Low-Resource Voice Synthesis Through Residual Style Encoding and Uncertainty-Based Domain Adaptation

Hitesh C. Sabharwal
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
hiteshcsabharwal@colostate.edu

Advika C. Muni
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
advika.muni451@missouri.edu

Abstract

Low-resource voice synthesis must generate intelligible and expressive speech for speakers and styles for which only limited transcribed or untranscribed audio is available. Conventional acoustic models trained on high-resource corpora often transfer poorly because target recordings differ in speaker identity, recording condition, prosody, and affect, while style encoders tend to entangle these factors with phonetic content. This paper develops a methodological framework that combines residual style encoding with uncertainty-based domain adaptation for low-resource voice synthesis. The framework represents a reference style as a residual vector between a reference-derived style code and a speaker-conditioned base style anchor; the residual is injected into an acoustic decoder through adaptive normalization and regularized by a conditional prior. Domain adaptation aligns source and target feature distributions using an uncertainty-weighted adversarial objective, in which per-sample heteroscedastic variance and ensemble disagreement determine the contribution of target frames and pseudo-labeled utterances. A reliability gate combines predictive variance, heteroscedastic uncertainty, and content consistency to suppress unstable target supervision. The framework yields a coherent training objective and explicit design trade-offs involving style disentanglement, adversarial alignment, pseudo-label noise, and low-resource overfitting. Its practical significance lies in providing a principled route for adapting expressive voice synthesis to under-resourced speakers without treating all target observations as equally trustworthy.

Keywords

low-resource speech synthesis; residual style encoding; uncertainty estimation; domain adaptation; acoustic modeling; prosody transfer

1. Introduction

Neural voice synthesis has advanced through autoregressive acoustic models [1], mel-spectrogram vocoders [2], feed-forward duration and prosody predictors [3], [4], and neural waveform generators [5]. Subsequent flow-based [6] and diffusion-based [7] acoustic models improved likelihood modeling and sample quality, while zero-shot multispeaker systems [8], neural codec language models [9], and latent diffusion synthesizers [10] extended synthesis to unseen speakers. These advances, however, depend on large corpora and broad phonetic

coverage. In low-resource settings, the target speaker may provide only a few hours or minutes of usable speech, and the available recordings may be noisy, weakly transcribed, or stylistically narrow. A model trained only on the source corpus can produce fluent speech but often fails to match the target speaker's timbre, rhythm, and expressive tendencies.

Style transfer and domain adaptation are two natural responses to this problem. Style transfer supplies a reference utterance whose prosody and affect can condition the acoustic model, while domain adaptation aligns representations learned from a high-resource source domain with unlabeled or sparsely labeled target speech. Yet both responses have failure modes when combined naively. Style encoders frequently absorb speaker identity, channel characteristics, and phonetic content into a single latent variable, so that transferring style also transfers unwanted speaker traits. Domain adaptation methods frequently treat all target samples as equally informative, which is problematic because low-resource target data contain heteroscedastic noise, unstable prosody, and out-of-distribution phonetic contexts. Pseudo-labels generated for such samples can reinforce early model errors.

This paper develops a unified methodological framework that addresses these issues through residual style encoding and uncertainty-based domain adaptation. The central contribution is a formal decomposition in which a reference style is represented as a residual relative to a speaker-conditioned base style anchor. The residual is intended to capture local prosodic, affective, and recording-condition variation that is not explained by the speaker's average style. It is injected into a shared acoustic decoder through adaptive normalization and regularized by a conditional prior, so that limited target data do not force the residual code to absorb speaker identity or phonetic content. The second contribution is an uncertainty-weighted adaptation objective that modulates adversarial alignment, pseudo-label acceptance, and consistency regularization using heteroscedastic variance and ensemble disagreement. Variational style latents [16], out-of-domain singing-voice style transfer [17], discrete acoustic codes [18], adaptive instance normalization [19], and self-ensembling [20] inform the design space, but the present framework is organized around low-resource spoken voice synthesis and the reliability of target-domain supervision.

The framework is analytical and design-oriented. It specifies variables, objectives, constraints, and failure conditions, and it derives implications for when residual style encoding is preferable to direct speaker embedding, when uncertainty weighting improves adaptation, and when it can discard useful hard examples. The next section reviews the relevant technical background. Section 3 formulates residual style encoding. Section 4 develops uncertainty-based domain adaptation. Section 5 states the training objective and inference procedure. Section 6 analyzes trade-offs and boundary conditions. Section 7 concludes.

2. Related Work

Neural voice synthesis has progressed from autoregressive sequence-to-sequence models [1] to non-autoregressive predictors [3], [4] that reduce inference latency and improve duration control. Parallel acoustic models require robust alignment and variance prediction, and their success depends on large annotated corpora. Neural vocoders [2], [5] convert predicted acoustic features into waveforms, while flow-based [6] and diffusion-based [7] models provide alternative generative formulations. Zero-shot multispeaker synthesis [8] and codec or latent-diffusion language models [9], [10] reduce the need for per-speaker training data, but they still rely on broad pretraining and do not directly solve the problem of adapting to a noisy, small, or stylistically unusual target corpus.

Domain adaptation provides a complementary route. Domain-adversarial training [11] aligns source and target feature distributions through a domain classifier and gradient reversal. Conditional adversarial adaptation [12] conditions alignment on discriminative information, and maximum classifier discrepancy [13] uses disagreement between classifiers to identify target samples that are poorly aligned. These methods are effective when target data are sufficiently abundant and when domain labels are meaningful. In low-resource voice synthesis, however, the target domain may be represented by a single speaker or a small group of speakers, so domain alignment can conflate speaker identity with recording condition and style.

Uncertainty estimation is relevant because target adaptation is inherently noisy. Monte Carlo dropout [14] approximates Bayesian predictive uncertainty through stochastic forward passes, and heteroscedastic regression [15] learns input-dependent observation noise. Such estimates can distinguish aleatoric noise, such as recording artifacts, from epistemic uncertainty caused by unfamiliar phonetic or prosodic contexts. The framework in this paper uses both sources of uncertainty to weight adaptation and to gate pseudo-label acceptance.

Style representation is the third background element. Variational autoencoders [16] provide a probabilistic latent space for acoustic variation. Out-of-domain singing-voice style transfer [17] addresses the related problem of separating reference style from speaker identity when the target style lies outside the training distribution. Discrete acoustic codes [18] offer a compact alternative to continuous latents, while adaptive instance normalization [19] supplies an efficient mechanism for injecting global style statistics into a decoder. Self-ensembling [20] provides a teacher-student consistency mechanism that is useful when target labels are absent. The proposed framework differs from these approaches by coupling a speaker-conditioned residual style code to an uncertainty-weighted domain adaptation objective designed for low-resource voice synthesis.

3. Residual Style Encoding

3.1 Problem Variables and Domain Setting

Let a source domain contain high-resource paired data consisting of phonetic sequences, acoustic targets, speaker identities, and reference style recordings. Let a target domain contain a small set of paired utterances and a larger set of unlabeled or weakly labeled utterances from one or more low-resource speakers. An utterance is represented by a phonetic sequence x, a speaker identity s, and an optional reference recording y_ref that carries prosodic and affective style. The acoustic model predicts a mel-spectrogram $\hat{M}$. The target acoustic representation is M.

The purpose of residual style encoding is to prevent the reference style code from duplicating speaker information already available through s. A speaker-conditioned base anchor A_ψ(s) represents the average style associated with the speaker or speaker cluster. A style encoder E_φ(y_ref) maps the reference recording to a latent vector. The residual style code is defined as the difference between the reference code and the base anchor.

$$r = E_{\varphi}(y_ref) - A_{\psi}(s)$$

Here r is the residual style vector, E_φ is a style encoder with parameters φ, A_ψ is a speaker-conditioned anchor network with parameters ψ, y_ref is the reference waveform or acoustic feature sequence, and s is the speaker identity or embedding. The subtraction is not claimed to guarantee disentanglement, but it provides an explicit inductive bias: the model

need only encode the part of the reference style that is not predicted by the speaker anchor. When no reference is available, r is set to zero, which recovers a neutral speaker-conditioned synthesis path.

3.2 Injection Through Adaptive Normalization

The residual code must influence prosody and timbre without overwriting phonetic content. Layer-wise adaptive normalization provides a controlled injection mechanism. For decoder hidden state H_l at layer l, the residual code produces scale and bias parameters through small multilayer perceptrons.

$$H_l' = (1 + \gamma(r)) \odot H_l + \beta(r)$$

Here H_l′ is the modulated hidden state, γ(r) and β(r) are residual-conditioned scale and bias vectors, and ⊙ denotes elementwise multiplication. The identity-centered form (1 + γ(r)) keeps the modulation close to identity when r is near zero, which is important in low-resource training because it prevents the style branch from destabilizing the shared content decoder. The modulation is applied at multiple decoder layers so that the residual can affect duration-related and pitch-related representations without replacing the phonetic encoder output.

A conditional prior regularizes the residual. The style encoder is trained with a variational objective in which the residual is encouraged to explain acoustic variation beyond the speaker anchor.

$$L_style = -E_q[\log p(M \mid r, x, s)] + \beta_KL\, KL(q(r \mid y_ref) \,\|\, p(r \mid s))$$

Here q(r | y_ref) is the approximate posterior produced by the style encoder, p(r | s) is a conditional prior centered at zero with speaker-dependent variance, β_KL controls the capacity of the residual, and M is the acoustic target. The reconstruction term encourages the modulated decoder to match the acoustic target, while the Kullback-Leibler term limits how far the residual can drift from the speaker-conditioned prior. In low-resource settings, a larger β_KL favors a conservative residual that captures only robust prosodic variation; a smaller β_KL permits expressive variation but increases the risk of encoding speaker identity or channel artifacts. Discrete residual codes [18] are a plausible alternative when the residual must be compact and robust, whereas continuous codes offer smoother interpolation for prosody.

4. Uncertainty-Based Domain Adaptation

4.1 Why Uniform Alignment Is Insufficient

Domain adaptation aligns source and target representations by minimizing a domain discrepancy or by training a domain classifier adversarially [11]. When target data are scarce, the domain classifier can exploit speaker identity, channel noise, or recording level rather than transferable phonetic or prosodic properties. A small number of atypical target utterances can then dominate the gradient. The framework therefore assigns each target frame or utterance a reliability weight derived from predictive uncertainty and content consistency.

The acoustic model outputs a mean $\hat{M}_i$ and a variance σ_i^2 for frame i. The heteroscedastic regression loss for frames with available or accepted pseudo-labels is formulated as follows.

$$L_unc = \sum_i \left(\|M_i - \hat{M}_i\|^2 / (2\sigma_i^2) + \tfrac{1}{2} \log \sigma_i^2 \right)$$

Here M_i is the reference acoustic target or accepted pseudo-label, $\hat{M}$_i is the predicted mean, and σ_i² is the predicted variance. The first term down-weights examples with high estimated noise, while the logarithmic term prevents the variance from growing without bound. Under domain shift, σ_i² can reflect both acoustic noise and model ignorance; it is therefore useful but not sufficient as a reliability measure.

4.2 Uncertainty-Weighted Adversarial Alignment

Let F_i be an intermediate feature representation, d_i the domain label, D a domain classifier, and ℓ_D the domain classification loss. Instead of weighting all frames equally, the adversarial objective uses inverse uncertainty weights.

$$L_adv = (\sum_i w_i \ell_D(F_i, d_i)) / (\sum_i w_i), w_i = 1 / (\sigma_i^2 + \varepsilon)$$

Here w_i is the reliability weight, ε is a small positive constant for numerical stability, and the summation runs over source and target frames used in the adaptation batch. Source frames have known domain labels and can be weighted by their supervised uncertainty; target frames are weighted by predictive uncertainty. The weighting reduces the influence of high-variance target frames, which are more likely to arise from segmentation errors, channel noise, or unseen phonetic contexts. A floor on w_i is necessary in practice, because otherwise the model may ignore difficult but informative target regions and reinforce a narrow alignment.

The domain classifier is trained adversarially with a gradient reversal layer or an equivalent min-max procedure. The feature extractor minimizes domain predictability, while the classifier maximizes it. Conditioning the alignment on phonetic or content labels, as in conditional adversarial adaptation [12], can reduce the risk that the domain classifier separates speakers rather than domains. Classifier discrepancy [13] offers a complementary signal: target frames with high disagreement between multiple acoustic heads are less reliable for alignment.

4.3 Reliability Gate for Pseudo-Labels

Pseudo-labeling is attractive in low-resource synthesis because unlabeled target speech is often more plentiful than transcribed speech. Its failure mode is confirmation bias: early predictions become training targets, and uncertain or incorrect predictions are amplified. The framework combines two uncertainty sources into a reliability score.

$$u_i = \lambda_v Var_T(\hat{M}_i) + \lambda_h \sigma_i^2$$

Here Var_T($\hat{M}$_i) is the predictive variance across stochastic forward passes or an ensemble of teacher models, σ_i² is the heteroscedastic variance, and λ_v and λ_h are nonnegative trade-off coefficients. The first term captures epistemic uncertainty arising from unfamiliar inputs; the second captures learned observation noise. A content consistency score c_i is computed from the agreement between content features under two perturbations, such as time masking and small pitch shifts. The acceptance rule is then defined as follows.

$$A_i = 1 \; if \; u_i \le \tau_u \; and \; c_i \ge \tau_c, \; else \; 0$$

Here A_i is an acceptance indicator, τ_u is an uncertainty threshold, and τ_c is a content-consistency threshold. Accepted frames contribute to pseudo-label supervision and consistency regularization; rejected frames contribute only to unsupervised alignment with reduced weight. The thresholds should be set from source-domain calibration and target-domain quantiles rather than from target labels. When the target domain is extremely small, a conservative threshold reduces confirmation bias but may leave too few accepted frames; a

permissive threshold increases coverage but raises the risk of learning speaker-specific artifacts.

5. Training Objective and Inference

The complete training objective combines supervised synthesis, residual style regularization, uncertainty-weighted adversarial alignment, and consistency regularization.

$$L = L_sup + \alpha\, L_style + \beta\, L_adv + \lambda_c\, L_cons + \rho\|\theta\|^2$$

Here L_sup is the supervised acoustic loss on source data and accepted target pseudo-labels, L_style is the residual style objective from Equation 3, L_adv is the uncertainty-weighted domain loss from Equation 5, L_cons is a consistency term between the student prediction and a slowly updated teacher prediction on accepted target frames, θ denotes trainable parameters, and α, β, λ_c, and ρ are nonnegative weights. The consistency term is defined in prose as a weighted mean squared error over accepted frames, with weights exp(−u_i/τ) so that low-uncertainty frames dominate. The teacher is an exponential moving average of the student, following the self-ensembling principle [20], and it stabilizes pseudo-labels during early adaptation.

Training proceeds in three coupled stages within each batch. First, source and target utterances are encoded, and the residual style code is computed for utterances with reference audio. Second, the acoustic decoder predicts mean and variance, and the domain classifier produces domain logits from intermediate features. Third, pseudo-label acceptance and uncertainty weights are computed, and gradients are applied to the supervised, style, adversarial, and consistency terms. The style encoder and speaker anchor are updated only when reference audio is available; otherwise the residual path is set to zero and the model reduces to a speaker-conditioned acoustic model.

At inference, the residual style code is computed once from the reference recording. The modulation parameters are applied to the decoder, and the predicted mel-spectrogram is passed to the vocoder. When the residual norm exceeds a calibrated radius, it is clipped to prevent out-of-distribution modulation. The uncertainty path is not used to reject user requests; it is used during adaptation to decide which unlabeled target frames should influence training. This separation is important because inference-time uncertainty may reflect expressive variation that is desirable rather than erroneous.

6. Analytical Assessment and Technical Discussion

The residual decomposition has a clear inductive bias but also a structural limitation. If the speaker anchor A_ψ(s) is too expressive or too flexible, the residual will be small, and the model will attribute most prosodic variation to the speaker identity rather than to the reference style. If the anchor is too weak, the residual must encode speaker identity, channel, and style simultaneously, which increases entanglement and overfitting in low-resource settings. The conditional prior in Equation 3 controls this trade-off through β_KL, but the appropriate capacity is speaker- and corpus-dependent. A useful design principle is to initialize the anchor as a speaker-level mean style and to allow the residual to grow only when the reference recording differs systematically from that mean. The adaptive normalization in Equation 2 further constrains the residual to multiplicative and additive modulation, which is less destructive than replacing decoder hidden states.

Uncertainty weighting changes the effective gradient assigned to target frames. Frames with large σ_i² receive small w_i in Equation 5, so their influence on domain alignment is

reduced. This is desirable when high variance indicates noise, but it can also suppress difficult phonetic contexts that are precisely the ones the target speaker needs. A lower bound on w_i and a phoneme-balanced sampler mitigate this risk. The heteroscedastic loss in Equation 4 is also not a substitute for calibration: a model can learn high variance for an entire target speaker and thereby exclude that speaker from adaptation. Monitoring variance by phoneme class and by speaker is therefore necessary. The reliability score in Equation 6 combines epistemic and aleatoric terms, but the weights λ_v and λ_h should reflect whether the dominant failure mode is model ignorance or recording noise. When unlabeled target data are clean but acoustically distant from the source, epistemic uncertainty should dominate; when target data are noisy, heteroscedastic variance should carry more weight.

Adversarial alignment introduces a second tension. A domain classifier can align target features to source features even when the target phoneme inventory or prosodic system differs. Conditional alignment [12] and classifier discrepancy [13] reduce this risk by focusing on class-consistent features, but they do not eliminate it. In voice synthesis, alignment should be applied to content and prosody representations separately, or at least with a gradient scale that is lower for speaker-sensitive layers. The residual style branch should not be aligned adversarially to the source style distribution, because the purpose of the residual is to represent target-specific variation. Instead, adversarial alignment should operate on the shared content encoder and the speaker-conditioned base path, leaving the residual free to encode target style within the prior radius.

The framework is most appropriate when three conditions hold. First, the target speaker's phoneme inventory is covered by the source model or by a shared multilingual phoneme set; otherwise, synthesis errors arise from missing units rather than from style or domain shift. Second, reference recordings are available for the expressive styles of interest; residual style encoding cannot invent an unseen style from text alone. Third, the target corpus contains enough unlabeled speech to estimate uncertainty and enough clean segments to anchor the speaker prior. When these conditions fail, a simpler fine-tuning or speaker-embedding approach may be preferable because the additional style and adaptation machinery introduces optimization cost without a corresponding benefit.

Compared with direct speaker embedding, residual style encoding adds an explicit pathway for reference-dependent prosody and affect. Compared with global style tokens, it conditions the residual on the speaker anchor and therefore reduces the chance that a reference utterance transfers its speaker identity. Compared with unweighted domain-adversarial training, uncertainty weighting protects the model from noisy target frames and unstable pseudo-labels. The cost is additional hyperparameters, a style encoder, a variance head, and a teacher model. These costs are justified when target data are scarce and heterogeneous, but they should be evaluated against parameter-efficient adaptation and simple data augmentation in any specific deployment.

7. Conclusion

This paper has developed a methodological framework for low-resource voice synthesis that combines residual style encoding with uncertainty-based domain adaptation. The residual style code is defined relative to a speaker-conditioned base anchor, injected through adaptive normalization, and regularized by a conditional prior so that limited target data do not force style to absorb speaker identity. Domain adaptation is formulated as an uncertainty-weighted adversarial and consistency objective, with heteroscedastic variance and ensemble disagreement controlling the influence of target frames and pseudo-labels. The framework

specifies how style representation, domain alignment, and reliability estimation interact, and it identifies the main trade-offs: anchor capacity versus residual expressiveness, uncertainty weighting versus retention of hard examples, and adversarial alignment versus preservation of target-specific style. These relationships provide a principled basis for adapting expressive voice synthesis to under-resourced speakers while maintaining explicit control over what is transferred and what is learned from the target domain.

References

1. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of Interspeech 2017 (pp. 4006–4010). https://doi.org/10.21437/Interspeech.2017-1452

2. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779–4783). https://doi.org/10.1109/ICASSP.2018.8461368

3. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech. In Advances in Neural Information Processing Systems (Vol. 32, pp. 3165–3174).

4. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. In International Conference on Learning Representations.

5. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17022–17033).

6. Kim, J., Kim, S., Kong, J., & Yoon, S. (2021). Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. In Advances in Neural Information Processing Systems (Vol. 33, pp. 8067–8077).

7. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., & Kudinov, M. (2021). Grad-TTS: A diffusion probabilistic model for text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (pp. 8599–8608).

8. Casanova, E., Weber, J., Shulby, C. D., Candido, A., Jr., Oliveira, E. C., & Ponti, M. A. (2022). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Proceedings of the 39th International Conference on Machine Learning (pp. 2709–2720).

9. Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., & Wei, F. (2023). Neural codec language models are zero-shot text to speech synthesizers. arXiv. https://arxiv.org/abs/2301.02111

10. Shen, K., Ju, Z., Yang, D., Tan, X., Liu, E., Leng, Y., He, L., Qin, T., Zhao, S., & Bian, J. (2023). NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv. https://arxiv.org/abs/2304.09116

11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

12. Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018). Conditional adversarial domain adaptation. In Advances in Neural Information Processing Systems (Vol. 31, pp. 1640–1650).

13. Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3723–3732). https://doi.org/10.1109/CVPR.2018.00392

14. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059).

15. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584).

16. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In International Conference on Learning Representations.

17. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

18. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6306–6315).

19. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 1501–1510). https://doi.org/10.1109/ICCV.2017.167

20. French, G., Mackiewicz, M., & Fisher, M. (2018). Self-ensembling for visual domain adaptation. In International Conference on Learning Representations.