
Ground Truth	4.46 ± 0.06	4.50	-	1.00	4.64	5.0	5.0	5.0	-	-
Input	3.44 ± 0.06	2.99	8.4	0.79	1.97	3.34	2.82	2.74	-	-
MetricGAN+ [48]	3.82 ± 0.06	3.90	8.5	0.83	**3.13**	4.12	3.16	3.62	2.7	-
ResUNet-Decouple+ [49]	3.94 ± 0.04	4.13	**18.4**	0.84	2.45	3.38	3.15	2.89	102.6	-
DEMUCS (non-caus.) [45]	4.06 ± 0.03	**4.37**	**18.5**	**0.87**	**3.03**	**4.36**	**3.51**	**3.72**	60.8	-
VoiceFixer [56]	4.10 ± 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 ± 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)	**4.24 ± 0.09**	4.34	17.9	0.86	2.88	4.25	3.40	3.57	**0.42**	4.39
FFC-AE-V1 (ours)	**4.33 ± 0.03**	**4.37**	17.5	**0.87**	2.96	**4.34**	**3.42**	**3.66**	**1.7**	16.33
FFC-UNet (ours)	**4.28 ± 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 ± 0.07	4.05	16.7	0.84	2.68	3.94	3.23	3.31	2.9	2.25
vanilla UNet	4.10 ± 0.07	4.11	17.2	0.85	2.73	3.94	3.28	3.34	20.7	11.2(x2)

Table 8: Speech denoising results on DNS dataset. * indicates results on DNS-BLIND. Best three results are highlighted in bold.

Model	MOS	MOS*	WV-MOS	WV-MOS*	SI-SDR	STOI	PESQ	CSIG	CBAK	COVL	# Params (M)	# GMAC on 16k
Ground Truth	4.40 ± 0.08	-	3.845	-	-	1.00	4.64	5.0	5.0	5.0	-	-
Input	2.75 ± 0.07	2.43 ± 0.08	1.195	0.80	-	0.69	1.49	2.59	2.32	1.99	-	-
DEMUCS [45]	3.52 ± 0.15	2.94 ± 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 ± 0.08	2.75 ± 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 ± 0.04	2.51 ± 0.08	2.94	1.86	14.78	0.81	2.09	2.82	3.06	2.43	102.6	-
FullSubNet [?]	3.73 ± 0.02	**3.08 ± 0.09**	2.90	2.41	**14.96**	0.82	2.43	3.59	**3.27**	3.0	5.6	-
FFC-AE-V0 (ours)	**3.92 ± 0.09**	2.88 ± 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)	**4.02 ± 0.05**	**3.10 ± 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)	**4.00 ± 0.06**	**3.11 ± 0.08**	**3.35**	**2.70**	**15.48**	**0.86**	**2.69**	**4.08**	**3.44**	**3.41**	7.7	19.81

4 Conclusion

This section presents a summary of the key contributions of our work. The main results of the work are efficient parameterizations and architecture modules for GAN generators for solving domain adaptation problem in computer vision and speech enhancement in signal processing.

1. In *HyperDomainNet* we proposed a new StyleGAN parametrization for domain adaptation, which has only 6 thousand trained parameters compared to 30 million weights in the conventional full parametrization. This parametrization is based on domain modulation technique, which allows efficient modification of the generator weights using a small training vector. In a series of extensive experiments for text-based and image-based domain adaptation, we have shown that this parametrization achieves the same quality as current methods that use the full parametrization of the StyleGAN generator. We also proposed a new HyperDomainNet that solves the problem of multi-domain adaptation. The idea is that from a textual description of a domain or an example of a domain picture, the hypernet predicts a domain vector that the generator adapts using domain modulation technique. This makes it possible to adapt on hundreds or thousands of new domains at once, without having to retrain the generator on each domain individually. In experiments, we have shown that HyperDomainNet allows adapting the generator to new domains in the same way as conventional methods that work in single domain adaptation. Additionally, this model showed promising generalisation results for unseen domains.

2. In *StyleDomain* we conduct a systematic analysis to address the adaptation of StyleGAN across domains. Our investigation unfolds in two parts: initially, we pinpoint which parts of StyleGAN require adaptation based on the similarity between source and target domains. For similar domains, fine-tuning only the affine layers often suffices, while more dissimilar domains necessitate optimizing additional parameters, though not the entire network, indicating the potential for more efficient parameterizations. In the second part, we introduce two novel parameterizations: for similar domains, we propose *StyleSpace*, which optimizes adaptation directions without fine-tuning all weights, and for more distant domains, we present *Affine+*, reducing trainable parameters significantly while maintaining quality. Further refinement with

AffineLight+ employs low-rank decomposition for affine layer weights, outperforming complex baselines in few-shot adaptation. Additionally, we explore the properties of *StyleDomain* directions, revealing their mixability and transferability, which can create new styles or be applied to other fine-tuned StyleGAN models. These findings are leveraged in various computer vision tasks, such as image-to-image translation and cross-domain morphing.

3. In the *HiFi++* paper, we introduce a novel HiFi++ generator architecture for bandwidth extension and speech enhancement tasks. This architecture incorporates new components, including spectral preprocessing (SpectralUnet), a convolutional encoder-decoder network (WaveUNet), and learnable spectral masking (SpectralMaskNet), enabling our generator to effectively address these challenges. Extensive experiments reveal that our model performs competitively with state-of-the-art solutions in BWE and SE, while being notably more lightweight and maintaining superior or equivalent quality. Additionally, in *FFC-SE* work we propose novel neural architectures based on the fast Fourier convolution (FFC) operator, originally designed for computer vision tasks. FFC's global receptive field is advantageous for complex spectrum prediction, particularly for handling periodic structures in spectrograms, aiding in producing coherent phases. Leveraging these insights, we design new neural architectures for direct complex-valued spectrogram estimation in speech enhancement, achieving state-of-the-art performance on VoiceBank-DEMAND and Deep Noise Suppression datasets with significantly fewer parameters than baseline methods.

References

[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.