applied sciences

MDPI

Article

Masked Style Transfer for Source-Coherent Image-to-Image Translation

Filippo Botti *, Tomaso Fontanini, Massimo Bertozzi and Andrea Prati

Department of Engineering and Architecture, University of Parma, 43124 Parma, Italy; tomaso.fontanini@unipr.it (T.F.); massimo.bertozzi@unipr.it (M.B.); andrea.prati@unipr.it (A.P.)

* Correspondence: filippo.botti@unipr.it

Abstract: The goal of image-to-image translation (I2I) is to translate images from one domain to another while maintaining the content representations. A popular method for I2I translation involves the use of a reference image to guide the transformation process. However, most architectures fail to maintain the input's main characteristics and produce images that are too similar to the reference during style transfer. In order to avoid this problem, we propose a novel architecture that is able to perform source-coherent translation between multiple domains. Our goal is to preserve the input details during I2I translation by weighting the style code obtained from the reference images before applying it to the source image. Therefore, we choose to mask the reference images in an unsupervised way before extracting the style from them. By doing so, the input characteristics are better maintained while performing the style transfer. As a result, we also increase the diversity in the generated images by extracting the style from the same reference. Additionally, adaptive normalization layers, which are commonly used to inject styles into a model, are substituted with an attention mechanism for the purpose of increasing the quality of the generated images. Several experiments are performed on the CelebA-HQ and AFHQ datasets in order to prove the efficacy of the proposed system. Quantitative results measured using the LPIPS and FID metrics demonstrate the superiority of the proposed architecture compared to the state-of-the-art methods.

Keywords: deep learning; style transfer; image-to-image translation; generative adversarial networks

check for updates

Citation: Botti, F.; Fontanini, T.; Bertozzi, M.; Prati, A. Masked Style Transfer for Source-Coherent Image-to-Image Translation. *Appl. Sci.* **2024**, *14*, 7876. https://doi.org/10.3390/app14177876

Academic Editor: Thomas Lindner

Received: 23 July 2024
Revised: 26 August 2024
Accepted: 28 August 2024
Published: 4 September 2024

1. Introduction

Image-to-image translation (I2I) aims to generate an output image with a different style while preserving the content information of the input [1]. More specifically, the goal of I2I is to convert an image x_A belonging to a *source* domain A into an image y_B belonging to a *target* domain B by preserving its intrinsic content belonging to the source domain and modifying its extrinsic content by making it as similar as possible to that characterizing the target domain.

A lot of frameworks that use generative models to perform I2I translation are emerging in a variety of areas: from face editing [2] to style transfer [3] and the automotive field [4]. Focusing on style transfer, StarGANv2 [5] introduces an innovative approach. Specifically, StarGANv2 incorporates a style encoder that is designed to extract the style characteristics of an image, which is referred to as the *reference image*. Subsequently, the extracted style is applied to the input image using a single generator that is able to perform image translation across multiple domains. StarGANv2 also has a mapping network that is in charge of generating styles for the generator from random noise. In their work, the creators of StarGANv2 introduce diversity as the characteristic of each image within a domain to be different, despite the images coming from the same domain. By this definition, the authors show how the output changes as the reference changes, even if the images are picked from the same domain. However, this architecture design tends to apply global changes to the entire input image without preserving its intrinsic content representation. This can be described as a heavy form of *reference-based* style transfer, which can be seen in the generated output images by extracting the style from the same reference. In such cases, the output images tend to closely mirror the reference image, and the generated images collapse to the reference image, as can be seen in Figure 1.

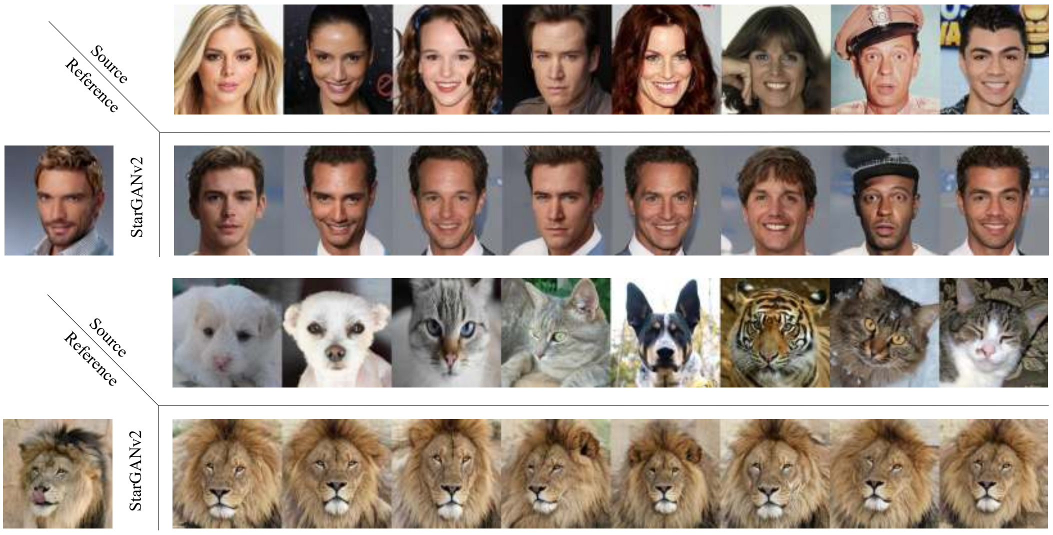

Figure 1. Results generated using StarGANv2. It can be seen how the generated images lose the intrinsic characteristics of the input like hair length or coat color and end up looking too similar to the reference, resulting in a lack of diversity.

To address this limitation, we present a novel architecture that is able to perform *source-coherent* I2I translation across multiple domains. Our solution involves adding a segmentation layer before the style encoder: this layer computes segmentation masks that are used to separate the subjects within the reference image and to select only the desired part of the image. In this way, we can remove all of the unnecessary content in the reference, like the background or out-of-domain parts. Ultimately, the style encoder extracts the style only from the relevant part of the reference image, and the generator produces images with the style of the reference image without collapsing to it.

Our network architecture takes inspiration from StarGANv2 [5], though it has some fundamental changes; in particular, we change the style application by using cross attention layers [6] and not adaptive instance normalization (AdaIN) [7]; then, we adapt the style encoder by feeding it with both the image and its corresponding mask.

Moreover, a crucial aspect lies in the utilization of an unsupervised architecture for extracting masks from reference images. Specifically, we choose to use the STEGO [8] model, which is an architecture that can perform unsupervised semantic segmentation, in order to produce masks. With STEGO, we produce binary images that are used in order to separate information regarding where to extract the style code inside the style encoder.

To summarize, the main contributions of the proposed work are as follows:

- Innovative architecture for style transfer: We introduce a novel architecture that is able to perform source-coherent I2I translation between multiple domains by preserving input details and increasing diversity during generation.
- Semantic style separation: The model utilizes an unsupervised segmentation architecture to produce masks in order to localize the style only for specific subjects of the images and to remove useless areas like backgrounds or out-of-domain details. By this weighing of the reference images, the model is able to focus only on the relevant parts and to better understand the characteristics of the image styles, resulting in more accurate style codes compared to the ones generated by state-of-art architectures.
- Transferring styles using cross attention: The proposed architecture also shows how attention mechanisms—more specifically, cross attention layers—are able to

improve the quality of style transfers compared to commonly used adaptive instance normalization layers.

2. Related Work

Image-to-image translation: Image-to-image translation was first introduced by [9] as the task of translating one possible representation of a scene into another given sufficient training data. Pix2Pix [10] was the first attempt to use GANs—in particular, conditional GANs (CGANs)—in order to translate an image from the source domain to the target domain and vice versa with paired datasets. Later, CycleGAN [3] improved Pix2Pix performance by removing the requirement of paired datasets and suggested a method for I2I translation on unpaired datasets by employing a cycle consistency loss that guarantees that an image should accurately replicate the source image when it is translated to the target domain and then reversed. MUNIT [11] was one of the first attempts to enhance the diversity of the generated images by feeding the generator with a style code that is randomly sampled from Gaussian noise. Later, MSGAN [12] tried to improve the diversity of generated images by maximizing the ratio of the distance of two images in the image space with respect to the distance of their corresponding latent code in the latent space. StarGAN [13] reached better performance in terms of both diversity and quality by using only a single generator to train between multiple domains. StarGANv2 [5] later improved the StarGAN architecture by introducing a style encoder that is in charge of learning new styles from images and then uses this style code in order to condition the output. Nevertheless, all of the cited architectures tend to share the same limitation of lack of diversity when using the same reference.

Diffusion probabilistic models (DPMs) [14] have recently showed impressive results in the generative field. Despite this, DPMs are still not at the same level as GANs for I2I translation problems. Architectures like ControlNet [15], BBDM [16], and Palette [17] show good results for primitive forms of I2I, but they lack the capacity to perform I2I from multiple domains. For this reason, in this paper, we chose to adapt the StarGANv2 architecture to perform our task.

Style transfer: Style transfer is a way to perform I2I translation by generating a sample with the same content as the input image but with another style. In this way, we can translate images between multiple domains and preserve the intrinsic characteristics of the input. One of the first application used conditional GAN [18] in order to perform style transfer, but it was based on a slow optimization process that iteratively updates the image to minimize the content and style losses. Later, adaptive instance normalization (AdaIN) [7] became the state-of-the-art in style transfer applications. AdaIN enables fast, arbitrary style transfers in real-time without being limited to a specific set of styles, as in previous works. Recently, transformers [19] have exhibited impressive results in NLP, and a lot of transformer-based architectures have been used across a multitude of vision-related tasks. In particular, the StyTr^2 [20] and latent diffusion [6] models have highlighted the power of transformers and cross attention layers when used to transfer styles from multimodal references like text, class labels, or images. For this reason, we selected cross attention layers in order to apply the domain style to the generated input. Additionally, recent approaches have leveraged the capability of latent diffusion to perform style transfer between pictures and paintings [21,22].

One of the main challenges during style transfer is to identify only the regions from which to extrapolate the style and to remove unnecessary regions like the background or other parts of the image. Ref. [23] introduced an attention layer in order to select the area from which to apply the style during I2I translation. Ref. [24] proposed cycle-consistent attention loss in order to train the model to apply changes in the same areas during translation and reconstruction by using a residual block activation map. Recently, SEAN [25] demonstrated that by using a mask that represents only the relevant area of the image, it is possible to perform an average pooling operation on the extracted features inside the style encoder and to produce more accurate style codes. Following a similar idea, we propose to modify the StarGANv2 style encoder by introducing mask multiplication and pooling.

Unsupervised semantic segmentation: Semantic segmentation aims to discover and localize the semantically meaningful categories present in an image. Typically, Mask R-CNN [26] or YOLO [27] are used in order to produce segmentation from an image, but they require labeled datasets, which is not always feasible and, in any case, is not scalable. Recently, several works have introduced semantic segmentation systems that can learn from weaker forms of labels such as classes, tags, bounding boxes, scribbles, or point annotations. The IIC system [28] focuses on maximizing the mutual information of patch-level cluster assignments between an image and its augmentations. It operates as an implicit clustering method, with the network directly predicting the (soft) clustering assignment for each pixel-level feature vector [8,28,29]. PiCIE [29] enhances the semantic segmentation outcomes achieved by IIC by leveraging invariance to photometric effects and equivariance to geometric transformations as an inductive bias. In PiCIE, the network aims to minimize the distances between features subjected to different transformations. The distance metric is determined through an in-the-loop k-means clustering process [8,29]. Conversely, STEGO [8] achieves impressive results in semantic segmentation without any kind of labeled dataset. STEGO shows that unsupervised deep network features have correlation patterns that are largely consistent with true semantic labels and uses these patterns to categorize every pixel of the image. Based on its valuable characteristics, STEGO is a perfect candidate for our purposes, and it represents the state-of-the-art in unsupervised semantic segmentation.

3. Proposed System

In the next sections, the proposed model, which is able to perform source-coherent translation by preventing the results from collapsing to the references, is described.

3.1. Network Architecture

As stated above, the network architecture of our system follows that of StarGANv2 and is composed of a generator G, a discriminator D, a style encoder E, and a mapping network M (see Figure 2). The style code s_{trg} can be generated both from images by using the style encoder or from random noise z by using the mapping network. During generation, the generator G takes both the source image x_{src} and the style code s_{trg} and generates the output $x_{trg} = G(x_{src}, s_{trg})$. In finer detail, G is designed as an encoder–decoder architecture featuring four downsampling residual blocks and four upsampling residual blocks, but it uses cross attention layers (instead of adaptive instance normalization layers) to apply the style. The discriminator D serves the role of evaluating which domain the generated samples belong to. D follows the StarGANv2 implementation and is a multitask discriminator that consists of multiple output branches. Finally, the style encoder E is a CNN with two residual blocks as feature extractors and an average pooling layer for the area covered by the mask, while M is an MLP that is in charge of generating style codes from noise.

Generating Styles from References

As previously stated, our style encoder is heavily modified compared to the one introduced by StarGANv2. In fact, we took inspiration from the SEAN style encoder [25] and adapted it to our goal. First, only two, instead of six, downsampling layers are used, following the implementation reported in [30], because we do not need shape information inside our style code. Then, we multiply the extracted features with the precomputed mask that is computed from the segmentation map obtained using STEGO (see Section 3.3). This allows us to delete information that irrelevant to the style computation, such as the background. Finally, an average pooling layer followed by a fully connected layer for each domain are applied (see Figure 3a).

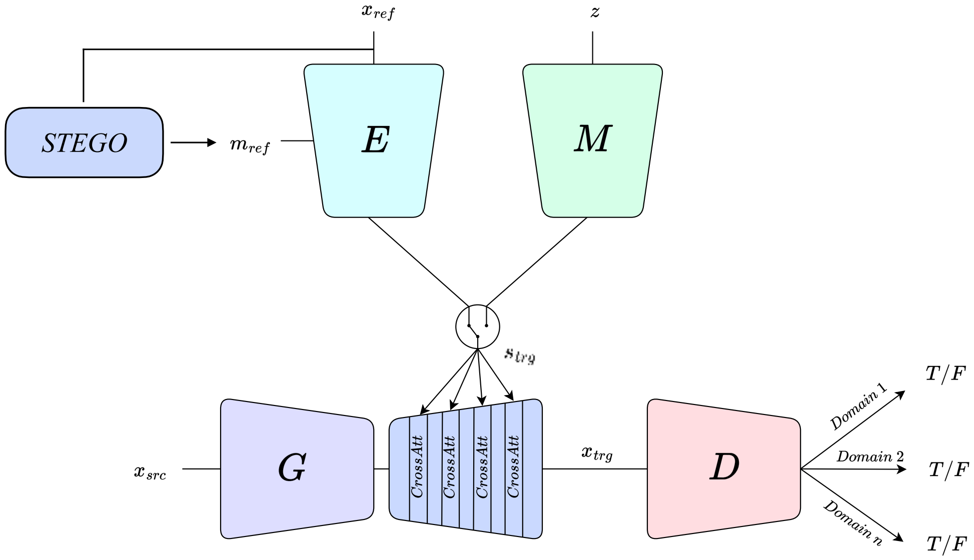

Figure 2. Overview of the proposed system. The generator takes the input image x_{src} and applies the style code, which is computed using reference (x_{ref}) and its correspondent mask (m_{ref}) or using random noise (z), with cross attention layers. Finally, the result x_{trg} passes through the discriminator.

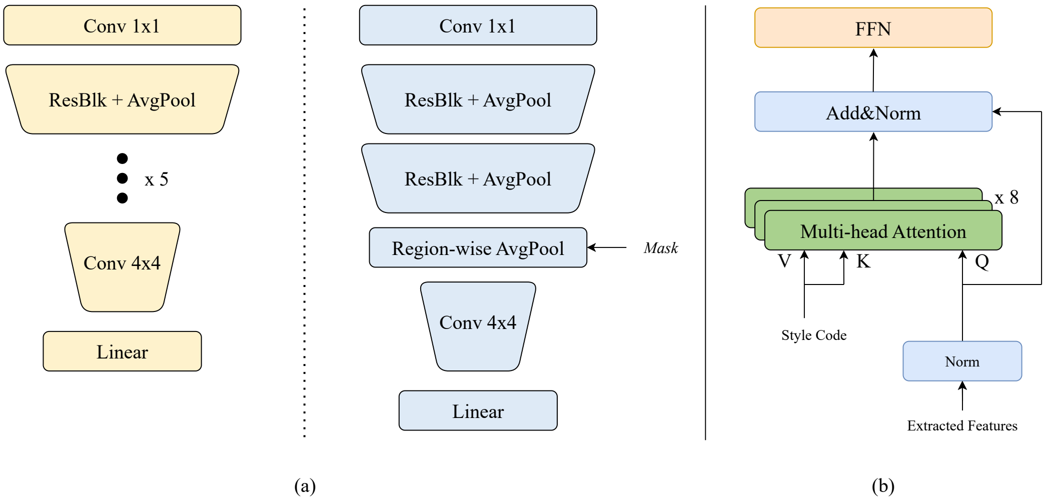

Figure 3. Overview of the style encoder architecture. (**a**) Comparison between the StarGANv2 style encoder (**left**) and the proposed style encoder (**right**). Different from StarGANv2, we utilize a mask for computing the style code. (**b**) Overview of the cross attention layers utilized for style transfer. We compute cross attention using extracted features from the input image as the query vector and the style code as the key and value vectors.

3.2. Transferring Style with Cross Attention

In order to apply style to the image, we use cross attention layers instead of AdaIN during the decoding phase of the generator. As shown in Figure 3b, we first extract features from source image x_{src} and style code s_{trg} extracted from x_{ref}. Subsequently, the features extracted from the source image are normalized using layer normalization, and then for every cross attention, the style code is injected as follows:

$$Att(x,s) = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V$$

where Q is the projection of the features extracted from x_{src}, K, V are projections of the style code s_{trg}, and d is the dimension of a single attention head. Finally, the resulting tensors are normalized with layer normalization and are linearly transformed with a feedforward layer.

3.3. Extracting Masks with STEGO

For the purpose of maintaining the network as fully unsupervised, the unsupervised semantic segmentation architecture STEGO is employed for extracting masks from the reference image before extracting the style from it. As described in [8], features f are first extracted from the reference image using the DiNo [31] feature extractor, then a STEGO segmentation head is devoted to extracting a non-linear projection and to learning patterns inside the image. Finally, the results are clustered and refined with a conditional random field (CRF) layer [32].

Only the semantic cluster of the style that needs to be transferred is selected, and the others are set to zero in the segmentation mask. More specifically, the selected semantic cluster/class is the one corresponding to the main subject of the image (e.g., animals in the AFHQ dataset [5] and persons in the CelebA-HQ dataset [33]).

3.4. Training and Losses

In order to train the proposed model, we choose to maintain the training phase of StarGANv2 without any changes. Therefore, the total loss is composed of four losses:

- The *adversarial loss* is used to learn the generation of realistic results:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{src}[\log D_{src}(x_{src})] + \mathbb{E}_{trg}\left[\log\left(1 - D_{trg}\left(G\left(x_{src}, s_{trg}\right)\right)\right)\right]$$

  where $x_{src} \in \mathcal{X}$ is the input image, and $G(\cdot)$ is the generator and takes x_{src} and s_{trg}, which is the style code extracted from the reference image x_{ref}.
- The *style reconstruction loss* is introduced in order to prevent the generator G from ignoring the style s_{trg} during the generation phase:

$$\mathcal{L}_{sty} = \mathbb{E}_{src,trg}\left[\left|\left|s_{trg} - E\left(x_{trg}\right)\right|\right|\right]$$

  where $E\left(x_{trg}\right)$ is the style code extracted from the generated image.
- The *style diversification loss* is used to differentiate the styles generated from two different images:

$$\mathcal{L}_{div} = \mathbb{E}_{src,trg_1,trg_2}\left[\left|\left|x_{trg_1} - x_{trg_2}\right|\right|\right]$$

- The *cycle consistency loss* maintains the domain-invariant characteristics of the generated image, like the pose and shape:

$$\mathcal{L}_{cyc} = \mathbb{E}_{src,trg}\left[\left|\left|x_{src} - G\left(x_{trg}, \tilde{s}_{src}\right)\right|\right|\right]$$

  where $\tilde{s}_{src}$ is the estimated style code extracted from the input.

  The final loss is, therefore, as follows:

$$\min_{G,M,E} \max_{D} \mathcal{L}_{\text{adv}} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{div}\mathcal{L}_{div} + \lambda_{cyc}\mathcal{L}_{cyc}$$

It is worth noting that during training, a reference image and random noise are used alternatively for generating the style code through the mapping network.

4. Experimental Results

This section reports details about the experimental results: both qualitative and quantitative.

4.1. Selected Baseline

Since our work is an extension of StarGANv2, we decided not to compare the results that we produce with those of other architectures, following the comparison in [30]. In fact, our work can be seen as an improved version of StarGANv2, with the objective of showing how to leverage the mistakes made by StarGANv2 and how to improve that network by adding our masked style encoder. As discussed, StarGANv2 style transfer is limited in terms of diversity and tends to generate images with the same style applied; on the contrary, we show a proper way to perform style transfer without losing diversity. Moreover, it is quite difficult to compare our architecture with others since, when considering diffusion models, for example, the style transfer is a totally different task and cannot be compared to our architecture. All of the hyperparameters and training strategies are the ones proposed in the original paper for StarGANv2.

4.2. Datasets

We tested our model on two datasets: CelebA-HQ, composed of 30 k images [33], and AFHQ, composed of 16 k images [5]. CelebA-HQ is organized in two domains (male and female), and AFHQ is organized in three domains (cat, dog, and wildlife animal). Binary masks are extracted using STEGO pretrained on COCOstuff [34] by selecting the "person" and "animal" attributes in order to identify the subjects of the images for the two datasets. No other information is employed during training or inference. We resize all images to 256 × 256 and all masks to 64 × 64 during training.

4.3. Implementation Details

During all the experiments, we train the network for 100 k iterations and we use Adam [35] as the optimizer. Learning rates of 10^{-4} for G, D, and E and 10^{-6} for M are used. Training took about 1 day using a single NVIDIA A100 GPU, which is the same amount of time as StarGANv2, proving that our approach does not add complexity in the training. For CelebA-HQ training, we weigh every loss equally; on the contrary, for AFHQ, we set λ_{div} to 2, while λ_{cyc} and λ_{sty} are 1, following the implementation in [5], in order to make an equal comparison with StarGANv2 and to show that the results obtained are better because of the architecture and not because of these hyperparameters.

4.4. Evaluation Metrics

In order to evaluate our model, we use the Frechét inception distance (FID) [36] for image quality and the learned perceptual image patch similarity (LPIPS) [37] to measure diversity in the generated results. More specifically, the FID metric measures the distance between two distributions, and in our case, it is used in order to measure the distances between generated images, e.g., generated cat images and the test set that contains real images, i.e., real cat images. So intuitively, a low value of FID means that two distributions are similar. Indeed, given two Gaussian distributions (m, C) and (m_w, C_w), the FID is computed as follows:

$$d^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + \mathrm{Tr}(C + C_w - 2(CC_w)^{\frac{1}{2}})$$

The learned perceptual image patch similarity (LPIPS) calculates the perceptual similarity between two images. The LPIPS essentially computes the similarity between the activations of two image patches for some pre-defined and pre-trained network. This measure has been shown to match human perception well. A low LPIPS score means that image patches are perceptually similar. Indeed, given two patches x and x_0, their distance is computed as follow:

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

where $\hat{y}^l_{hw}$ and $\hat{y}^l_{0hw}$ are the stacked features extracted from the patches. These features are normalized, and the distance between them is modulated by a learned weight vector w_l that adjusts the contributions of different feature channels [37].

Since the main contribution of our model is to perform source-coherent translation, which aims to improve diversity for images generated with the same reference, the evaluation is designed as follows:

- Firstly, we randomly select one image for every domain as reference;
- Secondly, given a set of source images, we generate samples with those reference images;
- Thirdly, we compute the FID and LPIPS (with consecutive pairs of images);

- Finally, we repeat this evaluation phase 10 times in order to remove randomness from the results.

It is worth emphasizing that we decided to not use the StarGANv2 FID algorithm, which calculates the FID by using ten references for each domain, because we want to improve the diversity of the results generated from a single reference. Therefore, the FID is computed with only one reference per domain in order to evaluate the quality of the images generated with our method.

4.5. Discussion

Depending on the dataset, different styles are transferred. For CelebA-HQ, *male2female* and *female2male* were chosen. For the AFHQ dataset, we transferred *cat-dog2wildlife* and *wildlife2cat-dog*. As can be seen from Figure 4, the proposed architecture can perform I2I translation between multiple domains similar to StarGANv2, but it gains the capability to preserve the intrinsic characteristics of the input during translation. Looking at the CelebA-HQ results, the proposed architecture maintains the input facial attributes but applies changes to the gender and hair color, which are taken from the reference images. In the AFHQ results, our method maintains the same expression and better preserves fur color during translation, but it changes the class of the animal. We introduced our work with the claim that it increases the diversity in the results generated using the same reference. This is shown clearly in Figures 5 and 6, where the results are compared with the ones obtained using StarGANv2. From these examples, it is evident that StarGANv2 tends to collapse to the reference image and loses the majority of the intrinsic attributes of the input except for the poses in AFHQ and the expressions in CelebA-HQ. In contrast, our results maintain many more of the original details, such as fur colors in AFHQ and ages and hair styles in CelebA-HQ. This leads to more diversity and variety in the generated images and to less reference-based generation.

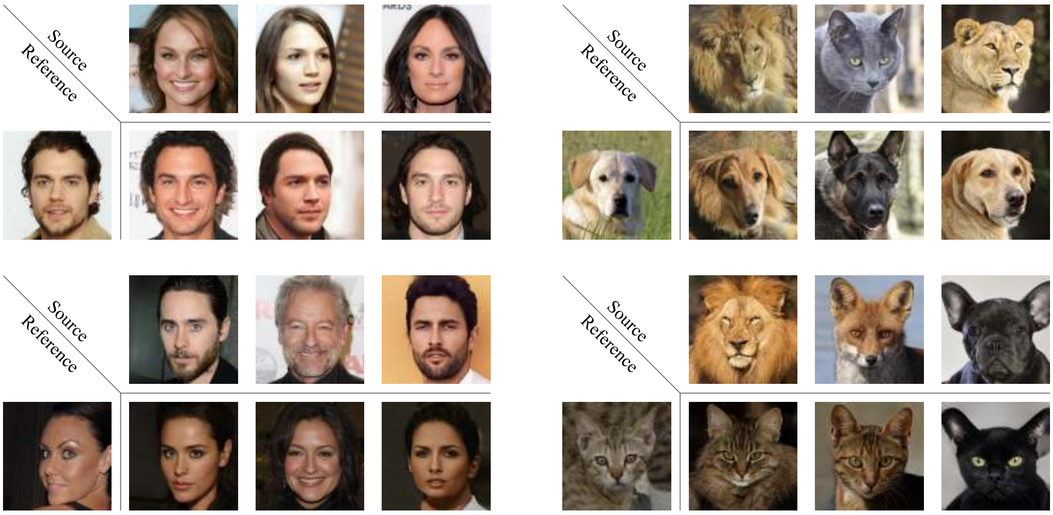

Figure 4. Main results obtained with our architecture. We can see how the results are indistinguishable from real images and how they maintain the intrinsic characteristics of the input, like the expression, age, or fur color.

More specifically, StarGANv2 seems to capture attributes from the reference, like hair color and length in Figure 5, and apply them to the input in a rigid scheme that does not maintain the input attributes. On the contrary, the proposed method, in addition to understanding what the main details are from the reference, also considers the details from the input image before applying the transformation. This leads, for instance, to a higher variety of hair lengths rather than a prefixed hair length, like with StarGANv2. This is also visible in Figure 6, where StarGANv2 produces the same animal with different poses. Our method, on the contrary, better understands the input characteristics and generates different breeds of dog/cat/wildlife animals based on the input breed.

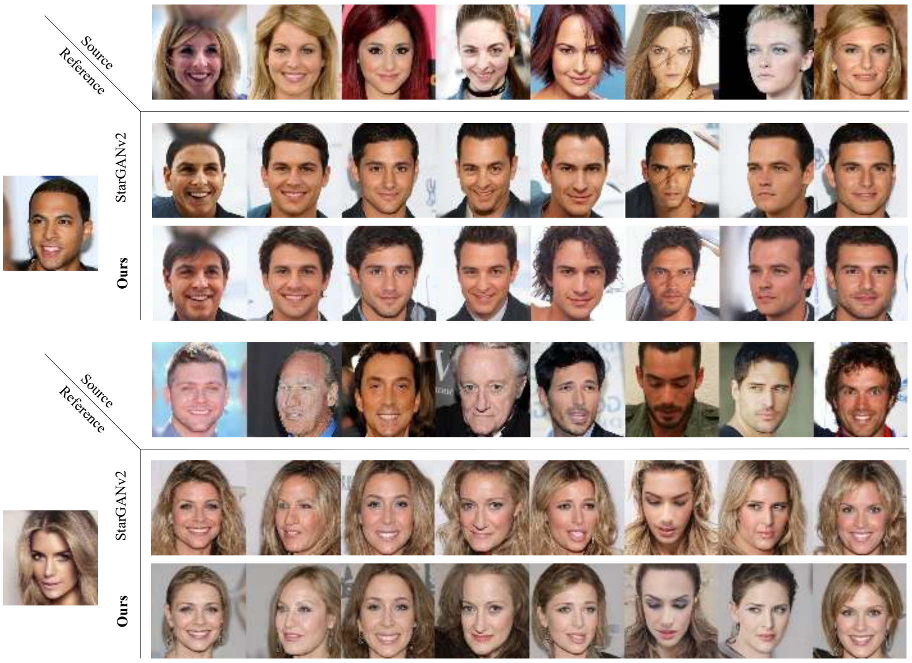

Figure 5. Comparison between StarGANv2 (first rows) and our architecture (second rows) on CelebA-HQ dataset.

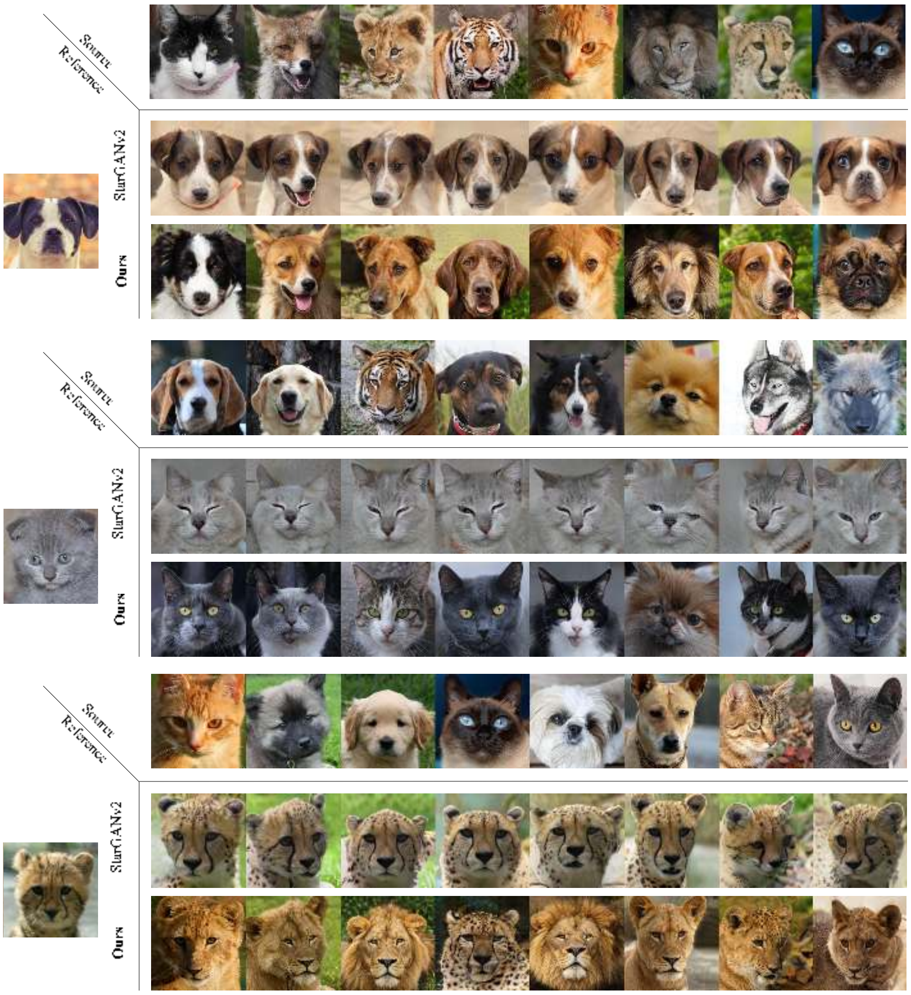

Figure 6. Comparison between StarGANv2 (first rows) and our architecture (second rows) on AFHQ dataset.

All the previous considerations are also valid when the style code s_{trg} is sampled from random noise by using the mapping network M. This is presented in Figure 7, where our architecture produces various and more source-coherent results than the ones generated by StarGANv2.

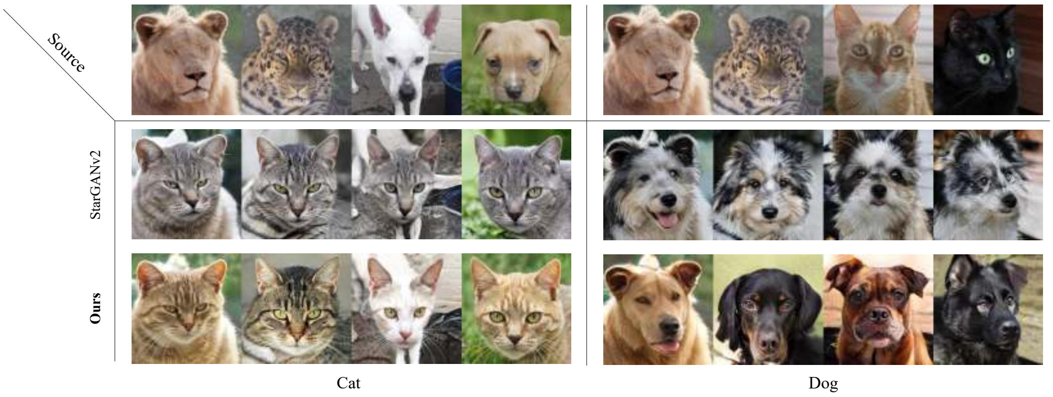

Figure 7. Comparison between StarGANv2 (first rows) and our architecture (second rows) on AFHQ dataset for results generated using random noise as the reference.

In order to support our claim, we also show how our method produces similar results when similar reference images are employed, as shown in Figure 8. This can be seen as a positive effect due to our source-coherent method that does not ignore input attributes.

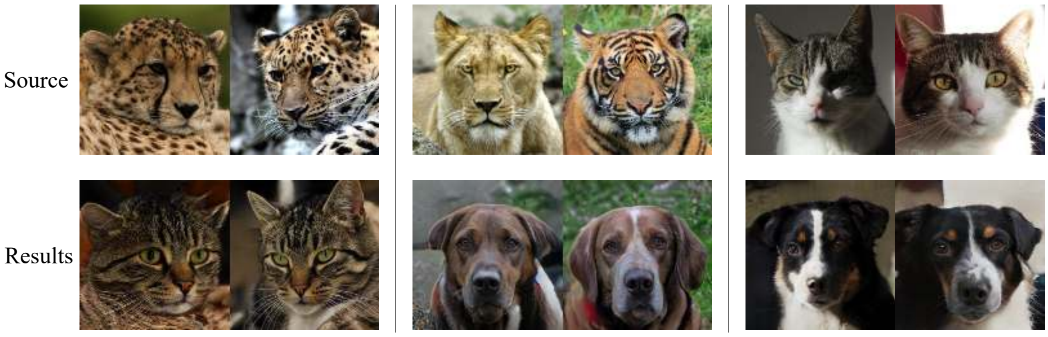

Figure 8. Similar inputs generate similar outputs due to the fact that we preserve input characteristics during translation.

Quantitative Results

The above considerations are reflected in the quantitative results reported in Table 1. The proposed architecture significantly improves the LPIPS results for both datasets. Furthermore, the FID results highlight how our architecture produces much higher-quality images.

Table 1. Quantitative comparison between StarGANv2 and our architecture. For our architecture, we also include a model with AdaIN instead of the cross attention layers.

Architecture	AFHQ		CelebA-HQ	
	FID ↓	**LPIPS ↑**	**FID ↓**	**LPIPS ↑**
StarGANv2 [2]	104.86	0.457	81.175	0.365
Ours (AdaIN)	76.15	0.523	57.67	0.420
Ours	67.72	0.517	54.12	0.425

As shown in Table 2, we also compute the FID using the StarGANv2 algorithm on CelebA-HQ, and we obtain opposite results. This is due to how the FID works (this is also explained in Section 4.4): given the alignment of the StarGANv2 generated images with the reference image shown in the qualitative results, the FID computed as in the StarGANv2 original paper is natively lower. Indeed, the FID tends to measure the difference between two distributions, and since the images generated by StarGANv2 have less diversity than the ones generated by our architecture, the FID score in this case is better for the StarGANv2 results. Nevertheless, in Section 4.4, we justified how is not fair to compute the FID in this way in order to consider the diversity in generated images. However, this comes at the cost of more limitations for StarGANv2 with respect to our architecture, such as losing input characteristics, a lack of diversity in generated results, and the results collapsing to the reference images.

Table 2. Quantitative comparison between StarGANv2 and our architecture using StarGANv2 FID algorithm.

Architecture	**CelebA-HQ** **FID ↓**
StarGANv2 [2]	29.88
Ours (AdaIN)	32.94
Ours	30.99

4.6. Ablation

Finally, we perform ablation studies to find the optimal configuration for our architecture. First, we try to transfer the style using AdaIN and not cross attention layers. As shown in Table 1 and Figure 9, using AdaIN leads to slightly better results compared to StarGANv2 in terms of diversity, but the network still does not maintain the input characteristics like our final configuration. Additionally, we also tested employing two or three downsampling layers inside our style encoder, as can be seen in the second and third rows in Figure 9. Indeed, the configuration with three downsampling layers tends to collapse more to the reference than the one with two downsampling layer, as can be seen from the fur style. Furthermore, by comparing these results with the ones produced by the final architecture, it is evident that the cross attention layers improve the quality of the generated results and better maintain the input characteristics. Finally, the results generated without masks are reported in the fourth row, proving that masks are necessary to identify major input information like fur color and ear pose.

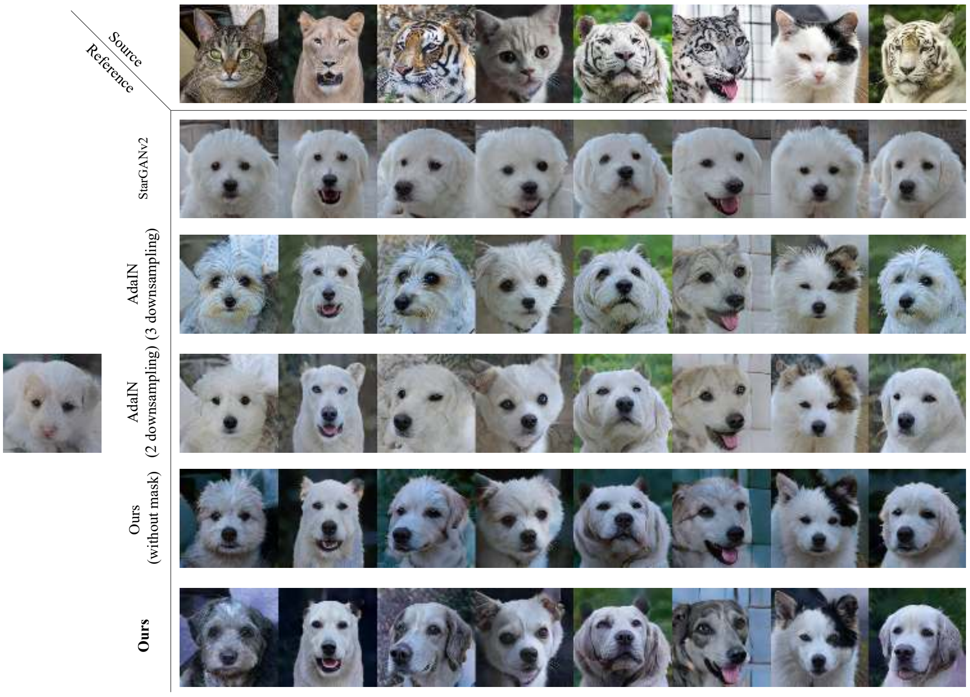

Figure 9. Differences between results generated with different architectures: the first row shows StarGANv2, the second row shows our style encoder and AdaIN for style transfer, the third row shows the same architecture as before but with 3 downsampling layers inside the style encoder, the fourth row shows cross attention layers for style transfer but without masks and with 3 downsampling layers in the style encoder, and the final row shows the proposed architecture.

5. Conclusions

For the task of I2I translation, StarGANv2 has shown limitations in preserving input details during translation. Additionally, StarGANv2 is not able to generate diverse samples when using the same reference image. For these reasons, this paper proposes a novel architecture for source-coherent image-to-image translation that preserves input characteristics and increases diversity in the generated results. More specifically, the reference images are masked in order for the model to focus only on the relevant information, and the styles extracted from these images are injected into the model using cross attention layers. By doing so, we manage to improve both the quantitative and qualitative results.

Future works could focus on improving the results generated by using two different references and only one source, since our method, by preserving the intrinsic characteristics of the input, tends to produce similar results when the same source image is utilized.

Author Contributions: Conceptualization, F.B. and T.F.; methodology, F.B. and T.F.; software, F.B.; validation, T.F.; formal analysis, F.B. and T.F.; investigation, F.B.; resources, F.B.; data curation, F.B.; writing—original draft preparation, F.B.; writing—review and editing, F.B., T.F., M.B. and A.P.; visualization, F.B.; supervision, T.F., M.B. and A.P.; project administration, F.B., T.F., M.B. and A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both the CelebA-HQ and AFHQ datasets can be downloaded from: https://github.com/clovaai/stargan-v2 (accessed on 23 July 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pang, Y.; Lin, J.; Qin, T.; Chen, Z. Image-to-Image Translation: Methods and Applications. *IEEE Trans. Multimed.* **2022**, *24*, 3859–3881. [CrossRef]
2. Li, Y.A.; Zare, A.; Mesgarani, N. StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion. *arXiv* **2021**, arXiv:2107.10394.
3. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv* **2017**, arXiv:1703.10593.
4. Zheng, Z.; Wu, Y.; Han, X.; Shi, J. ForkGAN: Seeing into the Rainy Night. In Proceedings of the IEEE European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
5. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J. StarGAN v2: Diverse Image Synthesis for Multiple Domains. *arXiv* **2019**, arXiv:1912.01865.
6. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv* **2021**, arXiv:2112.10752.
7. Huang, X.; Belongie, S.J. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *arXiv* **2017**, arXiv:1703.06868.
8. Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; Freeman, W.T. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. *arXiv* **2022**, arXiv:2203.08414.
9. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv* **2016**, arXiv:1611.07004.
10. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
11. Huang, X.; Liu, M.; Belongie, S.J.; Kautz, J. Multimodal Unsupervised Image-to-Image Translation. *arXiv* **2018**, arXiv:1804.04732.
12. Mao, Q.; Lee, H.; Tseng, H.; Ma, S.; Yang, M. Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis. *arXiv* **2019**, arXiv:1903.05628.
13. Choi, Y.; Choi, M.; Kim, M.; Ha, J.; Kim, S.; Choo, J. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv* **2017**, arXiv:1711.09020.
14. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239.
15. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv* **2023**, arXiv:2302.05543.
16. Li, B.; Xue, K.; Liu, B.; Lai, Y.K. Bbdm: Image-to-image translation with brownian bridge diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1952–1961.
17. Saharia, C.; Chan, W.; Chang, H.; Lee, C.A.; Ho, J.; Salimans, T.; Fleet, D.J.; Norouzi, M. Palette: Image-to-Image Diffusion Models. *arXiv* **2021**, arXiv:2111.05826.

18. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2414–2423. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
20. Deng, Y.; Tang, F.; Pan, X.; Dong, W.; Ma, C.; Xu, C. StyTrˆ2: Unbiased Image Style Transfer with Transformers. *arXiv* **2021**, arXiv:2105.14576.
21. Chung, J.; Hyun, S.; Heo, J.P. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 8795–8805.
22. Deng, Y.; He, X.; Tang, F.; Dong, W. Z*: Zero-shot Style Transfer via Attention Reweighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 6934–6944.
23. Mejjati, Y.A.; Richardt, C.; Tompkin, J.; Cosker, D.; Kim, K.I. Unsupervised Attention-guided Image to Image Translation. *arXiv* **2018**, arXiv:1806.02311.
24. Fontanini, T.; Botti, F.; Bertozzi, M.; Prati, A. Avoiding Shortcuts in Unpaired Image-to-Image Translation. In Proceedings of the Image Analysis and Processing—ICIAP 2022, Lecce, Italy, 23–27 May 2022; Sclaroff, S., Distante, C., Leo, M., Farinella, G.M., Tombari, F., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 463–475.
25. Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. *arXiv* **2019**, arXiv:1911.12861.
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870.
27. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
28. Ji, X.; Henriques, J.F.; Vedaldi, A. Invariant Information Distillation for Unsupervised Image Segmentation and Clustering. *arXiv* **2018**, arXiv:1807.06653.
29. Cho, J.H.; Mall, U.; Bala, K.; Hariharan, B. PiCIE: Unsupervised Semantic Segmentation using Invariance and Equivariance in Clustering. *arXiv* **2021**, arXiv:2103.17070.
30. Fontanini, T.; Ferrari, C. Would Your Clothes Look Good on Me? Towards Transferring Clothing Styles with Adaptive Instance Normalization. *Sensors* **2022**, *22*, 5002. [CrossRef] [PubMed]
31. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. *arXiv* **2021**, arXiv:2104.14294.
32. Krähenbühl, P.; Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *arXiv* **2012**, arXiv:1210.5644.
33. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
34. Caesar, H.; Uijlings, J.R.R.; Ferrari, V. COCO-Stuff: Thing and Stuff Classes in Context. *arXiv* **2016**, arXiv:1612.03716.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Klambauer, G.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. *arXiv* **2017**, arXiv:1706.08500.
37. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv* **2018**, arXiv:1801.03924.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.