TRUST: Token-dRiven Ultrasound Style Transfer for Cross-Device Adaptation

Nhat-Tuong Do-Tran*, Ngoc-Hoang-Lam Le*, Ian Chiu†, Po-Tsun Paul Kuo‡, and Ching-Chun Huang*
*National Yang Ming Chiao Tung University, Taiwan
†Advantech Company, Taiwan ‡Advanced Operational Development, Delta Electronics, Inc., Taiwan
*{tuongdotn.ee12, lengochoanglam.ee12, chingchun}@nycu.edu.tw, †Ian.Chiu@advantech.com.tw, ‡Paul.PT.Kuo@deltaww.com

***Abstract*—Ultrasound images acquired from different devices exhibit diverse styles, resulting in decreased performance of downstream tasks. To mitigate the style gap, unpaired image-to-image (UI2I) translation methods aim to transfer images from a source domain, corresponding to new device acquisitions, to a target domain where a frozen task model has been trained for downstream applications. However, existing UI2I methods have not explicitly considered filtering the most relevant style features, which may result in translated images misaligned with the needs of downstream tasks. In this work, we propose TRUST, a token-driven dual-stream framework that preserves source content while transferring the common style of the target domain, ensuring that content and style remain unblended. Given multiple styles in the target domain, we introduce a Token-dRiven (TR) module that operates from two perspectives: (1) a data view–selecting "suitable" target tokens corresponding to each source token, and (2) a model view–identifying "optimal" target tokens for the downstream model, guided by a behavior mirror loss. Additionally, we inject auxiliary prompts into the source encoder to match content representation with downstream behavior. Experimental results on ultrasound datasets demonstrate that TRUST outperforms existing UI2I methods in both visual quality and downstream task performance.**

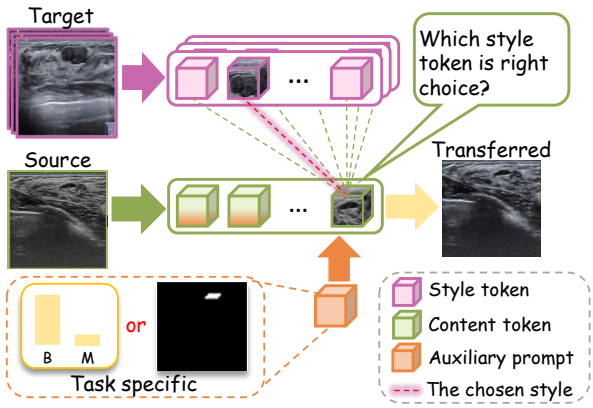


Fig. 1: TRUST selects the most appropriate style token for each content token by considering both the data view (e.g., texture, pattern) and the model view (e.g., classification, segmentation), the latter enabled by auxiliary prompts injected into the content encoder.

I. Introduction

Ultrasound (US) imaging is widely adopted in clinical diagnosis due to its efficiency, affordability, safety, and real-time capability [1], [2]. However, images acquired from different devices or medical centers often exhibit substantial variation in visual characteristics—such as intensity, contrast, resolution, and speckle patterns—due to differences in hardware configurations and acquisition protocols. These cross-device discrepancies pose serious challenges to both automated systems and human interpretation. Deep learning models, once trained on a specific source domain, tend to suffer significant performance degradation when deployed in unseen domains. Similarly, doctors—accustomed to images from familiar devices—may face difficulty interpreting cases from new devices that display unfamiliar style distributions. While retraining on new devices is a possible solution, it is often impractical: (1) For doctors, adapting to new device styles requires considerable effort in diagnosing new image characteristics. (2) For deployed AI models, fine-tuning is typically not feasible due to proprietary software constraints (i.e., cannot be reconfigured model parameters). To mitigate such domain shifts, Ultrasound Style Transfer (UST) has emerged as a promising solution. Instead of retraining, UST translates the source-domain images to match the style of the target domain while preserving their structural content. In this way, both doctors and downstream models operate in a target-style environment without altering their original behavior.

During the last decades, unpaired image-to-image translation methods have been explored, aiming to adapt images from the source domain to the target domain without requiring paired data. Among these, GAN-based methods [3]–[5] have been early solutions for bridging domain gaps, but are mainly effective for small shifts. Moreover, their shared discriminator often entangles content and style. Transformer-based models [6], [7] have shown strong capabilities in disentangling content and style by separating their respective branches, effectively preventing undesired blending. Their self-attention mechanism enables a global understanding of spatial structure, which helps preserve content integrity and maintain consistent patterns across layers [8], [9]. Despite these advantages, existing transformer-based methods often rely on pairwise style translation, which limits their ability to capture diverse styles, especially when the target domain exhibits multiple stylistic variations. Furthermore, transformer-based methods lack trustworthy token alignments, which are key to reliable

Equal contribution: Nhat-Tuong Do-Tran and Ngoc-Hoang-Lam Le.
Corresponding author: Ching-Chun Huang.

979-8-3315-7206-8/25/$31.00

downstream performance.

To address the above challenges, we propose the **TRUST**, a **T**oken-d**R**iven **U**ltrasound **S**tyle **T**ransfer for Cross-Device Adaptation. Our method begins with a disentangling strategy that separates the source and target into content and style branches, respectively. This strategy enables explicit extraction and independent processing of style and content features, allowing finer control over the transfer. Inspired by Einstein's mindset that **"not all information is knowledge!"**, TRUST recognizes that not all style features are beneficial for transfer. Naively integrating global style tokens can lead to artifacts, such as overly bright tumor regions or suppressed tissue contrast. To this end, TRUST employs a selective mechanism that searches through a *pool* of style tokens to identify those best aligned with each content token (Figure 1). Crucially, the selection of style features is guided not only by the content tokens but also by auxiliary prompts injected into the content encoder layers. These prompts help the TR module more effectively retain task-specific and highly relevant style features. Together, the TR module and auxiliary prompts strengthen the connection between content and style branches, both at the data and model levels, enabling more accurate knowledge transfer. enabling more accurate knowledge transfer. Additionally, to align the output with the downstream task, we construct a mimic model that replicates the behavior of the black-box downstream model. Then, we use the mimic model's predictions as supervision to align the styled image output through a behavior mirror loss. We summarize the key contributions of TRUST as follows:

1) The Token-dRiven (TR) module employs attention to identify trustworthy style tokens and selectively fuses them based on content relevance, enabling more precise and natural style integration.
2) We inject auxiliary prompts into the content encoder to enhance feature representation and more effectively minimize the domain gap.
3) We introduce a behavior mirror loss that uses a mimic downstream to replicate black-box behavior and supervise the styled image output.
4) Comprehensive experiments demonstrate that TRUST consistently outperforms baseline methods across multiple datasets, achieving outstanding results with well-preserved content structures and desirable style patterns.

II. Proposed method

A. Problem definition

We formulate the ultrasound style transfer for cross-device adaptation as an unpaired image-to-image translation problem. Let $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{N_{labeled}} \cup \{x_s^k\}_{k=1}^{N_{unlabeled}}$ represent the source domain dataset, where a **few** x_s^i denotes an labeled image with corresponding annotations y_s^i, while the majority of source samples x_s^k are unlabeled. Similarly, $\mathcal{D}_T = \{x_t^j\}_{j=1}^{N_T}$ represents the target domain dataset without annotations, where images x_t^j are acquired from different ultrasound scanners.

Our goal is to learn a translation function $G : X_S \to X_T$ capable of mapping images $x_s \in X_S$ to the target domain $\mathcal{D}_T$, thus addressing the domain shift caused by cross-device variations. The translated images $\hat{x}_t = G(x_s)$ should preserve the anatomical content structures inherent in x_s while effectively adapting to the style characteristics found in X_T.

To achieve this, we design our style transfer network $G(x_s)$, TRUST, as illustrated in Figure 2. TRUST comprises three key components: **feature extractors** (Section II-B) to encode representations from the source and target domains, a **Token-dRiven (TR)** module (Section II-C) that aligns one-source tokens with multiple-target tokens via data-view and model-view, and a **decoder** (Section II-D) that reconstructs stylized images while preserving source content and transferring target style. In addition, **auxiliary prompts** (Section II-E) are injected into the content encoder to enhance feature representations.

B. Feature extractors

Our feature extraction module consists of two Transformer-based branches: a **content branch** and a **style branch**. Both branches are constructed with multiple Transformer layers [10], leveraging the self-attention mechanism to effectively preserve structural information and progressively enhance semantic representations across layers.

Content branch. Focuses on capturing structural and anatomical details from the source domain. Transformer-based architectures are particularly suitable for this purpose because they keep resolution unchanged, while enriching the semantic information layer by layer, as demonstrated in [9]. Given a source image $x_s \in \mathbb{R}^{H\times W\times C}$, the image is first divided into non-overlapping patches of size $P \times P$, resulting in $N = \frac{H\cdot W}{P^2}$ patches. After linear projection, the patch tokens are added with *learnable positional embeddings* to encode spatial information. The source tokens are then processed by the transformer layers as follows:

$$F_s = E_s(x_s, P), \quad F_s \in \mathbb{R}^{N\times d}, \tag{1}$$

where E_s represents the source encoder, N is the number of patches, and d is the embedding dimension of each token. P is a set of auxiliary prompts that can be inserted into the token sequence, as described in Section II-E.

Style branch. Processes target domain images to extract target tokens that reflect the distributional and stylistic patterns of the target domain. Given a batch of target images $X_T^{batch} = \{x_t^1, x_t^2, \ldots, x_t^B\}$, where each $x_t^j \in \mathbb{R}^{H\times W\times C}$ and B is the batch size, the encoder produces target tokens:

$$F_t = E_t(X_T^{batch}), \quad F_t \in \mathbb{R}^{B\times N\times d}, \tag{2}$$

where E_t denotes the target encoder.

The use of stacked Transformer layers in both branches ensures that spatial information is preserved while semantic understanding is progressively improved, enabling effective feature extraction for both content and style domains [9].

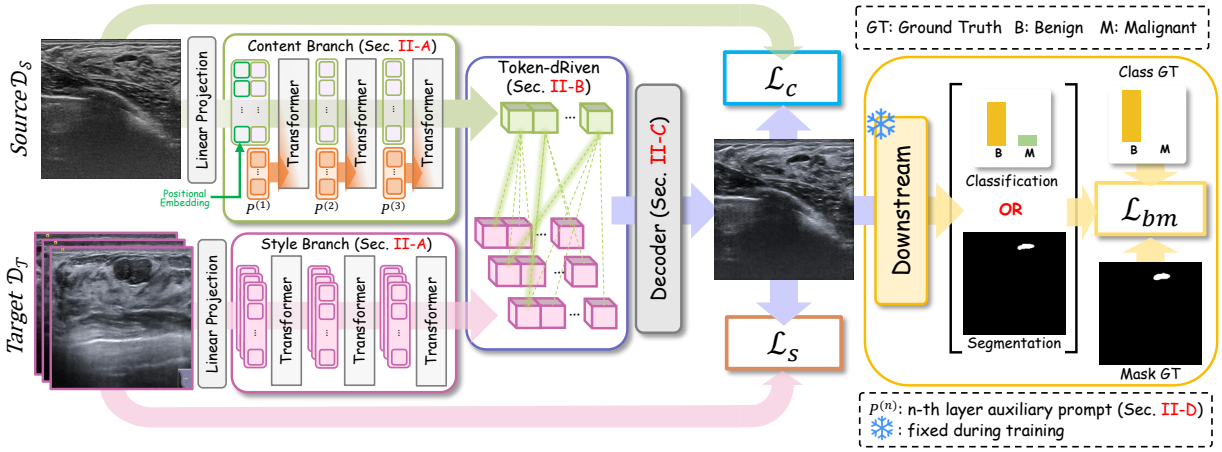


Fig. 2: **Overview of the proposed TRUST framework for ultrasound style transfer**. The architecture consists of a content branch with multi-layer auxiliary prompts for data adaptation, a style branch for extracting target style features, a Token-dRiven (TR) module that fuses latent source content tokens with target style tokens to perform style transfer, and a decoder that reconstructs the stylized ultrasound images. Given a frozen downstream task network, a behavior mirror loss ($\mathcal{L}_{bm}$) is introduced to provide task-specific supervision, encouraging the source content tokens to align with downstream objectives originally tailored for the target domain.

C. Token-dRiven module

To effectively select suitable and informative multiple-target tokens for one-source tokens, we propose the TR module via data-view and model-view.

Data-view. From a data perspective, we design the TR module to align each source token with target tokens from multiple samples by computing a correlation matrix that captures their pairwise relationships. Formally, given source tokens $F_s \in \mathbb{R}^{N\times d}$ and target tokens $F_t \in \mathbb{R}^{B\times N\times d}$, the correlation matrix $M \in \mathbb{R}^{N\times(B\cdot N)}$ is computed as:

$$M_{ij} = \text{dot}(F_s^i, F_t^j), \tag{3}$$

where $\text{dot}(\cdot,\cdot)$ represents the dot product between each source token and the aggregated target tokens across the batch.

The aligned features $F_{align} \in \mathbb{R}^{N\times d}$ are then obtained by aggregating the target tokens weighted by this correlation matrix, followed by a skip connection to preserve the original source content:

$$F_{align} = F_s + MF_t. \tag{4}$$

In this way, each source token selectively attends to the most relevant target tokens across different samples, enabling style adaptation while maintaining source content structure.

Model-view. To further enforce consistency with downstream tasks, we introduce the **behavior mirror loss** ($\mathcal{L}_{bm}$), which leverages supervision signals from a downstream model trained on the labeled source samples, $D_s^{labeled} = \{(x_s^i, y_s^i)\}_{i=1}^{N_{labeled}}$. The behavior mirror loss is formulated as:

$$\mathcal{L}_{bm}\big(G(x_s^i), y_s^i\big) = \ell_{down}\big(f_{downs}(G(x_s^i)), y_s^i\big), \tag{5}$$

where $f_{downs}(\cdot)$ is the downstream model, selected according to the deployment scenario (i.e., ViT-B/16 [11] for classification and SAMUS [12] for segmentation), and $\ell_{down}(\cdot,\cdot)$ represents the corresponding loss function. By minimizing $\mathcal{L}_{bm}$, the source tokens F_s are directly optimized toward the downstream objective, ensuring that the translated image $G(x_s^i)$ produces predictions consistent with the label y_s^i.

D. Decoder

The decoder reconstructs the stylized image from the aligned token features through a convolutional architecture following previous style transfer works [9], [13]. Given the aligned tokens $F_{align} \in \mathbb{R}^{N\times d}$, the decoder first reshapes them into a spatial feature map of size $H' \times W' \times d$, where H' and W' equal to $\sqrt{N}$. The final output is obtained via CNN-based decoder layers as follows:

$$\hat{x}_t = D(F_{align}), \tag{6}$$

where D denotes the decoder, and $\hat{x}_t \in \mathbb{R}^{H\times W\times 3}$ is the reconstructed stylized image.

E. Auxiliary prompts

To enhance content representations for better compatibility with the downstream model. We inspired of VPT [14] to inject a set of learnable prompt tokens $P = \{P^{(l)} \in \mathbb{R}^{L_p\times d}\}_{l=1}^{L}$ into the source encoder (Section II-B), where L is the number of transformer layers, L_p is the number of prompt tokens per layer, and d is the token dimension.

At each transformer layer, a unique set of prompt tokens $P^{(l)}$ is prepended to the input token sequence. The prompt-enhanced content extraction is thus formulated as:

$$F_s = E_s\big(\{P^{(l)} \oplus x_s^{(l)}\}_{l=1}^{L}\big), \tag{7}$$

where $x_s^{(l)}$ denotes the input tokens at layer l, and $\oplus$ represents the concatenation operation at the token level.

TABLE I: Performance comparison of style transfer methods on 6 cross-device transfer tasks for both classification and segmentation. Classification performance is evaluated by Accuracy (Acc) and Area Under the Curve (AUC), while segmentation performance is evaluated by Dice score (Dice) and Intersection over Union (IoU). The results are reported in the order: Acc / AUC / Dice / IoU. w/o ST indicates the baseline setting *without Style Transfer*. The top and second-best results are highlighted in **bold** and underline.

Method	UCLM → BUSI	BUSI → UCLM	UCLM → UDIAT	UDIAT → UCLM	BUSI → UDIAT	UDIAT → BUSI
w/o ST	70.0/74.8/77.1/66.5	65.6/68.8/77.7/68.0	63.8/68.2/82.2/72.1	75.5/77.1/85.1/75.6	73.3/73.2/79.5/70.6	85.7/91.9/84.3/74.8
CycleGAN [5]	51.3/64.6/65.2/52.4	69.2/72.4/75.7/65.5	56.3/56.9/74.7/63.9	67.4/69.2/80.4/69.2	70.8/72.0/79.7/70.4	79.5/81.6/82.5/72.6
DiscoGAN [15]	70.0/72.0/67.9/56.3	63.1/74.1/73.2/63.2	65.0/70.1/42.6/30.6	69.4/81.4/44.1/33.0	72.3/73.8/77.4/67.3	87.8/88.0/79.2/68.0
S2WAT [7]	72.5/75.5/76.2/65.9	62.6/57.4/77.4/67.5	65.0/45.7/80.5/70.0	75.5/71.6/84.6/75.2	71.8/73.4/80.1/71.0	85.7/93.4/81.8/72.1
TransColors [6]	72.5/77.3/75.9/65.4	64.1/65.0/77.7/67.9	62.5/68.4/82.6/72.6	77.6/73.9/85.7/76.2	69.2/71.1/80.4/71.4	83.7/92.1/83.4/74.0
TRUST	**80.0/81.9/79.9/69.7**	**71.8/77.9/78.6/69.0**	**72.5/76.7/83.3/73.4**	**85.7/89.7/86.9/77.9**	**79.5/84.5/80.7/71.8**	**89.8/94.9/85.4/76.5**

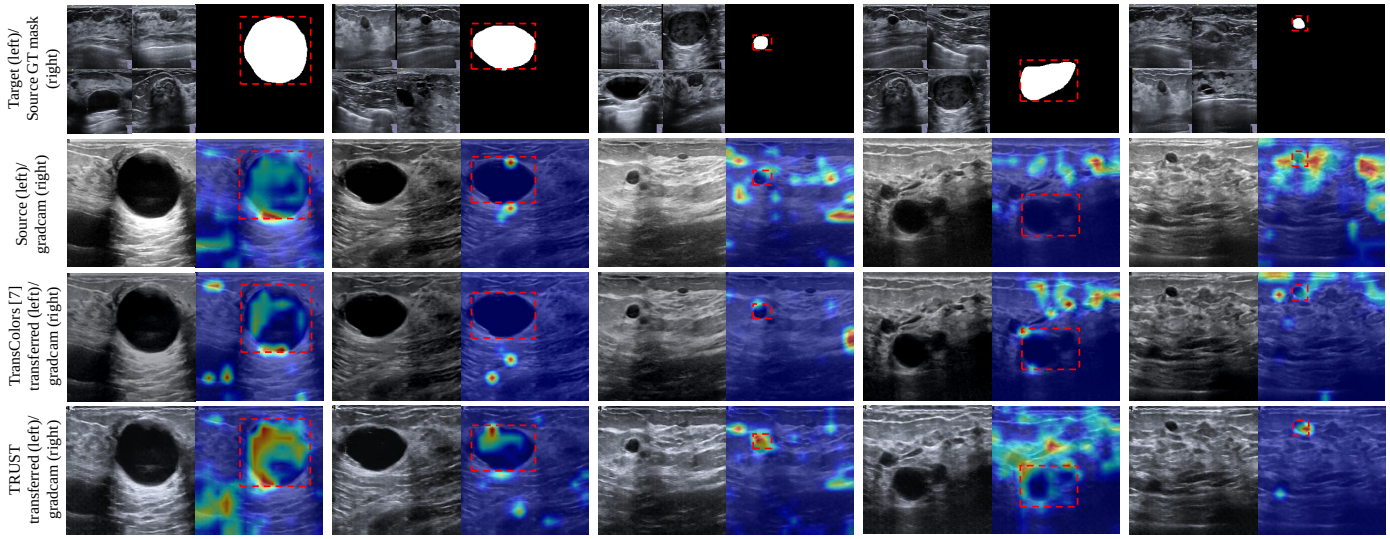


Fig. 3: **GradCAM [16] comparison of style transfer methods**, including Source without transfer (Source), TransColors [6], and our proposed TRUST. The GradCAM visualizations highlight the attention regions of a frozen downstream classifier on the BUSI→UCLM transfer task. The source mask (top-right) delineates the ground-truth tumor region that the model should attend to for accurate classification, while the target images (top-left) serve as style references for the style transfer methods. Notably, the transferred source images (bottom-left) produced by TRUST exhibit more concentrated attention within the tumor region, thereby enhancing downstream classification performance.

F. Training Scheme

Following previous style transfer works [7], [9], [13], we adopt a content loss ($\mathcal{L}_c$) to preserve source structure and a style loss ($\mathcal{L}_s$) to match target style statistics, both computed from the feature representations of a pre-trained VGG network. The objective function to optimize TRUST is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_c(\hat{x}_t, x_s) + \mathcal{L}_s(\hat{x}_t, x_t) + \mathcal{L}_{bm}(\hat{x}_t, y_s^i). \quad (8)$$

III. Experimental Results

A. Experimental setup

Dataset. We conduct experiments on three publicly available breast ultrasound datasets: BUSI [17], UDIAT [18], and UCLM [19]. Each dataset is randomly divided into training and testing subsets with a ratio of 7:3. In each experimental setting, one dataset (e.g., BUSI) is assigned as the *source domain*, $\mathcal{D}_S$, while another serves as the *target domain*, $\mathcal{D}_T$ (e.g., UDIAT or UCLM). The proposed TRUST framework is trained on the source training set with style information transferred from the target training set, and evaluated on the source testing set. The testing set of the target domain is reserved for selecting the best downstream model during validation. Notably, the target labels are only used to train the downstream model.

Implementation details. All experiments are conducted on a single NVIDIA RTX 4090 GPU with input images resized to 256×256. In the labeled source set $D_s^{labeled} = \{(x_s^i, y_s^i)\}_{i=1}^{N_{labeled}}$, only 20 samples per class are available. The entire TRUST framework is initialized using Xavier uniform initialization [20]. Both the source encoder E_s and target encoder E_t consist of 3 Transformer [10] layers with $L_p = 1024$ prompt tokens per layer and a patch size of $P = 8$. We adopt the Adam optimizer with an initial learning rate of 5×10^{-4}, following the warm-up adjustment strategy [21], and use a batch size of $B = 6$. The total number of iterations is

TABLE II: **Ablation study on different components of our framework**. "w/o ST" denotes the case without style transfer. CA refers to the baseline model employing Cross Attention, TR represents the Token-dRiven module, and TRUST integrates TR with auxiliary prompts for enhanced style adaptation.

Settings	BUSI → UCLM	UCLM → UDIAT	UDIAT → BUSI
w/o ST	65.6/68.8/77.7/68.0	63.8/68.2/82.2/72.1	85.7/91.9/84.3/74.8
CA	66.7/64.9/77.8/68.0	60.0/66.3/82.9/72.9	87.8/92.3/83.6/73.9
TR	68.7/76.7/78.5/68.6	71.3/76.2/83.1/72.9	89.8/92.5/85.4/76.2
TRUST	71.8/77.9/78.6/69.0	72.5/76.7/83.3/73.4	89.8/94.9/85.4/76.5

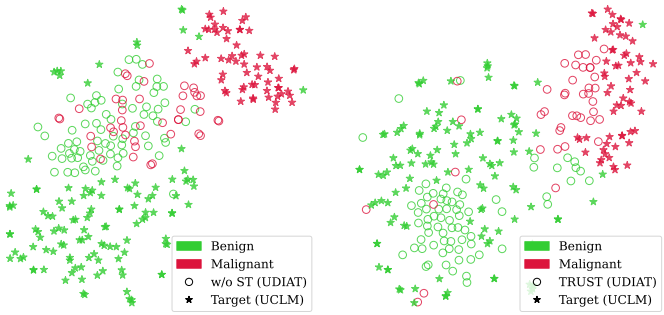
set to $20,000$. The classification downstream is ViT-B/16 [11], trained using SGD with a learning rate of 0.001, momentum of 0.9, weight decay of 0.0005, and batch size of 16. For the segmentation downstream, we employ SAMUS [12] with its default training configuration.

B. Comparison results

In this section, we provide quantitative and qualitative comparisons between the proposed **TRUST** and prior style transfer methods, including GAN-based approaches (CycleGAN [5], DiscoGAN [15]) and Transformer-based approaches (S2WAT [7], TransColors [6]). These comparisons demonstrate that TRUST effectively preserves the structural content of the source and injects suitable, task-optimal styles from the target domain. Surprisingly, TRUST achieves state-of-the-art results on 6 cross-device ultrasound tasks.

Quantitative results. Table I reports the classification (Accuracy/AUC) and segmentation (Dice/IoU) performance before and after applying style transfer. In particular, TRUST substantially shows improvements compared to without style transfer (w/o ST) and GAN-based works. Furthermore, compared to the Transformer-based method TransColors, TRUST achieves notable gains of $+10.0\%$ in Accuracy and $+8.3\%$ in AUC on the UCLM → UDIAT classification task. In segmentation, it outperforms S2WAT under the UDIAT → BUSI setting, with improvements of $+3.6\%$ in Dice and $+4.4\%$ in IoU. These results indicate that S2WAT or TransColors relies heavily on one-to-one (source-target image) pairings, limiting their ability to capture the common target styles. Moreover, their naive style injection often introduces unwanted noise into the source content, resulting in degraded performance—even below w/o ST—on tasks, such as BUSI → UCLM and UDIAT → BUSI.

Qualitative results. In Figure 3, we present visual comparisons and Grad-CAM [16] results on the test set, revealing several key observations. First, TRUST outperforms TransColors in capturing diverse styles within a specific target domain. Its outputs clearly reflect this strength, showing the closest resemblance to the target samples in color and brightness. Second, TRUST generates finer tissue details and clearer tumor boundaries, highlighting the advantage of using reliable information. Third, TRUST improves semantic cues for downstream classification, as shown by Grad-CAM maps that focus more precisely on tumors and their boundaries, indicating more effective attention to diagnostically relevant regions. Overall, TRUST achieves precise style transfer and task-focused attention, providing clear advantages over conventional approaches that treat all target information equally.



(a) Before Style Transfer (b) After Style Transfer

Fig. 4: t-SNE [22] visualization of feature distributions from the downstream classifier on the UDIAT → UCLM task. Each point represents a sample, where green and red denote benign and malignant classes, respectively. ⋆ marks target samples (UCLM), while ○ denotes source samples (UDIAT), either original (w/o ST) or translated (TRUST).

IV. Analysis

Ablation studies. Table II presents an ablation study across three cross-device settings to evaluate the effectiveness of each component in TRUST. Applying naive cross attention (CA) introduces noisy target tokens and can degrade performance—for example, on UCLM → UDIAT, classification accuracy drops from 63.8 (w/o ST) to 60.0. This supports our key motivation: *"Not all information is knowledge."* Replacing CA with our Token-dRiven (TR) module leads to consistent improvements across all tasks. Compared to w/o ST, TR improves classification accuracy by $+3.1\%$, $+7.5\%$, and $+4.1\%$ for BUSI → UCLM, UCLM → UDIAT, and UDIAT → BUSI, respectively. Similar trends are observed in segmentation, where TR enhances Dice scores by up to $+0.8\%$ and IoU by up to $+0.6\%$ on the BUSI → UCLM task. Finally, integrating TR with auxiliary prompts further refines content representations. TRUST achieves the best overall performance, notably improving AUC from 92.5 to 94.9 and IoU from 76.2 to 76.5 on UDIAT → BUSI. These results confirm the complementary benefits of prompt tuning in guiding source tokens toward downstream predictions.

Feature space. Figure 4 presents t-SNE [22] visualizations of feature distributions from the downstream classifier (ViT-B/16) under the UDIAT → UCLM task. Before style transfer (Figure 4a), a clear domain gap exists between source samples (○) and target samples (⋆), leading to potential misclassification in the absence of style adaptation (w/o ST). After applying TRUST (Figure 4b), the translated source features are better aligned with the target distribution. Through visualizing, we can conclude that TRUST facilitates domain-invariant feature learning while preserving class-discriminative structures essential for downstream tasks.

Computational cost. We further analyze the efficiency of TRUST in terms of floating-point operations (FLOPs). The overall complexity of TRUST amounts to 138.63G FLOPs for an input resolution of 256×256.

V. Conclusion

In this work, we present **TRUST**, a novel token-driven framework for structure-preserving style transfer in cross-device ultrasound analysis. TRUST introduces a content-aware mechanism that selectively integrates target style cues **without distorting** the anatomical structure of the source. Through extensive experiments on 6 cross-domain tasks, TRUST consistently outperforms existing GAN-based and Transformer-based methods across both classification and segmentation, achieving state-of-the-art performance. Furthermore, ablation studies and t-SNE visualizations confirm that TRUST produces domain-invariant and class-discriminative features. These results demonstrate the potential of TRUST as a reliable solution for real-world deployment across heterogeneous ultrasound devices.

Acknowledgement. This work was financially supported in part (project number: 112UA10019) by the Co-creation Platform of the Industry Academia Innovation School, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE) and industry partners in Taiwan. It also supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC-114-2218-E-A49 -024, - Grant NSTC-112-2221-E-A49-089-MY3, Grant NSTC-114-2425-H-A49-001, Grant NSTC-113-2634-F-A49-007, Grant NSTC-112-2221-E-A49-092-MY3, and in part by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and the Ministry of Education (MOE), Taiwan. It is also partly supported by MediaTek Inc., Hon Hai Research Institute, and Industrial Technology Research Institute.

References

[1] C. A. Linte, J. Moore, C. Wedlake, and T. M. Peters, "Evaluation of model-enhanced ultrasound-assisted interventional guidance in a cardiac phantom," *IEEE transactions on biomedical engineering*, vol. 57, no. 9, pp. 2209–2218, 2010.

[2] B. G. Ewigman, J. P. Crane, F. D. Frigoletto, *et al.*, "Effect of prenatal ultrasound screening on perinatal outcome," *New England journal of medicine*, vol. 329, no. 12, pp. 821–827, 1993.

[3] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *IEEE International Symposium on Biomedical Imaging*, IEEE, 2019, pp. 953–956.

[4] S. Kim and B. C. Song, "Us-gan: Ultrasound image-specific feature decomposition for fine texture transfer," *IEEE Access*, 2024.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2223–2232.

[6] Q. Liu, D. Zhao, L. Tang, and L. Xu, "Tanrscolour: Transformer-based medical image colourization with content and structure preservation," *IET Image Processing*, vol. 18, no. 10, pp. 2702–2714, 2024.

[7] C. Zhang, X. Xu, L. Wang, Z. Dai, and J. Yang, "S2wat: Image style transfer via hierarchical vision transformer using strips window attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 7024–7032.

[8] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2071–2081.

[9] Y. Deng, F. Tang, W. Dong, *et al.*, "Stytr2: Image style transfer with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 326–11 335.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] X. Lin, Y. Xiang, L. Yu, and Z. Yan, "Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 24–34.

[13] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5880–5888.

[14] M. Jia, L. Tang, B.-C. Chen, *et al.*, "Visual prompt tuning," in *European Conference on Computer Vision*, Springer, 2022, pp. 709–727.

[15] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*, Pmlr, 2017, pp. 1857–1865.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 618–626.

[17] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104 863, 2020.

[18] M. H. Yap, G. Pons, J. Martí, *et al.*, "Automated breast ultrasound lesions detection using convolutional neural

networks," *IEEE J. Biomed. Health Informatics*, vol. 22, no. 4, pp. 1218–1226, 2018.

[19] N. Vallez, G. Bueno, O. Deniz, M. A. Rienda, and C. Pastor, "Bus-uclm: Breast ultrasound lesion segmentation dataset," *Scientific Data*, vol. 12, no. 1, p. 242, 2025.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.

[21] R. Xiong, Y. Yang, D. He, *et al.*, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, PMLR, 2020, pp. 10 524–10 533.

[22] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.