HyperGAN-CLIP: A Unified Framework for Domain Adaptation, Image Synthesis and Manipulation

ABDUL BASIT ANEES, Koç University, Turkey
AHMET CANBERK BAYKAL, University of Cambridge, United Kingdom
MUHAMMED BURAK KIZIL, Koç University, Turkey
DUYGU CEYLAN, Adobe Research, United Kingdom
ERKUT ERDEM, Hacettepe University, Turkey
AYKUT ERDEM, Koç University, Turkey

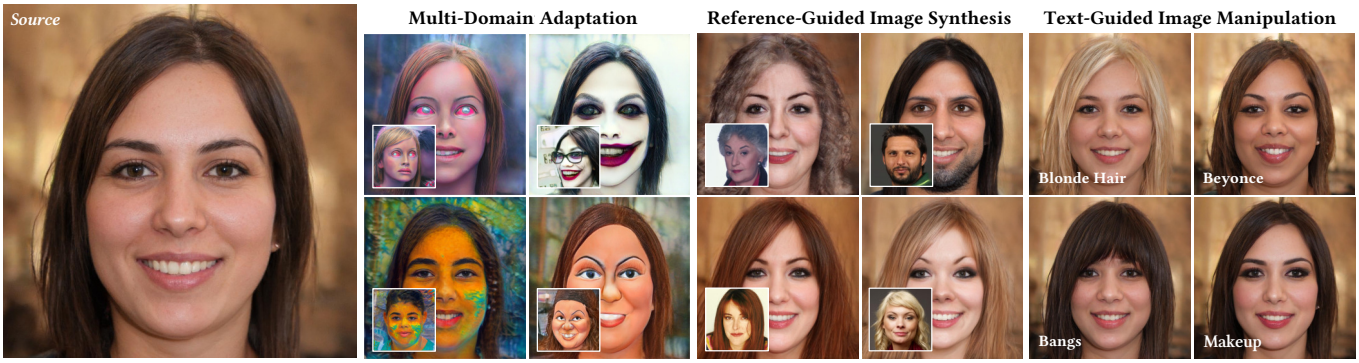


Fig. 1. **HyperGAN-CLIP and its Applications.** Introducing HyperGAN-CLIP, a flexible framework that enhances the capabilities of a pre-trained StyleGAN model for a multitude of tasks, including multiple domain one-shot adaptation, reference-guided image synthesis and text-guided image manipulation. Our method pushes the boundaries of image synthesis and editing, enabling users to create diverse and high-quality images with remarkable ease and precision.

Generative Adversarial Networks (GANs), particularly StyleGAN and its variants, have demonstrated remarkable capabilities in generating highly realistic images. Despite their success, adapting these models to diverse tasks such as domain adaptation, reference-guided synthesis, and text-guided manipulation with limited training data remains challenging. Towards this end, in this study, we present a novel framework that significantly extends the capabilities of a pre-trained StyleGAN by integrating CLIP space via hypernetworks. This integration allows dynamic adaptation of StyleGAN to new domains defined by reference images or textual descriptions. Additionally, we introduce a CLIP-guided discriminator that enhances the alignment between generated images and target domains, ensuring superior image quality. Our approach demonstrates unprecedented flexibility, enabling text-guided image manipulation without the need for text-specific training data and facilitating seamless style transfer. Comprehensive qualitative and quantitative evaluations confirm the robustness and superior performance of our framework compared to existing methods.

CCS Concepts: • **Computing methodologies** → **Image manipulation**.

Additional Key Words and Phrases: GANs, Domain Adaptation, Reference-Guided Image Synthesis, Text-Guided Image Manipulation

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12.
https://doi.org/10.1145/3680528.3687613

ACM Reference Format:
Abdul Basit Anees, Ahmet Canberk Baykal, Muhammed Burak Kizil, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. 2024. HyperGAN-CLIP: A Unified Framework for Domain Adaptation, Image Synthesis and Manipulation. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3680528.3687613

1 INTRODUCTION

Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] have dramatically advanced the synthesis of highly realistic images through novel ideas such as progressive growth [Karras et al. 2018] and style-based generators [Karras et al. 2021, 2019, 2020]. These techniques enable the training of cutting-edge GANs on large, high-resolution datasets by exploiting semantically rich latent spaces for precise style manipulation. However, their reliance on substantial training and large datasets poses significant challenges in data-scarce environments.

Addressing the data scarcity issue, traditional domain adaptation techniques for GANs typically involve fine-tuning pre-trained generators with limited samples from the target domain. While these methods enhance model applicability, they often struggle with a trade-off between the fidelity of domain-specific attributes and the quality of images generated from the source domain. Additionally, methods that utilize multi-modal CLIP embeddings for guided image generation and manipulation [Gal et al. 2022; Zhu et al. 2022] are constrained by the attributes present during training [Baykal

et al. 2023; Lyu et al. 2023; Wei et al. 2022], and they face difficulties with out-of-distribution images. Per-edit optimization techniques [Chefer et al. 2022; Patashnik et al. 2021; Xia et al. 2021], though highly flexible, incur substantial computational costs at inference.

In response to these challenges, we propose HyperGAN-CLIP, a unified framework that not only addresses the limitations of existing domain adaptation methods but also expands their functionality to include reference-guided image synthesis and text-guided image manipulation. This comprehensive framework utilizes a single example from each target domain to efficiently adapt pre-trained GAN models, eliminating the need for task-specific models. Central to HyperGAN-CLIP is a conditional hypernetwork that dynamically adjusts the generator's weights based on domain-specific embeddings from images or text, facilitated by CLIP embeddings.

The strategic use of our hypernetwork module design results in a duplicated generator network that produce domain-specific features via CLIP embeddings. These features are seamlessly integrated into the original generator through a residual feature injection mechanism, which not only preserves the identity of the source domain but also enhances the robustness of the generator by preventing mode collapse. This mechanism effectively addresses common challenges in domain adaptation, and enables our framework to adapt to different domains without requiring separate training sessions for each one. Unlike prior approaches, CLIP-oriented hypernetworks effectively understand and leverage the common characteristics shared among target domains during adaptation, leading to improved results. Moreover, they enhance our framework's capabilities by allowing the use of images and text prompts as guidance, making it well-suited for tasks like reference-guided image synthesis and text-guided image manipulation.

In summary, the key contributions of our work are as follows:

- We propose a conditional hypernetwork that effectively adapts a pre-trained StyleGAN generator to multiple domains with minimal data, maintaining high-quality synthesis image synthesis without increasing model size.
- Our novel design offers more flexibility and supports a wide range of synthesis and editing tasks, including reference-guided image synthesis and text-guided manipulation, without any need for training separate models for each task.
- We conduct extensive evaluations across multiple domains and datasets, demonstrating our framework's effectiveness and adaptability compared to existing methods.

Our code and models are publicly available at the project website: https://cyberiada.github.io/HyperGAN-CLIP.

2 RELATED WORK

2.1 State-of-the-art in GANs

Field of image synthesis and editing has experienced significant advances through the use of generative adversarial networks (GANs) [Goodfellow et al. 2014]. These advances have been by innovative architectural and training strategies that yield highly realistic images. Notably, PGGAN [Karras et al. 2018] introduces progressive resolution enhancement, while BigGAN [Brock et al. 2019] scales up image synthesis with larger batch sizes and introduces techniques like residual connections and the truncation trick for improved quality. StyleGAN [Karras et al. 2019] and its successors, StyleGAN2 [Karras et al. 2020] and StyleGAN3 [Karras et al. 2021], further enhance photorealism and reduce artifacts by using a generator inspired by style transfer literature [Gatys et al. 2015]. StyleSwin [Zhang et al. 2022a] and GANformer [Hudson and Zitnick 2021] incorporate transformers or bipartite structures to generate complex images with multiple objects.

StyleGAN is particularly acclaimed for its rich, semantically meaningful latent space, which enables users to finely manipulate image attributes. GAN inversion, a common technique to embed real images into this space, can be accomplished through methods such as direct optimization [Abdal et al. 2019, 2020; Creswell and Bharath 2019; Tewari et al. 2020], learning-based approaches [Alaluf et al. 2021; Bai et al. 2022; Bau et al. 2019a; Richardson et al. 2021; Tov et al. 2021; Zhu et al. 2020], or hybrids [Bau et al. 2019b; Zhu et al. 2016]. These techniques allow for exploration and manipulation of the latent space to discover and apply meaningful editing directions, often in an unsupervised manner [Härkönen et al. 2020; Shen and Zhou 2021; Voynov and Babenko 2020], or by leveraging image-level attributes [Abdal et al. 2021; Shen et al. 2020a; Wu et al. 2021].

2.2 Domain Adaptation for GANs

Few-shot GAN domain adaptation involves adjusting pre-trained models to new image domains with limited data, often leading to challenges such as overfitting and mode collapse. To address these challenges, several novel strategies have been implemented. Ojha et al. [2021] employ a cross-domain distance consistency loss to maintain diversity while transferring to new domains. Back [2021] fine-tunes StyleGAN2 by freezing initial style blocks and adding a structural loss to minimize deviations between the source and target domains. DualStyleGAN [Yang et al. 2022] employs distinct style paths for content and portrait style transfer, while RSSA [Xiao et al. 2022] compresses the latent space for better domain alignment. StyleGAN-NADA [Gal et al. 2022] uses CLIP embeddings for directional guidance during adaptation, enhancing the fidelity of transfers. Mind-the-Gap [Zhu et al. 2022] introduces regularizers to reduce overfitting. JoJoGAN [Chong and Forsyth 2022] learns a style mapper from a single example using GAN inversion and StyleGAN's style-mixing property. DiFa [Zhang et al. 2022b] leverages CLIP embeddings for both global and local-level adaptation, and employs selective cross-domain consistency to maintain diversity. OneshotCLIP [Kwon and Ye 2023] employs a two-step training strategy involving CLIP-guided latent optimization and generator fine-tuning with a novel loss function to ensure CLIP space consistency. DynaGAN [Kim et al. 2022a] modulates the pre-trained generator's weights for dynamic adaptation. HyperDomainNet [Alanov et al. 2022] employs hypernetworks to predict weight modulation parameters, combined with regularizers and a CLIP directional loss for multi-domain adaptation. Adaptation-SCR [Liu et al. 2023] proposes a spectral consistency regularizer to alleviate mode collapse and preserve diversity and granularity adaptive regularizer to balance diversity and stylization during domain adaptation. Our method extends these studies by using a hypernetwork to modulate a StyleGAN2 generator's weights, integrating missing domain-specific features into a frozen generator for better identity preservation

and minimal distortion. Unlike the direct tuning in DynaGAN, our approach uses CLIP embeddings to generate and inject features, significantly differing from StyleGAN-NADA's finetuning approach, which risks overfitting. Moreover, our hypernetwork is conditioned on multimodal CLIP embeddings, broadening our model's application from domain adaptation to reference-guided image synthesis and text-guided manipulation.

2.3 Reference-Guided Image Synthesis

Reference-guided image synthesis combines the content of one image with the style of another, a process that has evolved significantly from early neural style transfer techniques like [Gatys et al. 2015], which often suffered from style-artifacts due to inadequate handling of local semantic details. To improve upon these limitations, WCT^2 [Yoo et al. 2019] introduced wavelet-corrected transfers that better preserve structural integrity and local feature statistics. DeepFaceEditing [Chen et al. 2021] further refines this approach by using local disentanglement and global fusion to more effectively separate and combine geometric and stylistic elements. BlendGAN [Liu et al. 2021b] adopts a self-supervised method, developing a style encoder that integrates a weighted blending module for seamless style integration. TargetCLIP [Chefer et al. 2022] uses the StyleGAN2 latent space to identify desired editing direction that align with reference images, optimizing the CLIP similarity with the target. NeRFFaceEditing [Jiang et al. 2022] utilizes appearance and geometry decoders in a tri-plane-based neural radiance field, using an AdaIN-based approach for enhanced decoupling of appearance and geometry. Different from these methods, our HyperGAN-CLIP model uses CLIP embeddings to dynamically control the modulation weights and decode the StyleGAN2 latent vectors, offering a more enhanced flexibility and precision in synthesis process. With the growing interest in diffusion models, there have been efforts to guide the denoising diffusion process using reference images as well. For example, diffusion frameworks in [Balaji et al. 2022; Bansal et al. 2024] allow image generation to be steered by the style of a reference image, while the content is specified by a text prompt. MimicBrush [Chen et al. 2024] builds on these works by enabling local semantic edits on input images using a reference image. This is achieved by automatically extracting the semantic correspondence between the input and reference images.

2.4 Text-Guided Image Manipulation

Text-guided image manipulation modifies images based on textual descriptions while preserving their structure and incorporating the specified attributes. Recent studies leverage CLIP [Radford et al. 2021], which provides a shared latent space for images and text, enabling precise text-driven editing. StyleCLIP-LO [Patashnik et al. 2021] optimizes latent codes to generate target images aligned with textual prompts. StyleCLIP-LM [Patashnik et al. 2021] predicts residual latent codes based on the CLIP similarity of attributes and output images. StyleCLIP-GD [Patashnik et al. 2021] maps text prompts to global directions in the original StyleGAN space, while StyleMC [Kocasari et al. 2021] explores global directions within StyleGAN2's lower dimensional $\mathcal{S}$ space to enhance this alignment. HairCLIP [Wei et al. 2022] modulates latent codes for specific style attributes like hair color, using text for fine-grained control, optimizing similarity in the CLIP space. DeltaEdit [Lyu et al. 2023] trains latent mappers solely on images using semantically aligned Δ-CLIP space, enabling manipulations guided by reference textual descriptions or images. CLIPInverter [Baykal et al. 2023] conditions the inversion stage on textual descriptions, obtaining manipulation directions as residual latent codes through a CLIP-guided adapter module. In diffusion-based synthesis methods, DiffusionCLIP [Kim et al. 2022b] modifies input images by first converting them to noise through forward diffusion and then guiding the reverse diffusion process using CLIP similarity to obtain the final image. Plug-and-play [Tumanyan et al. 2022] enhances image synthesis by injecting image feature maps from a latent diffusion model into the denoising process guided by textual descriptions. Pix2Pix-Zero [Parmar et al. 2023] maintains the structure of the original image with cross-attention guidance and applies targeted edits using an edit-direction embedding to modify specific objects. InstructPix2Pix [Brooks et al. 2023] and MagicBrush [Zhang et al. 2023] enable semantic image editing based on user-provided textual instructions. ZONE [Li et al. 2024] extends these approaches to zero-shot local image editing, utilizing the localization capabilities within pre-trained instruction-guided diffusion models.

2.5 Hypernetworks

Hypernetworks [Ha et al. 2017] are neural networks designed to predict or modulate the weights of another network, known as the primary network. This ability enhances the flexibility and generalizability of models. For instance, HyperInverter [Dinh et al. 2022] employs hypernetworks to adjust encoder parameters, while HyperStyle [Alaluf et al. 2022] uses them to adapt the StyleGAN generator, improving representation of out-of-domain images. DynaGAN [Kim et al. 2022a] and HyperDomainNet [Alanov et al. 2022] use hypernetworks for dynamic weight modulation in few-shot domain adaptation. Building on these, our method enhances StyleGAN's adaptability by integrating hypernetworks with CLIP embeddings to modulate weights according to different modalities, letting our framework be used for both domain adaptation, reference-guided image synthesis and text-guided image manipulation.

3 APPROACH

HyperGAN-CLIP represents a unified architecture built upon StyleGAN2 [Karras et al. 2020], designed to address a wide range of generative tasks such as domain adaptation, reference-guided image synthesis, and text-guided image manipulation. In Sec. 3.1, we introduce the core components of HyperGAN-CLIP. Then, in Sec. 3.2, we describe the training procedures employed to deploy HyperGAN-CLIP across the various generative and editing tasks.

3.1 HyperGAN-CLIP

As shown in Fig. 2, our HyperGAN-CLIP framework dynamically adjusts the weights of a StyleGAN2 generator pre-trained on a source domain using input images or text prompts. These versatile inputs can represent a target domain for adaptation, serve as an in-domain reference for attribute transfer, or function as a textual description for editing. This flexibility allows our framework to generate images

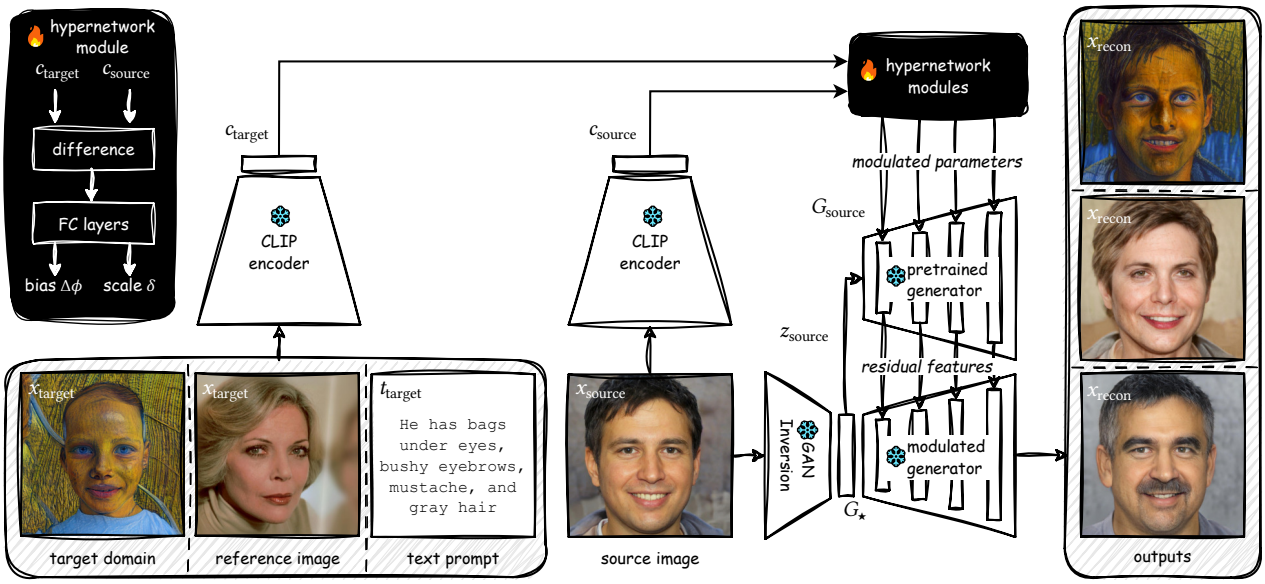


Fig. 2. **Overview of HyperGAN-CLIP**. This framework employs hypernetwork modules to adjust StyleGAN generator weights based on images or text prompts. These inputs facilitate domain adaptation, attribute transfer, or image editing. The modulated weights blend with original features to produce images that align with specified domains or tasks like reference-guided synthesis and text-guided manipulation, while maintaining source integrity.

that not only align with target domain characteristics but also support both reference-guided image synthesis and text-guided image manipulation, all while preserving the source domain's integrity.

At the core of HyperGAN-CLIP is a unified adaptation strategy that employs a single architecture to handle various generative tasks dynamically. This strategy centers around a hypernetwork module that interacts with each layer of a pre-trained StyleGAN generator to produce task-specific adaptations. However, rather than directly updating the original generator network, our approach involves updating the weights of a duplicated generator network. This network generates the missing features based on the provided CLIP [Radford et al. 2021] embeddings of the conditioning inputs. These features are then integrated into the original, frozen generator network via a residual feature injection module, ensuring the preservation of the source domain's integrity.

More formally, the final features of a layer i, denoted by $F_i^{'}$, are estimated by injecting the scaled down modulated features F_i^* into the original features F_i, as given below:

$$F_i^{'} = F_i + \eta \cdot F_i^* \,, \tag{1}$$

where η is the scaling parameter. By this way, the final features remain close to the original distribution at the beginning of the training process. The original intermediate features, F_i, are derived from the preceding layer's output $F_{i-1}^{'}$ using:

$$F_i = F_{i-1}^{'} \circledast \theta_i + b_i \,, \tag{2}$$

with θ_i and b_i respectively representing the layer weights and the layer bias of the pre-trained StyleGAN. Meanwhile, the modulated features, F_i^*, are computed using the weights θ_i^* modulated by the proposed CLIP-conditioned hypernetwork module as follows:

$$F_i^* = F_{i-1} \circledast \theta_i^* + b_i \,, \tag{3}$$

where the modulated weights, θ_i^* are defined as

$$\theta_i^* = \delta_i \cdot f(\phi_i + \Delta\phi_i, s_i) \,. \tag{4}$$

Here, f represents the composite function of cascaded modulation and demodulation operations, s_i is the style vector transformed from the latent code w of the source image, and ϕ_i denotes the convolutional weights of the pre-trained generator at layer i. Notably, the modulation parameters $\Delta\phi_i$ and δ_i, the task-specific weight bias and the channel-wise scale parameter, are dynamically predicted by our proposed CLIP-conditioned hypernetwork module $H_i(\cdot)$, as:

$$\Delta\phi_i, \delta_i = H_i(\Delta c) \,, \tag{5}$$

where Δc is the Δ-CLIP embedding [Lyu et al. 2023] representing the difference between the CLIP embedding of the conditioning input (an image or a text prompt) and the CLIP embedding of the source image. Each hypernetwork module is composed of two individual fully-connected layers that generate affine transformation parameters for each convolution layer, one for the weight bias matrix $\Delta\phi_i$ and the other for the weight scaling parameter δ_i, respectively. Hence, the number of parameters introduced by the hypernetwork module depends on the length of Δ-CLIP embeddings and the size of the corresponding convolutional layer, and often very less compared to the base generator network.

Previous studies have shown that CLIP embeddings are effective at capturing the stylistic elements of reference images [Balaji et al. 2022; Bansal et al. 2024]. Utilizing Δ-CLIP embeddings allows our model to focus solely on the attributes absent in the source domain,

thereby eliminating any redundant information. This approach centers the input embeddings to the hypernetwork around zero, simplifying the training process. Moreover, our findings suggest that using raw CLIP embeddings directly can significantly change the identity and noticeably degrade image quality. A detailed analysis is given in the Supplementary Material. Another key outcome of using CLIP embeddings is that it allows for adapting the pre-trained generator to multiple domains with just a single network model.

3.2 Training HyperGAN-CLIP

Consider x as a synthetic image generated from noise or a natural image from the source domain $\mathcal{D}_{\text{source}}$. In the context of StyleGAN's architecture, x is produced by the mapping $x = G_{\text{source}}(z)$, where z is a latent vector either sampled from a noise distribution or derived using a GAN inversion technique. HyperGAN-CLIP is designed to adapt the pre-trained generator G_{source} into a modulated generator $G_\star$. This adaptation enables $G_\star$ to handle multiple tasks: multiple domain adaptation, reference-guided image synthesis, and text-guided image manipulation. It accomplishes this by leveraging additional inputs, which may be specific images or text prompts, to customize the generator's output to the requirements of these varied applications. We train our HyperGAN-CLIP framework by minimizing a multi-task loss $\mathcal{L}$, defined as:

$$\begin{aligned}\mathcal{L} = \lambda_1\mathcal{L}_{\text{CLIP}} + \lambda_2\mathcal{L}_{\text{CLIP-Across}} + \lambda_3\mathcal{L}_{\text{CLIP-Within}} + \lambda_4\mathcal{L}_{\text{cGAN}} \\ +\lambda_5\mathcal{L}_{\text{Contrastive}} + \lambda_6\mathcal{L}_{\text{ID}} + \lambda_7\mathcal{L}_{\text{L2}} + \lambda_8\mathcal{L}_{\text{LPIPS}} \end{aligned} \tag{6}$$

where λ_* depicts the corresponding regularization coefficients.

3.2.1 CLIP-based Losses. For domain adaptation, the core objective is to align the semantics of the adapted domain images with those of a target domain image x_{target}. We define z_{source} as the latent code corresponding to x_{target} inverted to the source domain, where it generates x_{fixed}, the source domain equivalent of x_{target}. The adapted generator aims to use the same z_{source} to produce an adapted image x_{recon}. Leveraging the CLIP embeddings of the target images, we enforce semantic consistency through the CLIP similarity loss:

$$\mathcal{L}_{\text{CLIP}} = 1 - \langle c_{\text{recon}}, c_{\text{target}} \rangle \,, \tag{7}$$

where c_{target} and c_{recon} represent the CLIP embeddings of x_{target} and x_{recon}, respectively, and $\langle \cdot, \cdot \rangle$ denotes the cosine similarity.

Global CLIP losses can lead to mode collapse and content loss [Gal et al. 2022]. Hence, as explored in [Zhu et al. 2022], we additionally adopt the following directional CLIP losses that measure the semantic shift within and across domains in CLIP space:

$$\mathcal{L}_{\text{CLIP-Across}} = 1 - \langle \Delta c_{\text{sample}}, \Delta c_{\text{fixed}} \rangle \,, \tag{8}$$

$$\mathcal{L}_{\text{CLIP-Within}} = 1 - \langle \Delta c_{\text{source}}, \Delta c_{\text{target}} \rangle \,. \tag{9}$$

To compute these losses, we begin by generating an image x_{sample} using the frozen generator G_{source} from a randomly sampled latent code. This image is then adapted to the target domain using $G_\star$, resulting in x_{trained}. Semantically, we anticipate that the differences between the source and target domains, captured by the Δ-CLIP embeddings $\Delta c_{\text{sample}} = \text{CLIP}(x_{\text{trained}}) - \text{CLIP}(x_{\text{sample}})$ and $\Delta c_{\text{fixed}} = \text{CLIP}(x_{\text{target}}) - \text{CLIP}(x_{\text{fixed}})$, should align as they represent the transformation induced by domain adaptation. Additionally, to ensure the adaptation preserves essential semantic features across the transformation, the differences between source and adapted images, as measured by $\Delta c_{\text{source}} = \text{CLIP}(x_{\text{fixed}}) - \text{CLIP}(x_{\text{sample}})$ and $\Delta c_{\text{target}} = \text{CLIP}(x_{\text{target}}) - \text{CLIP}(x_{\text{trained}})$), should also align.

For reference-guided image synthesis, HyperGAN-CLIP utilizes a refined methodology with in-domain data, adjusting StyleGAN's weights to faithfully replicate the style of target images. By leveraging pairs of source and target images from the source dataset, we effectively cover a broad distribution of CLIP embeddings, ensuring robust alignment between the CLIP space and StyleGAN image space. Specifically, we redefine $\mathcal{L}_{\text{CLIP-Across}}$ using the average StyleGAN image as the anchor image x_{fixed}, departing from the use of inverted target images typical in domain adaptation. During training, x_{target} and x_{sample} are randomly sampled. Furthermore, for $\mathcal{L}_{\text{CLIP-Within}}$, we substitute x_{target} with x_{recon} to enhance identity and content preservation. Please refer to the Supplementary Material for the graphical illustrations of these directional losses.

Notably, HyperGAN-CLIP trained for reference-guided image synthesis is also capable of performing text-guided image editing by using the Δ-CLIP embedding $\Delta c_{\text{text}} = \text{CLIP}(t_{\text{target}}) - \text{CLIP}(t_{\text{source}})$ to modulate the generator weights, with t_{target} representing the input text prompt and t_{source} denoting any text matching the source image. In our experiments, we use a generic prompt like "*face*" for t_{source}, but it can be replaced with a more fine-grained one.

3.2.2 CLIP-conditioned discriminator loss. To preserve sample quality during domain adaptation, we introduce an adversarial loss $\mathcal{L}_{\text{cGAN}}$ with a discriminator conditioned on CLIP embeddings. This discriminator, modeled after [Kang et al. 2023; Kumari et al. 2022], uses a frozen CLIP vision transformer backbone and only trains the outermost head layers. It dynamically measures the difference between source and target domain distributions. To deal with the data scarcity (we only have a single image per each target domain), we use differentiable augmentation [Zhao et al. 2020]. The conditioning of the discriminator on CLIP embeddings, implemented using a projection discriminator [Miyato and Koyama 2018], ensures that the generated images align with the target domain characteristics and accelerates training convergence and prevents mode collapse.

3.2.3 Contrastive Adaptation Loss. To ensure that images generated from a target domain distinctly differ from those of other domains, we employ an adaptation loss $\mathcal{L}_{\text{Contrastive}}$ encouraging the network to learn domain-specific transformations. Inspired by [Kim et al. 2022a], this contrastive loss enhances similarity relationships, ensuring positive pairs (same domain) show higher similarity, while negative pairs (different domains) show less. Formally, it is given as:

$$\mathcal{L}_{\text{Contrastive}} = -\log \frac{\exp(l_{\text{pos}})}{\exp(l_{\text{pos}}) + \Sigma_j \mathbf{1}_{[j \neq k]} \exp(l^j_{\text{neg}})} \tag{10}$$

with l_{pos}, l^j_{neg} representing the cosine similarities of positive and negative pairs, respectively:

$$l_{\text{pos}} = \left\langle \text{CLIP}(x^k_{\text{target}}), \text{CLIP}(x^k_{\text{recon}}) \right\rangle \tag{11}$$

$$l^j_{\text{neg}} = \left\langle \text{CLIP}(\text{Aug}(x^j_{\text{target}})), \text{CLIP}(x^k_{\text{recon}}) \right\rangle \tag{12}$$

where Aug(·) applies horizontal-flip and color-jitter augmentations to enhance training stability [Liu et al. 2021a]. This loss is calculated over a minibatch of 4 target domains for diverse domain learning.

3.2.4 Identity Loss. To preserve source identity when adapting to a target domain, we implement an identity similarity loss designed to maximize the cosine similarity between the image features from the source and target domains:

$$\mathcal{L}_{\text{ID}} = 1 - \langle R(x_{\text{sample}}), R(x_{\text{trained}}) \rangle, \tag{13}$$

where $R(\cdot)$ extracts deep features using the ArcFace model [Deng et al. 2022], specifically trained for face recognition.

3.2.5 Perceptual and Reconstruction Losses. To complement the CLIP loss $\mathcal{L}_{\text{CLIP}}$, we align x_{recon} with x_{target} using the L2 and LPIPS losses:

$$\mathcal{L}_{\text{L2}} = \|x_{\text{target}} - x_{\text{recon}}\|_2 \tag{14}$$

$$\mathcal{L}_{\text{LPIPS}} = \|F(x_{\text{target}}) - F(x_{\text{recon}})\|_2 \tag{15}$$

where $F(\cdot)$ represents AlexNet [Krizhevsky et al. 2012] features.

4 EXPERIMENTS

4.1 Training and Implementation Details

We use the Adam optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.99$. We set the learning rate to 0.002 and the batch size to 4. For CLIP based losses, we use ViT-B/16 and ViT-B/32 CLIP encoder models and add their results as done in MTG. We use the ViT-B/16 CLIP encoder while modulating the generator. The scaling parameter for the modulated features is set as $\eta = 0.1$ to prevent a large shift in feature distribution of the pretrained generator, ensuring stable training from the start. We empirically set the weights for the individual loss terms as $\lambda_1 = 30$, $\lambda_2 = 1.5$, $\lambda_3 = 0.5$, $\lambda_4 = 0.2$, $\lambda_5 = 1.0$, $\lambda_6 = 3.0$, $\lambda_7 = 8.0$, and $\lambda_8 = 12.0$. Each minibatch includes 4 randomly sampled target domain images x_{target} and 4 source images x_{trained}. For domain adaptation and reference guided image synthesis, to find x_{fixed} in the source domain corresponding to a target image, we use e4e inversion [Tov et al. 2021]. However, instead of using the inversion directly, we bring it closer to the mean latent by applying latent truncation. This prevents the inversion to lie in an out-of-distribution region and avoids x_{fixed} and x_{target} to be too close, and thus limiting meaningful editing directions.

4.2 Domain Adaptation

We conduct two distinct experiments. First, we adapt a StyleGAN2 model, pre-trained on the FFHQ dataset [Karras et al. 2019], to 101 new domains introduced in the expanded version of StyleGAN-NADA [Gal et al. 2022]. The training data was generated using the extended StyleGAN2 model provided by the authors of Domain Expansion [Nitzan et al. 2023][1]. For each target domain, we sample a single image using the extended model, and use these sampled images to train our HyperGAN-CLIP model for multiple domain adaptation. Second, we use the AFHQ dataset to expand a StyleGAN2 model pre-trained on Cat images to 52 other animal domains (including 22 dog breeds and 30 wildlife animals represented by 7 cheetah, 6 tiger, 6 lion, 7 fox and 4 wolf images). For each target domain, we select a single image and use these samples to train HyperGAN-CLIP accordingly. We compare HyperGAN-CLIP to state-of-the-art GAN domain adaptation models, including Mind-the-GAP [Zhu et al. 2022], StyleGAN-NADA [Gal et al. 2022], HyperDomainNet [Alanov et al. 2022], DynaGAN [Kim et al. 2022a], and Adaptation-SCR [Liu et al. 2023]. Each model is trained in the one-shot setting using the same training data. Notably, Mind-the-GAP, StyleGAN-NADA, and Adaptation-SCR require separate models for each target domain, whereas HyperDomainNet, DynaGAN, and HyperGAN-CLIP can model multiple domains with a single unified model. To quantitatively assess the quality and fidelity of the generated images, we adopt the widely used Fréchet Inception Distance (FID) score [Heusel et al. 2017] along with the Quality and Diversity metrics suggested in [Alanov et al. 2022]. Details of these evaluation metrics are given in the Supplementary Material.

In Fig. 3, we present sample images generated by the evaluated domain-adaptation techniques on the AFHQ and FFHQ datasets. Each sample includes the source image, the corresponding target domain training image and the synthesized outputs. Mind-the-Gap struggle to fully capture the visual characteristics of the target domains, often producing visually poor results. HyperDomainNet appears to have failed in learning very diverse domains, which leads to low-fidelity outcomes. While StyleGAN-NADA and Adaptation-SCR achieve better quality, they tend to slightly overfit to specific features of the representative target domain. DynaGAN shows improved performance over these models but sometimes generates unnatural and slightly distorted results, particularly in animal domains. It fails to fully reflect key features of the target domain, e.g., it does not generate desired small animal ears in the first row. Compared to DynaGAN, HyperGAN-CLIP better preserves source content. By leveraging CLIP-guided hypernetwork modules, it produces images with remarkable visual fidelity and effectively captures the essence of the target domains, as validated by the FID scores in Table 1. Moreover, the Diversity scores highlight that our approach demonstrates higher variability among the adapted images. Additional demonstrations of our model's ability to blend domains and perform semantic edits are given in Fig.4. In the Supplementary Material, we provide additional comparisons, explore controllable image generation in more detail, and present an ablation study. Moreover, we demonstrate that our approach can perform zero-shot domain adaptation relatively well on novel domains that are not semantically very different from the domains used during training.

4.3 Reference-Guided Image Synthesis

In this experiment, our objective is to synthesize a new image that combines the identity of a source image with the style of a target image, as represented by its CLIP embedding. For quantitative analysis, we use the test set of the CelebA-HQ dataset [Lee et al. 2020], which comprises a total of 6000 diverse images, as the source and the target images. We assign a different target image to each source image by making sure that the same image is not used as source and target. We invert the source images to the latent space using an e4e encoder [Tov et al. 2021] pre-trained on the FFHQ dataset.

[1]The NADA-expanded model used in our experiments is available at https://github.com/adobe-research/domain-expansion/tree/main.

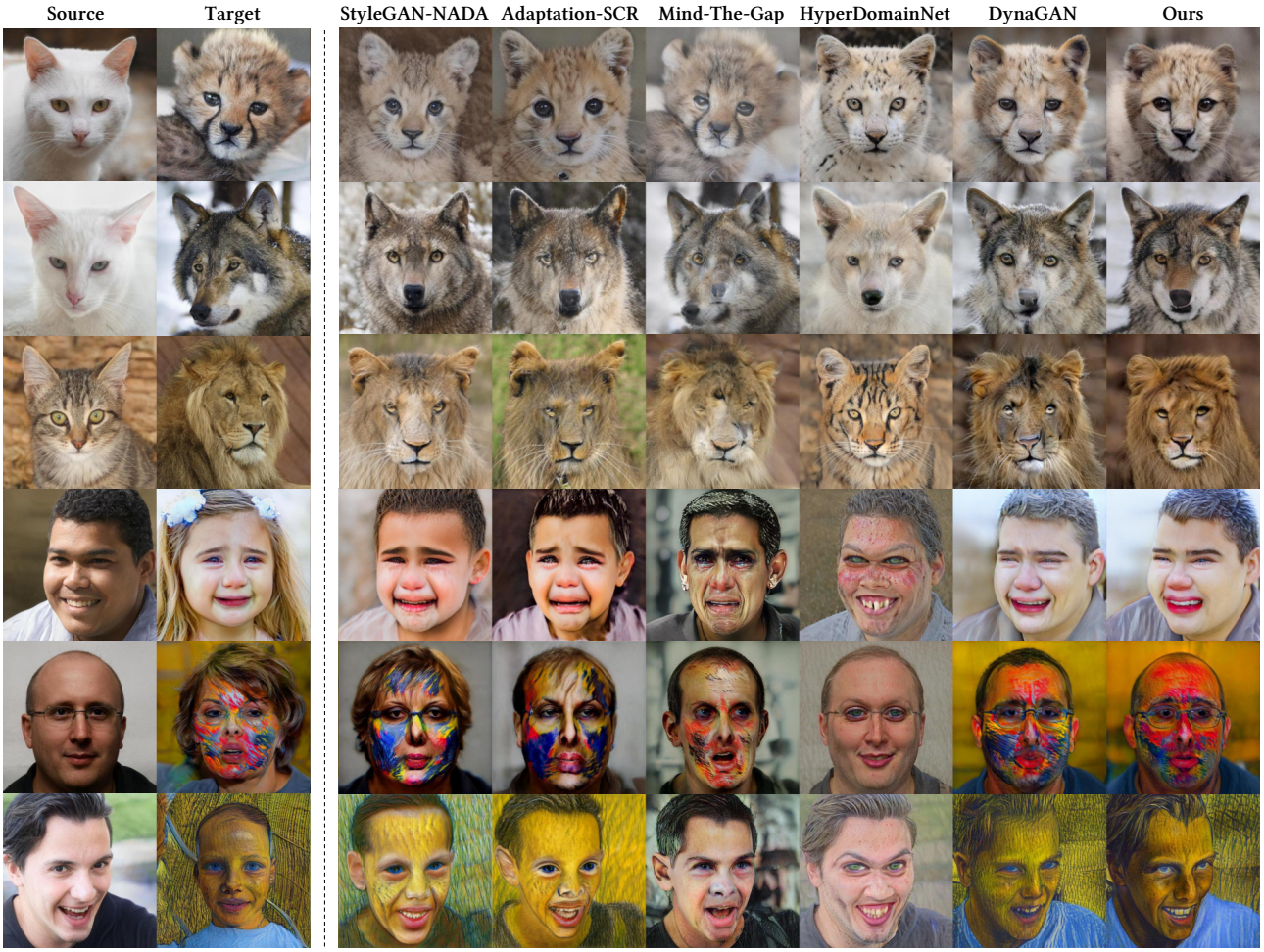
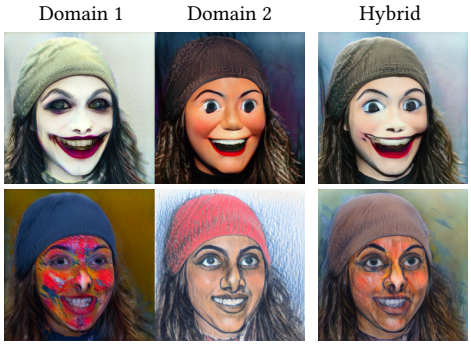
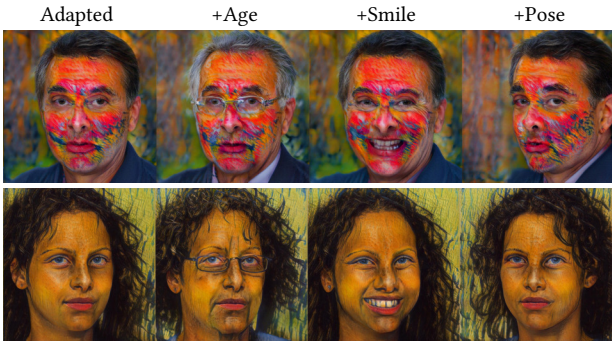


Fig. 3. **Comparison against the state-of-the-art few-shot domain adaptation methods**. Our proposed HyperGAN-CLIP model outperforms competing methods in accurately capturing the visual characteristics of the target domains.



(a) **Domain mixing.** Our approach can fuse multiple domains to create novel compositions. By averaging and re-scaling the CLIP embeddings of two target domains, we can generate images that blend characteristics from both.



(b) **Semantic editing in target domains.** Since latent mapper is kept intact, our approach allows for using existing latent space discovery methods to perform semantic edits. We manipulate two sample face images from adapted domains by playing with age, smile, and pose using InterfaceGAN [Shen et al. 2020b].

Fig. 4. **Capabilities of HyperGAN-CLIP in blending domains and performing semantic edits within adapted domains.**

Table 1. **Quantitative results for multi domain adaptation.** HyperGAN-CLIP demonstrates strong performance in adapting characteristics of multiple target domains with a single model. The best and second best models are indicated in bold and underlined, respectively.

	AFHQ			FFHQ		
Method	FID↓	Qual↑	Div↑	FID↓	Qual↑	Div↑
Mind-The-Gap	72.90	0.93	0.04	45.93	0.73	0.10
StyleGAN-NADA	71.15	0.93	0.04	49.48	0.90	0.04
Adaptation-SCR	**70.84**	0.92	0.03	45.88	0.59	0.06
HyperDomainNet	105.90	0.78	**0.05**	100.92	0.67	0.11
DynaGAN	72.16	**0.94**	0.02	28.94	**0.83**	0.14
Ours	71.93	**0.94**	0.04	**24.74**	0.81	**0.16**

Table 2. **Quantitative results for reference-guided image synthesis.** HyperGAN-CLIP outperforms the existing models, generating high-quality images. It effectively preserves source identity while transferring the semantic details of the target images. The best and second-best models are highlighted in bold and underlined, respectively.

Method	FID↓	ID (source)↑	ID (target)↓	CLIP Sim.↑
BlendGAN	14.54	34.58±9.91	**2.63±9.53**	77.08±7.17
TargetCLIP-O	11.26	50.77±16.61	17.78±10.54	77.16±9.71
TargetCLIP-E	29.48	41.51±11.61	26.94±10.40	72.41±8.01
MimicBrush	37.06	11.19±10.43	65.91±14.65	82.69±7.29
Ours	**8.73**	**78.73±6.01**	10.51±10.04	**90.78±3.80**

The inverted latents are fed to our framework along with the CLIP embedding obtained from the target image to synthesize the final output. We compare HyperGAN-CLIP against BlendGAN [Liu et al. 2021b], TargetCLIP-O [Chefer et al. 2022], TargetCLIP-E [Chefer et al. 2022], and MimicBrush [Chen et al. 2024]. While BlendGAN and TargetCLIP-E are encoder-based approaches, TargetCLIP-O employs a direct optimization scheme, and MimicBrush is a diffusion based approach (the whole image region is used as the input mask). Our approach, apart from these studies, is based on modulating the StyleGAN generator via CLIP-guided hypernetworks.

In Fig. 5, we present sample qualitative comparisons. Sample source-target pairs show a diverse range of visual characteristics in terms of gender, age, hair color, ethnicity. BlendGAN tends to produce cartoon-like outputs that lack naturalness. Optimization-based TargetCLIP-O shows superior performance compared to its encoder-based counterpart TargetCLIP-E in maintaining identity while incorporating the desired style changes depicted in the target image. MimicBrush directly copies the target face onto the source pose, failing to transfer just the style and often resulting in unrealistic outputs. Notably, HyperGAN-CLIP gives superior performance in seamlessly transferring the attributes from the chosen target faces to the source faces while preserving identity to a greater extent than the competing methods. These results affirm the effectiveness of our approach in generating visually compelling outputs with enhanced fidelity and plausibility. Table 2 shows the quantitative results. Our method achieves competitive results in terms of FID, better than TargetCLIP-O, which performs latent optimization for each target. This highlights our method's ability to generate high-quality and faithful images. Moreover, our approach outperforms competing methods in preserving the identity of the source image, as indicated by the ID similarity scores. Additionally, our method excels in CLIP semantic similarity, affirming its capability to capture the semantics of the target image in the synthesized results. Overall, our approach strikes a favorable balance across multiple evaluation metrics, showing its effectiveness in photo-realistic image synthesis and preserving key visual attributes.

One key limitation of both our proposed method and the competitive approaches is that, in some cases, they struggle to transfer fine attributes from reference images because their global image embeddings lack the specificity needed to capture these details. To address this issue, we explore a strategy that combines the CLIP embeddings of reference images with those of text prompts designed to capture specific target attributes. By leveraging CLIP's capability to encode both visual and textual data, we refine the reference image embedding by incrementally adding the embedding of the target attribute, modulated by an α parameter, following the formula $\text{CLIP}(x_{\text{target}}) + \alpha\ \text{CLIP}(t_{\text{target}})$. As demonstrated in Fig. 6, this strategy enhances the editing process by allowing fine-tuned adjustments to specified attributes, resulting in more accurate and detailed image modifications based on the reference image.

4.4 Text-Guided Image Manipulation

In this experiment, we show the versatility of our proposed framework by demonstrating its ability to manipulate input images based on target textual descriptions. For the quantitative analysis, we leverage the CelebA dataset's test set [Liu et al. 2015] along with its attribute annotations. We select attributes that are absent from the images and construct target descriptions that prompt the desired attribute manipulation. Leveraging a pre-trained e4e model [Tov et al. 2021], we perform an image-to-latent-space inversion, generating latent representations of the input images. These inverted images serve as inputs to our framework. To condition the synthesis process, we utilize Δ-CLIP embeddings, which capture the discrepancy between the CLIP embeddings of the target description and the input image. We perform a comprehensive comparison of our method against several state-of-the-art text-guided image manipulation approaches. These include TediGAN-B [Xia et al. 2021], StyleCLIP-LO [Patashnik et al. 2021], StyleCLIP-GD [Patashnik et al. 2021], HairCLIP [Wei et al. 2022], DeltaEdit [Lyu et al. 2023], and CLIPInverter [Baykal et al. 2023] as representative GAN-based methods. Among these, DeltaEdit is the only model that utilizes text-free training like our method. Additionally, we also compare against diffusion-based approaches, namely DiffusionCLIP [Kim et al. 2022b], Plug-and-Play [Tumanyan et al. 2022], and InstructPix2Pix [Brooks et al. 2023]. Among these, the method most similar to ours is DeltaEdit in the sense that it is also solely trained on image data and does not utilize any text data during training. By evaluating our method against these diverse approaches, we provide a comprehensive analysis of its performance and highlight its distinct advantages in text-guided image manipulation. To evaluate the approaches quantitatively, we employ Fréchet Inception Distance (FID) [Heusel et al. 2017], Attribute Manipulation Accuracy (AMA),

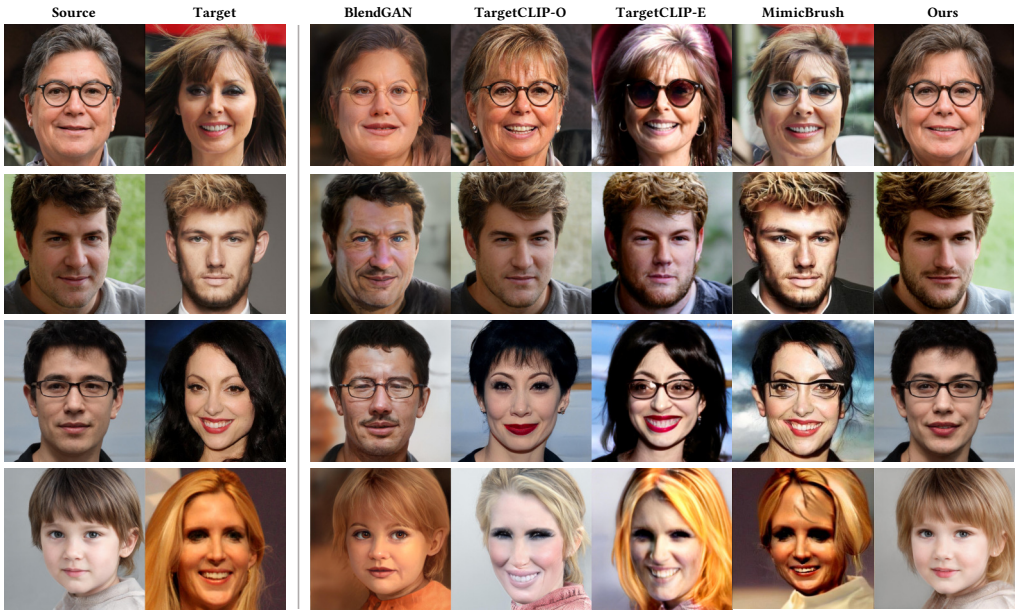


Fig. 5. **Comparison with state-of-the-art reference-guided image synthesis approaches.** Our approach effectively transfers the style of the target image to the source image while effectively preserving identity compared to competing methods.

Fig. 6. **Reference-guided image synthesis with mixed embeddings**. Each row shows the input image, the initial result with the CLIP image embedding, the refined result with a mixed embedding that incorporates the target attribute with $\alpha = 0.5$, and the reference image, respectively. Target text attributes are *"beard"* (top row), *"black hair"* (middle row), and *"smiling"* (bottom row). Incorporating mixed modality embeddings results in more accurate and detailed image modifications.

and CLIP Manipulative Precision (CMP) following the methodology introduced by CLIPInverter [Baykal et al. 2023]. Please refer to the supplementary material for more details on the evaluation metrics.

Fig. 7 presents text-guided image manipulation results of our proposed approach along with several competing methods across various textual descriptions. TediGAN-B and DeltaEdit struggle to effectively manipulate the images, often resulting in images similar to the input. While StyleCLIP-LO, StyleCLIP-GD and HairCLIP perform better, they still exhibit limitations when manipulating all specified attributes.CLIPInverter performs well when explicit attribute manipulations are specified in the descriptions (first two rows), but it falls short when encountering novel descriptions unseen during its training, such as "*surprised*" or "*Elsa from Frozen*". DiffusionCLIP [Kim et al. 2022b] generates images with noticeable artifacts, leading to poor output quality. While Plug-and-play [Tumanyan et al. 2022] successfully applies most manipulations, the resulting images often lack realism, appearing cartoonish and with unintended attribute modifications. In contrast, our model, even trained without any textual data, successfully applies single or multiple attribute changes while better preserving the identity of the input images compared to the competing approaches.

Table 3 presents the quantitative results. Here, we group our approach and DeltaEdit together to distinguish these works from the others which utilize additional text data during training. We evaluate manipulation accuracy and precision using AMA (Single) for single attribute changes and AMA (Multiple) for multiple attribute changes. Remarkably, our model achieves comparable or even better performance in manipulation accuracy and precision compared to leading text-guided image manipulation models, including StyleCLIP, and DiffusionCLIP. In terms of FID, the diffusion-based models, DiffusionCLIP and Plug-and-play, excel as compared to GAN-based approaches due to their high-quality generation capabilities. Even though we do not use textual data during training, our model finds a good balance between the metrics and consistently delivers competitive performance. It effectively handles descriptions involving multiple attribute changes. More importantly, as compared to DeltaEdit,

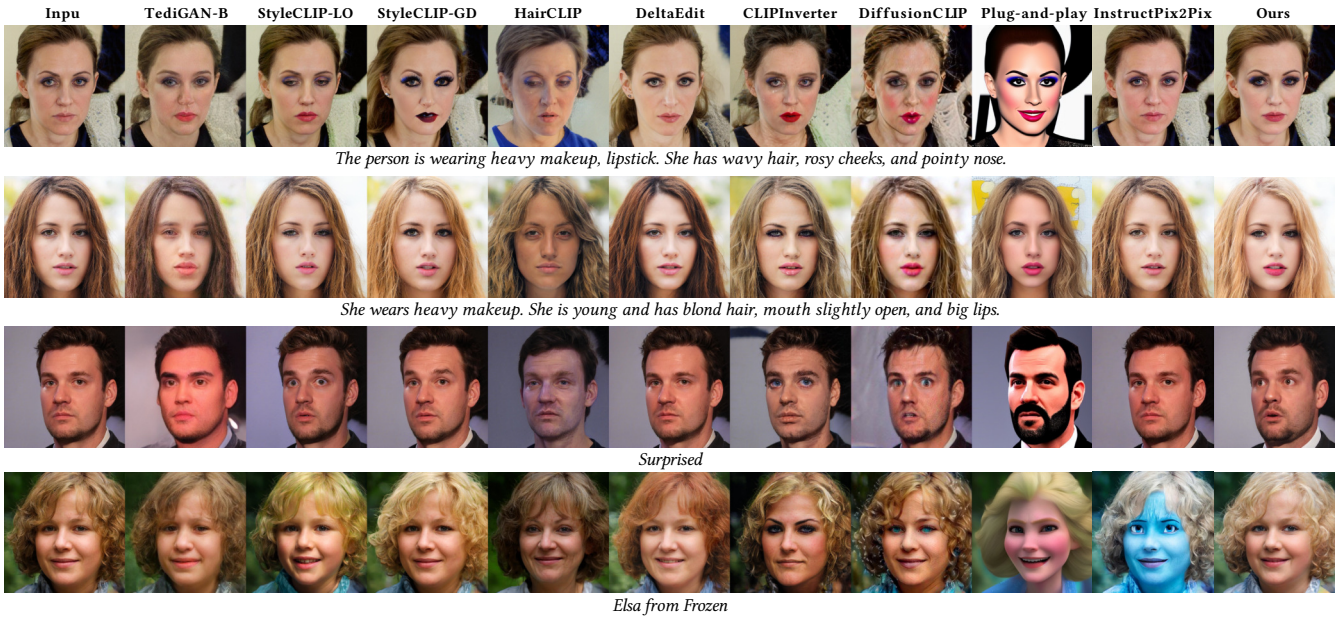


Fig. 7. **Comparisons with state-of-the-art text-guided image manipulation methods.** Our model shows remarkable versality in manipulating images across a diverse range of textual descriptions. The results vividly illustrate our model's ability to accurately apply changes based on target descriptions encompassing both single and multiple attributes. Compared to the competing approaches, our model preserves the identity of the input much better while successfully executing the desired manipulations.

Table 3. **Quantitative results for text-guided image editing.** Even without explicit training on textual descriptions, HyperGAN-CLIP achieves results competitive with the state-of-the-art methods. The best and second best models are highlighted in bold and underlined, respectively.

	FID↓	CMP↑	AMA↑ (Sng.)	AMA↑ (Mult.)
TediGAN-B	55.424	**0.285**	11.286	1.142
StyleCLIP-LO	80.833	0.210	15.857	3.429
StyleCLIP-GD	82.393	0.191	33.143	11.429
HairCLIP	93.523	0.218	<u>41.571</u>	15.149
CLIPInverter	97.210	0.221	**61.429**	**41.714**
DiffusionCLIP	**29.280**	<u>0.243</u>	26.000	4.857
Plug-and-play	68.287	0.199	27.429	7.143
InstructPix2Pix	<u>47.531</u>	0.173	40.571	<u>19.714</u>
DeltaEdit	80.316	0.171	8.857	0.571
Ours	87.851	0.189	25.143	10.000

the other text-guided image manipulation method with text-free training, our HyperGAN-CLIP gives much superior performance.

In the Supplementary Material, we provide further visual comparisons and example results on the CUB-Birds dataset for reference-guided image synthesis and text-guided image manipulation tasks. In addition to the quantitative analyses, we conducted a user study using Qualtrics with 16 participants to evaluate the performance of the models for all three tasks. We focused on methods that have similar characteristics to ours: all-in-one models for multiple domain adaptation and text-based editing methods with text-free training. In our human evaluation, we randomly generated 25 questions for each task and asked participants to rank the models based on their performance. The rankings showed that our HyperGAN-CLIP model, using a single unified framework, achieves highly competitive results, often outperforming or matching the existing models. For more details, please refer to the Supplementary Material.

5 CONCLUSION

We present HyperGAN-CLIP, a flexible framework for addressing domain adaptation challenges in GANs, also supporting both reference-guided image synthesis and text-guided image manipulation. Our efficient hypernetwork modules adapt a pre-trained StyleGAN generator to handle both image and text inputs. By utilizing residual feature injection and a conditional discriminator, it preserves source identity and image diversity while effective transferring target domain characteristics to produce high-fidelity images. Extensive evaluations show that HyperGAN-CLIP outperforms existing domain adaptation methods, excels in text-guided editing, and competes strongly in reference-guided image synthesis. While our framework handles various tasks, some require distinct training processes. Future research could seamlessly incorporate a mixture-of-experts approach to train a single model equipped with routing mechanisms.

ACKNOWLEDGMENTS

This work was supported by KUIS AI Fellowships to ABA, ACB and MBK, Cambridge Trust & Computer Science Premium Scholarship to ACB, TUBA GEBIP 2018 Award to EE, BAGEP 2021 Award to AE, and an Adobe research gift.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4431–4440. https://doi.org/10.1109/ICCV.2019.00453

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2StyleGAN++: How to Edit the Embedded Images?. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/CVPR42600.2020.00832

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. *ACM Trans. Graph.* 40, 3, Article 21 (May 2021), 21 pages. https://doi.org/10.1145/3447648

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. 2022. HyperStyle: StyleGAN Inversion With HyperNetworks for Real Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18511–18521.

Aibek Alanov, Vadim Titov, and Dmitry Vetrov. 2022. HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Jihye Back. 2021. Fine-Tuning StyleGAN2 For Cartoon Face Generation. *CoRR* abs/2106.12445 (2021). arXiv:2106.12445 https://arxiv.org/abs/2106.12445

Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujiu Yang, and Yujun Shen. 2022. High-fidelity GAN inversion with padding space. In *European Conference on Computer Vision*. Springer, 36–53.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Universal Guidance for Diffusion Models. In *International Conference on Learning Representations (ICLR)*.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019a. Semantic Photo Manipulation with a Generative Image Prior. *ACM Trans. Graph.* 38, 4, Article 59 (jul 2019), 11 pages. https://doi.org/10.1145/3306346.3323023

David Bau, Jun-Yan Zhu Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. 2019b. Inverting Layers of a Large Generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ahmet Canberk Baykal, Abdul Basit Anees, Duygu Ceylan, Erkut Erdem, Aykut Erdem, and Deniz Yuret. 2023. CLIP-Guided StyleGAN Inversion for Text-Driven Real Image Editing. *ACM Trans. Graph.* 42, 5, Article 172 (aug 2023), 18 pages.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. 2022. Image-Based CLIP-Guided Essence Transfer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII* (Tel Aviv, Israel). 695–711.

Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L. Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. 2021. DeepFaceEditing: Deep Generation of Face Images from Sketches. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2021)* 40, 4 (2021), 90:1–90:15.

Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. 2024. Zero-shot Image Editing with Reference Imitation. *arXiv preprint arXiv:2406.07547* (2024).

Min Jin Chong and David Forsyth. 2022. JoJoGAN: One Shot Face Stylization. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Antonia Creswell and Anil Anthony Bharath. 2019. Inverting the Generator of a Generative Adversarial Network. *IEEE Transactions on Neural Networks and Learning Systems* 30, 7 (2019), 1967–1974. https://doi.org/10.1109/TNNLS.2018.2875194

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (oct 2022), 5962–5979. https://doi.org/10.1109/tpami.2021.3087709

Tan M. Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. 2022. HyperInverter: Improving StyleGAN Inversion via Hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Trans. Graph.* 41, 4, Article 141 (jul 2022), 13 pages.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A Neural Algorithm of Artistic Style. *ArXiv* abs/1508.06576 (2015).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. HyperNetworks. In *International Conference on Learning Representations*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

Drew A Hudson and C. Lawrence Zitnick. 2021. Generative Adversarial Transformers. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021* (2021).

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Kaiwen Jiang, Shu-Yu Chen, Feng-Lin Liu, Hongbo Fu, and Lin Gao. 2022. NeRFFaceEditing: Disentangled Face Editing in Neural Radiance Fields. In *ACM SIGGRAPH Asia 2022 Conference Proceedings* (Daegu, Korea) *(SIGGRAPH Asia'22)*. Association for Computing Machinery, New York, NY, USA.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up GANs for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In *Proc. NeurIPS*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022b. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2426–2435.

Seongtae Kim, Kyoungkook Kang, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022a. DynaGAN: Dynamic Few-shot Adaptation of GANs to Multiple Domains. In *Proceedings of the ACM (SIGGRAPH Asia)*.

Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. 2021. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. In *WACV*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. 1097–1105.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2022. Ensembling Off-the-shelf Models for GAN Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gihyun Kwon and Jong Chul Ye. 2023. One-Shot Adaptation of GAN in Just One CLIP. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12179–12191.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xuhui Liu, Jiaming Liu, Li Lin, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. 2024. ZONE: Zero-Shot Instruction-Guided Local Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. 2021b. BlendGAN: Implicitly GAN Blending for Arbitrary Stylized Face Generation. In *Advances in Neural Information Processing Systems*.

Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. 2021a. FuseDream: Training-Free Text-to-Image Generation with Improved CLIP+GAN Space Optimization. *arXiv preprint arXiv:2112.01573* (2021).

Zhenhuan Liu, Liang Li, Jiayu Xiao, Zheng-Jun Zha, and Qingming Huang. 2023. Text-Driven Generative Domain Adaptation with Spectral Consistency Regularization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. 2023. DeltaEdit: Exploring Text-Free Training for Text-Driven Image Manipulation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 6894–6903.

Takeru Miyato and Masanori Koyama. 2018. cGANs with Projection Discriminator. In *International Conference on Learning Representations*.

Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. 2023. Domain Expansion of Image Generators. (2023).

Utkarsh Ojha, Yijun Li, Cynthia Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot Image Generation via Cross-domain Correspondence. In *CVPR*.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot Image-to-Image Translation. In *SIGGRAPH*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2021. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020a. Interpreting the Latent Space of GANs for Semantic Face Editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020b. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *TPAMI* (2020).

Yujun Shen and Bolei Zhou. 2021. Closed-Form Factorization of Latent Semantics in GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. PIE: Portrait Image Embedding for Semantic Control. *ACM Trans. Graph.* 39, 6, Article 223 (nov 2020), 14 pages.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an Encoder for StyleGAN Image Manipulation. *ACM Trans. Graph.* 40, 4, Article 133 (jul 2021), 14 pages. https://doi.org/10.1145/3450626.3459838

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572* (2022).

Andrey Voynov and Artem Babenko. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*. PMLR, 9786–9796.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2022. Hairclip: Design your hair by text and reference image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12863–12872.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. 2022. Few Shot Generative Model Adaption via Relaxed Spatial Structural Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11204–11213.

Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. In *CVPR*.

Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. 2019. Photorealistic Style Transfer via Wavelet Transforms. In *International Conference on Computer Vision (ICCV)*.

Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. 2022a. StyleSwin: Transformer-Based GAN for High-Resolution Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11304–11314.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Yabo Zhang, Mingshuai Yao, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. 2022b. Towards Diverse and Faithful One-shot Adaption of Generative Adversarial Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. 2020. Differentiable Augmentation for Data-Efficient GAN Training. arXiv:2006.10738 [cs.CV]

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN Inversion for Real Image Editing. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. 2016. Generative Visual Manipulation on the Natural Image Manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. 2022. Mind the Gap: Domain Gap Control for Single Shot Domain Adaptation for Generative Adversarial Networks. In *International Conference on Learning Representations*.

Supplementary Material: HyperGAN-CLIP: A Unified Framework for Domain Adaptation, Image Synthesis and Manipulation

ABDUL BASIT ANEES, Koç University, Turkey
AHMET CANBERK BAYKAL, University of Cambridge, United Kingdom
MUHAMMED BURAK KIZIL, Koç University, Turkey
DUYGU CEYLAN, Adobe Research, United Kingdom
ERKUT ERDEM, Hacettepe University, Turkey
AYKUT ERDEM, Koç University, Turkey

CCS Concepts: • **Computing methodologies** → **Image manipulation**; **Neural networks**.

ACM Reference Format:
Abdul Basit Anees, Ahmet Canberk Baykal, Muhammed Burak Kizil, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. 2024. Supplementary Material: HyperGAN-CLIP: A Unified Framework for Domain Adaptation, Image Synthesis and Manipulation. *ACM Trans. Graph.* 1, 1 (September 2024), 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

The purpose of this document is to provide extra material to complement the main paper. Section 1 presents graphical illustration of the directional CLIP losses for a better understanding of these losses and their roles. Section 2 provides details about the evaluation setups used in domain adaptation, reference-guided image synthesis, and text-guided image manipulation experiments. Section 3 gives the details of the user study that was carried out to assess the model performances on

Section 4 explores the controllable generation of images by manipulating the scaling of CLIP embeddings and style mixing. Section 6 presents the results of our ablation study, highlighting the contribution of each component of our model to overall performance. Section 7 demonstrates qualitative comparisons between using raw CLIP embeddings and Δ-CLIP embeddings for text-guided editing. Section 8 presents additional visual comparisons between our proposed model and existing approaches in few-shot domain adaptation, reference-guided image synthesis, and text-guided image manipulation. Section 9 and Section 10 discusses the limitations and ethical implications of our work, respectively.

Authors' addresses: Abdul Basit Anees, abdulbasitanees98@gmail.com, Koç University, Turkey; Ahmet Canberk Baykal, canberk.baykal1@gmail.com, University of Cambridge, United Kingdom; Muhammed Burak Kizil, mkizil19@ku.edu.tr, Koç University, Turkey; Duygu Ceylan, duygu.ceylan@gmail.com, Adobe Research, United Kingdom; Erkut Erdem, erkut@cs.hacettepe.edu.tr, Hacettepe University, Turkey; Aykut Erdem, aerdem@ku.edu.tr, Koç University, Turkey.

0730-0301/2024/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

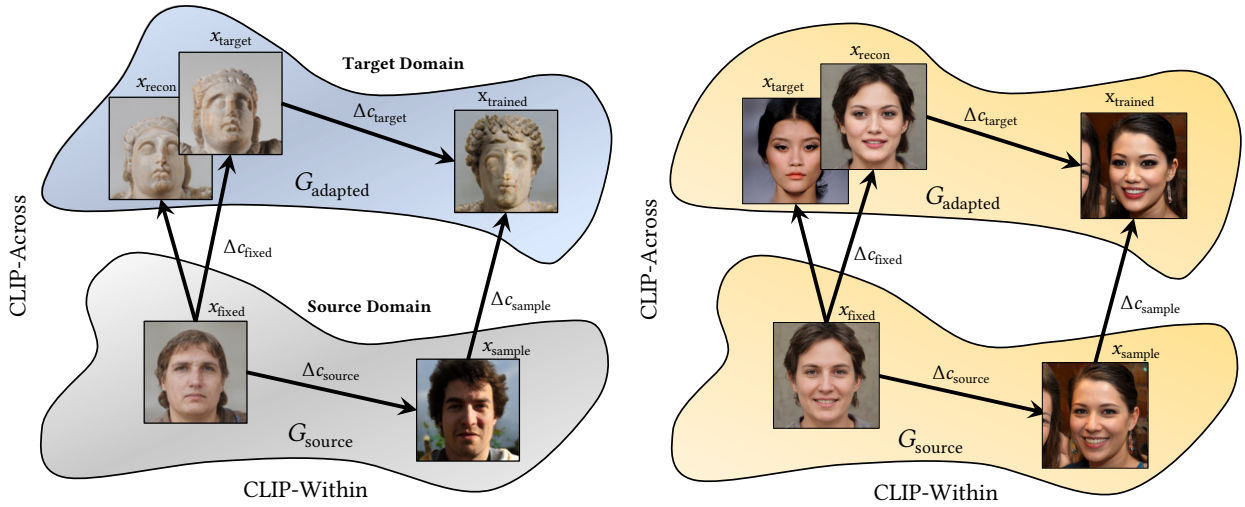
1 DIRECTIONAL CLIP LOSSES

Together, these losses ensure that crucial semantic information is preserved while capturing variations unique to each domain. Their definitions differ slightly between domain adaptation and reference-guided image synthesis due to the varying nature of source and target data, yet both share the core objective of measuring the semantic shifts to enhance diversity and preserve content. CLIP-Across captures the directional relationship between source and target/reference samples to guide adaptation, while CLIP-Within ensures that transformations maintain internal consistency within the adapted domains or transformations. These losses are instrumental in refining the generator's ability to retain identity and style information effectively, aligning generated outputs with the desired target characteristics. We provide graphical illustrations in Fig. 1 to clarify their definitions and distinctions across domain adaptation and reference-guided image synthesis tasks.

2 EVALUATION DETAILS

Domain Adaptation Experiments. To quantitatively assess the quality and fidelity of the generated images, we used the widely used Fréchet Inception Distance (FID) score [Heusel et al. 2017] and the Quality and Diversity metrics suggested in [Alanov et al. 2022]. The FID score provides a measure of the statistical distance between the distributions of real and generated images. The Quality metric evaluates how closely the adapted images align with the text description of the target domain. This is computed as the mean cosine similarity between the CLIP embeddings of the images and the CLIP embedding of the text description. The Diversity metric, on the other hand, measures the variability among the adapted images. This is quantified as the mean pairwise cosine distance between the CLIP embeddings of all the adapted images. In our evaluation, we generate a set of 1K images for each target domain using the NADA-expanded Domain Expansion model, treating these images as real. For the FID evaluation, we compare the distribution of these images with that of images generated by the evaluated methods. Specifically, we randomly sample 100 images from each target domain to represent the generated image distribution.

Reference-Guided Image Synthesis Experiments. We use FID [Heusel et al. 2017] to measure the quality and the fidelity of the synthesized images as a lower FID score indicates that the synthesized images are closer to the FFHQ domain, which is the original domain the StyleGAN2 is trained on. We use the ID similarity [Deng et al. 2022] to measure identity preservation. Ideally, we want the ID similarity



(a) **For domain adaptation.** We encode the images generated by the original and the modulated generators, representing the source and target domains, in the CLIP space. CLIP-Across loss, involving Δc_{sample} and Δc_{fixed}, captures the differences between the source and target domains. On the other hand, CLIP-Within loss, computed using Δc_{source} and Δc_{target}, preserves the semantic information that is unrelated to the domain gap.

(b) **For reference-guided image synthesis.** In reference-guided image synthesis, source and target domains are the same, and thus it involves in-domain adaptation. CLIP-Across loss uses the mean StyleGAN image as the anchor image x_{fixed}. On the other hand, CLIP-Within loss utilizes the reconstructed image x_{recon} to better preserve facial identity and image content.

Fig. 1. **Visualization of the directional CLIP losses.** (a) for domain adaption. (b) for reference-guided image synthesis.

with the source image to be high and ID similarity with the target image to be low as we want to preserve the identity of the source image while only transferring the attributes of the target image to the source. Finally, we use the CLIP embedding space to measure the semantic similarity of the target and output images to evaluate how well the semantics of the target image are transferred.

Text-Guided Image Manipulation Experiments. To evaluate the approaches quantitatively, we employ multiple quantitative metrics, namely Fréchet Inception Distance (FID) [Heusel et al. 2017], Attribute Manipulation Accuracy (AMA), and CLIP Manipulative Precision (CMP) following the methodology introduced by CLIPInverter [Baykal et al. 2023]. FID serves as a measure of the quality and fidelity of the synthesized images.

Attribute Manipulation Accuracy (AMA) [Baykal et al. 2023] measures how well a single manipulation is applied. To calculate the AMA score of a model, for each attribute (such as *blonde hair*), we first select 50 images that the attribute is not present in. Then, we edit these images with a corresponding caption, such as *The person has blonde hair*. Finally, we use pre-trained attribute classifiers to measure the manipulation accuracy on the output images. We average the accuracy accross the attributes to obtain the final AMA score. We trained attribute classifiers for each of the 40 attributes that are present in the CelebA [Liu et al. 2015] dataset. We used the 15 attributes that achieve 90% or higher validation accuracies to calculate the AMA scores. Here is a full list of attributes we used for the CelebA dataset:

- *blonde hair*
- *bushy eyebrows*
- *chubby*
- *double chin*
- *eyeglasses*
- *goatee*
- *gray hair*
- *heavy makeup*
- *male*
- *mouth slightly open*
- *mustache*
- *rosy cheeks*
- *smiling*
- *wearing lipstick*
- *wearing necktie*

In order to quantify the alignment between the output images and the target captions, while preserving the contents of the input image, we employ CMP, which is defined as CMP = $(1 - \text{diff}) \cdot \text{sim}$, with diff denoting the L1 pixel difference between the input and output images, and sim denotes the CLIP semantic similarity between the output image and the target description.

3 USER STUDY

To further assess our approach and compare it with other competing approached across all three tasks, we conduct a user study using Qualtrics. In the domain adaptation task, we compare our model with DynaGAN [Kim et al. 2022a] and HyperDomainNet [Alanov et al. 2022], both of which also facilitate adaptation across multiple domains with a single model architecture, akin to ours. For reference-guided image synthesis, our comparisons include TargetCLIP-E [Chefer et al. 2022], TargetCLIP-O [Chefer et al. 2022], and BlendGAN [Liu et al. 2021]. For text-guided image manipulation, we evaluate our framework against DeltaEdit [Lyu et al. 2023], which similarly does not utilize textual textual data during its training phase.

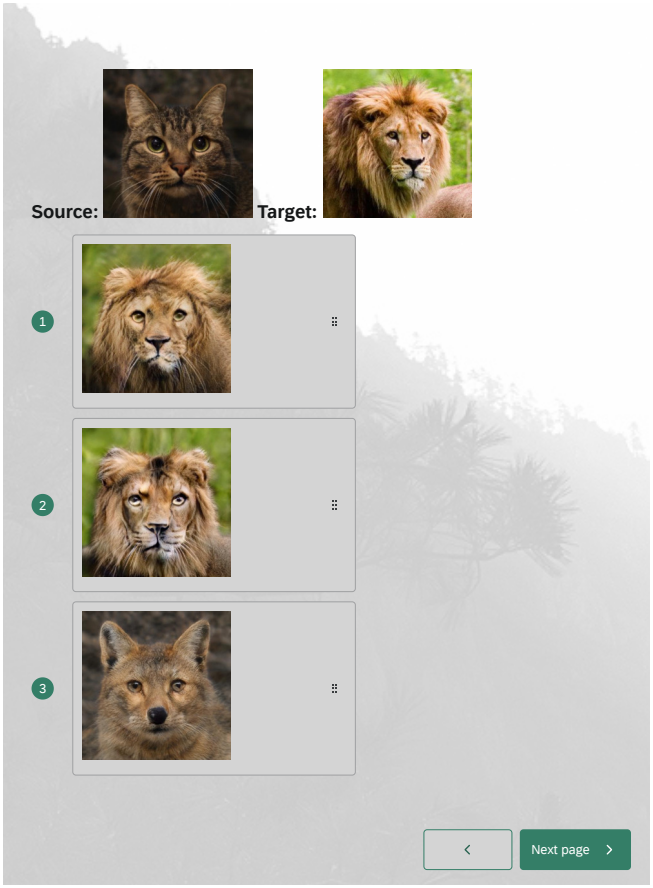


Fig. 2. **A sample question from the user study.** The participants rank the options from best to worst.

The human evaluation comprises three sections, each dedicated to one of our tasks, with 25 questions per section. Within each part, users are shown a random source image alongside a target image (or target text in the case of text-guided editing). In the domain adaptation and reference-guided editing sections, participants rank the results based on each model's performance by arranging the images in order of preference, with the top position indicating the best result. For text-guided editing section, participants choose the superior output between our model and DeltaEdit. To mitigate any bias in the evaluation process, the order in which results are displayed is randomized. An example question from the user study is illustrated in Fig. 2.

In Table 1, we present the average human rankings of the methods for all three tasks. As shown, for text-guided image manipulation, nearly all participants consistently preferred the results of HyperGAN-CLIP over those of DeltaEdit. In the domain adaptation task, HyperGAN-CLIP and DynaGAN received similar rankings, indicating comparable performance. For reference-guided image synthesis, the task appears more subjective, as all average rankings are above 2, with HyperGAN-CLIP showing competitive results against the TargetCLIP models.

Table 1. **User study results.**

Multiple Domain Adaptation	
Method	Ranking
HyperDomainNet	2.77
DynaGAN	1.58
Ours	1.65

Reference-Guided Image Synthesis	
Method	Ranking
BlendGAN	3.24
TargetClip-O	2.20
TargetClip-E	2.13
Ours	2.43

Text-Guided Image Manipulation	
Method	Ranking
DeltaEdit	1.93
Ours	1.07

4 CONTROLLABLE MANIPULATION

We observe that scaling the CLIP embeddings translates roughly to scaling of the modulation weights, and consequently, feature maps. By adjusting the scaling ratio of the residual target domain features injected into the source domain features by a factor β, we can control the degree of adaptation during inference. Furthermore, in line with previous GAN domain adaptation studies, we can enhance the style quality by employing style mixing over the latent codes. This involves interpolating the original latent code w with the target domain's latent code w_t as $\hat{w} = \alpha * w + (1 - \alpha) * w_t$, where α is a scalar between 0 and 1, controlling the level of style mixing. The resulting interpolated latent code $\hat{w}$ can be then used as input to the generator for image synthesis. We demonstrate the impact of these control parameters on the generated images in Fig. 3.

Moreover, we observe that even if we additionally scale the amount of features injected into the original image features by some factor β, our approach gives very plausible results. It does not change the features that not related to the target. We compare it with our baseline model with conditional discriminator trained with Δ-CLIP embeddings, where we use the same β to scale the modulation parameters to control the degree of adaptation (as in DynaGAN). Since, it does not use original domain features and weight modulation is responsible for preserving both original image characteristics and transferring target style content from CLIP embeddings, it fails to scale well with amount of the feature scaling parameter, as demonstrated in Fig. 4. This highlights the importance of using domain specific features as residuals to the original features instead of directly generating the overall combined features.

5 ZERO-SHOT DOMAIN ADAPTATION

The use of CLIP conditioning in the design of our proposed hypernetwork module has critical advantages over the prior methods. In this way, during training the model can not only exploit the common characteristics shared among target domains, but it also allows for zero-shot domain adaptation, especially well when the novel target domain not seen during training is semantically close to the target domains in training data. In Fig.5, we provide example results on target dog breed images from the AFHQ dog dataset not used in the training.

6 ABLATION STUDY

We conduct an ablation study to assess the impact of each component in our model on domain adaptation performance. The qualitative and quantitative results are presented in Fig. 6 and Table 2, respectively. The baseline network uses only the features given by the target-domain modulated generator, and ignores the source domain features. This approach results in the loss of the source identity and is prone to overfitting to the provided target image. Adding a conditional discriminator loss helps to mitigate the problems to some extent and enhances image quality. Considering residual features scheme that employs target features alongside with the source features preserves the facial identity better than the baseline, but falls short in terms of image quality. Finally, our full model, HyperGAN-CLIP, which utilizes residual feature scheme together with a conditional discriminator effectively preserves identity while capturing target style and maintaining high image quality.

Fig. 3. **Controllable Manipulation.** In our approach, we can vary the amount of residual features injected as well as the amount of target style latent, which gives users the ability to control degree of adaptation with respect to style consistency vs data fidelity.

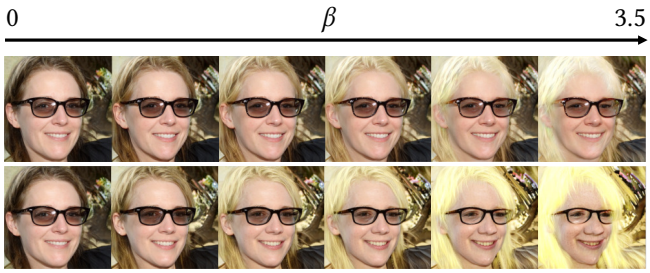


Fig. 4. **Scaling residual features.** The top row shows the results obtained with our approach, whereas the bottom on corresponds to the results by the baseline model with the discriminator. Several artifacts instantly start to appear in the baseline results when scaling beyond the training value of 1 (third column corresponds to $\beta = 1$). On the other hand, our approach works relatively seamlessly and preserves the identity better.

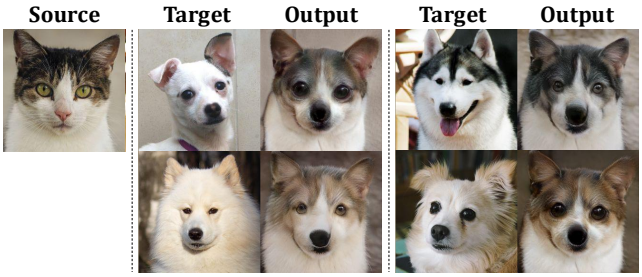


Fig. 5. **Zero-Shot Domain Adaptation.** Our model can perform domain adaption quite reasonably well on target domains not seen during training. Here we provide results on target dog breed images from the AFHQ dog dataset not used in the training.

Table 2. **Quantitative analysis of the ablation study.** The baseline model that employs target-domain modulated features gives the worst score. However, incorporating CLIP-conditioned discriminator and leveraging residual features scheme introduce notable improvements. Our full HyperCLIP-GAN model, utilizing all these components achieves the best score.

Component	FID ↓
Baseline (B)	43.43
Baseline + Cond. Disc. (B + CD)	33.76
Baseline + Res. Features (B + RF)	33.43
HyperGAN-CLIP (B + CD + RF)	**30.55**

7 IMPACT OF Δ-CLIP EMBEDDINGS

As stated in the main paper, the Δ-CLIP space provides a semantic embedding space that offers improved alignment between text and image modalities compared to the original CLIP space. This distinction becomes particularly evident when examining the text-guided image manipulation task. In Fig. 7 and Fig. 8, we show the impact of utilizing these spaces within our hypernetworks module adjusting the weights of the pre-trained StyleGAN generator. The

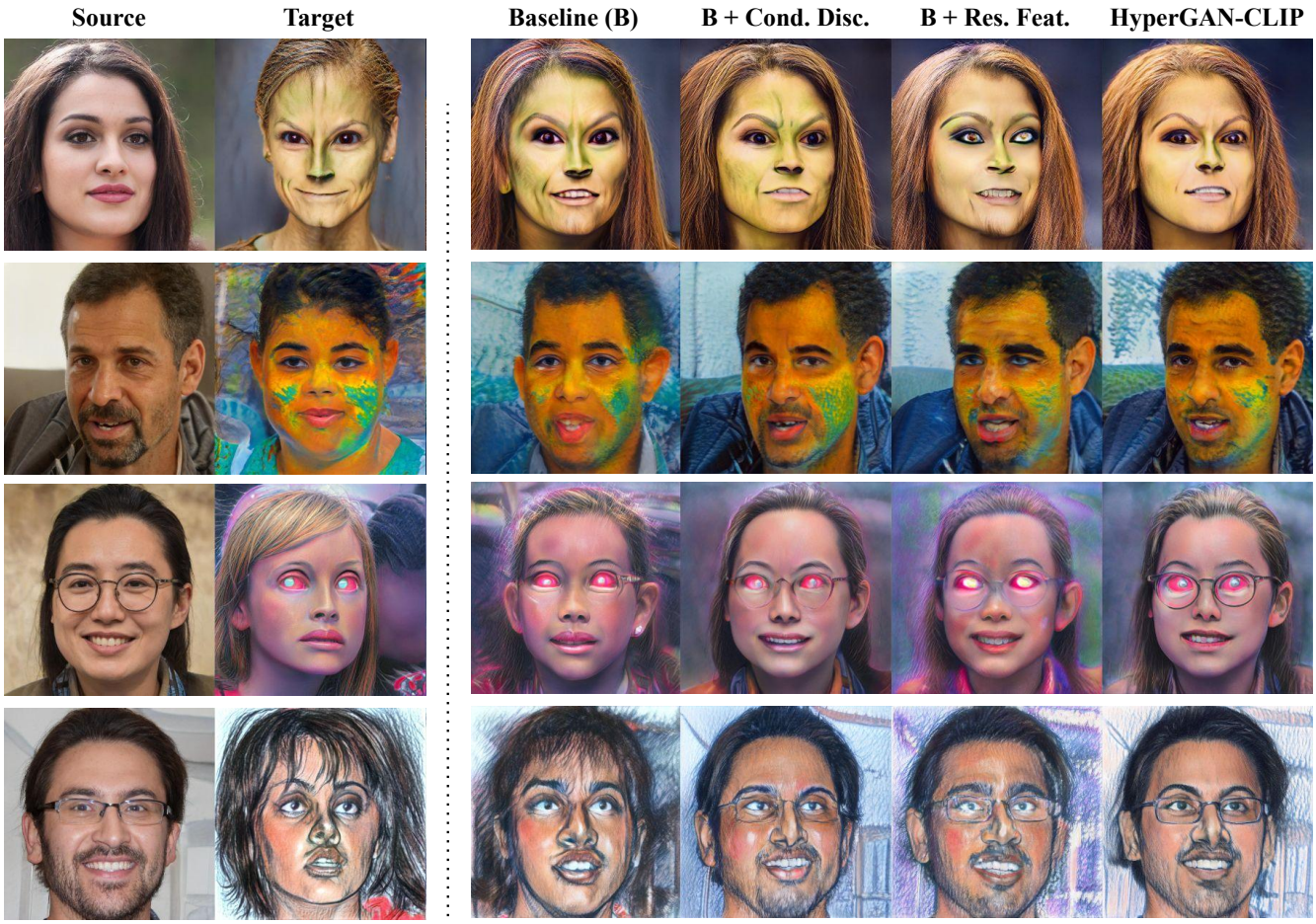


Fig. 6. **Qualitative results for the ablation study.** Baseline network does not preserve the facial identity of the source image, giving an outcome closely resembling to the target image. When CLIP-conditioned discriminator is incorporated to the baseline, the image quality is improved. Using residual features scheme preserves the facial identity better. Our full model gives the best results in terms of both identity and image quality.

results demonstrate that the Δ-CLIP embeddings enable highly precise text-based editing while preserving image quality and identity fidelity.

8 ADDITIONAL PERFORMANCE COMPARISONS

In Fig. 9 and Fig. 10, we present additional comparisons in multple domain adaptation against Mind-the-GAP [Zhu et al. 2022], StyleGAN-NADA [Gal et al. 2022], HyperDomainNet [Alanov et al. 2022], DynaGAN [Kim et al. 2022a], and Adaptation-SCR [Liu et al. 2023] on AFHQ and FFHQ datasets, respectively. In Fig. 11, we give additional comparisons of our approach against the BlendGAN [Liu et al. 2021], TargetCLIP-O [Chefer et al. 2022] and TargetCLIP-E [Chefer et al. 2022] models in reference-guided editing. Finally, in Fig. 12, we provide additional qualitative comparisons in text-driven manipulation against TediGAN-B [Xia et al. 2021], StyleCLIP-LO [Patashnik et al. 2021], StyleCLIP-GD [Patashnik et al. 2021], HairCLIP [Wei et al. 2022], DeltaEdit [Lyu et al. 2023], CLIPInverter [Baykal et al. 2023], DiffusionCLIP [Kim et al. 2022b] and plug-and-play [Tumanyan et al. 2022].

We trained our HyperCLIP-GAN on the CUB-Birds dataset [Wah et al. 2011] as well to demonstrate the generalization capabilities of our approach. When training these models, we use the same losses as described in the main paper except for the identity preservation loss, where we alternatively employ a ResNet50 [He et al. 2015] network trained with MOCOv2 [Chen et al. 2020]. In Fig. 13, we provide various text-guided editing results, and in Fig. 14, we provide several reference-guided synthesis results for the CUB-Birds dataset.

We also perform an additional analysis covering our approach and HyperDomainNet, two hypernetworks-based multi-domain adaptation approaches. In particular, we train our framework on a much smaller, less diverse set of domains involving one image per domain and consisting of 20 different domains from [Alanov et al. 2022]. In Fig. 15, we provide sample side-by-side qualitative comparisons using the pre-trained model provided by the authors.

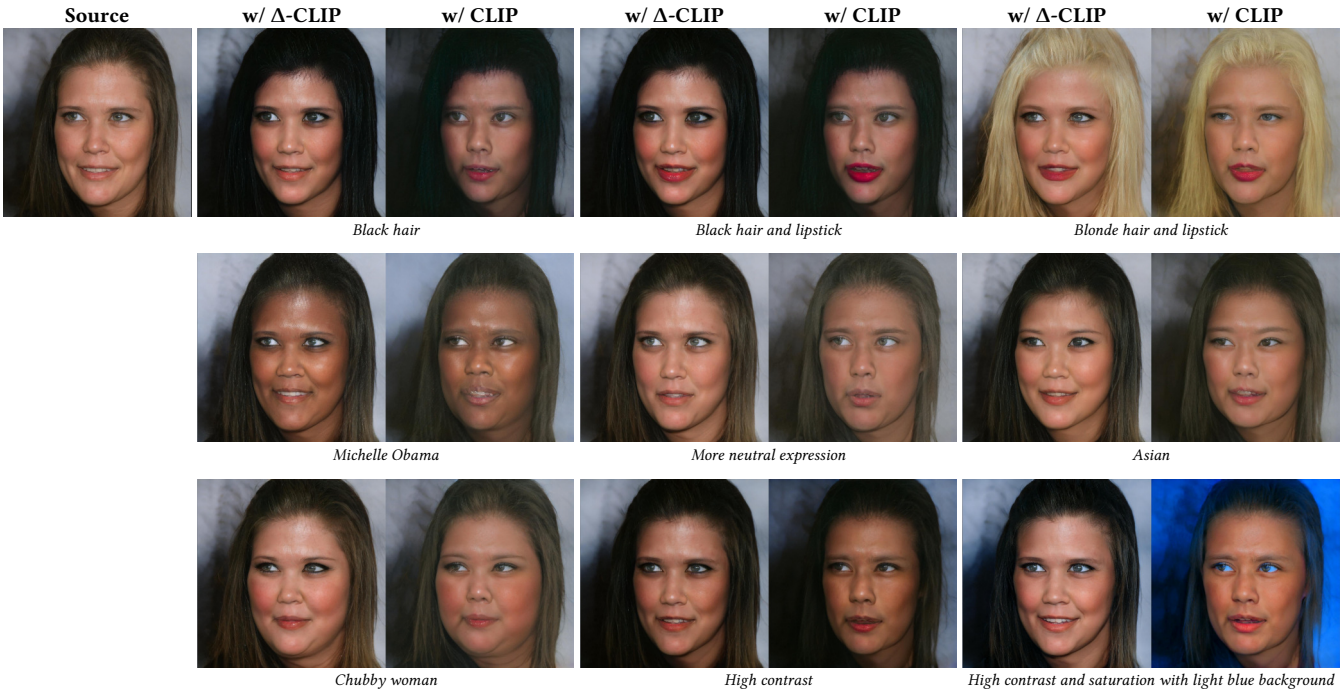


Fig. 7. **Impact of Δ-CLIP Embeddings.** Our model equipped with Δ-CLIP embeddings performs semantic edits that are better aligned with the provided textual descriptions as compared to the version of our model that employs original CLIP embeddings.

Overall, the results show that our method performs either better or on par with HyperDomainNet on this more limited set of domains.

9 LIMITATIONS

HyperGAN-CLIP performs optimally when the target domain shares a resemblance with the source domain in terms of content. However, when there is a significant domain gap, it struggles to adapt the pre-trained generator to the target domain. Additionally, for reference-guided image synthesis and text-guided image manipulation, HyperGAN-CLIP can produce visually plausible results only for concepts encountered during training or those that are semantically similar. It fails to generalize to entirely different, unseen concepts.

10 ETHICAL STATEMENT

The transformative capabilities of Generative Adversarial Networks (GANs) in image editing, particularly in the realm of human face manipulation present not only technological advancements but also ethical implications that merit careful consideration. Prominent concerns involve the potential for the creation of deceptive or harmful content, exemplified by the emergence of deepfakes, which can be exploited for malicious purposes such as disseminating misinformation or perpetuating identity theft. Moreover, biases encoded in training data may perpetuate societal prejudices. This study emphasizes responsible research practices, advocating for transparent disclosure of limitations and risks. Open dialogue within the research community is crucial for addressing these ethical implications. Implementing safeguards, including content detection methods and adherence to ethical guidelines, is essential for the responsible development and deployment of face editing technologies, ensuring a positive societal impact.

REFERENCES

Aibek Alanov, Vadim Titov, and Dmitry Vetrov. 2022. HyperDomainNet: Universal Domain Adaptation for Generative Adversarial Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Ahmet Canberk Baykal, Abdul Basit Anees, Duygu Ceylan, Erkut Erdem, Aykut Erdem, and Deniz Yuret. 2023. CLIP-Guided StyleGAN Inversion for Text-Driven Real Image Editing. *ACM Trans. Graph.* 42, 5, Article 172 (aug 2023), 18 pages.

Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. 2022. Image-Based CLIP-Guided Essence Transfer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII* (Tel Aviv, Israel). 695–711.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297 [cs.CV]

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (oct 2022), 5962–5979. https://doi.org/10.1109/tpami.2021.3087709

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Trans. Graph.* 41, 4, Article 141 (jul 2022), 13 pages.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022b. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2426–2435.

Seongtae Kim, Kyoungkook Kang, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022a. DynaGAN: Dynamic Few-shot Adaptation of GANs to Multiple Domains. In *Proceedings of the ACM (SIGGRAPH Asia)*.

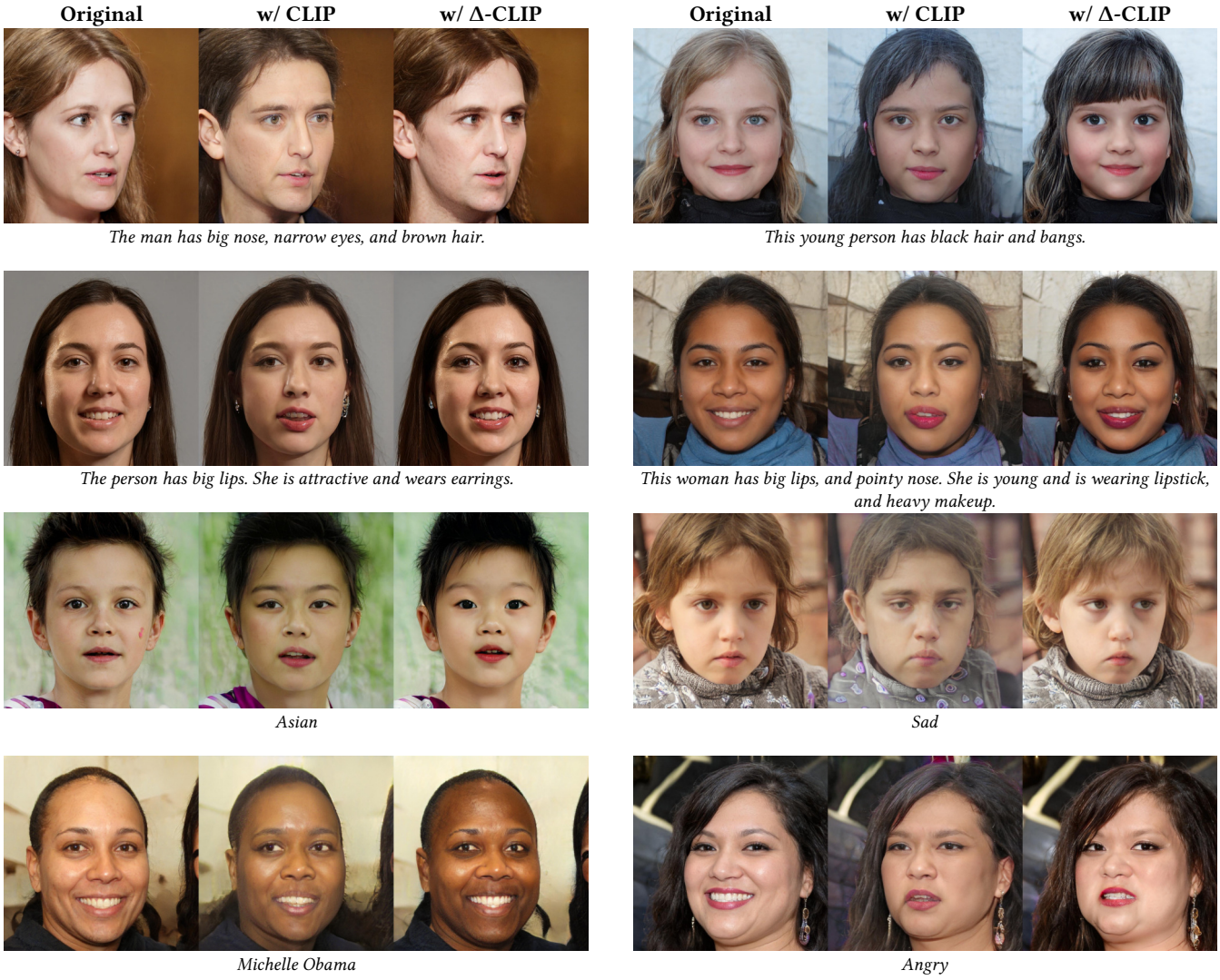


Fig. 8. **Impact of Δ-CLIP Embeddings.** As observed, we obtain much accurate manipulations while preserving the quality and fidelity when Δ-CLIP embeddings are used.

Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. 2021. BlendGAN: Implicitly GAN Blending for Arbitrary Stylized Face Generation. In *Advances in Neural Information Processing Systems*.

Zhenhuan Liu, Liang Li, Jiayu Xiao, Zheng-Jun Zha, and Qingming Huang. 2023. Text-Driven Generative Domain Adaptation with Spectral Consistency Regularization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. 2023. DeltaEdit: Exploring Text-Free Training for Text-Driven Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6894–6903.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572* (2022).

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.

Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2022. Hairclip: Design your hair by text and reference image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. 2022. Mind the Gap: Domain Gap Control for Single Shot Domain Adaptation for Generative Adversarial Networks. In *International Conference on Learning Representations*.

Fig. 9. **Additional qualitative comparison against the state-of-the-art few-shot domain adaptation methods on AFHQ dataset.** Our proposed HyperGAN-CLIP model outperforms competing methods in accurately capturing the visual characteristics of the target domains. The synthesized images exhibit a higher degree of fidelity and realism, demonstrating the effectiveness of our approach.

Fig. 10. **Additional qualitative comparison against the state-of-the-art few-shot domain adaptation methods on AFHQ dataset.** Our proposed HyperGAN-CLIP model outperforms competing methods in accurately capturing the visual characteristics of the target domains. The synthesized images exhibit a higher degree of fidelity and realism, demonstrating the effectiveness of our approach.

Fig. 11. **Additional qualitative comparison with state-of-the-art reference-guided image synthesis approaches.** Our approach effectively transfers the style of the target image to the source image while effectively preserving identity compared to competing methods.

Fig. 12. **Additional qualitative comparisons with state-of-the-art text-guided image manipulation methods.** Our model shows remarkable versality in manipulating images across a diverse range of textual descriptions. The results vividly illustrate our model's ability to accurately apply changes based on target descriptions encompassing both single and multiple attributes. Compared to the competing approaches, our model preserves the identity of the input much better while successfully executing the desired manipulations.

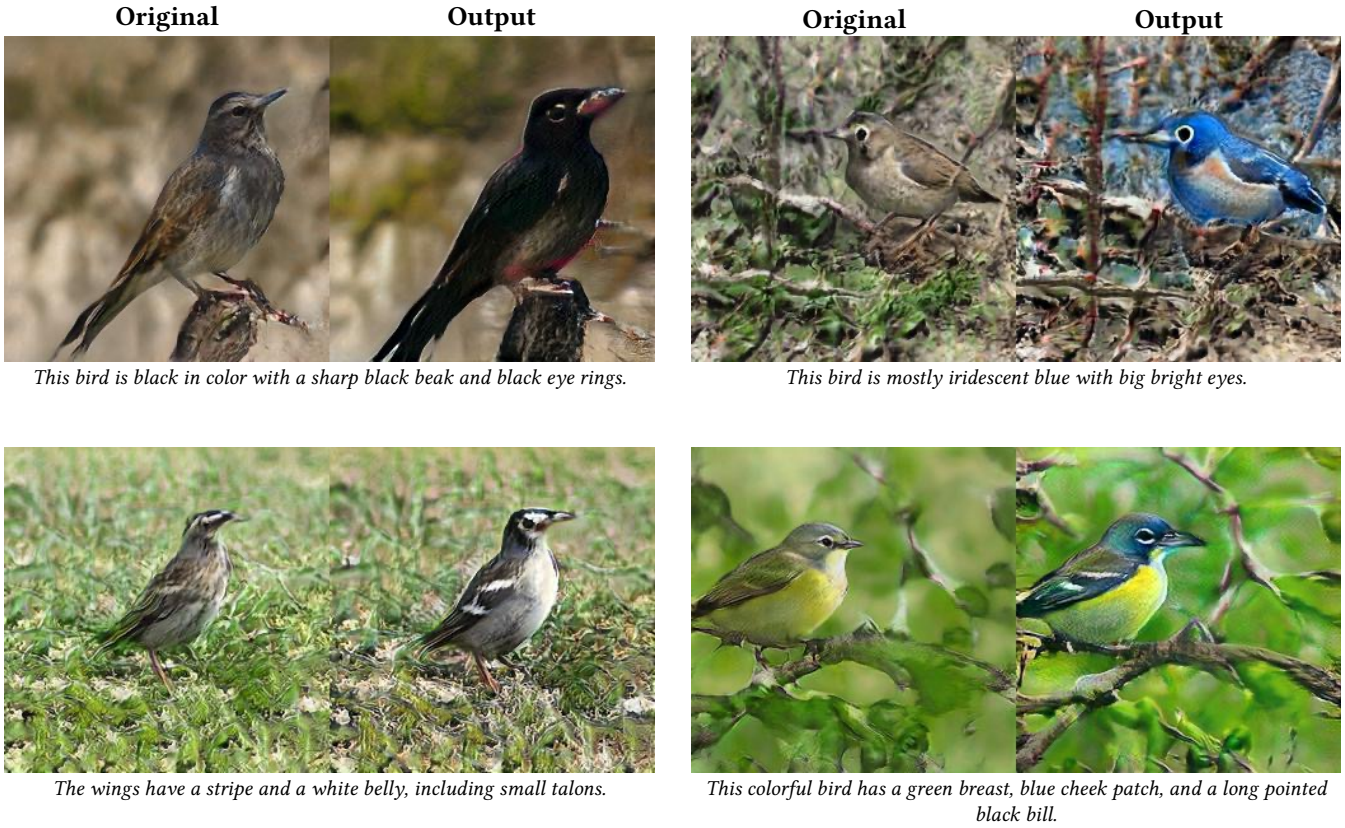


Fig. 13. **Text-guided editing results for the birds dataset.** Our approach generalizes to other domains, such as the bird images. We demonstrate zero-shot text-guided image editing results.

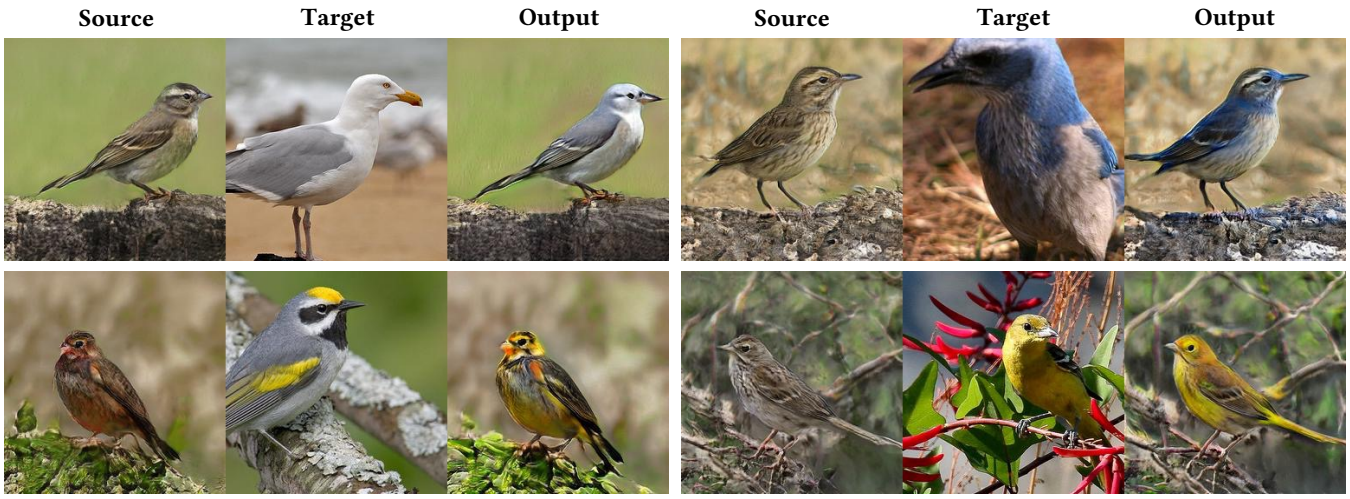


Fig. 14. **Reference-guided editing results for the birds dataset.** Our reference-guided synthesis generalizes to the birds domain, illustrated by the various targets we provide.

Fig. 15. **Additional qualitative comparison with HyperDomainNet.** The comparisons on a smaller set of domains shows that our proposed HyperGAN-CLIP model performs comparably or better than HyperDomainNet.