Politecnico di Torino

Politecnico di Torino

Computer Engineering - Data Analysis and Artificial Intelligence - LM-32

A.Y. 2024/2025

Graduation Session December 2025

Towards meaningful Image Anonymization using Semiotic Analysis

Supervisors:

Prof. Lia MORRA

Candidate:

Silvano NANETTI

Abstract

In everyday life, we come in contact with an ever growing amount of data, making its every related aspect a matter of concern: where to retrieve data, where to store it, and how to organize it are some notable mentions, but also only a fraction of a much longer list. Among these challenges, respecting the privacy of the individuals captured in a media deserves the right attention. In our work, we focus on the Image Anonymization task, which aims to preserve the personal information of people depicted inside photographs by modifying them. There are many techniques to achieve this goal: the main ones include covering the people's faces, sometimes even entire bodies, with blur, solid colors, generated masks, or by recreating the entire picture from scratch, but these approaches come with the issues of maintaining the meaning of the original image, not compromising the overall quality when editing identifiable details, and achieving a good level of anonymization. Our architecture is an extension of the CAMOUFLaGE-Light model, where we make use of pretrained models to extract different types of information from a picture and generate an anonymized version based on a portion of the original and the obtained features: we employ FRESCO and RelTR to analyze the starting input and produce a data structure containing every information deemed relevant, IP-Adapter and T2I-Adapter to learn from the extracted features, and Stable Diffusion to reconstruct the de-personalized photograph. Particular attention is paid to distill information such as ethnicity, age, gaze direction, body pose, and similar characteristics that help to define the semiotic of the picture we want to obfuscate. We execute different types of tests to assess the quality of the final output, where image quality, re-identification rate, semiotic analysis, and downstream-task model performance are used as metrics to compare the output scores with other state-of-the-art methods. Finally, we discuss possible future developments that could bring further advances in the efficacy of the proposed method.

Acknowledgements

I would like to express my gratitude to Professor Lia Morra, who guided and encouraged me during this thesis work, and PhD candidate Pietro Basci, who helped me with his insights at every step of this research. I would also like to voice my thanks to my friends and family, who proved to be an essential support throughout these academic years.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today, pictures surround us in everyday life more than ever before: through social media, ads, the possibility to snap a picture just by extracting the phone from a pocket, self-driving vehicles, and many more ways to interact with such media. Visual data is the main support to how we experience our lives, thus making the regulation of this type of information a field of ever-growing interest as we face the raise of concern on how to protect the associated privacy issues: some examples of regularization in digital privacy are the European Union's General Data Protection Regulation (GDPR) [1], China's Personal Information Protection Law (PIPL) [2] and Canada's Digital Charter Implementation Act (DCIA) [3]. To correctly follow these laws, the scientific community is working on methodologies that allow researchers and industries to protect personal data without compromising useful information unrelated to the subjects' identity. In particular, the task this thesis is concerned with is called Image Anonymization: as the name suggests, its goal is to hide personal information from a picture while keeping as much non-sensible data as possible. Some of the approaches used thus far involve inpainting (such as producing masks used to cover a picture's sub-area) [4, 5], and generating the image from scratch [6, 7]. The first method is based on modifying only a portion of the original input, making most of the time an output picture only slightly different from the original, and thus prone to be re-identified. Using the second approach, the final result is entirely generated by a model, which means that the produced image is less likely to be linked to the original data, but in this case artifact generation is more common and the overall utility can decrease.

1.1 Current work's limitations

Today's state-of-the-art anonymization models have to find the right balance between maintaining the utility of the synthesized images, while reducing the

possibility to isolate the source image when analyzing the de-identified one. This target has proved a difficult one: in order to distance original and generated images, it is needed to act not only on the depicted persons but also on the environment around them, as CAMOUFLaGE-Base [8] surpasses DeepPrivacy2 [4] at image level re-identification by modifying multiple areas of the picture. It is also important to note that the more data gets modified, the more it could get damaged and not be as useful.

1.2 Goal

This thesis aims to define an architecture focused on generating an output that, after the anonymization process, maintains a good level of utility for the training and testing of downstream task models. In order to achieve this goal, the model has been designed to take into account the semantic information contained in an image without forgetting about plastic-level data. Going more into the details, with semantic information is meant to address the meaning of what is displayed in the picture, which subjects are interacting with each other, their peculiar characteristics and how the interactions are happening, while plastic-level data is information concerning the picture's structure: for example the position of said subjects into the frame, their depth with respect to the camera, the poses assumed by the captured individuals, and many more. Thus, this methodology aims to use the original picture only to condition the generated output, reducing re-identification rates by, ideally, synthezising a picture starting from pure noise. The thesis work focuses on extending the architecture defined in [8] as CAMOUFLaGE-Light, which uses two adapters to get a combination of image and text features to guide a picture generation process. The extension consists in adopting many diverse image analysis techniques to obtain different kinds of data from a starting photograph, then use them to distill ad hoc data maps, composed of plastic attributes, and text features, consisting of a caption and relational descriptions, and employ them to condition image synthesis.

1.3 Outline

The next thesis' chapters are structured as follows:

- *Chapter 2* - **Related work**: the second chapter explores some of the existing state-of-the-art approaches regarding Image Anonymization, highlighting differences, strengths and shortcomings of such methods, with the addition of a brief summary on the metrics used to measure their performance.

- *Chapter 3* - **Background**: the third chapter includes the background knowledge needed to realize the architecture proposed, from defining the origin of Diffusion Models to the examination of core elements employed in the new method.
- *Chapter 4* - **Methodology**: the fourth chapter exposes the structure of the architecture by describing the process the initial picture undergoes to be transformed into its anonymized counterpart, delving into the details of every intermediate step.
- *Chapter 5* - **Experimental environment** the fifth chapter explains the training and testing experiments performed, including the datasets used through their characteristics, the hyperparameter settings with their motivations, and the evaluation metrics used to assess the model's performance.
- *Chapter 6* - **Results**: the sixth chapter defines which methods are confronted with the one proposed, the quantitative outcomes of the evaluations and their corresponding analysis.
- *Chapter 7* - **Ablation study**: the seventh chapter explores the impact of the text feature regarding relational descriptions by training and testing a model without their use, then confronting image quality and re-identification rate with the ones of the complete architecture.
- *Chapter 8* - **Conclusions**: the eighth and final chapter summarizes the results achieved in this thesis and proposes possible future directions of development for the architecture designed, including update of existing modules and insertion of brand new ones.

Chapter 2

Related work

As stated previously, this thesis focuses on the Image Anonymization task: it aims to obfuscate personal information from pictures while keeping their utility, defined by Jordon et al. [9] as the property measurable by the change in evaluation metrics, such as accuracy or prediction confidence, when synthetic images are used instead of their original counterparts. Most of the current state of the art in Image Anonymization is focused on producing de-identified faces, while a minority of projects extend their attention to full-body masking and sometimes even environmental modifications. These goals are achieved by approaches as inpainting, where the obfuscation happens directly on the original picture by substituting only some of its parts with synthesized ones that usually lower data quality, or by generating completely new images while retaining some of the input information. The models analyzed have been implemented usually with GANs [4, 6, 10], while recently also diffusion models have been employed [11, 12, 13, 7, 14].

2.1 GAN based models

Some of the oldest but still highly performing architectures commonly use GANs [15], where a Generator $\mathcal{G}$ module is trained to produce realistic images, while a Discriminator $\mathcal{D}$ module learns to discern between authentic pictures and the ones generated by $\mathcal{G}$. At the end of the training process $\mathcal{D}$ is discarded and $\mathcal{G}$ is kept to synthesize new images.

An example of a GAN-based model is DeepPrivacy2 [4], which acted as the best performing model in the field and a benchmark to evaluate other architectures: the general idea is to detect where persons are located in the original picture, generate compatible masks to cover said subjects, and obtain the final output by applying the generated data on the input image. FALCO [6], on the other hand, uses a pre-trained StyleGAN2 [16] to create, starting from a real dataset, a fake one. From

this synthesized collection the most similar image to the input one is selected to mix parts of their latent code and generate the final picture.

2.2 Diffusion based models

Lately, models are instead relying on diffusion [17], a process where the architecture learns to remove more and more noise from a picture to restore it; when the training phase has ended, a diffusion model is able to transform pure noise into an image. This setting is generally preferred, as it is able to synthesize more realistic images.

Deep Identity Distraction [11] and IDDiffuse [7] are examples of diffusion-based architectures: given a photo containing a face, the models extract its visual features and mix them with the ones of similar candidate images drawn from a feature space; then those features get combined with the input's identity-independent data to condition the final output generation. Full-body Anonymization using Diffusion Models [12] adapts diffusion to the more conventional approach of masking: in a first step it detects persons and other objects inside the photograph, then it generates masks to cover the subjects, and finally it inpaints the generated data to the initial input. Rendering-Refined Stable Diffusion [14] adopts a very similar approach to [12], but trades the inpainting of only human subjects with doing it in a more advanced manner: it detects 3D meshes used to estimate the body pose, it renders an avatar based on said data which allows to produce a mask with enhanced fidelity to the original context represented in the photograph.

2.3 Limitations

Most of the mentioned methods focus on inpainting in the original picture a generated mask, thus achieving anonymization only regarding human faces, entire bodies, and sometimes other detected objects: since the picture's structure remains unchanged, re-identification would be possible, for example, by matching the segmentation maps obtainable from anonymized and original images. Another issue worth of notice is that by simply generating masks based on a person's segmentation or body pose, the final output could lack the meaning the original input represented: data like ethnicity, gender, age, or emotion add meaning to what a photo conveys to the watcher, that being a human being or a machine learning architecture. **Add an image that highlights the invariance of elements as background and bodies from SOTA methods.**

2.4 Quantitative evaluation metrics

The Image Anonymization task makes it important to tackle three different issues about the output images: they have to be as realistic as possible while removing useless identifiable visual data and keeping all the useful information needed to train models on downstream tasks. This would mean that a picture representing a person performing a certain action has to be modified enough not to be recognizable by their facial features, clothes, and surroundings, but those details do not have to be altered to the extent of changing the meaning conveyed by said action being performed. Fréchet Inception Distance (FID) [18] is widely adopted when evaluating Computer Vision models: it confronts the data distribution of real images and generated ones to evaluate how dissimilar they are, but this method presents some performance issues as stated by [19]. Other examples of image quality measurement are Structural Similarity Index Measure (SSIM) [20], computed with a simple function at the picture's pixel level, Learned Perceptual Image Patch Similarity (LPIPS) [21] and VisualDNA [22], which confront neuron activations during data elaboration. As in the VisualDNA case, tests of this kind are executable both at dataset-level, where the whole collections are opposed after collecting data distribution properties, or at image-level, where each pair of original and synthesized images is confronted singularly before averaging out the results. In order to assess the efficacy of anonymization, Morra et al. [23] introduced the FRESCO score, which makes it possible to estimate image similarity by considering each property that FRESCO.v1 outputs upon classification. Furthermore, it is important to measure the effect of identity obfuscation of the mentioned approaches: this metric is defined by Barattin et al. [6] as the number of images whose identity is still detected in the anonymized version, over the total number of images. Finally, it is important to evaluate the generated dataset's utility by how a downstream model's performance changes when trained on anonymized data instead of the original one.

Chapter 3

Background

3.1 Diffusion Models

The fundamentals of Diffusion Models have been set by Sohl-Dickstein et al. [17], in which the authors define the diffusion process as a Markov chain where, starting from a distribution, it is possible to convert it into another one as seen in Figure 3.1. More specifically: while the encoder module transforms data through a series of noise-adding steps to the initial data distribution $q(\boldsymbol{x}^{(0)})$ gradually mutating it into a Gaussian distribution by applying forward diffusion (Equation 3.1) where $\mathcal{N}$ is the Normal distribution, $\sqrt{1-\beta_t}$ makes the data points tend towards the origin while $\mathbf{I}\beta_t$ noise is added and t represents the timestep; the decoder network learns how to remove noise and recover the original information using the backward diffusion operation (Equation 3.2), where $p(\boldsymbol{x}^{(t)})$ describes the distribution at the t-th step of the reverse diffusion process, $\mathbf{f}_\mu$ and $\mathbf{f}_\Sigma$ are the mean and covariance learned functions that regulate the denoising.

$$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)}) = \mathcal{N}(\boldsymbol{x}^{(t)}; \boldsymbol{x}^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t) \forall t \in \{1, ..., T\} \tag{3.1}$$

$$p(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}) = \mathcal{N}(\boldsymbol{x}^{(t-1)}; \mathbf{f}_\mu(\boldsymbol{x}^{(t)}, t), \mathbf{f}_\Sigma(\boldsymbol{x}^{(t)}, t)) \forall t \in \{1, ..., T\} \tag{3.2}$$

Later, Denoising Diffusion Probabilistic Models [25] employed a U-Net [26] backbone and improved the diffusion process by simplifying noise scheduling, with β_t following a linearly increasing sequence and $\mathbf{f}_\Sigma$ set to be dependent only on β_t instead of being a learnable function; these changes resulted in more stable training and better sample quality. Another notable addition to diffusion-based models is defined by Dhariwal and Nichol [27] introduced the concept of Classifier Guided Diffusion which, by training a classifier $f_\phi(y|x_t, t)$ on noisy data x_t, makes it possible to condition the diffusion process using gradients $\nabla_{x_t} \log f_\phi(y|x_t, t)$ and thus produce an output based on the class label y, resulting in higher image

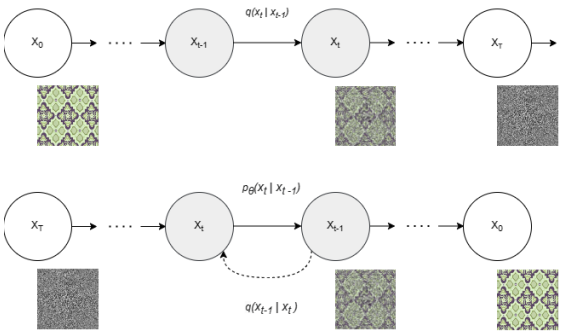

Figure 3.1: Representation of a diffusion process (above) and corresponding reverse operation (below). Image by [24].

fidelity while decreasing sample diversity. Ho and Salimans [28] improved diffusion conditioning by defining Classifier-Free Guidance: it removes the need to train a classifier, but maintains the advantages discussed above by using scores derived from a conditioned and an unconditioned diffusion model to simulate the presence of a classifier's gradient.

3.2 Stable Diffusion

Based on Rombach et al. [29] Latent Diffusion Model, Stable Diffusion is an generative architecture that produces images. This model is composed by various modules: a Variational Autoencoder [30] which employs an encoder $\mathcal{E}$ that compresses the input sample to a smaller and more efficient latent space and a decoder $\mathcal{D}$ that converts the denoised latent into the final picture, a U-Net [26] that removes the noise applied right after the latent space-conversion of the data, and a text encoder that conditions the synthesis. In detail: an initial RGB image x of height H and width W is processed by the pretrained VAE encoder to get the latent $z = \mathcal{E}(x)$ of dimensions $h < H$, $w < W$ and $c = 4$ channels; the latent code obtained undergoes a diffusion process, then it is passed through the U-Net that is learning to reverse the previously defined noise addition step, as explained in Subsection 3.1. The backbone is enriched by cross-attention [31] layers, which help with the interpretation of conditional data y as text, segmentation maps or other pictures and thus influence the image generation: the conditioning information is first passed through a domain-specific encoder τ_θ that projects y to be correctly processed by cross-attention layers defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^\intercal}{\sqrt{d_k}})V \tag{3.3}$$

where d_k is the dimension of K, $Q = W_Q^{(i)} \varphi_i(z_t)$, $K^\intercal$ is the transposed matrix of $K = W_K^{(i)} \tau_\theta(y)$ and $V = W_V^{(i)} \tau_\theta(y)$ are the result of latent code z_t and conditional input $\tau_\theta(y)$ in the space of learned projection matrices $W_Q^{(i)}$, $W_K^{(i)}$ and $W_V^{(i)}$ representing respectively the query, key and value triplet. Once the backward diffusion process has terminated, the resulting latent data z is brought back in pixel space by the decoder $\mathcal{D}$, obtaining the final output $\bar{x} = \mathcal{D}(z)$.

3.3 CAMOUFLaGE-Light

Starting from the assumption that from an image x it is possible to divide sensitive and non-sensitive representations into R_s and $\neg R_s$ respectively; CAMOUFLaGE [8] models aim to perform the decomposition and use the non-personal details $\neg R_s$ to reconstruct a de-personalized picture, while maintaining its usefulness and meaning. In particular, CAMOUFLaGE-Light performs Image Anonymization by analyzing the individuals depicted in a photograph using IP-Adapter [32] and FACER [33], as visible in Figure 3.2; the first extracts both image features, by employing a pre-trained FaRL encoder [33] instead of ViT-H/14 from CLIP mentioned on the original IP-Adapter, and, during training, text features distilled from a corresponding caption by a pre-trained CLIP text encoder [34], these embeddings pass through a decoupled cross-attention layer in order to influence the image generation process. The FACER module is instead used to extract specific

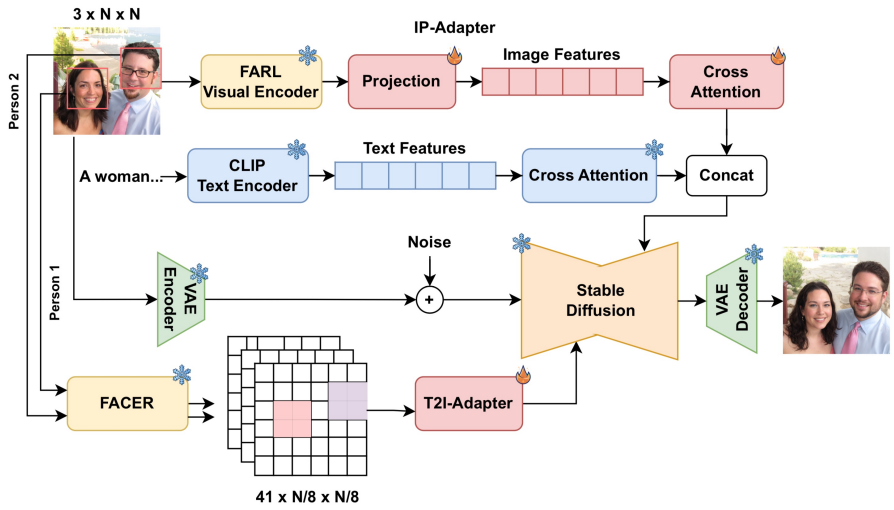

Figure 3.2: CAMOUFLaGE-Light architecture by [8], formed of an IP-Adapter at the upper branch, Stable Diffusion embedded in the center and feature maps interpreting T2I-Adapter below.

visual characteristics, such as hair color, presence of makeup and accessories like glasses, for a total of 40 different facial attributes taken from the individuals in the picture and organize them in the corresponding 40 data maps, to which is added another one containing faces' keypoints. The resulting feature matrix is passed through a trainable T2I-Adapter [35]. In order to create the final image, noise is added to the starting picture, which is then converted to the corresponding latent data and fed into a pre-trained Stable Diffusion model, as the one defined in Subsection 3.2, together with image encoding, text embeddings and data maps to manage the noise reversion process and obtain a new photo with the same set of non-personal data $\neg R_s$, but different sensitive information R_s than the image used at the beginning.

3.4 FRESCO

The FRESCO architecture [23] aims to provide an in-depth analysis of images, which is divided into 3 levels: plastic (lines, shapes and colors), figurative (objects and concepts) and enunciation (primarily about the observer's point of view). The FRESCO.v1 prototype, which is composed of a variety of classification models that perform image caption, object identification, depth estimation, panoptic segmentation, and many other different analyses, manages to accomplish a thorough measure of the three levels defined earlier, and compiles an image identikit that contains all the results about the examination performed.

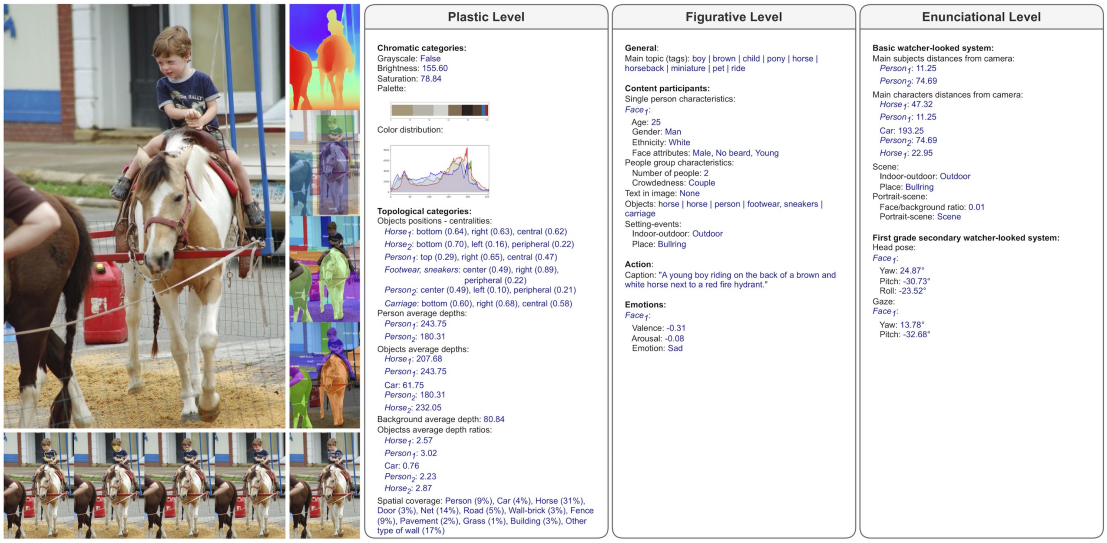

Figure 3.3: Example of annotation maps and identikit of the picture to the left. Image by [23].

3.5 Scene Graphs

Scene Graphs provide an ulterior kind of analysis about visual data: by representing the captured objects as nodes and relations as edges. The final scheme condenses both the subjects' attributes and how the actors portrayed relate to each other. This information allows distilling the general semantics of the picture, without referring to a unique photograph, showing potential in tasks such as image generation [36, 37, 38], image retrieval [39] and image captioning [40, 41, 42]. The modules appointed to perform Scene Graph Generation inside these architectures can be divided into two-stage and one-stage methods, where the former refers to performing object detection and graph generation in separate and consecutive steps as in [43], while the latter consists in inferring detection and generation in a single pass [40, 41].

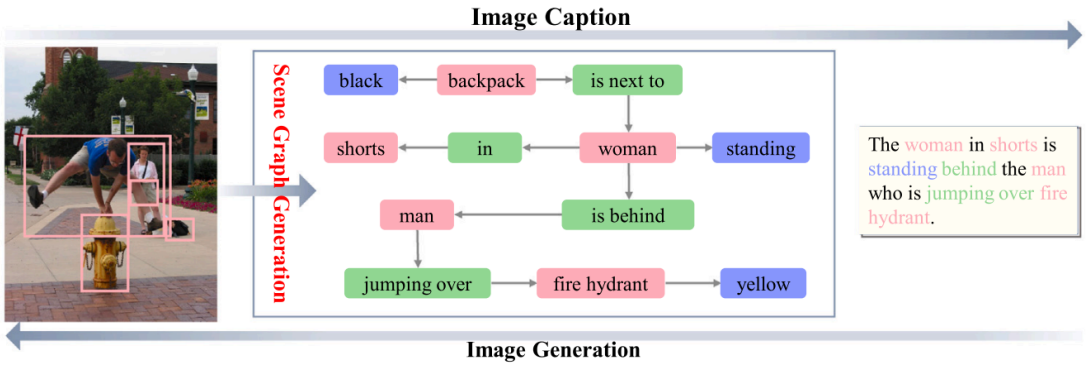

Figure 3.4: Correspondence between an image, its possible scene graph and the resulting caption. Image by [44]

3.6 RelTR

Relation Transformer [40] is a one-stage model trained on the Visual Genome dataset [45] that, given a photograph representing multiple objects, produces the corresponding scene graph. As shown in Figure 3.5, this architecture is composed of a Detection Transformer [46] and a Triplet Decoder module. DETR is an entity detection framework made of a CNN used to extract the image's features Z_0, which will be passed, together with the positional encoding E_p, to the encoder's self-attention layer that generates a feature context Z used by the decoder to extract from a set of learned entity queries the entity representations Q_e. The Triplet Decoder is split into subject and object branches that combine subject/object encodings E_s/E_o, triplet encodings E_t and subject/object queries Q_s/Q_o, respectively. The results get concatenated into $Q = K = [Q_s + E_s + E_t, Q_o + E_o + E_t]$ and processed by the same Coupled Self-Attention module,

which computes the triplets' context and dependencies between subject and objects $[Q_s, Q_o] = Att_{CSA}(Q, K, [Q_s, Q_o])$. Q_s and Q_o are split apart and passed to two separate Decoupled Visual Attention modules that, using the feature context Z mentioned previously, update Q_s and Q_o with the influence of Z: in the subject branch case the queries become $Q = Q_s + E_t$ and keys $K = Z + E_p$ and the output is computed as $Q_s = Att^{sub}_{DVA}(Q, K, Z)$; the object branch mirrors the same functioning using Q_o. Additionally, DVA computes the attention heat maps M. The third attention module used is Decoupled Entity Attention, which is provided with entity representations Q_e and thus unifies the output of DETR and Triplet Decoder by adjusting $Q_s = Att^{sub}_{DEA}(Q_s + E_t, Q_e, Q_e)$ and in the same way updating Q_o. Finally, triplets are defined by processing queries Q_s and Q_o via two different FFNs and the attention maps M through a CNN to obtain spatial feature vectors V_{spa} and the final predictions are computed as $\hat{p}_{prd} = softmax(MLP([Q_s, Q_o, V_{spa}]))$, representing subject and object's bounding boxes, class labels and predicate label.

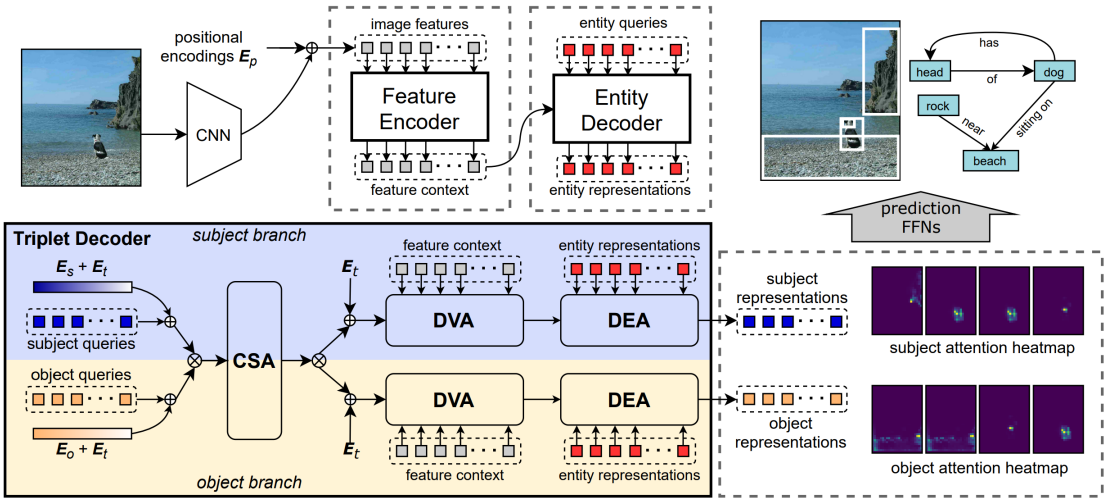

Figure 3.5: RelTR model's schema by [40], composed of DETR in the upper-left section, the Triplet Decoder in the lower-left corner and the FFN finalizing the output to the right side.

Chapter 4

Methodology

Given the partial attention regarding visual semiotics in most of today's state-of-the-art models, this thesis' work is focused on extending the use of this type of analysis to condition Stable Diffusion's image generation in [8]'s CAMOUFLaGE-Light architecture with the help of detailed analysis performed by FRESCO [23] and a scene graph of relations produced by RelTR [40], with the final goal of enriching CAMOUFLaGE-Light with more text features while using less visual data, which is the main cause of re-identification.

4.1 Extended Scene Graph

As previously stated in Section 3, the results of the image analysis performed by FRESCO [23] are concentrated into a JSON file, called identikit, and visualization maps depicting panoptic segmentation, depth data, edge detection, and others. Of these different types of data, a subset of them will be used to extract information that will be employed to compose the Extended Scene Graph (ESG). In particular, ESG will contain three different sections: a "scene" section composed of the image caption produced by CLIP and the picture's dimensions, useful for scaling the bounding boxes to the resolution in which the pipeline works, the "objects" list containing the instances detected with their related characteristics, and the "relationships" section representing the interactions between objects. From the information stored in the identikit file, objects are defined as follows:

- id: unique identifier;
- type: label of the detected object;
- position: contains the upper-left and lower-right points of the object's bounding box;

- depth: estimates the distance of the object from the camera.

The "depth" attribute deserves a brief in-detail explanation: the value assigned is computed by intersecting panoptic segmentation's pixels and object detection's bounding box, which, when the designated area of pixels encloses more than the *completely_contained* threshold of an instance's total amount of pixels, allows to determine which instance's mask of the panoptic segmentation corresponds to the detected object. The most present instance within the bounding box under examination is defined as:

$$i^* = argmax(\frac{\sum^B p_i}{\sum^B p}) \tag{4.1}$$

where B is the bounding box mentioned, p_i symbolizes a pixel of the i-th instance's mask inside B, and p is a generic pixel of B. Once the best candidate instance for a bounding box is found by confronting the corresponding $\frac{\sum^B p_{i^*}}{\sum^B p}$ with the *completely_contained* threshold, its panoptic mask is used to take into consideration only the same pixels from the depth map and compute on that data the average depth of the instance in the following manner:

$$depth_{i^*} = \frac{\sum^B d_{i^*}}{\sum^B p_{i^*}} \tag{4.2}$$

where d_{i^*} is the value of a point in the depth map corresponding to i^* contained in B. When the best candidate cannot be found because the *completely_contained* threshold is not being met, the instance closer to the camera's point of view is selected, with the idea that the most important object of the detection could be placed in front of other less relevant things. Objects' bounding boxes are sorted by ascending area in order to process instances composed of a low amount of pixels first, which could tamper with larger boxes that could completely contain multiple object masks. For the list of detectable objects, the reader is referred to Prismer [47], the model used for object detection. Since the main goal of this work is to generate anonymized images without losing useful information, general data about persons' faces are linked to "human face" objects, which are detected by FACER's RetinaFace module [33]:

- face attributes scores: a list of the subject's facial attributes, the same that CAMOUFLaGE-Light [8] computes using FACER [33];
- age: estimates how old the subject is;
- gender scores: a list of the subject's gender scores;
- ethnicity scores: a list of the subject's ethnicity scores;

- emotion scores: a list of the subject's emotion scores;
- head pose: describes the instance's head position by evaluating yaw, pitch and roll;
- gaze direction: describes the instance's gaze direction by evaluating eyes' XY positions, yaw, and pitch.

Face attributes data come with a bounding box which defines the corresponding detected face. In order to match this area with a "human face" object previously detected, its bounding box and the one corresponding to the face attributes are matched through Intersection over Union (IoU), then confronted with a lower limit *bbox_match* threshold. In the "gaze direction" case, only the eyes' position is provided, thus it is sufficient to check in which bounding box those coordinates are comprehended. After the nodes have been established, edges that represent relations are outlined in this fashion:

- source: source object's id;
- target: target object's id;
- type: label of the relation.

Some of the possible relationships are assigned by confronting two objects: based on their relative position it is possible to define the general composition of the image:

- in front of;
- behind;
- next to;
- below;
- above;
- to the right of;
- to the left of;
- below-right of;
- below-left of;
- above-right of;
- above-left of.

In particular: "in front of", "behind", and "next to" are depth-related spatial interactions, where the first two are assigned when source and target objects' depths ratio $d_r = \frac{d_{source}}{d_{target}}$ is above the *depth_upper_limit* parameter or below the *depth_lower_limit* one, respectively, while every other case of overlapping bounding boxes falls into the last type, and "below", "above", "to the right of" and the rest concern only non-overlapping objects and the distance between source and target's boxes centroid, which has to be above the *positional_relation_tolerance* threshold. "human face" objects and "person" objects get linked together based on which "person" instance, in the panoptic segmentation computed by Prismer's module Mask2Former [47] contains the central pixel of the "human face" bounding box. Finally, additional relationships are imported by RelTR's generated scene graph: triplets in the form of "<subject> <relation> <object>" with the addition of subject and object's detected bounding boxes are confronted with the objects' position attribute; when both subject (source node) and object (target node) find a matching object based on bounding boxes' IoU greater than a configurable threshold, a new relation is preposed to the spatial ones defined above, since the interactions identified by RelTR are more meaningful.

4.2 Data maps

The first use of the extended scene graph is to compose 61 data maps, at 1/8 of the photo's resolution, which are then passed to T2I-Adapter [35], a lighter alternative to ControlNet [48], which contributes to conditioning the final image. In order to compose this feature matrix, attribute scores of the listed objects are used: from non-"human face" entities depth data is employed for the corresponding data map, while from "human face" objects every additional attribute is used to compile 60 features, including depth. The final and 61st map contains the body pose of "human" subjects. The complete list of feature maps is the following:

- depth;
- facial attributes scores:
  - 5 o clock shadow;
  - arched eyebrows;
  - attractive;
  - bags under eyes;
  - bald;
  - bangs;
  - big lips;
  - big nose;
  - black hair;
  - blond hair;
  - blurry;
  - brown hair;
  - bushy eyebrows;
  - chubby;
  - double chin;
  - eyeglasses;
  - goatee;

  - gray hair;
  - heavy makeup;
  - high cheekbones;
  - male;
  - mouth slightly open;
  - mustache;
  - narrow eyes;
  - no beard;
  - oval face;
  - pale skin;
  - pointy nose;
  - receding hairline;
  - rosy cheeks;
  - sideburns;
  - smiling;
  - straight hair;
  - wavy hair;
  - wearing earrings;
  - wearing hat;
  - wearing lipstick;
  - wearing necklace;
  - wearing necktie;
  - young;
- age;
- gender scores:
  - gender score woman;
  - gender score man;
- ethnicity scores:
  - asian;
  - indian;
  - black;
  - white;
  - middle eastern;
  - latino hispanic;
- emotion scores:
  - neutral;
  - happy;
  - sad;
  - surprise;
  - fear;
  - disgust;
  - anger;
  - contempt;
- gaze direction;
- head pose;
- body pose.

As Figure 4.1 shows, data maps can be quite diverse; "depth" information is composed by using the values of every entity, computed as explained in Subsection 4.1 and confined into the objects' bounding box area, while scores regarding "facial attributes", "age", "gender", "ethnicity" and "emotion" make use of "human face" entities' areas. The "gaze direction" feature traces the position of the eyes and lines symbolizing the angle that connects them to the point observed by the subject. The "head pose" map is defined by drawing the three axes of estimated yaw, pitch, and roll, starting from the center of the target "human face" object and by using

grayscale in order to differentiate them. Finally, the matrix produced by OpenPose [49] containing body poses of the depicted individuals highlights how keypoints are linked, thus constructing subjects' skeleton, hands and faces. Each feature map is computed separately, exception made for body pose information, and aggregated at a later time, with the intention of keeping only the maximum value in case of overlapping data, as it is visible in "depth" and "blond hair" features depicted in Figure 4.1, since adding them together would affect the normalization of said data.

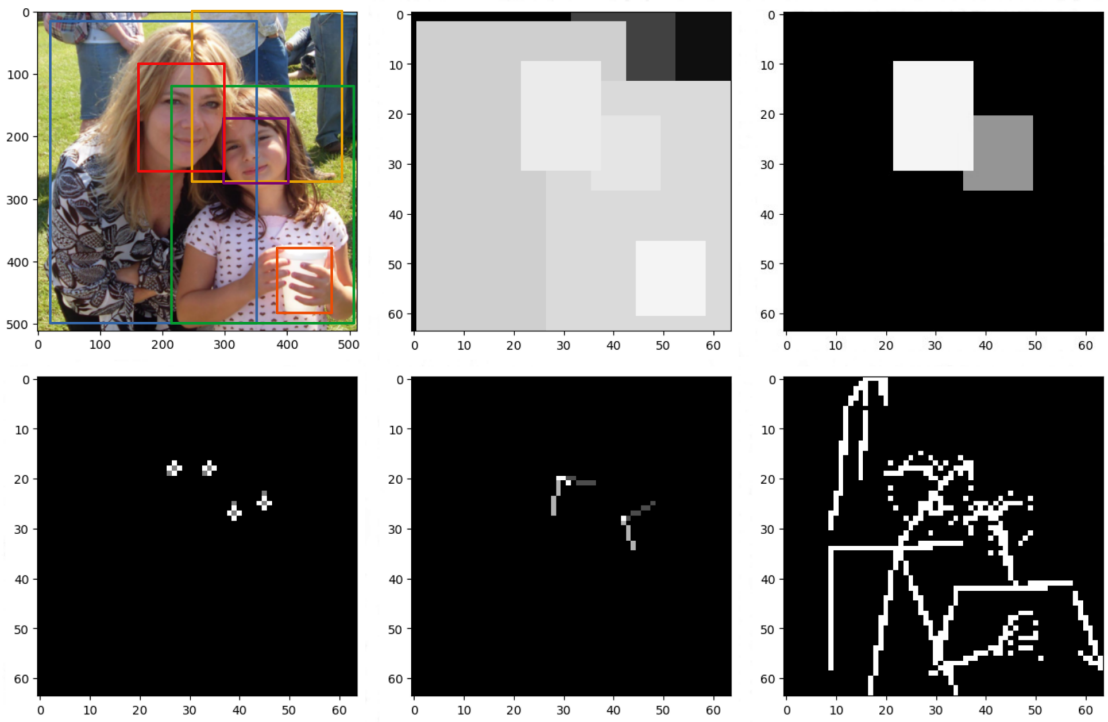

Figure 4.1: Upper left: original image with object detection annotation (person, human face, person, human face, cup, person), indicative of which area belongs to each detected entity; upper middle: depth data; upper right: blond hair feature's score; lower left: gaze direction; lower middle: head pose; lower right: body pose. Every data map is used to condition the model's generation process.

4.3 Relation triplets

With the goal of adding context and meaning to the text prompt used by Stable Diffusion, in parallel to the image's caption, the second use of ESG is to create a collection of the relationships detected between object pairs: triplets in "<subject> <relation> <object>" format are built by using the relationship entries of source

object, relation type and target object; such triplets get concatenated until the upper limit of 77 tokens manageable by CLIP [34] is reached. Within the ESG, interactions predicted by RelTR are the first ones to be listed, while spatial relations come right after, with the idea that when the token limit is reached, only less descriptive interactions remain unused. The final string is then processed by the text encoder, and the resulting embeddings are concatenated to the caption encoding in order to guide the image generation process. An example of the relation triplets composed starting from an ESG is:

```
cup in front of person, cup to the right of person, cup below
person, cup below-left of footwear, sneakers, cup below person,
person next to person, person in front of person, person in
front of footwear, sneakers, person in front of person, person
in front of person, person below-left of footwear, sneakers,
person below-left of person
```

4.4 Final architecture

In summary, the main differences between CAMOUFLaGE-Light and the proposed method are:

- the use of image captions during inference, before it was employed only at training time;
- the use of a secondary caption composed of relation triplets;
- the 20 additional data maps passed to T2I-Adapter.

Now that the main implementations have been discussed, the complete training process can be defined. As Figure 4.2 displays, the original image is used by different models in parallel: IP-Adapter uses FaRL to extract visual features, VAE's encoder produces an embedding into the latent space, FRESCO computes the in-depth image analysis, and RelTR defines a fitting scene graph; the last two data structures are refined and combined into an ESG as previously explained, from which image caption and relation triplets are kept to obtain text features via CLIP's text encoder, and the objects list is used to compose the data maps. Once this is all set, the caption encoding and triplets embedding are processed by frozen cross-attention modules, while image features are refined by trainable ones followed by a Resampler Q-Former module[1], which performs image projection to obtain tokens. The resulting outputs are concatenated to be received by Stable

[1]https://github.com/shan18/Perceiver-Resampler-XAttn-Captioning

Diffusion in order to condition image denoising, while the data maps are processed by T2I-Adapter and subsequently provided to Stable Diffusion for regulating the generation process. To the image's latent embedding is added Gaussian noise, then it is passed to SD's UNet so to begin the denoising phase. The optimized loss function is based on the Mean Squared Error between the noise added to the image's latent code and the predicted one, the latter being used by VAE's decoder to produce the final anonymized image.

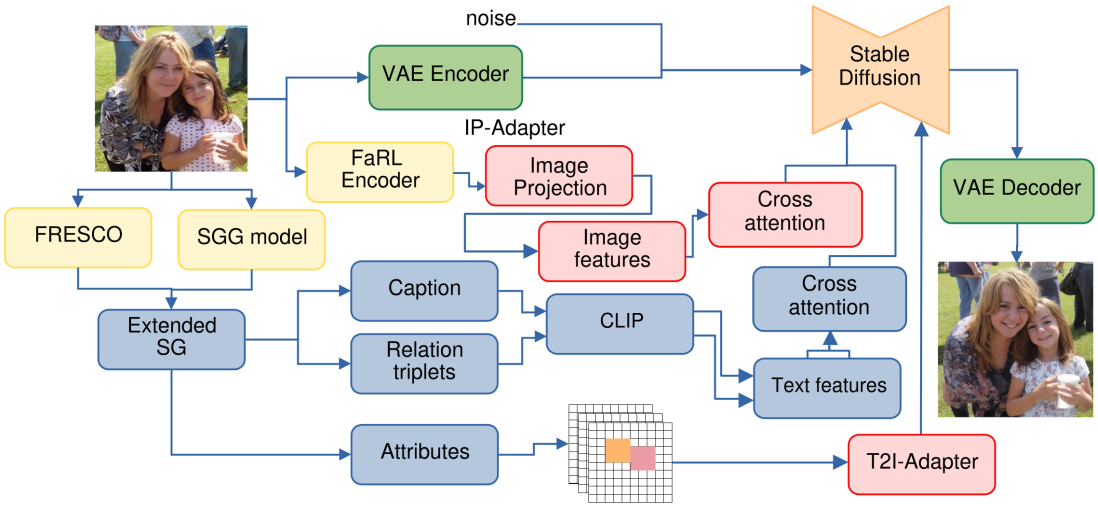

Figure 4.2: Schema of the complete architecture.

Chapter 5

Experimental environment

This chapter defines how the proposed method has been trained and tested, comprehending the datasets used, the hyperparameters configured, the metrics used to measure its performance and some output examples.

5.1 Datasets

During the training phase, it has been used the Flickr-Faces-HQ-in-the-wild (FFHQ-itw) dataset proposed by Karras et al. [16] to train StyleGAN and later established as a pillar of training and testing generative image models. It is composed of 70,000 images in 1024×1024 resolution, focused on portrait images with high variability of age, ethnicity, background, and accessories. Original images were rescaled to the size of 512×512 pixels prior to use. The test results have been computed on a subset of CelebA-HQ [50], composed of 1,000 images in 512×512 resolution, each depicting the face of a famous person, thus making this set a good choice for evaluating how the designed model performs when data are mainly focused on visage, and an OpenImages-v7 [51] subset of 2,002 pictures in the same 512×512 definition, which represent scenes with more variability, making it possible to perform inference on pictures containing more than one person.

5.2 Training

As the first step, the FFHQ-itw dataset has been preprocessed to obtain the images' identikits, with FRESCOv1 [23], and scene graphs, using RelTR [40]. The model configuration of FRESCO is the default one featured in its own GitLab repository[1],

[1] `https://gitlab.com/grains2/fresco`

and in the same way RelTR has been executed using the pretrained weights in *checkpoint0149.pth* with the default hyperparameters, both can be found in the official implementation's codebase[2]. In order to aggregate identikits and scene graphs into ESGs, *min_objects_for_relations* has been set to 2, making possible the presence of the relation list only when at least 2 objects have been recognized, *bbox_match* at 0.75 as the threshold to be reached by the Intersection over Union between two bounding boxes in order to consider them as referring to the same object, *completely_contained* to 0.95 as the lower limit above which the proportion of pixels belonging to a segmentation mask is considered to be entirely included into a bounding box, *depth_overlap* at 0.0 the minimum IoU value at which two objects are considered eligible for a depth-related spatial relation, *depth_upper_limit* and *depth_lower_limit* respectively set to 1.1 and 0.9 as thresholds that regulate at which depth ratio one object can be considered as "in front" or "behind" another one, and *positional_relation_tolerance* at 0.5 to manage the limit above which the distance between two bounding box centroids has to be in order to be eligible for a directional relation. The learning phase has been divided into two separate steps: the first one using only IP-Adapter with image and text features, both caption and relation triplets, for 120,000 optimization steps, while the second phase included both IP-Adapter and T2I-Adapter to add data maps-conditioning for ulterior 100,000 optimization steps. Both phases account for a total of 220,000 steps, the whole FFHQ-itw dataset was used as defined before, using *AdamW* optimizer, a learning rate of 1e-4, weight decay set to 1e-2 and the batch size was 8. The image encoder being *Green-Sky/FaRL-Base-Patch16-LAIONFace20M-ep64*, and the Stable Diffusion model used was *stablediffusionapi/realistic-vision-v51.*

5.3 Inference

Once the model had completed training, it has been used to generate images based on CelebA-HQ and OpenImages v7 data. Again, the first step is to preprocess with FRESCOv1 and RelTR the entire datasets, and then merge the results into ESGs; the preprocessing configurations have been maintained the same as the ones used during training. This method's inference can be customized by acting on the hyperparameters dimension d, strength s, guidance scale g and timesteps T. d is the size of the final picture and can be greater than the starting one thanks to rescaling, s regulates how much of the original picture will be overwritten with noise before using it as a starting point for the generation, g regulates how much the final output should adhere to the conditioning data other than a condition-free generation, and T is the amount of steps used to complete the denoising phase. The

[2] `https://github.com/yrcong/RelTR`

same hyperparameter values have been set regardless of which dataset was used as the starting point: it has been tested d of both 512 and 768 pixels in combination with s of 0.6 and 0.8 to evaluate multiple levels of image obfuscation, while cfg and T have been fixed to 3.0 and 30, respectively. Additionally, it has been set a fixed Random Number Generator seed of 12345 to facilitate the reproducibility of this work. Unlike the case of CAMOUFLaGE-Light [8], both caption and relation triplets are provided during inference.

Figure 5.1: Inference example with different resolutions and noise proportions. Left: original image; upper middle: $d = 512$, $s = 0.6$; upper right: $d = 512$, $s = 0.8$; lower middle: $d = 768$, $s = 0.6$; lower right: $d = 768$, $s = 0.8$.

5.4 Evaluation metrics

To quantitatively evaluate the performance of the proposed model, Fréchet Inception Distance (FID) has been used [18] to assess image quality by estimating the distance of the features, extracted by an Inception-v3 [52] neural network, between the original data and the anonymized one. Visual Distributions of Neuron Activation (VisualDNA) [22] was also employed to measure the neuron activation differences of a Mugs-ViT-B [53] model between the original and anonymized datasets with Earth

Mover's Distance (EMD) [54]. As an additional test, it has been used VisualDNA to confront original and generated images in pairs and then compute the average the score to identify the distance between the data. Re-identification rate tests have been conducted at both face and image level. The former is based on using Multitask Convolutional Neural Network (MTCNN) to extract face crops from the pictures. Then, a FaceNet architecture pre-trained on VGGFace2 and CASIA WebFace datasets, computes for each anonymized face its K-Nearest Neighbors (K-NN) based on their extracted features. Instead, image-level anonymization evaluation adopts a CLIP visual encoder to extract features from pictures and then determines a set of K-NN photos. The presence of the original input inside the sets of K-NN define an high, and worse, re-identification rate. It has been measured Re-ID@K with ranks $K = [1, 5, 10]$ to show how the anonymization rate changes the more top-K neighbors are considered, and mAP@50. To conclude the performance assessment, it has been evaluated how much a pretrained model output would differ when provided with generated data instead of the original: by testing the FACER [33] model with pictures from CelebA-HQ and the anonymized counterparts, it has been computed the distance between the original facial features annotated for the dataset and the actual output of FACER.

5.5 Baselines

The architectures confronted with the proposed model are the following:

- DeepPrivacy2 [4]: full-body anonymization realized by inpainting synthesized masks on the original image, resulting in images of 250×250 pixels;
- FALCO [6]: focused on face anonymization, generates a completely new 1024×1024 picture by mixing the original image's latent code with the one belonging to a picture selected from a synthesized dataset;
- CAMOUFLaGE-Base [8]: analyzes the photo with a series of task-specific modules to condition the reconstruction of a heavily obfuscated version of the initial image ($s = 0.9$) to obtain a 768×768 image;
- CAMOUFLaGE-Light [8]: similarly to the Base version, extracts visual information with lighter modules that allow a lower rate of obfuscation $s = 0.6$, but generates 768×768 pictures in a shorter time.

Being CAMOUFLaGE-Light the predecessor and direct competitor, in terms of architecture similarity, of the new method, their confrontation is the main focus of the following sections.

Chapter 6

Results

In the following, the results of performance metrics defined in the previous chapter have been reported and they are then confronted with other state-of-the-art architectures. Finally, a discussion regarding the reasons of such results is presented.

6.1 Inference samples

Figure 6.1 shows how the same images have been generated by the mentioned architectures and the defined method using different configurations, with CelebA-HQ images as starting points, where photos are portraits of only one person each, while Figure 6.2 represents the synthesis of pictures inferred from the OpenImages v7 dataset, which contains more complex scenarios where multiple subjects come into play. It possible to notice a degree of similarity between CAMOUFLaGE-Light and the architecture proposed with parameters $s = 0.6$ and $d = 768$, as both models are composed of very similar modules and share the same s and d values during inference.

6.2 Image quality and fidelity

Table 6.1 shows that in terms of picture quality, when evaluating the synthesized datasets with FID, every configuration of the proposed model performs better than the GAN-based methods DeepPrivacy2 and FALCO, which score 49.4 and 41.2, respectively, while CAMOUFLaGE-Base shortens the distance when the parameter *anonymization scale* is set to 1.0, reaching 35.4. CAMOUFLaGE-Light gets the best result of 28.7, slightly better than its extended version, which scores 30.8 when dimension $d = 768$ and noise strength $s = 0.6$. Similar results are achieved by evaluating the data distribution using VisualDNA, where CAMOUFLaGE-Light achieves the best score of 4.6, followed by DeepPrivacy2 at 5.0 and this thesis'

Figure 6.1: Images generated by the evaluated models based on the CelebA-HQ dataset. From top row to bottom: original images, DeepPrivacy2, FALCO, CAMOUFLaGE-Base, CAMOUFLaGE-Light, the proposed model at 512p and 60% noise, 512p and 80% noise, 768p and 60% noise, 768p and 80% noise.

model at both 512 and 768 resolution, scoring 5.8 and 6.0 respectively. Instead, when computed on pairs of original photo and corresponding anonymized image, VisualDNA measures image fidelity, highlighting a different trend: DeepPrivacy2, as an inpainting approach where the amount of modified pixels is small, scores 10.4 ± 1.3, overtaking by a meaningful margin CAMOUFLaGE-Light's 12.3 ± 1.4 as the second best and the proposed architecture at 768p and 60% noise with slightly higher 13.3 ± 1.4. It is also possible to notice the difference in image fidelity when, in the new approach, the noise strength s has been set at 60% and 80%: in the latter both metrics register substantially greater dissimilarity between the

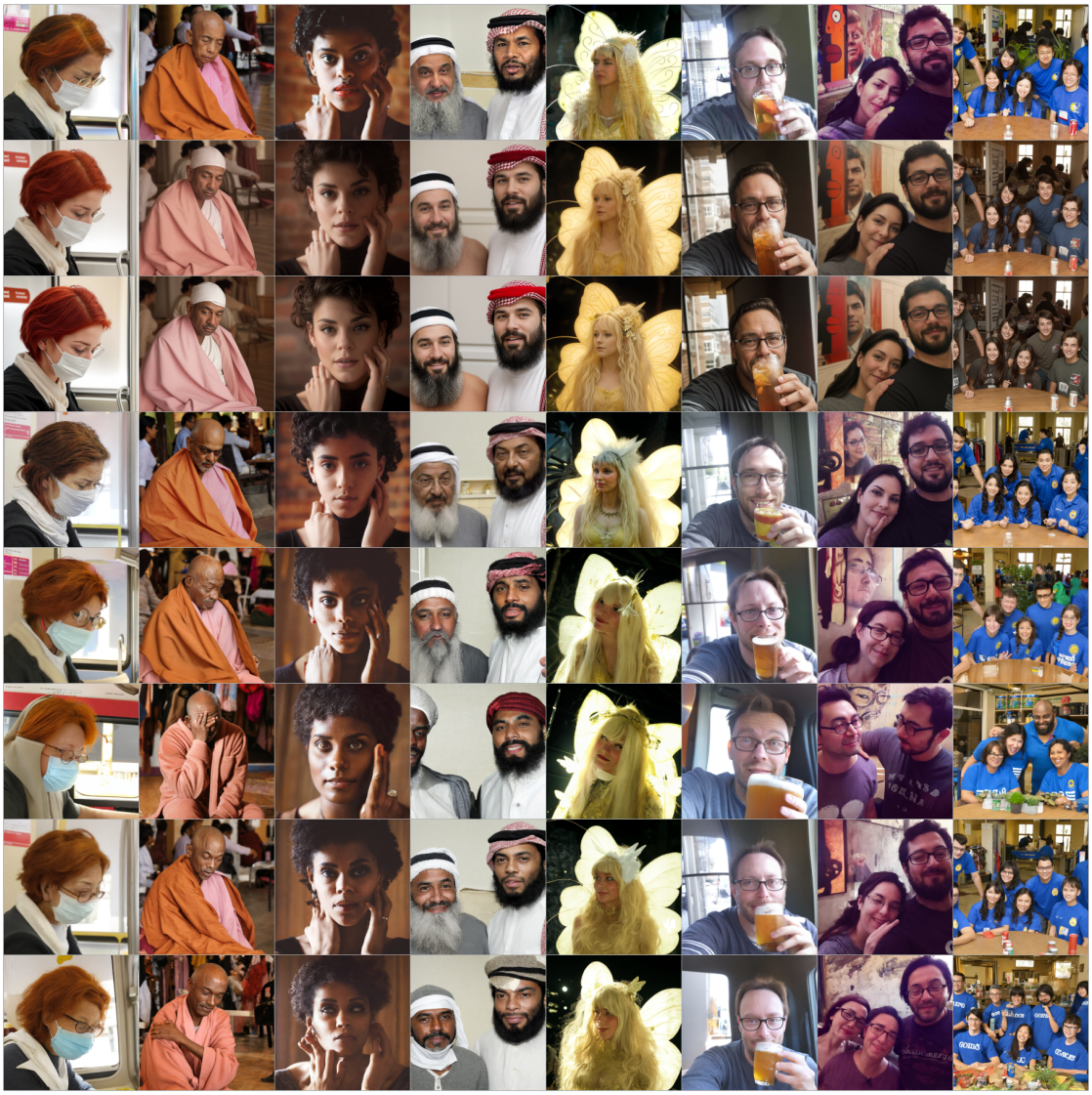

Figure 6.2: Images generated by the evaluated models based on the OpenImages v7 dataset. From top row to bottom: original images, CAMOUFLaGE-Base $a_s = 1.0$, CAMOUFLaGE-Base $a_s = 1.25$, CAMOUFLaGE-Light, the proposed model at 512p and 60% noise, 512p and 80% noise, 768p and 60% noise, 768p and 80% noise.

original data and the synthesized one, which is the expected behavior when a larger portion of the image gets obfuscated. The numerical results show that the best overall configuration for the designed model confirms the settings applied to CAMOUFLaGE-Light: dimension $d = 768$ and noise strength $s = 0.6$.

Method	Celeba-HQ	OpenImages v7
	FID	
DeepPrivacy2	49.4	-
FALCO	41.2	-
CAMOUFLaGE-Base ($a_s = 1.0$)	35.4	31.5
CAMOUFLaGE-Base ($a_s = 1.25$)	41.5	38.3
CAMOUFLaGE-Light	**28.7**	<u>30.1</u>
Proposed method $(512, 0.6)$	32.2	**28.6**
Proposed method $(512, 0.8)$	39.2	34.2
Proposed method $(768, 0.6)$	<u>30.8</u>	**28.6**
Proposed method $(768, 0.8)$	40.7	35.8
	VisualDNA (dataset-level)	
DeepPrivacy2	<u>5.0</u>	-
FALCO	6.3	-
CAMOUFLaGE-Base ($a_s = 1.0$)	8.9	6.9
CAMOUFLaGE-Base ($a_s = 1.25$)	9.9	7.8
CAMOUFLaGE-Light	**4.6**	<u>4.2</u>
Proposed method $(512, 0.6)$	5.8	**3.7**
Proposed method $(512, 0.8)$	6.7	<u>4.2</u>
Proposed method $(768, 0.6)$	6.0	4.3
Proposed method $(768, 0.8)$	6.5	4.8
	VisualDNA (image pair-level)	
DeepPrivacy2	**10.4±1.3**	-
FALCO	15.2 ± 2.1	-
CAMOUFLaGE-Base ($a_s = 1.0$)	15.9 ± 2.0	16.5 ± 2.7
CAMOUFLaGE-Base ($a_s = 1.25$)	17.7 ± 2.1	18.5 ± 2.9
CAMOUFLaGE-Light	<u>12.3 ± 1.4</u>	<u>14.9 ± 2.8</u>
Proposed method $(512, 0.6)$	13.7 ± 1.3	15.4 ± 2.3
Proposed method $(512, 0.8)$	16.1 ± 1.6	19.1 ± 3.0
Proposed method $(768, 0.6)$	13.3 ± 1.4	**14.7±2.5**
Proposed method $(768, 0.8)$	16.0 ± 1.7	18.8 ± 3.1

Table 6.1: FID computed on dataset distribution and VisualDNA executed at dataset level and original-anonymized image pair level test results across different models (lower is better).

6.3 Anonymization

Table 6.2 shows that the best result of 17.6% when tested for re-identification rate at rank 1 with the environmental visual features extracted by CLIP's ViT-B/32, is

achieved by the proposed model using resolution $d = 512$ and obfuscation $s = 0.8$ during inference, and falls behind CAMOUFLaGE-Base's 8.7%, when using the anonymization scale parameter $a_s = 1.25$, and FALCO's 10.9%, which impact more heavily on the totality of the image as FALCO generates it from scratch and CAMOUFLaGE-Base uses a noise strength $s = 0.9$, in both cases the portionn of original data is lower than what our model uses. When testing the anonymization focused on face features, the newly proposed model reaches a Re-ID@1 of 3.1%, still worse than FALCO's 0.2% when testing with FaceNet trained on VGGFace2, while it shortens the distance to 2.6% against 0.4% while measuring with the same architecture trained on the CASIA WebFace dataset. The high re-identification rate could be caused by using too much of the starting image, with $s = 0.6$, as a foundation to generate the output, and combining it with the higher number of features extracted from the Extended Scene Graph puts this model behind of CAMOUFLaGE-Light scores, which operates with an obfuscation of $s = 0.6$ and performs a face-swap before regenerating the image, thus achieving better results.

6.4 FACER inference

It will now be discusses how a pretrained model interacts with de-personalized pictures with respect to the original ones. FACER [33] is a state-of-the-art face-related toolkit, able to detect faces, segment their regions and categorize some of their characteristics. Because of the last capability, FRESCOv1 adopts it to implement facial attribute classification, ready to be utilized by the method defined in this thesis' work for the composition of data maps. Before performing the test, it has been used the CelebA-HQ subset of 1000 images together with their face attribute annotations to compute FACER's accuracy on this dataset: it achieved an accuracy score of 91.35, meaning that the model's predictions correspond, to an acceptable extent, to the truth stored in the annotations. Then, a closer look as been paid to the distance between ground truth and the score of predicted features when inferencing on CelebA-HQ original images and anonymized pictures at different resolutions and obfuscation: Figure 6.3 shows how every single facial attribute is perceived differently from its actual presence on average, representing in a negative value how much that characteristic is being recognized where it isn't supposed to, and in positive the case in which the attribute has not been found. Characteristics as "wearing hat" and "gray hair" present very low distances in every set of pictures, meaning that their presence is accurately confirmed by FACER, while "wearing necklace" is not being recognized as much as it should, given that in every image collection FACER does not find it at least 10% of the time. On the contrary, the "oval face" character is recognized more than the labels would suggest, as FACER detects it almost 20% more than the cases annotated when looking at

Method	Re-ID@1	Re-ID@5	Re-ID@10	mAP@50
	CLIP ViT-B/32			
DeepPrivacy2	0.306	0.440	0.497	0.123
FALCO	0.109	0.217	0.278	0.055
CAMOUFLaGE-Base ($a_s = 1.0$)	0.211	0.343	0.411	0.084
CAMOUFLaGE-Base ($a_s = 1.25$)	**0.087**	**0.162**	**0.205**	**0.037**
CAMOUFLaGE-Light	0.399	0.571	0.638	0.157
Proposed method $(512, 0.6)$	0.457	0.625	0.689	0.184
Proposed method $(512, 0.8)$	0.176	0.312	0.390	0.086
Proposed method $(768, 0.6)$	0.574	0.735	0.790	0.230
Proposed method $(768, 0.8)$	0.201	0.342	0.429	0.095
	VGGFace2			
DeepPrivacy2	0.008	0.023	0.036	0.007
FALCO	**0.002**	**0.008**	**0.015**	**0.003**
CAMOUFLaGE-Base ($a_s = 1.0$)	0.096	0.192	0.250	0.065
CAMOUFLaGE-Base ($a_s = 1.25$)	0.018	0.046	0.067	0.014
CAMOUFLaGE-Light	0.046	0.115	0.146	0.040
Proposed method $(512, 0.6)$	0.103	0.202	0.257	0.075
Proposed method $(512, 0.8)$	0.031	0.077	0.113	0.029
Proposed method $(768, 0.6)$	0.185	0.289	0.361	0.123
Proposed method $(768, 0.8)$	0.043	0.086	0.117	0.034
	CASIA			
DeepPrivacy2	0.008	0.024	0.037	0.006
FALCO	**0.004**	**0.011**	**0.021**	**0.003**
CAMOUFLaGE-Base ($a_s = 1.0$)	0.100	0.200	0.260	0.054
CAMOUFLaGE-Base ($a_s = 1.25$)	0.019	0.049	0.070	0.012
CAMOUFLaGE-Light	0.036	0.118	0.160	0.030
Proposed method $(512, 0.6)$	0.090	0.183	0.254	0.057
Proposed method $(512, 0.8)$	0.026	0.081	0.117	0.022
Proposed method $(768, 0.6)$	0.152	0.303	0.374	0.090
Proposed method $(768, 0.8)$	0.035	0.080	0.124	0.023

Table 6.2: Re-ID test results across different models. The images used come from a CelebA-HQ subset of 1,000 images (lower is better).

the official CelebA-HQ subset. This test proved that, in general, the facial features present in the initial data are kept after anonymization, preserving the usefulness of the information; in fact, FACER achieves an accuracy score between 87% and 89% when inferencing on de-personalized data. The comparison of the proposed method's best configuration with CAMOUFLaGE-Base and CAMOUFLaGE-Light, shown in Figure 6.4, highlights how the new architecture tends to maintain more attributes with respect to the other models: FACER achieves an accuracy of 87.1% when tested on the data generated by CAMOUFLaGE-Light, comparable with the 87.4% score when inferring on CAMOUFLaGE-Base with parameter anonimization

scale set to 1.0. The worst accuracy of 85.9% is registered after testing with CAMOUFLaGE-Base coupled with $a_s = 1.25$, 4% lower than the one achieved with the architecture suggested in this thesis' work, making it the best performing in terms of data utility preservation.

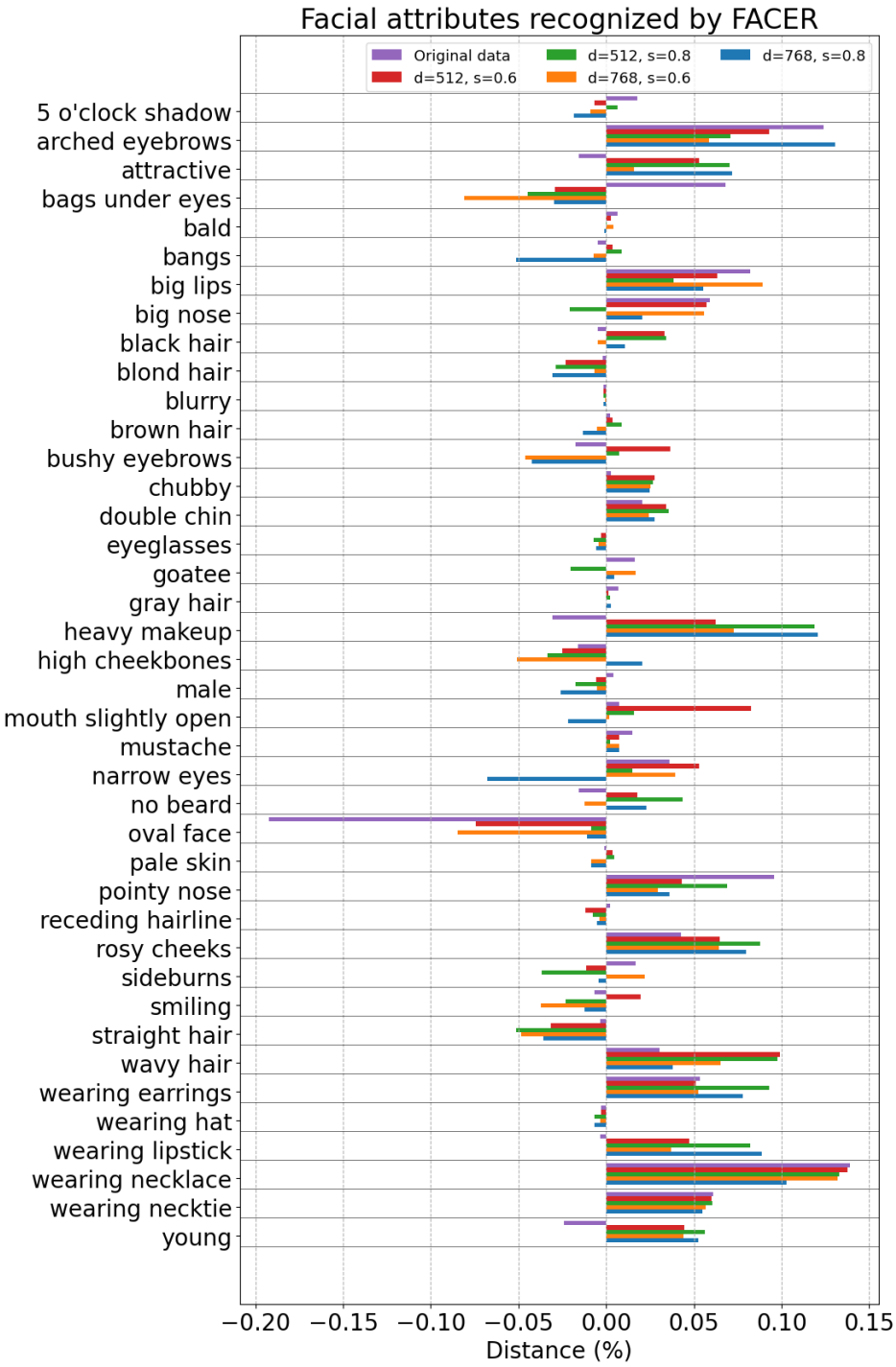

Figure 6.3: Distances between attributes ground truth and detected by FACER with different model configurations. Distances > 0 mean false negatives, distances < 0 symbolize false positives (lower absolute value is better).

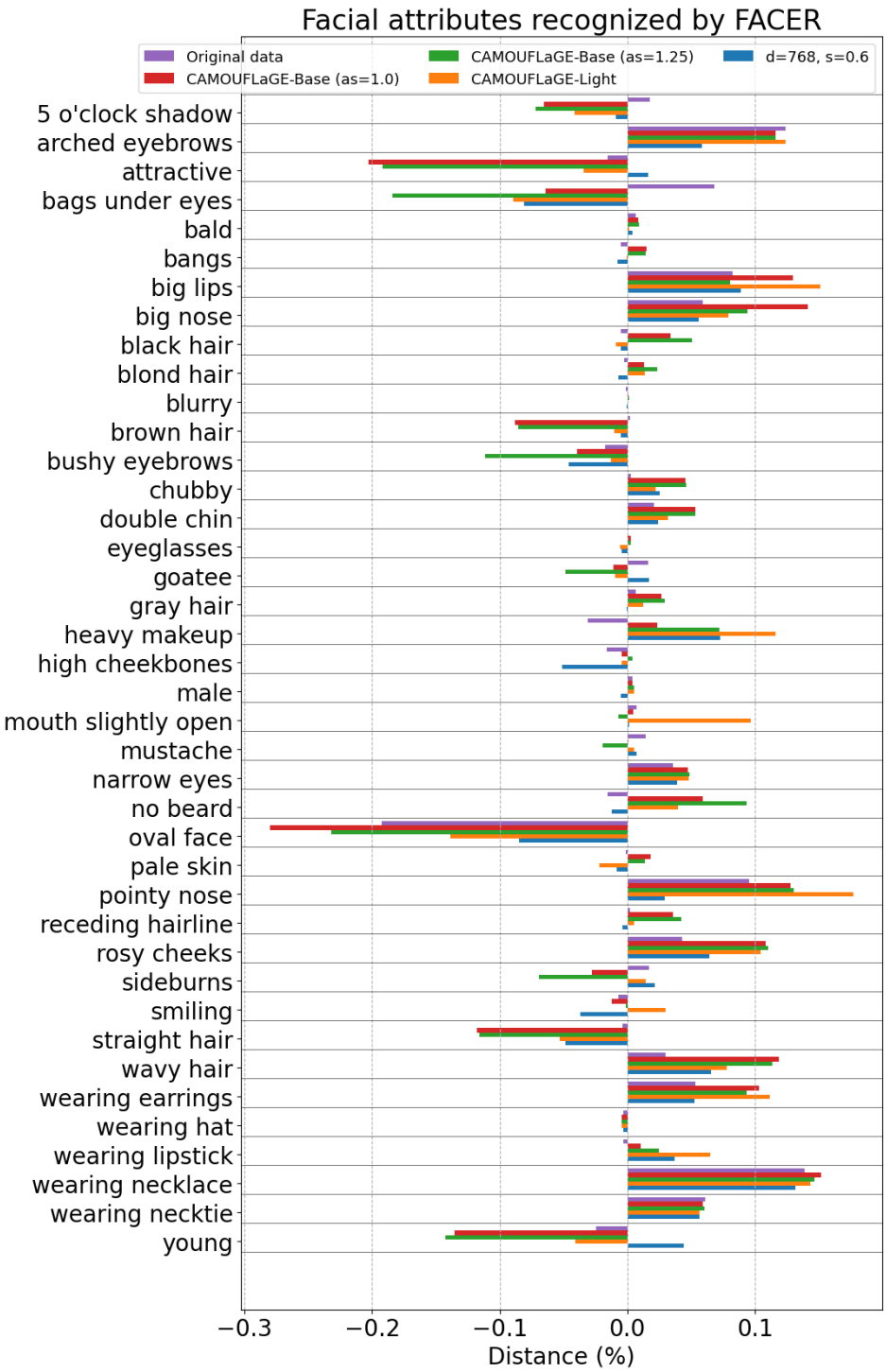

Figure 6.4: Distances between ground truth and attributes detected by FACER with different models. Distances > 0 mean false negatives, distances < 0 symbolize false positives (lower absolute value is better).

Chapter 7

Ablation study

This chapter focuses on the real contribution of integrating relation triplets into training and testing, with the goal of assessing if the model performs better without them: it is first evaluated how much the image quality and fidelity changes, then it is measured the impact on the anonymization effectiveness.

7.1 Image quality and fidelity

The degree in which the generated pictures are different from the original ones has been measured in the same way as explained during Section 5.4: using FID [18] and VisualDNA [22], the dissimilarity between the features extracted by the two methods has been tested across the whole subset of data. VisualDNA was also used to compare each pair of authentic and synthesized pictures. Following the previous chapter, the data used is from CelebA-HQ and OpenImages v7 datasets, and the parameters used during training and inference are the same as the ones already discussed, with an exception made for the relation triplets distilled from the Extended Scene Graph. Table 7.1 reports how both FID and VisualDNA highlight the same behaviors. The former makes it possible to notice that adding noise lowers the picture's fidelity and highlights that, when considering CelebA-HQ, the difference between the best performing model with $d = 768$ and $s = 0.6$ is only 0.7 points lower than the second best that uses the same configuration without relation triplets, while inferring on OpenImages v7 rewards $d = 512$ and $s = 0.6$ without interaction triplets with a score of 28.2 and its counterpart with triplets achieves a slightly worse 28.6, on par of the same configuration at $d = 768$. VisualDNA also shows that the absence of object interactions does not translate into different performance, where, at dataset level, the model that does not use triplets performs overall 0.1-0.3 better than the model that employs them, on both datasets, while, when analyzing single pairs, having triplets is slightly better when inferencing on

Figure 7.1: Images generated by the evaluated models based on the CelebA-HQ and OpenImages v7 datasets. From top row to bottom: original images, proposed model at 512p and 60% noise, 512p and 80% noise, 768p and 60% noise, 768p and 80% noise. The proposed method's images show, grouped in 2 rows, the same configuration of noise and resolution with and without the use of relation triplets.

CelebA-HQ and narrowly worse when testing on OpenImages v7.

Method	Celeba-HQ	OpenImages v7
	FID	
Proposed method $(512, 0.6)$	32.2	28.6
Proposed method $(512, 0.8)$	39.2	34.2
Proposed method $(768, 0.6)$	**30.8**	28.6
Proposed method $(768, 0.8)$	40.7	35.8
Proposed method $(512, 0.6$, no triplets$)$	34.5	**28.2**
Proposed method $(512, 0.8$, no triplets$)$	41.7	33.8
Proposed method $(768, 0.6$, no triplets$)$	<u>31.5</u>	<u>28.4</u>
Proposed method $(768, 0.8$, no triplets$)$	44.5	36.2
	VisualDNA (dataset-level)	
Proposed method $(512, 0.6)$	<u>5.8</u>	<u>3.7</u>
Proposed method $(512, 0.8)$	6.7	4.2
Proposed method $(768, 0.6)$	6.0	4.3
Proposed method $(768, 0.8)$	6.5	4.8
Proposed method $(512, 0.6$, no triplets$)$	**5.7**	**3.5**
Proposed method $(512, 0.8$, no triplets$)$	6.4	4.0
Proposed method $(768, 0.6$, no triplets$)$	5.9	4.1
Proposed method $(768, 0.8$, no triplets$)$	6.6	4.7
	VisualDNA (image pair-level)	
Proposed method $(512, 0.6)$	13.7 ± 1.3	15.4 ± 2.3
Proposed method $(512, 0.8)$	16.1 ± 1.6	19.1 ± 3.0
Proposed method $(768, 0.6)$	**13.3±1.4**	<u>14.7 ± 2.5</u>
Proposed method $(768, 0.8)$	16.0 ± 1.7	18.8 ± 3.1
Proposed method $(512, 0.6$, no triplets$)$	13.8 ± 1.4	15.3 ± 2.3
Proposed method $(512, 0.8$, no triplets$)$	16.5 ± 1.8	19.0 ± 2.8
Proposed method $(768, 0.6$, no triplets$)$	<u>13.3 ± 1.5</u>	**14.6±2.4**
Proposed method $(768, 0.8$, no triplets$)$	16.5 ± 2.1	18.8 ± 2.9

Table 7.1: FID computed on dataset distribution and VisualDNA executed at dataset level and at original-anonymized image pair level test results across different hyperparameter combinations, with and without the use of relation triplets (lower is better).

7.2 Anonymization

In order to evaluate the performance in terms of de-personalization, it has been adopted the same approach described in Chapter 5. Again, the re-identification

rate was measured by the amount of times the original picture gets placed in the top K-NN of a generated image and the mAP@50. In this case, the models have been tested only on the CelebA-HQ subset. As shown in Table 7.2, the architecture trained and tested without the use of relation triplets performs consistently, by a small margin, better than the one that uses the additional text features: the reason could be attributed to the specificity some relations can provide to the generation process, which could make Stable Diffusion mimic with more fidelity the initial input. The best performing configurations are the ones that involve an obfuscation rate of $s = 0.8$, which regulates how many pixels of the original image have been substituted by noise, and makes the proposed model reach a score of 7.5% when testing with CLIP Re-ID at rank 1 and a mAP@50 of 3.9%, reaching roughly the same results of CAMOUFLaGE-Base with $a_s = 1.25$, previously noted as a Re-ID@1 of 8.7% and mAP@50 of 3.7%. When measuring identification at face level with FaceNet, the same hyperparameter setting described earlier achieves the best results of re-identification rate at rank 1 of 0.6% when using the model trained on VGGFace2 and 0.8% when FaceNet has learned from CASIA WebFace, comparable to the results achieved by DeepPrivacy2 0.8% and 0.8% when evaluated in the same manner.

7.3 Findings

The tests executed reveal that image quality is not noticeably affected by the presence of the additional text features containing interactions among objects, the main cause would be the presence of a considerable portion of the original image pixels (40% when $s = 0.6$ and 20% when $s = 0.8$) which define the picture's structure, generating very similar outputs regardless of the relation triplets use. On the other hand, anonymization obtains slightly better re-identification rates when not employing relations: it is possible that the amount of information provided through additional text features conditions the synthesis process to be more adherent to the input photo.

Method	Re-ID@1	Re-ID@5	Re-ID@10	mAP@50
	CLIP ViT-B/32			
Proposed method $(512, 0.6)$	0.457	0.625	0.689	0.184
Proposed method $(512, 0.8)$	0.176	0.312	0.390	0.086
Proposed method $(768, 0.6)$	0.574	0.735	0.790	0.230
Proposed method $(768, 0.8)$	0.201	0.342	0.429	0.095
Proposed method $(512, 0.6$, no triplets$)$	0.338	0.513	0.581	0.146
Proposed method $(512, 0.8$, no triplets$)$	**0.075**	**0.163**	**0.214**	**0.039**
Proposed method $(768, 0.6$, no triplets$)$	0.491	0.661	0.721	0.193
Proposed method $(768, 0.8$, no triplets$)$	0.109	0.221	0.293	0.055
	FaceNet + VGGFace2			
Proposed method $(512, 0.6)$	0.103	0.202	0.257	0.075
Proposed method $(512, 0.8)$	0.031	0.077	0.113	0.029
Proposed method $(768, 0.6)$	0.185	0.289	0.361	0.123
Proposed method $(768, 0.8)$	0.043	0.086	0.117	0.034
Proposed method $(512, 0.6$, no triplets$)$	0.070	0.137	0.187	0.050
Proposed method $(512, 0.8$, no triplets$)$	**0.006**	**0.030**	**0.045**	**0.009**
Proposed method $(768, 0.6$, no triplets$)$	0.109	0.214	0.271	0.080
Proposed method $(768, 0.8$, no triplets$)$	0.018	0.041	0.069	0.016
	FaceNet + CASIA WebFace			
Proposed method $(512, 0.6)$	0.090	0.183	0.254	0.057
Proposed method $(512, 0.8)$	0.026	0.081	0.117	0.022
Proposed method $(768, 0.6)$	0.152	0.303	0.374	0.090
Proposed method $(768, 0.8)$	0.035	0.080	0.124	0.023
Proposed method $(512, 0.6$, no triplets$)$	0.065	0.149	0.195	0.041
Proposed method $(512, 0.8$, no triplets$)$	**0.008**	**0.032**	0.057	**0.007**
Proposed method $(768, 0.6$, no triplets$)$	0.118	0.232	0.290	0.069
Proposed method $(768, 0.8$, no triplets$)$	0.016	0.042	**0.055**	0.013

Table 7.2: Re-ID test results across different model configurations, with and without the use of interaction triplets. The images used come from a CelebA-HQ subset of 1,000 images (lower is better).

Chapter 8

Conclusions

This final chapter briefly summarizes what this thesis work has achieved and it defines the possible directions that this new method could take in the future.

8.1 Findings

As discussed in Chapter 1, today's state-of-the-art efforts in the Image Anonymization task, of which some are introduced in Chapter 2, are mostly focused on modifying sensitive data inside pictures by means of inpainting the key details limited to persons' faces, sometimes their entire bodies, while neglecting the surrounding environment. Then, in Chapter 4 it has been proposed a revised version of the CAMOUFLaGE-Light [8] architecture, which employs the combination of the in-depth image analysis performed by FRESCOv1 [23] with the identification of visual relationship inferred by RelTR [40] to define an Extended Scene Graph used to condition a Stable Diffusion 1.5 [29] pipeline to synthesize images that maintain the original context and their key characteristics, but have most of the less meaningful details modified in order to deter re-identification of the subjects involved. It has been empirically demonstrated in Chapter 6 that the suggested implementation attains image quality scores comparable with the existing state-of-the-art approaches when anonymizing portrait pictures where less elements could affect the image generation process, while it achieves the best performance where photos capture more complex scenarios. Additional tests have also confirmed that anonymized images preserve their utility for a downstream task model focused on facial attribute detection. Finally, the difference in performance when using the list of interactions provided by the Extended Scene Graph compared to when not operating with them was explored in Chapter 7, and it has been noticed how image fidelity does not present substantial change, but the re-identification rate is better when relation triplets are not applied: the reason being in the more information

provided during synthesis conditioning, which aids the new image to adhere more closely to the starting input and makes it more recognizable.

8.2 Future development

In Chapter 6 it is possible to notice that the designed model competes on the same level as the best performing architectures that the presented analysis took into consideration in terms of picture credibility: the subjects maintain their facial traits, their poses and their locations in frame, but the presence of visual artifacts, which is a well known issue regarding Stable Diffusion, are sometimes observable in the generated images; a possible solution could be to adopt a more recent Latent Diffusion model. Another possible development could be to change how the Extended Scene Graph is interpreted by the architecture: with the use of a Graph Convolutional Network (GCN) designed to harness graph-structured information would make possible to manage the information collected in a more compact manner, the use of interaction information would make a more noticeable contribution to the image synthesis, and the use of this module could allow operating pure noise as a starting point to obtain the final output, thus attaining even better result in terms of anonymization. Finally, the insertion of more information derived from visual analysis could extend this method's capabilities: text detection could be used to remove characters from the photo or accurately report them based on the downstream task's needs, color information could be valuable to maintain the photograph's intended atmosphere, the position from which the picture has been taken, and any other data deemed semiotically valuable could be inserted to add non-identifiable information, thus tilting the intended balance between utility and personal information obfuscation.

Bibliography

[1] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. May 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj (cit. on p. 1).

[2] Rogier Creemers and Graham Webster. *Translation: Personal Information Protection Law of the People's Republic of China – Effective Nov. 1, 2021*. 2021. URL: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/ (cit. on p. 1).

[3] *Bill C-27 summary: Digital Charter Implementation Act*. 2022. URL: https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter/bill-summary-digital-charter-implementation-act-2020 (cit. on p. 1).

[4] Håkon Hukkelås and Frank Lindseth. «DeepPrivacy2: Towards Realistic Full-Body Anonymization». In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 1329–1338. DOI: 10.1109/WACV56688.2023.00138 (cit. on pp. 1, 2, 4, 24).

[5] Zikui Cai, Zhongpai Gao, Benjamin Planche, Meng Zheng, Terrence Chen, M. Salman Asif, and Ziyan Wu. *Disguise without Disruption: Utility-Preserving Face De-Identification*. 2023. arXiv: 2303.13269 [cs.CV]. URL: https://arxiv.org/abs/2303.13269 (cit. on p. 1).

[6] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. «Attribute-preserving Face Dataset Anonymization via Latent Code Optimization». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8001–8010 (cit. on pp. 1, 4, 6, 24).

[7] Muhammad Shaheryar, Jong Lee, and Soon Jung. «IDDiffuse: Dual-Conditional Diffusion Model for Enhanced Facial Image Anonymization». In: Dec. 2024, pp. 426–442. ISBN: 978-981-96-0910-9. DOI: 10.1007/978-981-96-0911-6_25 (cit. on pp. 1, 4, 5).

[8] Luca Piano, Pietro Basci, Fabrizio Lamberti, and Lia Morra. *Latent Diffusion Models for Attribute-Preserving Image Anonymization*. 2024. arXiv: 2403.14790 [cs.CV] (cit. on pp. 2, 9, 13, 14, 23, 24, 39).

[9] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel Cohen, and Adrian Weller. *Synthetic Data – what, why and how?* May 2022. DOI: 10.48550/arXiv.2205.03257 (cit. on p. 4).

[10] Majed Helou, Doruk Cetin, Petar Stamenkovic, Niko Huber, and Fabio Zünd. «VerA: Versatile Anonymization Applicable to Clinical Facial Photographs». In: Feb. 2025, pp. 127–138. DOI: 10.1109/WACV61041.2025.00023 (cit. on p. 4).

[11] Zhenzhong Kuang, Xiaochen Yang, Yingjie Shen, Chao Hu, and Jun Yu. *Facial Identity Anonymization via Intrinsic and Extrinsic Attention Distraction*. 2024. arXiv: 2406.17219 [cs.CV]. URL: https://arxiv.org/abs/2406.17219 (cit. on pp. 4, 5).

[12] Pascal Zwick, Kevin Roesch, Marvin Klemp, and Oliver Bringmann. «Context-Aware Full Body Anonymization». In: *Computer Vision – ECCV 2024 Workshops*. Ed. by Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi. Cham: Springer Nature Switzerland, 2025, pp. 36–52. ISBN: 978-3-031-92591-7 (cit. on pp. 4, 5).

[13] Han-Wei Kung, Tuomas Varanka, Sanjay Saha, Terence Sim, and Nicu Sebe. «Face Anonymization Made Simple». In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. Feb. 2025, pp. 1040–1050 (cit. on p. 4).

[14] Kartik Patwari, David Schneider, Xiaoxiao Sun, Chen-Nee Chuah, Lingjuan Lyu, and Vivek Sharma. *Rendering-Refined Stable Diffusion for Privacy Compliant Synthetic Data*. 2024. arXiv: 2412.06248 [cs.CV]. URL: https://arxiv.org/abs/2412.06248 (cit. on pp. 4, 5).

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. «Generative Adversarial Nets». In: *Neural Information Processing Systems*. 2014. URL: https://api.semanticscholar.org/CorpusID:261560300 (cit. on p. 4).

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. «Analyzing and Improving the Image Quality of StyleGAN». In: *Proc. CVPR*. 2020 (cit. on pp. 4, 21).

[17] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. «Deep Unsupervised Learning using Nonequilibrium Thermodynamics». In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2256–2265. URL: https://proceedings.mlr.press/v37/sohl-dickstein15.html (cit. on pp. 5, 7).

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. «GANs trained by a two time-scale update rule converge to a local nash equilibrium». In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6629–6640. ISBN: 9781510860964 (cit. on pp. 6, 23, 34).

[19] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. «An Improved Evaluation Framework for Generative Adversarial Networks». In: *arXiv e-prints*, arXiv:1803.07474 (Mar. 2018), arXiv:1803.07474. DOI: 10.48550/arXiv.1803.07474. arXiv: 1803.07474 [cs.CV] (cit. on p. 6).

[20] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli. «Image quality assessment: from error visibility to structural similarity». en. In: *IEEE Trans Image Process* 13.4 (Apr. 2004), pp. 600–612 (cit. on p. 6).

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. «The Unreasonable Effectiveness of Deep Features as a Perceptual Metric». In: *CVPR*. 2018 (cit. on p. 6).

[22] B Ramtoula, M Gadd, P Newman, and D De Martini. «Visual DNA: representing and comparing images using distributions of neuron activations». In: IEEE, 2023, pp. 11113–11123 (cit. on pp. 6, 23, 34).

[23] Lia Morra, Antonio Santangelo, Pietro Basci, Luca Piano, Fabio Garcea, Fabrizio Lamberti, and Massimo Leone. «For a semiotic AI: Bridging computer vision and visual semiotics for computational observation of large scale facial image archives». In: *Computer Vision and Image Understanding* 249 (2024), p. 104187. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2024.104187. URL: https://www.sciencedirect.com/science/article/pii/S1077314224002686 (cit. on pp. 6, 10, 13, 21, 39).

[24] One Octadion, Novanto Yudistira, and Diva Kurnianingtyas. *Synthesis of Batik Motifs using a Diffusion – Generative Adversarial Network*. July 2023. DOI: 10.48550/arXiv.2307.12122 (cit. on p. 8).

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. «Denoising Diffusion Probabilistic Models». In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf (cit. on p. 7).

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4 (cit. on pp. 7, 8).

[27] Prafulla Dhariwal and Alexander Quinn Nichol. «Diffusion Models Beat GANs on Image Synthesis». In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. URL: https://openreview.net/forum?id=AAWuCvzaVt (cit. on p. 7).

[28] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: 2207.12598 [cs.LG]. URL: https://arxiv.org/abs/2207.12598 (cit. on p. 8).

[29] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. «High-Resolution Image Synthesis with Latent Diffusion Models». In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10674–10685. URL: https://api.semanticscholar.org/CorpusID:245335280 (cit. on pp. 8, 39).

[30] Diederik P. Kingma and Max Welling. «Auto-Encoding Variational Bayes». In: *CoRR* abs/1312.6114 (2013). URL: https://api.semanticscholar.org/CorpusID:216078090 (cit. on p. 8).

[31] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. «Attention is All you Need». In: *Neural Information Processing Systems*. 2017. URL: https://api.semanticscholar.org/CorpusID:13756489 (cit. on p. 8).

[32] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. «IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models». In: (2023) (cit. on p. 9).

[33] Yinglin Zheng et al. «General facial representation learning in a visual-linguistic manner». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18697–18709 (cit. on pp. 9, 14, 24, 29).

[34] Alec Radford et al. «Learning Transferable Visual Models From Natural Language Supervision». In: *International Conference on Machine Learning*. 2021. URL: https://api.semanticscholar.org/CorpusID:231591445 (cit. on pp. 9, 19).

[35] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. «T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models». In: *arXiv preprint arXiv:2302.08453* (2023) (cit. on pp. 10, 16).

[36] Justin Johnson, Agrim Gupta, and Li Fei-Fei. *Image Generation from Scene Graphs*. 2018. arXiv: 1804.01622 [cs.CV]. URL: https://arxiv.org/abs/1804.01622 (cit. on p. 11).

[37] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. «Diffusion-based scene graph to image generation with masked contrastive pre-training». In: *arXiv preprint arXiv:2211.11138* (2022) (cit. on p. 11).

[38] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. *Learning Canonical Representations for Scene Graph to Image Generation*. 2020. arXiv: 1912.07414 [cs.CV]. URL: https://arxiv.org/abs/1912.07414 (cit. on p. 11).

[39] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. «Image retrieval using scene graphs». In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3668–3678. DOI: 10.1109/CVPR.2015.7298990 (cit. on p. 11).

[40] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. «RelTR: Relation Transformer for Scene Graph Generation». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2022), pp. 11169–11183. URL: https://api.semanticscholar.org/CorpusID:246294640 (cit. on pp. 11–13, 21, 39).

[41] Jinbae Im, JeongYeon Nam, Nokyung Park, Hyungmin Lee, and Seunghyun Park. «EGTR: Extracting Graph from Transformer for Scene Graph Generation». In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), pp. 24229–24238. URL: https://api.semanticscholar.org/CorpusID:268856676 (cit. on p. 11).

[42] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. *From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models*. 2024. arXiv: 2404.00906 [cs.CV] (cit. on p. 11).

[43] Bo Dai, Yuqi Zhang, and Dahua Lin. «Detecting Visual Relationships with Deep Relational Networks». In: (2017) (cit. on p. 11).

[44] Hongsheng Li et al. «Scene Graph Generation: A comprehensive survey». In: *Neurocomputing* 566 (2024), p. 127052. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.127052. URL: https://www.sciencedirect.com/science/article/pii/S092523122301175X (cit. on p. 11).

[45] Ranjay Krishna et al. «Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations». In: *International Journal of Computer Vision* 123.1 (May 2017), pp. 32–73. ISSN: 1573-1405. DOI: 10.1007/s11263-016-0981-7. URL: https://doi.org/10.1007/s11263-016-0981-7 (cit. on p. 11).

[46] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. «End-to-End Object Detection with Transformers». In: Nov. 2020, pp. 213–229. ISBN: 978-3-030-58451-1. DOI: 10.1007/978-3-030-58452-8_13 (cit. on p. 11).

[47] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. «Prismer: A Vision-Language Model with An Ensemble of Experts». In: *arXiv preprint arXiv:2303.02506* (2023) (cit. on pp. 14, 16).

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023 (cit. on p. 16).

[49] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. «OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cit. on p. 18).

[50] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. «Progressive Growing of GANs for Improved Quality, Stability, and Variation». In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=Hk99zCeAb (cit. on p. 21).

[51] Alina Kuznetsova et al. «The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale». In: *IJCV* (2020) (cit. on p. 21).

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. «Rethinking the Inception Architecture for Computer Vision». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on p. 23).

[53] Pan Zhou, Yichen Zhou, Chenyang Si, Weihao Yu, Teck Khim Ng, and Shuicheng Yan. «Mugs: A Multi-Granular Self-Supervised Learning Framework». In: *arXiv preprint arXiv:2203.14415*. 2022 (cit. on p. 23).

[54] Y. Rubner, C. Tomasi, and L.J. Guibas. «A metric for distributions with applications to image databases». In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. 1998, pp. 59–66. DOI: 10.1109/ICCV.1998.710701 (cit. on p. 24).