StructFormer: Structure-Consistent Face De-Identification under Strong Privacy Constraints*

Haini Zhu[1], Deepak K. Jain[1,*], Xudong Zhao[1], Muyu Li[1], Vitomir Štruc[2], Sumarga K. Sah Tyagi[3]
[1]Dalian University of Technology, [2]University of Ljubljana,
[3]Florida Agricultural and Mechanical University

Abstract

The widespread use of face data in computer vision raises significant concerns about privacy and identity leakage. Conventional anonymization techniques, such as blurring or masking, often degrade facial structure and expression, while existing generative methods may still retain identifiable cues when evaluated against strong face recognition systems. To address these limitations, we propose ***StructFormer****, a structure-consistent face de-identification framework based on a Transformer–GAN generator. StructFormer adopts a dual-stream design in which facial landmarks and masks provide explicit structural priors that are fused with appearance features through a Structure-Aware Attention Fusion module. This enables the preservation of head pose, facial layout, and expression while modifying identity-related appearance. A privacy control coefficient further allows continuous adjustment of anonymization strength without architectural changes. Experiments on LFW, CelebA, and CelebA-HQ demonstrate that StructFormer achieves a favorable balance between visual fidelity and privacy protection, maintaining high face detection rates (approximately* 99%–100%*) and competitive FID scores, while substantially reducing re-identification performance under strong FaceNet and ArcFace attackers, with match rates as low as* 0.03% *on CelebA-HQ.*

1. Introduction

The increasing availability of large-scale face data has enabled rapid progress in face analysis and recognition systems, while simultaneously raising serious concerns about privacy and identity leakage. Regulatory frameworks such as the GDPR, along with the removal of several widely used face datasets due to privacy considerations [11, 15], reflect a growing recognition that facial identity constitutes sensitive biometric information. As a result, there is a pressing need for effective face de-identification methods that suppress identity information while preserving visual content required for legitimate analysis and evaluation.

A central challenge in face de-identification stems from the strength of modern face recognition (FR) models. Even on high-quality, well-aligned face images, state-of-the-art FR systems can reliably extract identity cues from subtle combinations of texture, geometry, and expression. As a result, effective de-identification must operate under realistic attack scenarios, in which naive privacy protections are insufficient. At the same time, facial data often remains valuable only if key semantic properties, such as pose, expression, and overall facial structure, are preserved. Face de-identification therefore entails a fundamental trade-off, i.e., suppressing identity-related information while retaining structural and semantic facial attributes required for downstream analysis and evaluation [40].

Traditional anonymization techniques, such as blurring [1], occlusion [6], or downsampling [26], attempt to address this trade-off through coarse visual obfuscation. While simple to deploy, such approaches offer limited robustness against modern FR systems and learning-based reconstruction attacks [45]. More importantly, these transformations indiscriminately degrade both identity-related appearance and structure-related cues, frequently distorting facial geometry and expression, and substantially reducing the utility of the anonymized data for downstream vision tasks. These limitations have motivated the adoption of generative approaches that aim to explicitly modify identity while preserving other facial attributes.

Recent advances in deep generative modeling, including StyleGAN-based architectures [20–22] and diffusion models [8, 14, 44], have demonstrated impressive capacity for high-fidelity face synthesis. These models provide new opportunities for fine-grained anonymization by replacing the original identity with synthesized surrogates. However, directly applying large generative models to face de-identification remains challenging. Diffusion-based methods typically require costly iterative sampling, while large generative models may exhibit memorization effects, reproducing identity traits or stylistic patterns from the training data and thereby increasing the risk of identity leakage or unauthorized mimicry [3]. More broadly, many existing

**Corresponding author
Supported by the ARIS Research Programme P2-0250

Figure 1. **Illustration of the impact of the privacy parameter (pp) of StructFormer.** Each row shows de-identification results for one test subject under different privacy parameter settings. Higher pp values remove more identity information, while still preserving the overall facial structure and expression relatively well. The four columns (from left to right) correspond to $pp = 0$, 0.3, 0.5, and 0.8, respectively.

generative anonymization methods struggle to reliably disentangle identity-related appearance from facial structure, particularly under strong recognition attacks [48] leading to significant re-identification risk.

Convolutional GAN-based anonymization models offer the advantage of efficient single-step inference, but often lack explicit mechanisms to enforce global geometric consistency when identity cues are intentionally altered. Due to their limited receptive fields, such models may introduce structural distortions, including misaligned facial components or corrupted expressions, which degrade visual quality and compromise the usability of anonymized faces. These observations highlight the need for de-identification frameworks capable of enforcing holistic geometric consistency and explicitly preserving structure-related information, such as pose and expression, while modifying identity-related texture and appearance.

In this work, **we propose StructFormer**, a novel structure-consistent face de-identification framework that meets these challenges. StructFormer is built on a Transformer–GAN generator and incorporates explicit structural priors in the form of facial landmarks and masks. Through an innovative Structure-Aware Attention Fusion (SAAF) module, geometric information is injected into the generative process of StructFormer to constrain identity transformation and preserve facial layout and expression. Additionally, we also introduce a coordinated loss design that balances content preservation with identity suppression and enables continuous control over anonymization strength, as illustrated in Figure 1. Extensive experiments on LFW, CelebA, and CelebA-HQ demonstrate that StructFormer achieves a favorable trade-off between visual fidelity and privacy protection, substantially reducing re-identification rates under strong attackers, while maintaining high face detectability and structural consistency of the de-identified facial images.

In summary, our main contributions are threefold:

- We introduce StructFormer, a Transformer–GAN based face de-identification framework that integrates explicit structural priors to address identity leakage under strong face recognition attacks.
- We propose a Structure-Aware Attention Fusion (SAAF) mechanism that injects landmark- and mask-based geometric information into the generative process via cross-attention, enabling global coordination between structure and appearance during identity suppresion.
- We present a coordinated learning objective that supports continuous control over the privacy-utility trade-off. Moreover, we conduct a comprehensive experimental evaluation on three standard face benchmarks against state-of-the-art FR models to quantify identity suppression, structural fidelity, and face detectability.

2. Related Work

In this section, we briefly survey closely related work on visual privacy and face de-identification, needed to provide context for our work. For a more comprehensive coverage of this field, we refer the reader to some of the existing surveys available in the open literature, e.g., [28, 40, 56].

2.1. Visual Privacy and Face De-identification

The widespread use of social media platforms and mobile cameras has led to the large-scale collection and dissemination of visual data, while simultaneously amplifying the risk of privacy leakage. As a result, many contemporary vision datasets contain identifiable facial imagery [35, 47], prompting the introduction of data protection regulations such as GDPR, PDPA, and PIPA [9, 15, 53]. Since facial identity constitutes a highly sensitive biometric attribute [9], protecting faces has become a central concern in privacy-preserving computer vision. Face de-identification (de-ID) has emerged as a practical approach for balancing privacy protection with data utility. In contrast to completely discarding faces or applying irreversible masking,

de-ID aims to suppress identity-related information, while preserving attributes such as pose, expression, visual fidelity and coarse appearance that are relevant for downstream analysis. The well-documented limitations of classical anonymization techniques, such as blurring, pixelation, and occlusion, have motivated increasing interest in generative de-ID methods, which provide a more flexible and controllable privacy–utility trade-off [40].

2.2. Traditional Anonymization of Faces

Early privacy-aware systems primarily relied on simple image transformations, such as blurring, pixelation, occlusion, or additive noise, to obscure facial identity [1, 6, 26, 38]. While easy to implement, these methods indiscriminately degrade facial information, including pose, expression, and skin tone, and provide limited robustness against modern face recognition models, which can often (also through so-called parrot-attacks) re-identify faces that appear visually anonymized [42]. An alternative line of work draws inspiration from k-anonymity, exemplified by the K-Same framework [49]. K-Same replaces an input face with the average of its $k-1$ nearest neighbors in a predefined feature space, ensuring that at least k identities share the same de-identified representation. Although more structured than heavy masking, these averaging-based methods frequently introduce visual artifacts and struggle to preserve fine-grained facial structure or semantic attributes [19]. Overall, classical anonymization techniques are simple to deploy but suffer from limited robustness, degraded image quality, and poor compatibility with downstream analysis, motivating the transition to generative de-identification approaches.

2.3. Generative Face Anonymization

Generative face de-identification methods replace the original identity with a synthesized surrogate while aiming to preserve utility in the form of pose, expression, and scene context, allowing anonymized data to remain useful for later analysis [39]. Most existing approaches are based on generative adversarial networks (GANs) [10]. CIAGAN [37], for instance, generates identity-substituted faces conditioned on the source and reinserts them into the original image, and subsequent works extend this paradigm in various directions. AdaDeID [36] introduces controllable anonymization strength, similarly to [41], A3GAN and RB-GAN [53, 55] emphasize semantic attribute preservation, Barattin et al. [2] anonymize entire datasets via latent-space optimization, RIDDLE [30] enables reversible and diversified identities, and semantic-aware models [23] selectively modify identity-sensitive regions. While these approaches improve realism and attribute retention, many rely on convolutional architectures that offer limited global coordination, and identity leakage under strong recognition attacks remains insufficiently characterized.

To address global structural modeling, recent works incorporate self-attention and Transformer-based components, including TransGAN [18], ViT-based GAN variants [29], and masked generative encoders [31]. Diffusion-based anonymization methods, such as Diff-Privacy and NullFace [12, 27], further improve visual fidelity but require iterative sampling and incur substantial computational overhead. In contrast, our approach adopts a Transformer–GAN generator augmented with explicit structural priors and structure-aware attention, enabling stable facial geometry and expression preservation, while maintaining efficient single-step inference for face de-identification.

3. Method

In this section, we introduce the main novelty of this work,i.e., StructFormer, a structure-consistent face deidentification framework designed to suppress identity information, while maintaining consistent facial geometry and expression. Below, we first provide an overview of the proposed approach, followed by an in-depth description of its' key components and corresponding learning objective.

3.1. Overview of StructFormer

Generator Architecture. The generator of StructFormer, illustrated in Figure 2, follows a Transformer–GAN style encoder–decoder design and incorporates a Structure-Aware Attention Fusion (SAAF) module. It operates on two complementary inputs: (1) *an appearance stream*, consisting of the source face concatenated with a binary mask, and (2) a *structural stream* represented by a landmark map. The SAAF module fuses geometric priors with appearance features in the latent space, after which the decoder progressively upsamples the fused representation to synthesize a de-identified face. This dual-stream formulation explicitly conditions generation on facial structure, helping to preserve expression and fine-grained geometric details during identity transformation.

After SAAF-based feature alignment, a residual upsampling decoder restores the spatial resolution. The decoder comprises five stages, each consisting of a main branch, upsampling followed by two convolutional layers, and a lightweight shortcut connection, whose outputs are summed. This design improves geometric consistency and mitigates artifacts commonly associated with transposed convolutions. A self-attention layer is inserted at an intermediate resolution ($B \times 256 \times 64 \times 64$), where features encode global semantic information, while remaining computationally efficient. By propagating long-range spatial interactions, the attention mechanism reinforces coherent facial layout and expression under strong identity suppression constraints. Ablation results, presented later in the experimental section, confirm that removing this component leads to increased structural drift.

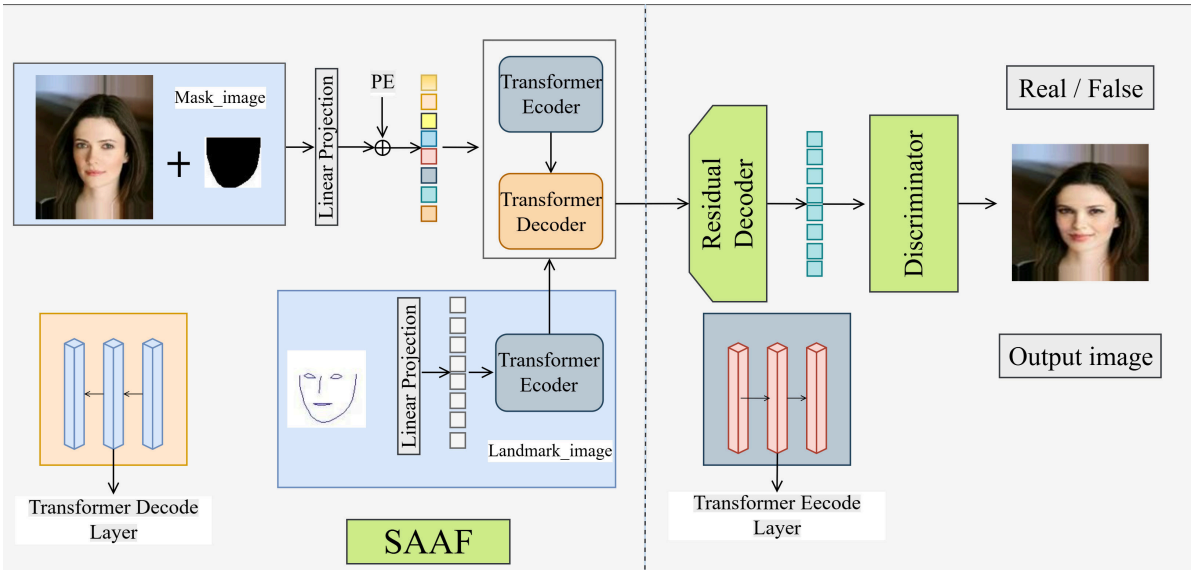

Figure 2. **Overview of the proposed Transformer–GAN generator used in StructFormer.** The architecture processes an *appearance stream* (source face and mask) and a *structural stream* (facial landmarks), which are fused in the latent space via the proposed Structure-Aware Attention Fusion (SAAF) module. The fused representation is then decoded by a residual upsampling decoder equipped with lightweight self-attention blocks to synthesize a de-identified face while preserving global facial structure and expression.

Multi-scale Discriminator. StructFormer employs a multi-scale PatchGAN discriminator [50], illustrated in Figure 3, in place of a U-Net-style design. Multiple PatchGAN discriminators operate at different image resolutions and jointly evaluate the realism of generated faces. High-resolution discriminators focus on local appearance details around salient facial regions, while low-resolution discriminators capture global facial shape and alignment, supporting stable synthesis when identity-related cues are modified. This multi-scale design encourages consistency across both fine-grained texture and global facial structure, which is critical when identity information is deliberately altered.

Joint training objectives. Beyond the adversarial objective, the model is trained with two complementary groups of auxiliary losses: (1) content-preserving losses and (2) identity-hiding losses. The content-preserving component consists of a reconstruction loss and an edge-based loss that stabilize overall appearance and geometric structure. The identity-hiding component includes a target-guided Perceptual Loss and an enhanced de-identification loss, which promote separation between the source and generated identities in perceptual and embedding spaces, respectively. This joint objective formulation enables controlled identity suppression while preventing unintended degradation of facial structure and visual fidelity. Detailed formulations and the roles of these objectives are presented in Section 3.3.

3.2. Structure-Aware Attention Fusion (SAAF)

Dual-stream Structural Priors. The SAAF module operates on two complementary inputs: a landmark-based structural stream and a masked appearance stream. Facial landmarks provide an explicit geometric prior encoding head pose, contour, and coarse expression, while containing minimal identity-specific information. We rasterize a subset of 41 facial keypoints [24, 37] into sparse heatmaps to form the structural stream. The appearance stream is constructed by masking the facial region in the source image, which preserves background consistency and confines identity modification to the face. For images containing multiple faces, each face is detected and processed independently within its corresponding bounding box.

Self-attention within Dual Streams. Both streams are flattened into token sequences and processed independently using multi-head self-attention to capture long-range dependencies within each modality. Given an input sequence $\mathbf{X}$, we compute

$$\mathbf{Q} = \mathbf{X}W_q, \quad \mathbf{K} = \mathbf{X}W_k, \quad \mathbf{V} = \mathbf{X}W_v, \tag{1}$$

followed by standard multi-head attention and a position-wise feed-forward network, as in standard Transformer blocks. Here, where W_q, W_k, and W_v denote learnable linear projection matrices that map input tokens to query, key, and value representations.

Cross-attention Fusion. To inject geometric information into appearance features, SAAF introduces a cross-attention layer that uses appearance tokens as queries and structural tokens as keys and values:

$$\mathbf{X}^{\text{fuse}} = \text{MHAtt}\big(\mathbf{Q}_{\text{app}}, \mathbf{K}_{\text{str}}, \mathbf{V}_{\text{str}}\big). \tag{2}$$

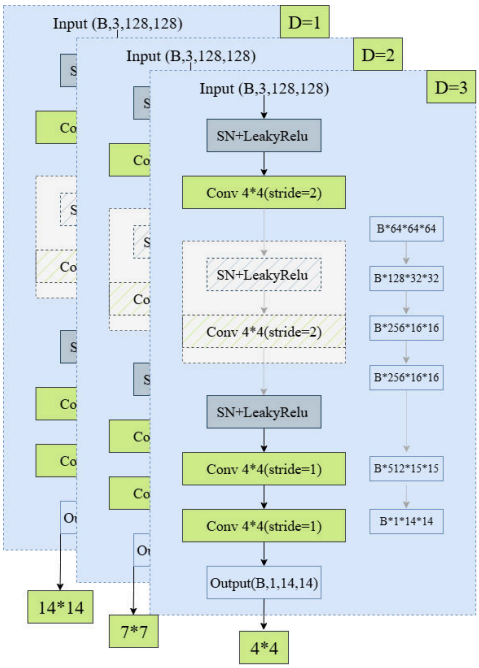

Figure 3. **Multi-scale PatchGAN discriminator.** Architecture of the discriminator composed of three scale-specific branches, denoted as D_1–D_3, which respectively produce 14×14, 7×7, and 4×4 patch-wise real/fake score maps. This multi-scale formulation enforces visual realism across both local details and global facial structure.

The resulting update is added to the appearance stream via a residual connection, after which the fused sequence is reshaped into a feature map and propagated to the next stage of the generator. As spatial resolution decreases and channel dimensionality increases, SAAF is applied repeatedly to progressively reinforce landmark-guided geometry.

Discussion. In contrast to simple input-level concatenation of facial masks and landmarks, SAAF performs explicit alignment between appearance and structure at multiple stages of the generator. This design enables identity-related texture to be reshaped under explicit geometric constraints, helping preserve facial contours and expressions even under strong identity suppression. Empirically, this reduces structural artifacts such as distorted mouths or collapsed eye regions, as confirmed by our ablation studies.

3.3. Training Objectives

To jointly preserve facial structure and suppress identity information during de-identification, we design the training objective around two complementary loss components: (1) a *content preservation loss*, which stabilizes geometry and appearance inherited from the source face, and (2) an *identity-hiding loss*, which explicitly discourages similarity to the source identity while enabling controlled identity suppression. The two losses are described in detail below.

Content Preservation Loss. Given the source face x_s, the target conditioning x_t, and the generated image $\hat{x} = G(x_s, x_t)$ with a binary face mask m, we encourage $\hat{x}$ to follow both the global appearance and the local boundary structure of x_s. To this end, we use an edge-aware boundary loss $\mathcal{L}_{\text{edge}}$ and a masked reconstruction loss $\mathcal{L}_{\text{rec}}$.

- **Edge-aware Boundary Loss.** We first compute an edge-strength map $E(x)$ from horizontal and vertical image gradients (*e.g.*, Sobel filters). Starting from the face mask m, a thin contour band b is obtained by one dilation and one erosion, so that b_{ij} is non-zero only around the facial boundary. The edge loss compares edge magnitudes of $\hat{x}$ and x_s within this band:

$$\mathcal{L}_{\text{edge}} = \frac{\left\| b \odot \left(E(\hat{x}) - E(x_s) \right) \right\|_1}{\|b\|_1 + \epsilon}. \tag{3}$$

- **Masked Reconstruction Loss.** Over the whole facial region, we additionally match the pixels of $\hat{x}$ to those of x_s inside the mask:

$$\mathcal{L}_{\text{rec}} = \frac{\left\| m \odot \left(\hat{x} - x_s \right) \right\|_1}{\|m\|_1 + \epsilon}. \tag{4}$$

- **Overall Objective.**
  The final content preservation term is a weighted sum of the two components:

$$\mathcal{L}_{\text{content}} = \lambda_{\text{edge}} \, \mathcal{L}_{\text{edge}} + \lambda_{\text{rec}} \, \mathcal{L}_{\text{rec}}, \tag{5}$$

  where λ_{edge} and λ_{rec} balance boundary sharpness and overall reconstruction quality.

Identity-hiding Loss. In the identity-hiding branch, we denote the source face by x_s, the target image by x_t, and the anonymized output by $\hat{x} = G(x_s, x_t)$. The loss combines a VGG-based perceptual term $\mathcal{L}_{\text{perc}}$ and an ArcFace-based identity term $\mathcal{L}_{\text{id}}$.

- **Perceptual Loss.** We use a pretrained VGG19 network as a fixed feature extractor and take activations $\Phi_l(\cdot)$ from a set of shallow, mid-level, and high-level layers. For a privacy parameter $pp \in [0, 1]$, the target feature at layer l is

$$T_l = \begin{cases} \Phi_l(x_s), & l \in \mathcal{L}_{\text{low-mid}}, \\ (1 - pp)\, \Phi_l(x_s) + pp\, \Phi_l(x_t), & l \in \mathcal{L}_{\text{high}}, \end{cases} \tag{6}$$

  so that low/mid layers preserve pose and local structure from x_s, while high layers smoothly interpolate identity cues from x_t. In all experiments, pp is treated as a user-controlled parameter that is fixed during training and inference. A resized face mask M_l restricts the penalty to the facial region, and the perceptual loss is computed as a weighted, masked L_1 distance,

$$\mathcal{L}_{\text{perc}} = \sum_l w_l \left\| M_l \odot \left(\Phi_l(\hat{x}) - T_l \right) \right\|_1, \tag{7}$$

encouraging sharp, structurally plausible faces while reducing direct alignment to the source identity in high-level features.

- **Identity Loss.** The identity loss is defined on unit-normalized embeddings from a pretrained ArcFace/MobileFaceNet backbone $f_{\text{id}}(\cdot)$. We measure cosine similarity between two images a and b as

$$s(a, b) = f_{\text{id}}(a)^{\top} f_{\text{id}}(b). \tag{8}$$

The identity term combines a "pull" component that encourages $\hat{x}$ to approach the target identity x_t and a margin-based "push" component that forces $\hat{x}$ away from the source x_s:

$$\mathcal{L}_{\text{id}} = \lambda_{\text{pull}}\big(1 - s(\hat{x}, x_t)\big) + \lambda_{\text{push}} \max\big(0,\ s(\hat{x}, x_s) - \gamma\big), \tag{9}$$

where γ controls the minimum allowed similarity to the source, and λ_{pull}, λ_{push} balance attraction to the target and repulsion from the source. In a pure de-identification setting, one can disable the pull term by setting $\lambda_{\text{pull}} = 0$, so that $\mathcal{L}_{\text{id}}$ only penalizes large similarity to x_s.

- **Overall Objective.** The total identity-hiding loss is a weighted sum of the perceptual and identity components:

$$\mathcal{L}_{\text{hide}} = \alpha_{\text{perc}}\, \mathcal{L}_{\text{perc}} + \alpha_{\text{id}}\, \mathcal{L}_{\text{id}}, \tag{10}$$

where α_{perc} and α_{id} control the trade-off between structural fidelity and identity suppression.

4. Experiments

In this section, we evaluate the proposed StructFormer against representative state-of-the-art face de-identification methods using standard evaluation methodology. We report both qualitative and quantitative results across multiple face datasets to assess visual fidelity, structural consistency, and identity suppression. We also perform a series of ablation studies to analyze the influence of key design factors, including the number of training identities, the privacy control coefficient, and the perceptual loss, on the behavior of the model.

4.1. Experimental Setup

Datasets. The **CelebA** dataset [33] consists of 202,599 face images spanning 10,177 identities. We use the aligned version, in which each image is registered to the eye midpoint, padded, and resized to 178×218 while preserving the facial aspect ratio, each identity contains at most 35 images. Facial landmarks used as structural priors are extracted using the HOG-based alignment method [5]. To examine higher-resolution settings, we additionally evaluate on **CelebA-HQ** [34], a high-quality variant of CelebA, and report results on its 256×256 subset to illustrate the behavior of the proposed method at increased image resolution. The **Labeled Faces in the Wild (LFW)** dataset [16] comprises over 13,000 unconstrained face images of 5,749 identities, among which 1,680 identities have at least two samples. We use LFW to evaluate de-ID performance and structural preservation under in-the-wild imaging conditions.

Implementation Details. All models are implemented in PyTorch [43] and trained on a single NVIDIA RTX 4090 GPU. During training, we use a batch size of 8 and optimize the networks using Adam [25] with momentum parameters $\beta_1 = 0.0$ and $\beta_2 = 0.9$. The learning rate follows a warm-up followed by cosine annealing: the generator is trained with a maximum learning rate of 4×10^{-5}, while the discriminator uses a max learning rate of 1×10^{-5}; both are decayed to a minimum learning rate of 1×10^{-6}. Training is performed for 100 epochs, with 3,044 iterations per epoch.

The overall objective is a weighted combination of the loss terms described in Section 3.3. All loss weights are fixed across experiments and ablation studies, with $\lambda_{\text{edge}} = 1.5$ and $\lambda_{\text{rec}} = 3$, and $\alpha_{\text{perc}} = 0.8$ and $\alpha_{\text{id}} = 0.5$.

4.2. Baselines and Performance Metrics

Baselines. We perform both quantitative and qualitative comparisons against representative face de-identification methods. For the **quantitive evaluation** of image quality and de-identification performance, we report results on CelebA, CelebA-HQ, and LFW, and compare against A3GAN [53], STGAN [32], Attribute-pre [19], RID-DLE [30], L2M-GAN [51], as well as the generative baselines DeepPrivacy [17] and CIAGAN [37]. These methods span a range of design choices for identity suppression and attribute preservation. For the **qualitative evaluation**, we present visual comparisons on CelebA-HQ and LFW, with a focus on structural and expression preservation alongside effective identity removal. In this setting, we include DeepPrivacy [17] and CIAGAN [37] as representative state-of-the-art generative de-identification approaches.

Performance Metrics. We evaluate the proposed method using metrics that capture both privacy protection and visual fidelity. To quantify identity removal, we first employ FaceNet [46], pretrained on CASIA-WebFace [52] and VGGFace2 [4], to compute a *re-identification rate*, defined as the proportion of anonymized faces that can still be matched to their original identities. In addition, we report an ArcFace-based de-identification score [7], where a pretrained ArcFace model is evaluated in a closed-set setting by matching anonymized embeddings against a gallery of original faces. To assess whether anonymized outputs remain structurally plausible and usable for downstream processing, we measure a *face detection rate* using MTCNN [54], defined as the fraction of generated images in which a face is successfully detected. An effective anonymization method should therefore maintain a detection rate close to 100% while driving re-identification met-

Table 1. Privacy and image quality results on LFW.

Method	FID↓	Detection(%)		Face re-ID(%)	
		dlib↑	MTCNN↑	CASIA↓	VGG↓
CIA-GAN	22.07	98.14	99.89	**0.17**	**0.91**
DeepPrivacy	23.46	96.70	99.57	2.74	1.52
Attribute-pre	27.45	**100.00**	**100.00**	2.07	1.58
Ours (0.8)	**8.35**	99.65	**100.00**	1.26	1.33

Table 2. Privacy and image quality results on CelebA.

	DeepPrivacy	CIA-GAN	L2M-GAN	STGAN
FID↓	30.12	34.95	18.83	20.14
Detection(%)↑	87.48	91.60	92.05	91.26
	AdaDeID	Ours (0.0)	Ours (0.3)	Ours (0.8)
FID↓	**2.19**	10.62	11.56	11.69
Detection(%)↑	95.90	96.94	**97.68**	97.65

Figure 4. **Qualitative results on the LFW dataset.** For each sample, the original image (left) and the anonymized result (right) at $pp = 0.8$ are shown, illustrating robust de-identification under unconstrained conditions.

rics toward zero. Finally, we compute the Fréchet Inception Distance (FID) [13] over all generated images as a global measure of visual realism and distributional quality.

4.3. Comparison to SOTA

Image Quality and De-identification. On LFW (Tab. 1), our method achieves a favorable quality–privacy trade-off compared with CIAGAN, DeepPrivacy, and Attribute-pre. Face detectability remains near perfect (99.65% with Dlib and 100.00% with MTCNN), while FID is reduced to 8.35, substantially lower than all baselines, indicating realistic anonymized faces with preserved structure. In the CASIA and VGG feature spaces, using a verification threshold of 0.7, our re-identification rate is slightly higher than CIAGAN but clearly lower than the remaining methods, resulting in reduced identity leakage for comparable visual quality. On CelebA (Tab. 2), DeepPrivacy, CIAGAN, L2M-GAN, and STGAN obtain FIDs between 18.83 and 34.95 with detection rates around 87%–92%. Across different

Figure 5. **Qualitative results on CelebA-HQ.** Each group shows an original face (left) and the corresponding de-identified output (right) generated with a privacy parameter of $pp = 0.8$.

Table 3. Privacy and image quality results on CelebA-HQ.

Method	FID↓	Detection(%)		Face re-ID(%)	
		dlib↑	MTCNN↑	CASIA↓	VGG↓
CIA-GAN	37.94	95.10	99.82	2.19	0.37
DeepPrivacy	32.99	92.82	99.85	3.91	1.05
Attribute-pre	29.93	98.58	**100.00**	2.8	1.67
RiDDLE	**5.39**	99.10	**100.00**	1.9	0.3
Ours (0.8)	7.51	**99.16**	**100.00**	**0.03**	**0.04**

privacy-control settings (e.g., $pp = 0.3$ and $pp = 0.8$), our variants maintain FID values around 11 while increasing face detection above 97%, demonstrating stable visual quality under a larger and more diverse identity distribution. Compared with AdaDeID, our FID is higher, but we achieve higher face detectability and provide an explicit mechanism for controlling the privacy level. On CelebA-HQ (Tab. 3), CIAGAN, DeepPrivacy, and Attribute-pre achieve FIDs above 29, whereas RIDDLE attains the lowest FID of 5.39. Our method reaches an FID of 7.51 at $pp = 0.8$, remaining competitive while achieving near-perfect face detection (99.16% with Dlib and 100.00% with MTCNN). At a similarity threshold of 0.6, existing methods exhibit de-identification rates between 0.27% and 3% in the CASIA and VGG feature spaces, with RIDDLE still yielding 1.9% / 0.3%. In contrast, our method reduces these rates to 0.03% / 0.04%, indicating substantially stronger identity suppression. Across LFW, CelebA, and CelebA-HQ, StructFormer consistently balances visual quality, face detectability, and identity removal, producing anonymized faces that remain structurally intact while exhibiting minimal linkage to the original identities under strong recognition models.

Qualitative Evaluation. Figs. 4 and 5 present qualitative results on LFW and CelebA-HQ. In each group, the left image shows the original face and the right image the corresponding anonymized output. Across both datasets, the generated faces remain visually plausible: head pose, coarse facial layout, and expression are preserved, while

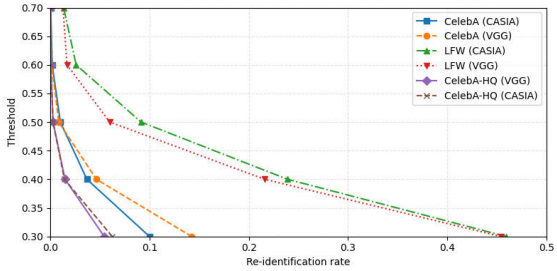

Figure 6. **Re-identification rate versus verification threshold.** Re-identification performance evaluated on CelebA, LFW, CelebA-HQ, and CASIA using FaceNet embeddings pretrained on CASIA-WebFace and VGGFace2. The x-axis denotes the cosine-similarity threshold and the y-axis the corresponding re-identification rate. Across all datasets and feature spaces, the re-identification rate drops rapidly and approaches zero for thresholds above 0.5, indicating effective de-identification.

Table 4. Ablation results on CelebA under threshold 0.6. We vary the number of identities (IDs) and the loss design.

Group	Metric	IDs		Loss	
		800	1200	content	id
Quality	FID↓	11.72	**10.72**	12.06	12.56
Detection↑	dlib	97.13	**96.94**	96.10	95.84
	MTCNN	99.97	99.97	99.97	99.95
Re-ID↓	CASIA	0.12	0.12	0.05	**0.03**
	VGG	0.23	0.18	0.09	**0.05**
	ArcFace	0.06	0.06	**0.04**	**0.04**

fine-grained appearance cues such as facial shape, texture, and local details are clearly altered, resulting in a distinct identity. Compared with LFW, CelebA-HQ samples exhibit sharper contours and cleaner textures, consistent with their higher resolution and lower FID scores. Nevertheless, even on the more challenging in-the-wild LFW images, StructFormer produces natural-looking anonymized faces with coherent structure and a clear visual identity shift relative to the source. Fig. 6 shows the re-identification rate as a function of the cosine-similarity threshold on CelebA, LFW, CelebA-HQ, and CASIA using FaceNet features trained on CASIA and VGGFace2. Across all datasets and feature spaces, the re-identification rate decreases rapidly with increasing threshold and approaches zero for thresholds above 0.5, indicating strong identity suppression.

4.4. Ablation study

We conduct an ablation study on CelebA to analyze the effects of three factors: (1) the number of training identities (IDs), (2) the anonymization strength pp, and (3) the loss design. Unless stated otherwise, all settings are fixed and evaluation is performed using a FaceNet attacker with a threshold of 0.6. For the results reported in Table 4, we set $pp = 0$ for both 800 and 1200 IDs.

Effect of the Number of Identities. Using the full loss, increasing the number of training identities from 800 to 1200 leads to a modest improvement in image quality, with FID decreasing from 11.72 to 10.72, while face detectability remains stable (approximately 97% with dlib and 100% with MTCNN). De-identification rates under CASIA, VGG, and ArcFace attackers change only marginally (0.12/0.23/0.06 vs. 0.12/0.18/0.06), indicating limited sensitivity to identity pool size. Based on this trade-off, we use 1200 identities as the default configuration.

Effect of Anonymization Strength. To illustrate the effect of anonymization strength, we vary $pp \in \{0, 0.3, 0.5, 0.8\}$ and visualize results on randomly selected test subjects in Fig. 1. As pp increases, the similarity to the source identity decreases progressively, while facial structure and expression remain largely stable. This demonstrates that pp provides effective and continuous control over identity suppression without compromising structural consistency.

Effect of the Loss Design. Fixing the number of identities to 1200, we compare the full objective with two reduced variants: content-only and identity-only losses. Removing either component degrades performance: FID increases to 12.06 (content-only) and 12.56 (identity-only), compared to 10.72 for the full model, and face detectability decreases slightly. De-identification rates under CASIA/VGG/ArcFace attackers also worsen (0.05/0.09/0.04 and 0.03/0.05/0.04, respectively), confirming that content-preserving losses are necessary for visual quality, while identity losses are essential for suppressing identity cues. Their combination yields the best balance between image quality and privacy protection.

5. Conclusion

We introduced StructFormer, a structure-aware face de-identification framework based on a Transformer–GAN generator. By decoupling structural priors from appearance features and integrating them via structure-aware attention, StructFormer preserves head pose, facial layout, and expression while effectively suppressing identity-related cues. An explicit privacy control coefficient further enables continuous adjustment of anonymization strength without architectural changes. Extensive evaluations on LFW, CelebA, and CelebA-HQ demonstrate that StructFormer achieves a strong balance between visual fidelity and privacy protection. Compared with classical obfuscation methods and recent generative approaches, it consistently improves image quality and face detectability while substantially reducing re-identification rates under strong face recognition models.

References

[1] Prachi Agrawal and PJ Narayanan. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):299–310, 2011. 1, 3

[2] 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*, pages 8001–8010, 2023. 3

[3] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023. 1

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 6

[6] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman Sadeh, and Mahadev Satyanarayanan. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1387–1396. IEEE, 2017. 1, 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[9] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E O'Connor. How important are faces for person re-identification? In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6912–6919. IEEE, 2021. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[11] Adam Harvey and Jules LaPlace. Megapixels: origins, ethics, and privacy implications of publicly available face recognition image datasets. *Megapixels*, 1(2):6, 2019. 1

[12] Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[13] 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. *Advances in neural information processing systems*, 30, 2017. 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[15] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019. 1, 2

[16] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 6

[17] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer, 2019. 6

[18] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021. 3

[19] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*, pages 278–285. IEEE, 2015. 3, 6

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 1

[23] Hyeongbok Kim, Zhiqi Pang, Lingling Zhao, Xiaohong Su, and Jin Suk Lee. Semantic-aware deidentification generative adversarial networks for identity anonymization. *Multimedia Tools and Applications*, 82(10):15535–15551, 2023. 3

[24] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 4

[25] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Kento Kobayashi, Keiichi Iwamura, Kitahiro Kaneda, and Isao Echizen. Surveillance camera system to achieve privacy protection and crime prevention. In *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 463–466. IEEE, 2014. 1, 3

[27] Han-Wei Kung, Tuomas Varanka, Terence Sim, and Nicu Sebe. Nullface: Training-free localized face anonymization. *arXiv preprint arXiv:2503.08478*, 2025. 3

[28] Lamyanba Laishram, Muhammad Shaheryar, Jong Taek Lee, and Soon Ki Jung. Toward a privacy-preserving face recognition system: A survey of leakages and solutions. *ACM Computing Surveys*, 57(6):1–38, 2025. 2

[29] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021. 3

[30] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. *arXiv preprint arXiv:2303.05171*, 2023. 3, 6

[31] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152, 2023. 3

[32] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019. 6

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6

[35] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 2

[36] Tianxiang Ma, Dongze Li, Wei Wang, and Jing Dong. Adadeid: Adjust your identity attribute freely. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 210–216. IEEE, 2022. 3

[37] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020. 3, 4, 6

[38] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016. 3

[39] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. Face deidentification with generative deep neural networks. *IET Signal Processing*, 11(9):1046–1054, 2017. 3

[40] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021. 1, 2, 3

[41] Blaž Meden, Manfred Gonzalez-Hernandez, Peter Peer, and Vitomir Štruc. Face deidentification with controllable privacy protection. *Image and Vision Computing*, 134:104678, 2023. 3

[42] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016. 3

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[45] Peter Rot, Klemen Grm, Peter Peer, and Vitomir Štruc. Privacyprober: Assessment and detection of soft–biometric privacy–enhancing techniques. *IEEE Transactions on Dependable and Secure Computing*, 21(4):2869–2887, 2023. 1

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[47] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[48] Jaisidh Singh, Harshil Bhatia, Mayank Vatsa, Richa Singh, and Aparna Bharati. Synthprov: Interpretable framework for profiling identity leakage. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4746–4756, 2024. 2

[49] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002. 3

[50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 4

[51] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2951–2960, 2021. 6

[52] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 6

[53] Liming Zhai, Qing Guo, Xiaofei Xie, Lei Ma, Yi Estelle Wang, and Yang Liu. A3gan: Attribute-aware anonymization networks for face de-identification. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5303–5313, 2022. 2, 3, 6

[54] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded

convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 6

[55] Yaofang Zhang, Yuchun Fang, Yiting Cao, and Jiahua Wu. Rbgan: Realistic-generation and balanced-utility gan for face de-identification. *Image and Vision Computing*, 141: 104868, 2024. 3

[56] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. Visual content privacy protection: A survey. *ACM Computing Surveys*, 57(5):1–36, 2025. 2