Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 12(4): 1091-1101 (2024)

Hierarchical Encoding for Image Inpainting with StyleGAN Inversion

Aysegul DÜNDAR[1*]

[1]*Bilkent University, Department of Computer Science, Ankara, Turkey*

Article Info

Research article
Received: 11/10/2024
Revision: 12/11/2024
Accepted: 12/12/2024

Keywords

Image Inpainting
Generative Adversarial Networks

Makale Bilgisi

Araştırma makalesi
Başvuru: 11/10/2024
Düzeltme: 12/11/2024
Kabul: 12/12/2024

Anahtar Kelimeler

Resim boyama
Üretici Karşıt Ağlar

Graphical/Tabular Abstract (Grafik Özet)

This work proposes a hierarchical encoder for image inpainting, seamlessly encoding visible and missing features. / Bu çalışma, görsel ve eksik özellikleri yüksek kalitede kodlayan bir hiyerarşik kodlayıcı öneriyor.

***Figure A**: Inpainting results of our method. Our method is able to control the inpainting style. /**Şekil A:** Yöntemimizin inpainting sonuçları. Yöntemimiz, inpainting tarzını kontrol edebilme yeteneğine sahiptir.*

Highlights (Önemli noktalar)

- *A novel hierarchical encoder is proposed to seamlessly encode both visible and missing image features for more effective inpainting. / Görünür ve eksik görsel özellikleri sorunsuz bir şekilde kodlamak için yeni bir hiyerarşik kodlayıcı önerilmiştir.*
- *The work introduces a single-stage architecture that encodes both low-rate and high-rate latent features, optimizing the use of StyleGAN for inpainting tasks. / StyleGAN için iç boyama görevlerini optimize etmek amacıyla, düşük ve yüksek oranlı latent özellikleri kodlayan tek aşamalı bir mimari tanıtılmıştır.*
- *Extensive experiments show improvements over state-of-the-art models. / Kapsamlı deneyler, önerilen yöntemin mevcut en son modellere göre önemli iyileştirmeler sağladığını göstermektedir.*

***Aim (Amaç):** The aim of this work is to improve image inpainting by proposing a novel approach that seamlessly encodes both visible and missing image features. / Bu çalışmanın amacı, görsel ve eksik özellikleri yüksek kalitede kodlayan yeni bir yaklaşım önererek iç boyama işlemini geliştirmektir.*

***Originality (Özgünlük):** The originality of this work lies in the hierarchical encoder and single-stage architecture. / Bu çalışmanın özgünlüğü, hiyerarşik kodlayıcı ve tek aşamalı mimaride yatmaktadır.*

***Results (Bulgular):** The results of this work show significant improvements in image inpainting performance compared to state-of-the-art models. / Bu çalışmanın sonuçları, mevcut en son modellere göre iç boyama performansında önemli iyileştirmeler göstermektedir.*

***Conclusion (Sonuç):** The conclusion of this work is that the proposed hierarchical encoder and single-stage architecture significantly enhance the effectiveness of image inpainting. / Bu çalışmanın sonucu, önerilen hiyerarşik kodlayıcı ve tek aşamalı mimarinin, iç boyama etkinliğini önemli ölçüde artırdığıdır.*

*Corresponding author, e-mail: adundar@cs.bilkent.edu.tr
DOI: 10.29109/gujsc.1563933

GU J Sci, Part C, 12(4): 1091-1101 (2024)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Hierarchical Encoding for Image Inpainting with StyleGAN Inversion

Aysegul DÜNDAR[1*]

[1]*Bilkent University, Department of Computer Science, Ankara, Turkey*

Article Info

Research article
Received: 11/10/2024
Revision: 12/11/2024
Accepted: 12/12/2024

Keywords

Image Inpainting
Generative Adversarial Networks

Abstract

Image inpainting, the process of removing unwanted pixels and seamlessly replacing them with new ones, poses significant challenges requiring algorithms to understand image context and generate realistic replacements. With applications ranging from content generation to image editing, image inpainting has garnered significant interest. Traditional approaches involve training deep neural network models from scratch using binary masks to identify regions for inpainting. Recent advancements have shown the feasibility of leveraging well-trained image generation models, such as StyleGANs, for inpainting tasks. However, effectively embedding images into StyleGAN's latent space and addressing the challenges of diverse inpainting remain key obstacles. In this work, we propose a hierarchical encoder tailored to encode visible and missing features seamlessly. Additionally, we introduce a single-stage architecture capable of encoding both low-rate and high-rate latent features used by StyleGAN. While low-rate latent features offer a comprehensive understanding of images, high-rate latent features excel in transmitting intricate details to the generator. Through extensive experiments, we demonstrate significant improvements over state-of-the-art models for image inpainting, highlighting the efficacy of our approach.

StyleGAN Tersine Çevirisi ile Görsel İç Boyama için Hiyerarşik Kodlama

Makale Bilgisi

Araştırma makalesi
Başvuru: 11/10/2024
Düzeltme: 12/11/2024
Kabul: 12/12/2024

Anahtar Kelimeler

Resim boyama
Üretici Karşıt Ağlar

Öz

Görsel iç boyama, istenmeyen piksellerin kaldırılması ve bunların yerini yeni piksellerle sorunsuz bir şekilde doldurma süreci, algoritmaların görsel bağlamı anlamasını ve gerçekçi yerine koymalar üretmesini gerektiren önemli zorluklar sunar. İçerik üretiminden görsel düzenlemeye kadar pek çok uygulama alanı bulunan görsel iç boyama, önemli bir ilgi görmüştür. Geleneksel yaklaşımlar, iç boyama için bölgeleri belirlemek amacıyla ikili maskeler kullanarak derin sinir ağı modellerinin sıfırdan eğitilmesini içerir. Son gelişmeler, iyi eğitilmiş görsel üretim modellerinin (örneğin, StyleGAN'ler) iç boyama görevlerinde kullanılabilirliğini göstermiştir. Ancak, görselleri StyleGAN'ın latent uzayına etkin bir şekilde yerleştirme ve çeşitli iç boyama zorluklarını aşma, hala ana engelleri oluşturmaktadır. Bu çalışmada, görünür ve eksik özellikleri sorunsuz bir şekilde kodlamak için tasarlanmış hiyerarşik bir kodlayıcı önermekteyiz. Ayrıca, StyleGAN tarafından kullanılan düşük ve yüksek oranlı latent özellikleri kodlayabilen tek aşamalı bir mimari tanıtmaktayız. Düşük oranlı latent özellikler, görsellerin kapsamlı bir şekilde anlaşılmasını sağlarken, yüksek oranlı latent özellikler, karmaşık detayların üreticiye iletilmesinde mükemmel sonuçlar elde etmektedir. Bu makalede, kapsamlı deneylerle, iç boyama için mevcut en son modellere göre önemli iyileştirmeler sağladığımızı ve yaklaşımımızın etkinliğini vurgulamaktayız.

1. INTRODUCTION (GİRİŞ)

Image inpainting involves the removal of unwanted pixels and their replacement with new ones in a manner that renders the alterations indistinguishable. This task presents significant challenges, requiring the algorithm to comprehend the context of the image based on the available unerased partial data and generate new pixels that seamlessly blend with the surrounding content. Due to its broad range of applications spanning from content generation to image editing, image inpainting algorithms have garnered considerable interest. The complexity of the task, coupled with its potential for enabling various applications, has made it a popular subject of research [1–9].

Traditionally, deep neural network models have been trained from scratch for image inpainting tasks. Binary masks are employed to delineate the

*Corresponding author, e-mail: adundar@cs.bilkent.edu.tr
DOI: 10.29109/gujsc.1563933

regions to be erased, after which images are multiplied by these masks to nullify the unwanted pixels. Leveraging paired data comprising erased and original images, networks are trained using various loss objectives, including pixel-wise reconstruction and adversarial losses [1–5, 10]. While previous models were randomly initialized and trained from scratch for this task, recent approaches demonstrate the feasibility of leveraging well-trained image generation models for inpainting tasks [6, 9]. These image generation models are trained on large-scale image datasets with substantial computational resources [7, 11– 15]. They possess the capability to generate images with realistic details, indicating rich feature representations and a robust implicit understanding of images, rendering them suitable candidates for inpainting tasks.

Among generative models, StyleGANs [11, 13] have been extensively explored for both image editing [16–23] and inpainting tasks [9, 24, 25]. Successfully utilizing StyleGAN for editing and inpainting presents a key challenge: correctly embedding a given image into StyleGAN's latent space so that the input image can be reconstructed via StyleGAN from the embedded vector. In the case of inpainting, there is an additional challenge: embedding the erased image into StyleGAN's latent space. While previous methods only embedded erased images and inpainted them with StyleGAN in a deterministic manner [24, 25], a recent method demonstrates the possibility of achieving diverse results by augmenting the embedded latent vectors with sampled ones [9]. Therefore, for diverse inpainting, another challenge arises: the encoder must encode the visible features of the input image while also being aware of the missing ones, allowing the new sampled codes to complete the features. For instance, when erasing a person's hair, the encoder must encode all the facial features except the hair and be able to incorporate hair features from the sampled codes. This enables different hair colors and styles to be encoded among different samples, resulting in the generation of various images, as illustrated in Figure 1.

Figure 1. Inpainting results of our method. Our method is able to control the inpainting style via InterFaceGAN directions [18] for StyleGAN. (Yöntemimizin iç boyama sonuçları. Yöntemimiz, StyleGAN için InterFaceGAN yönlendirmeleri [18] aracılığıyla iç boyama tarzını kontrol edebilme yeteneğine sahiptir.)

In this work, we introduce a hierarchical encoder tailored to the intricate task of encoding visible features while seamlessly integrating missing features from sampled ones. Furthermore, we propose a single-stage architecture that encompasses encoding both low-rate and high-rate latent features utilized by StyleGAN. While low-rate latent features possess a comprehensive understanding of images, they may not capture all intricate details due to an inherent information bottleneck. While high-rate latent features may lack robust feature extraction capabilities necessary for understanding the context required for realistic inpaintings, they excel in transmitting intricate image details to the generator.

In summary, our main contributions are as follows:

- We present a novel hierarchical encoder designed specifically for the complex task of encoding visible features while seamlessly integrating missing features obtained from sampled data. This encoder is tailored to address the challenges inherent in inpainting tasks, where the reconstruction of images necessitates a

comprehensive understanding of both visible and missing features.

- We propose a single-stage architecture capable of encoding both low-rate and high-rate latent features utilized by StyleGAN. While low-rate latent features offer a thorough understanding of images, they may lack the capacity to capture fine details due to inherent limitations of the low dimension. Conversely, high-rate latent features, while potentially deficient in robust feature extraction capabilities necessary for contextual understanding, excel in transmitting high-frequency image details to the generator, thereby enhancing the realism of the inpainted images.
- We conduct comprehensive experiments to evaluate the effectiveness of our framework. Our results demonstrate significant improvements over state-of-the-art models for image inpainting, showcasing the efficacy of our approach in addressing the challenges inherent in this task.

2. RELATED WORK (İLGİLİ ÇALIŞMALAR)

Image inpainting is a widely studied area owing to its diverse applications in image editing, object removal, and image extension [3, 26–29]. Effective inpainting requires a thorough understanding of context, prompting researchers to propose various techniques such as contextual encodings and semantic attention modules. These methods aim to guide the generation of erased pixels based on the valid (unerased) pixels [1, 4, 28, 30–32]. Additionally, specialized convolutional layers have been introduced to handle valid and invalid (erased) pixels differently, enhancing the encoding of valid pixel information [2, 3, 5]. Various approaches, including sketches, brush strokes, and semantic masks, have also been explored to guide image inpainting algorithms [3, 10, 26, 27]. These methods typically employ an encoder-decoder structure trained from scratch for the inpainting task.

In recent years, significant progress has been made in image generation methods, specifically by GANs, which involve sampling a point from a Gaussian distribution and learning to map it to a realistic image [7, 11–15]. This progress has sparked interest in utilizing the image generation capabilities of these networks for inpainting tasks.

Initially, architectural inspiration was drawn from these methods. For instance, CoModGAN [33] proposed a StyleGAN-like architecture trained for inpainting, demonstrating successful diverse inpainting results while still being trained from scratch for the task. Subsequently, well-trained image generation algorithms, especially StyleGANv2, have been directly applied to inpainting [9, 24, 25]. Among these approaches, Yu et al. [24] learned an encoder to project images into the W+ space of StyleGAN. The encoder predicts W+ from erased images, from which an image is generated via StyleGAN, minimizing pixel-wise and feature-wise distances between the generated and unerased original image. This leads to deterministic outputs since only one W+ code is encoded for erased images. Wang et al. [25] adopted a similar approach, embedding the image into a deterministic latent code using an encoder.

Our work is primarily related to diverse inpainting with GAN inversion [9]. Yildirim et al. [9] demonstrated that encoded latent codes can be mixed with sampled ones to achieve diversity in inpainting results. While their method uses a feed-forward architecture, we employ a hierarchical approach to achieve diverse inpainting with higher-quality results. Additionally, our single stage architecture is more efficient and achieves better results.

3. MATERIALS AND METHODS (MATERYAL VE METOD)

3.1. Architecture (Mimari)

In this work, we develop a StyleGAN-based image inpainting model. Our approach incorporates two key mechanisms to enhance the model's performance: augmenting missing features through the mapping network and facilitating seamless inpaintings by allowing high-rate latent features to bypass from the encoder directly to the StyleGAN generator. This section provides an in-depth explanation of both contributions.

Firstly, like previous inpainting methods [5, 9], we generate binary masks M with 1s defining the valid pixels and 0s defining the pixels we would like to erase. An input image, I, is erased by replacing the pixels we would like to erase with 0s. This is achieved by simply taking the dot product of mask and input image, $I_e = M \odot I$. We follow the approach of Yildirim et. al. [9] and use StyleGAN2's [13] mapping and generator networks to sample latent codes and synthesize images, respectively. These modules are pretrained and we do not tune them in our trainings.

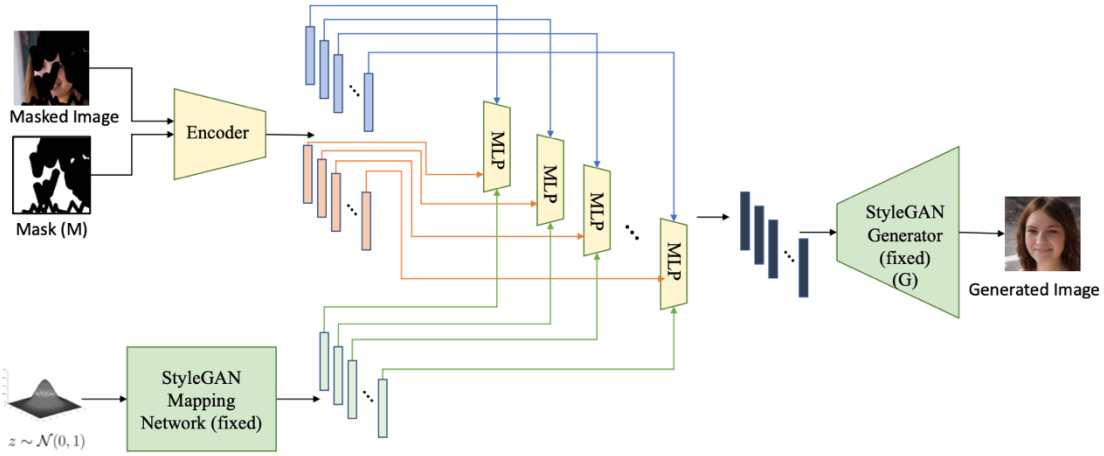

Figure 2. Overview of the proposed hierarchical image encoding framework. Through a hierarchical mixing mechanism, our model integrates visible information from masked images with missing details derived from mapped latent codes. The method sequentially generates output styles, ensuring awareness of previously determined styles. (Önerilen hiyerarşik görsel kodlama çerçevesinin genel bakışı. Hiyerarşik bir karıştırma mekanizması aracılığıyla, modelimiz maskelenmiş görsellerden görünür bilgiyi, haritalanmış latent kodlardan türetilen eksik detaylarla birleştirir. Yöntem, çıktıları sırasıyla üretir ve daha önce belirlenen stillerin farkında olmayı sağlar.)

Our encoder receives the erased image along with its corresponding mask, as illustrated in Fig. 2. We employ a straightforward encoder architecture based on [34] to map our erased images onto the latent space W^+. Nonetheless, merely encoding the erased images might not suffice for effective inpainting. For instance, if the erased regions correspond to essential facial features such as eyes or hair, the encoded features may lack crucial information about these elements. Hence, we establish a secondary pathway to supplement the absent features from the erased image. This secondary pathway leverages StyleGAN's mapping network to sample random z vectors and derive corresponding W^+ codes, which naturally align with StyleGAN's latent space. The encoded latent code, W^{enc}, and the mapped latent code, W^{rand}, are both inputted into the mixing mechanism as shown in Figure 2.

In this paper, we propose a novel hierarhical mixing mechanism different than previous works. We anticipate the mixing network to integrate the visible information from the masked image with the absent details derived from the mapped latent codes. To mix the 14 × 512 sampled W^{rand} with the 28 × 512 encoded W^{enc}, we set 14 MLP (Multi-layer perceptron) layers. We employ a 28 × 512 feature encoding scheme with a specific purpose in mind. Among these features, half (14 × 512) are dedicated to representing the image features themselves, while the remaining half are tasked with indicating which features have been accurately encoded and which have been affected by image erasure. Therefore, for each style code embedding, we use a single channel from W^{rand} and two channels from W^{enc}. Initially, we produce W^{out} for the coarse layers, representing the style codes directed towards StyleGAN's lowest resolution feature layers. Precisely, we first generate the W_1^{out} intended for the initial adaptive instance normalization layer of StyleGAN via a MLP that takes the first channel of W^{rand} and first two channels of W^{enc}, W_1^{rand} and $W_{1:2}^{enc}$, respectively. After that, in addition to the second channel of W^{rand} and the next two channels of W^{enc}, W_2^{rand}, and $W_{3:4}^{enc}$, we feed W_1^{out} to the next linear layer. By proceeding in this manner, the mixing mechanism remains cognizant of the styles that have been determined thus far. We generate the subsequent output styles sequentially using this approach. In our experiments, we show that using this hierarchical encoding greatly improves the results. MLP consists of two-layer linear networks. The first linear layer receives 3 × 512 input features and outputs 1 × 512 features. There is a ReLU activation following the first linear layer and the second linear layer has input and output dimensions set to 1 × 512.

Our second contribution is to achieve seamless inpaintings by allowing high-rate latent features to bypass from the encoder directly to the StyleGAN generator. Previous methods leveraging StyleGAN for image editing [22, 34] and inpainting [9] recognize the limitations of using low-rate latent codes of W^+. These codes often lack sufficient information to fully represent the image for the generator. The W^+ codes, being of size 14 × 512,

are considerably smaller than the input image size of 3 × 256 × 256. This information bottleneck often leads to the loss of many details in the final generation if only W+ codes are employed. To address this issue, we utilize a skip encoder network, depicted in Figure 3, to convey high-detail image information into the generation pipeline. Previously, to address the same issue, DivInv [9] proposed taking the generated image from StyleGAN and feeding it into a second encoder and generation pipeline to achieve the final results. This second encoder and generation process incorporated skip connections. In this study, we introduce a more streamlined architecture, delivering high-rate features to the StyleGAN generator within a single-stage framework. Unlike previous methods, we eliminate the need to generate images first and encode them again. Instead, we leverage the pretrained encoder and invoke the StyleGAN generator only once. We do that via a UNet architecture of Skip encoder operating within a spatial dimension of 64 × 64. By simultaneously inputting both the encoded and generated features, the Skip encoder can identify the absent high-rate details from the image generation and incorporate them into the generation process. This feature space is referred to as F^+ in the ablation study. The output of the skip encoder serves as the features for StyleGAN. We achieve that by replacing the 64 × 64 generated features with the output of the Skip encoder.

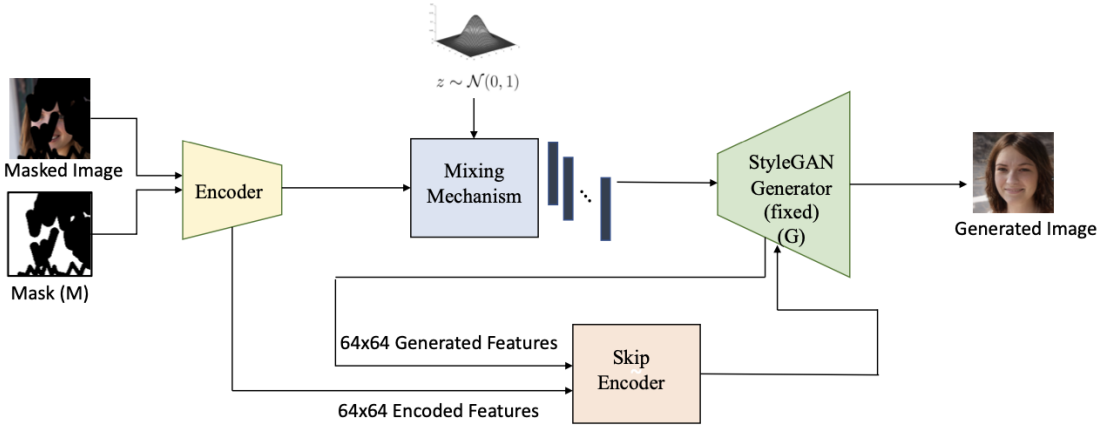

Figure 3. Architecture overview: Our study introduces an architecture, delivering high-rate features to the StyleGAN generator in a single-stage framework. Unlike prior methods, we bypass the generation-then-encoding process, instead leveraging a pretrained encoder and invoking the StyleGAN generator once. This is facilitated by the Skip encoder, operating within a spatial dimension of 64 × 64. By simultaneously inputting both encoded and generated features, the Skip encoder detects missing high-rate details in image generation, integrating them into the process. The output of the Skip encoder replaces the 64 × 64 generated features, serving as input features for StyleGAN. The detailed depiction of the Mixing mechanism in this figure has been omitted for brevity. For a comprehensive understanding of the Mixing mechanism, please refer to Figure 2. (Mimari Genel Bakışı: Çalışmamız, StyleGAN üreticisine yüksek oranlı özellikleri tek aşamalı bir çerçevede sunan bir mimari tanıtmaktadır. Önceki yöntemlerden farklı olarak, üretim-sonra-kodlama sürecini atlıyor ve bunun yerine önceden eğitilmiş bir kodlayıcı kullanarak StyleGAN üreticisini bir kez çağırıyoruz. Bu, 64 × 64 boyutunda bir uzaysal boyutta çalışan Skip kodlayıcı ile sağlanmaktadır. Hem kodlanmış hem de üretilmiş özellikleri aynı anda girdiler olarak vererek, Skip kodlayıcı, görsel üretimindeki eksik yüksek oranlı detayları tespit eder ve bunları sürece entegre eder. Skip kodlayıcısının çıktısı, 64 × 64 boyutundaki üretilmiş özelliklerin yerine geçer ve StyleGAN için giriş özellikleri olarak kullanılır. Bu şekildeki Mixing mekanizmasının ayrıntılı tasviri kısalık açısından çıkarılmıştır. Mixing mekanizmasının kapsamlı bir şekilde anlaşılması için lütfen Şekil 2'ye bakınız.)

3.2. **Training Objectives** (Mimari)

We train the framework with a combination of reconstruction and adversarial losses. Our framework outputs a generated image, I^o. Additionally, we obtain a final image by $I^f = M \odot I + (1 - M) \odot I^o$. The resulting image guarantees that unerased pixels remain unchanged throughout the process, as they are directly sourced from the input image. We adhere to the training pipeline outlined by Yildirim et al. [9], which comprises two image generation stages. Initially, we sample a latent code z and generate an image, which is then subjected to erasure before being fed into the encoder. Subsequently, we either utilize the same z for the mapping network destined for the mixing mechanism or sample a new z. In the first scenario, where the same z is used, the model has access to the image features and is tasked with faithfully reproducing the original image pixel by pixel. Conversely, in the second scenario, a different z is sampled, and the model is solely expected to

accurately generate the unerased pixels while producing realistic overall images. We refer to the first and second settings network generations as I_g^o and I_r^o, respectively.

Reconstruction Losses. To ensure accurate pixel reconstruction, we employ a combination of L2 and perceptual losses. Specifically, we utilize perceptual losses derived from VGG (Φ) across various feature layers (j) when comparing images. For I_g^o, where the objective is to generate overall images, losses are applied to each pixel individually. Conversely, for I_r^o, which is tasked solely with faithfully reproducing unerased pixels, a mask is applied to exclude losses contributed from erased pixels as given below.

$$L_{rg} = ||I_g^o - I_g||_2 + ||\phi_j(I_g^o) - \phi_j(I_g)||_2$$

$$L_{rr} = ||(M \odot I_r^o) - I_g^e||_2 + ||\phi_j(M \odot I_r^o) - \phi_j(I_g^e)||_2$$

Adversarial Losses. We anticipate that these final images should exhibit realism, therefore, we employ adversarial guidance on both I_g^f and I_r^f. To achieve this, we utilize the pretrained discriminator from StyleGAN training, denoted as D, and train it alongside the encoder and mixing network.

$$L_{adv} = 2\,log\,D(I_g) + log(1 - D(I_g^f)) + log(1 - D(I_r^f))$$

Full Objective. Our final objective is the weighted sum of the adversarial and reconstruction losses as given below.

$$\lambda_a L_{adv} + \lambda_{rg} L_{rg} + \lambda_{rr} L_{rr}$$

We use the same training hyperparameters as Yildirim et al. [9] without any tuning, in order to emphasize the improvements resulting solely from architectural enhancements. The parameters are λ_a= 8 × 10−2, λ_{rg}= 1, and λ_{rr} = 1.

3. EXPERIMENTS (DENEYLER)

Dataset and Metric. We utilize the FFHQ human face dataset by Karras et al. [11], employing both their train and validation splits.

To evaluate the models, we use masks of varying sizes to control the percentage of the image that is erased. A mask with a range of 0 means the input image is not erased at all, while a range of 1.0 indicates the entire image is erased. The mask size determines the difficulty of the task: inpainting images with smaller erased regions is easier, while larger masks make the task more challenging. To assess different scenarios, we use three mask settings: an "easy" setting with mask ranges from 0.0 to 0.4, a "hard" setting with ranges from 0.4 to 1.0, and a third setting where the full range (0.0-1.0) is used. We generate the masks one time for the validation set and use them in all our evaluations.

For evaluation metrics, we employ the Frechet Inception Distance (FID) [35] to assess realism, comparing the distribution of target images with inpainted images. If the inpainting is successful, there should not be visible boundaries between the erased and unerased pixels and the images should look like realistic faces since they are trained on FFHQ. FID is an important metric to assess the model's performance on these inpainting requirements.
We also evaluate the performance using the Learned Perceptual Image Patch Similarity (LPIPS) score [37], which compares the ground-truth original images with the inpainted images in feature level.

Baselines. To begin, we benchmark our method against state-of-the-art image inversion techniques including pSp [34], HFGI [17], and HyperStyle [36]. Utilizing the authors' released code, we train these models for inpainting tasks, augmenting the input with an additional channel for masks. The pSp model generates W+ predictions for image generation. In contrast, HFGI and HyperStyle employ a two-stage training approach. Initially, an encoder produces W+ predictions, followed by a second encoder that processes both the input image and StyleGAN-generated image with W+ predictions. The objective is to encode missing information into higher-rate latent codes. Subsequently, we conduct experiments with state-of-the-art image inpainting models for further comparisons. We perform inferences using CoModGAN's pretrained models [33], which propose training a StyleGAN-like model from scratch with co-modulation and skip connections tailored for inpainting tasks. Additionally, we utilize InvertFill [24] and DivInv [9] which are based on pretrained StyleGAN models. While DivInv serves as the closest comparison to our work, we surpass its performance with our enhanced hierarchical encoding architecture and streamlined single-stage high-rate feature bypassing. Our approach achieves superior results compared to theirs.

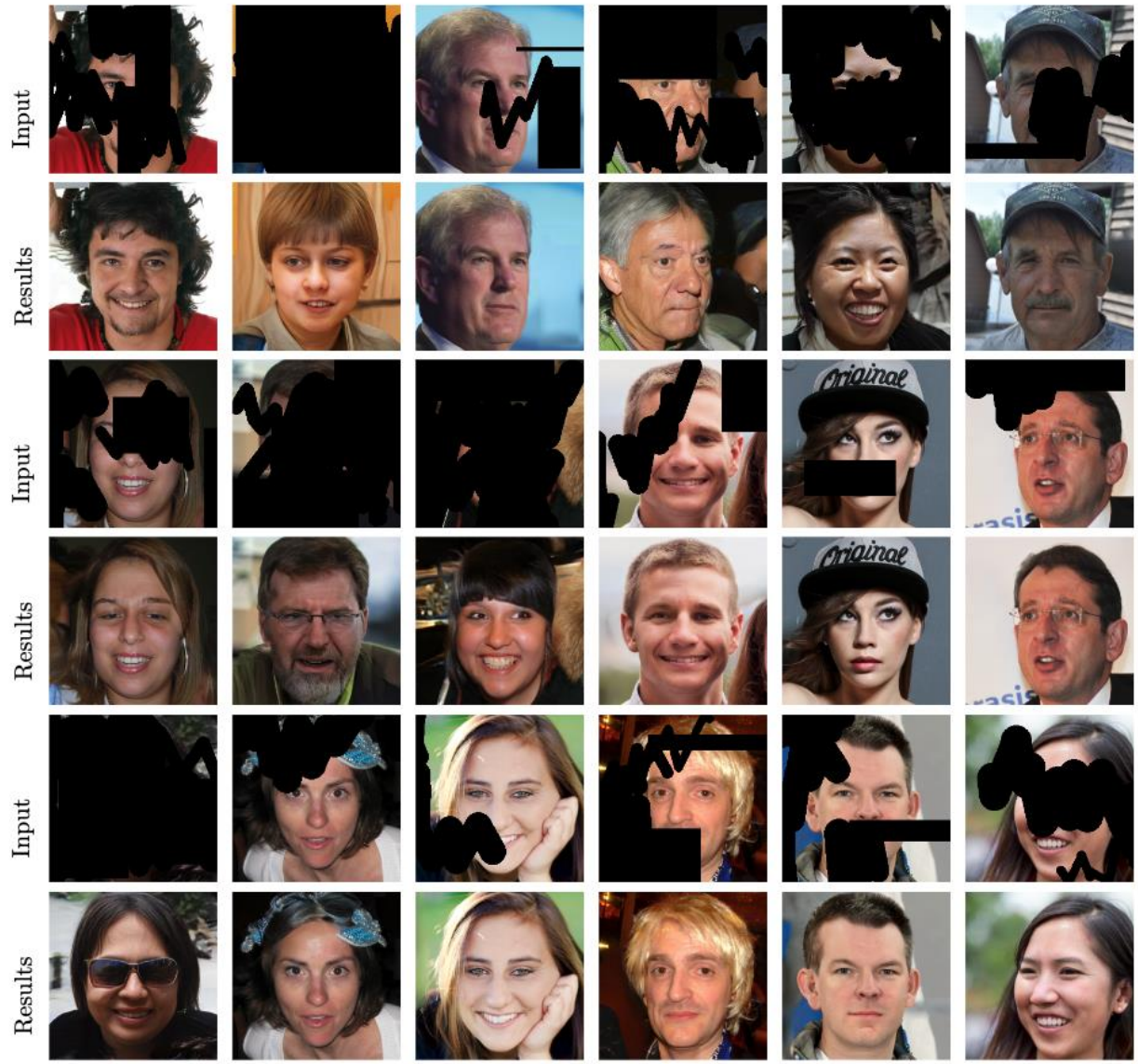

Figure 4. Inpainting results of our method. (Yöntemimizin iç boyama sonuçları.)

Qualitative Results. Fig. 4 shows the results of our methods on the FFHQ valdiation dataset. Our method achieves realistic inpainting results with small and large masks even when the mask is almost as large as the full image as shown in the second example from first row.

We also present inpainting and editing results in Figure 1. The images in the second row are generated by inpainting, following the framework outlined in Figure 2. The encoder and mapping networks produce, W^{enc} and W^{rand}, respectively. These codes pass through the MLP layers to generate the final W^{out} , which is used in the StyleGAN generator to produce the resulting images. In the third row, we apply edits to these W^{out} codes. Specifically, we leverage the directions learned by InterFaceGAN [18], which are derived from an SVM trained in the W+ latent space of StyleGAN for attributes like hair color. The SVM is fitted using latent samples from images pf people with blonde and non-blonde hair, yielding a direction vector that can modify the hair color to blonde. In the third row's results, we apply a step in this direction by adding the direction vector to the W^{out} from the second row. As a result, the hair color changes to blonde. Other than the addition we perform on W^{out}, we follow the exact same set-ups, and the features go through the skip encoder in the same way. Since we use a pretrained StyleGAN, our method seamlessly incorporates its existing editing capabilities within the inpainting pipeline.

Quantitative Results. We present the quantitative results in Table 1, where we compare our model with competing methods across three evaluation scenarios, as detailed in the Dataset and Metrics section. When considering all mask types—both

easy and difficult—the models designed for GAN inversion (pSp, HFGI, HyperStyle) perform worse than others, despite being trained for the inpainting task. On the other hand, models specifically designed for inpainting, such as CoModGAN and InvertFill, perform better, though still significantly worse than our approach. DivInversion is the closest competitor to our method, but we achieve nearly a 20% improvement, reducing the FID score from 5.92 to 4.85. A similar improvement is observed in the LPIPS score.

When the erased area is small (i.e., for easier masks), deterministic models like pSp, HFGI, HyperStyle, and InvertFill perform well, as they can recover most of the information from the unerased regions. In this scenario, InvertFill outperforms our model, with both models surpassing CoModGAN and DivInversion. Our method achieves the second-best result, coming close to InvertFill's performance. However, as the difficulty level increases, deterministic models begin to struggle, while our approach continues to outperform them by significant margins.

Table 1. Quantitative results of our and competing methods on FFHQ validation dataset. Best results are highlighted in bold. (FFHQ doğrulama veri kümesindeki bizim ve rakip yöntemlerin nicel sonuçları. En iyi sonuçlar kalın yazı ile vurgulanmıştır.)

	All Masks (0.0-1.0)		Easy Masks (0.0-0.4)		Difficult Masks (0.4-1.0)	
Models	FID	LPIPS	FID	LPIPS	FID	LPIPS
pSp[34]	8.23	0.272	3.15	0.164	9.63	0.374
HFGI [17]	7.66	0.214	2.24	0.152	8.87	0.352
HyperStyle [36]	7.46	0.208	3.08	0.159	8.67	0.337
CoModGAN [33]	7.35	0.151	4.68	0.170	7.13	0.230
InvertFill [24]	7.45	0.152	**1.13**	**0.123**	9.58	0.235
DivInversion [9]	5.92	0.153	2.26	0.145	6.23	0.223
Ours	**4.85**	**0.144**	1.87	0.135	**5.72**	**0.209**

Ablation Study. We present the results of our ablation study in Table 2. Our work starts with DivInversion [9] and proposes a hierarchical encoder and single-stage architecture. We start presenting the results of DivInversion - First stage model which only encodes features in W+ space. This set-up is comparable with our hierarchical W+ encoding. Hierarchical encoder improves the FID from 16.65 to 13.61. Next, we compare the methods that also incorporate feature encodings in F+ space. DivInversion as well as many other methods propose a two-stage architecture, which goes through StyleGAN generator twice. First, they generate an image from W+ encoding, and then the second encoder takes this generated image and erased image to also predict F+ features, and final image is generated via StyleGAN again. First, we compare our hierarchical encoding in the two-stage architecture to validate the effectiveness of this encoding mechanism. As shown in Table 2, with this encoding FID's improve from 5.92 to 5.20. Next, we replace the two-stage pipeline with our single stage one that predicts the W+ and F+ features with a single pass in the encoder which further improves the FID to 4.85.

Table 2. Ablation study conducted on all masks. (Tüm maskeler üzerinde yapılan ablation çalışması.)

Models	FID
DivInversion - First Stage	16.65
Hierarchical - W+ encoding	13.61
DivInversion - Two Stage	5.92
Hierarchical - Two Stage	5.20
Hierarchical - W+ and F+ encoding	4.85

4. CONCLUSIONS (SONUÇLAR)

In conclusion, image inpainting is vital for numerous applications, from editing to object removal. While traditional methods start from scratch, recent advances exploit pretrained models like StyleGANs. Our work is also built on pretrained StyleGAN because of its rich internal representations. In this work, we introduce a hierarchical encoder and single-stage architecture

that tackle the complexities of encoding visible and missing features. Our experiments confirm substantial enhancements over existing models. Our model is able to fill the erased areas even when they are as large as the whole image. Additionally, by using the editing directions explored via InterFaceGAN [18], we can edit images during inpainting as given in Figure 1.

ACKNOWLEDGMENTS (TEŞEKKÜR)

This work has been funded by The Scientific and Technological Research Council of Turkey (TUBITAK), 3501 Research Project under Grant No 121E097.

Bu çalışma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından 3501 Araştırma Projesi kapsamında 121E097 numaralı proje desteğiyle finanse edilmiştir.

DECLARATION OF ETHICAL STANDARDS (ETİK STANDARTLARIN BEYANI)

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

AUTHORS' CONTRIBUTIONS (YAZARLARIN KATKILARI)

***Aysegul DUNDAR*:** She conducted the experiments, analyzed the results and performed the writing process.

Deneyleri yapmış, sonuçlarını analiz etmiş ve maklenin yazım işlemini gerçekleştirmiştir.

CONFLICT OF INTEREST (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

REFERENCES (KAYNAKLAR)

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536–2544, 2016.

[2] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in Proceedings of the European conference on computer vision (ECCV), pp. 85–100, 2018.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 4471–4480, 2019.

[4] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7760–7768, 2020.

[5] G. Liu, A. Dundar, K. J. Shih, T.-C. Wang, F. A. Reda, K. Sapra, Z. Yu, X. Yang, A. Tao, and B. Catanzaro, "Partial convolution for padding, inpainting, and image synthesis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 5, pp. 6096–6110, 2022, https://doi.org/10.1109/TPAMI.2022.3209702.

[6] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11461–11471, 2022.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

[8] A. B. Yildirim, V. Baday, E. Erdem, A. Erdem, and A. Dundar, "Inst-inpaint: Instructing to remove objects with diffusion models," arXiv preprint arXiv:2304.03246, 2023.

[9] A. B. Yildirim, H. Pehlivan, B. B. Bilecen, and A. Dundar, "Diverse inpainting and editing with gan inversion," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23120–23130, 2023.

[10] H. Sivuk and A. Dundar, "Diverse semantic image editing with style codes," arXiv preprint arXiv:2309.13975, 2023.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410, 2019.

[12] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in International conference on machine learning, pp. 7354–7363, PMLR, 2019.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and

improving the image quality of stylegan," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110– 8119, 2020.

[14] N. Yu, G. Liu, A. Dundar, A. Tao, B. Catanzaro, L. S. Davis, and M. Fritz, "Dual contrastive loss and attention for gans," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6731–6742, 2021.

[15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," Advances in Neural Information Processing Systems, vol. 35, pp. 36479–36494, 2022.

[16] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for style- gan image manipulation," ACM Transactions on Graphics (TOG), vol. 40, no. 4, pp. 1–14, 2021, https://doi.org/10.1145/3450626.3459838.

[17] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High-fidelity gan inversion for image attribute editing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11379–11388, 2022.

[18] Y.Shen, J. Gu, X.Tang , and B.Zhou, "Interpreting the latent space of gans for semantic face editing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9243–9252, 2020.

[19] E. Ha ̈rko ̈nen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," Advances in Neural Information Processing Systems, vol. 33, 2020.

[20] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2085–2094, 2021.

[21] Z. Chen, R. Jiang, B. Duke, H. Zhao, and P. Aarabi, "Exploring gradient-based multi-directional controls in gans," in European Conference on Computer Vision, pp. 104–119, Springer, 2022.

[22] H. Pehlivan, Y. Dalva, and A. Dundar, "Styleres: Transforming the residuals for real image editing with stylegan," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1828–1837, 2023.

[23] A. B. Yildirim, H. Pehlivan, and A. Dundar, "Warping the residuals for image editing with stylegan," arXiv preprint arXiv:2312.11422, 2023.

[24] Y. Yu, L. Zhang, H. Fan, and T. Luo, "High-fidelity image inpainting with gan inversion," in Computer Vision– ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI, pp. 242–258, Springer, 2022.

[25] W. Wang, L. Niu, J. Zhang, X. Yang, and L. Zhang, "Dual-path image inpainting with auxiliary gan inversion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11421–11430, 2022.

[26] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 1745–1753, 2019.

[27] W. Luo, S. Yang, H. Wang, B. Long, and W. Zhang, "Context-consistent semantic image editing with style-preserved modulation," in European Conference on Computer Vision, pp. 561–578, Springer, 2022.

[28] Y. Wang, X. Tao, X. Shen, and J. Jia, "Wide-context semantic image extrapolation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1399–1408, 2019.

[29] Y.-C. Cheng, C. H. Lin, H.-Y. Lee, J. Ren, S. Tulyakov, and M.-H. Yang, "Inout: Diverse image outpainting via gan inversion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11431–11440, 2022.

[30] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4170–4179, 2019.

[31] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5505–5514, 2018.

[32] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirec- tional attention maps," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 8858–8867, 2019.

[33] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large scale image completion via co- modulated generative adversarial networks," In International Conference on Learning Representations, 2021.

[34] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2287–2296, 2021.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Advances in neural information processing systems, vol. 30, 2017.

[36] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, "Hyperstyle: Stylegan inversion with hypernetworks for real image editing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18511–18521, 2022.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018