Learning Visual Styles from Audio-Visual Associations

Tingle Li[1,3], Yichen Liu[1],
Andrew Owens[2], and Hang Zhao[1,3]

[1]Tsinghua University [2]University of Michigan [3]Shanghai Qi Zhi Institute
https://tinglok.netlify.com/files/avstyle

A Appendix

A.1 *Into the Wild* dataset

We introduce the *Into the Wild* dataset, a set of egocentric hiking videos for our proposed audio-driven image stylization (ADIS), because hiking is featured with a strong audio-visual association of nature.

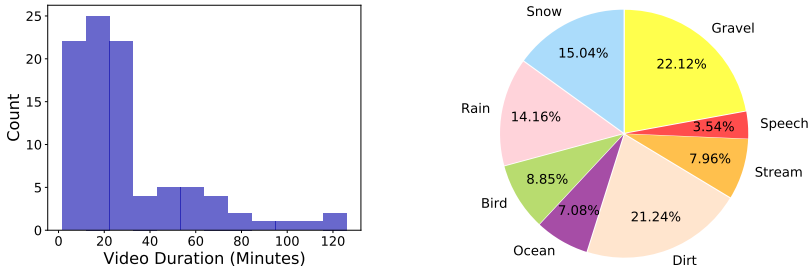
We collected these videos on YouTube by searching for the keywords like hike+POV, hike+footsteps, hike+ASMR, and hike+binaural. We employ YAMNet [13] to tag each associated soundtrack to ensure that they play the actual sound and are not replaced by any other sounds, such as background music.

The duration statistics of the *Into the Wild* dataset are shown in Figure 1a. Specifically, it contains 94 untrimmed videos, some of which are already presented in Figure 4 of the main paper. Please note that the category labels of these videos are not labeled by humans, but acquired from the YAMNet [13] predictions, which roughly consist of 8 categories: crunching snow, gravel, and dirt; rain; birds chirping; ocean; stream and human speech. The detailed categorical distribution is illustrated in Figure 1b.

A.2 Training Details

Training Setting Except for the batch size and audio network, we intentionally match the architecture and hyperparameter settings with CycleGAN [18] and CUT [11]. We employ ResNet-based generator [8] with 9 residual blocks, PatchGAN discriminator [7], Least Square GAN loss [10], ResNet18-based audio encoder [5], with the batch size of 16, and the Adam optimizer [9] with 0.002 learning rate. Both λ and μ in Eq.(4) of the main paper are set to 0.5.

Our model is trained for 50 epochs, with the learning rate remaining constant for the first 30 epochs and linearly decaying to zero over the last 20 epochs. The encoder G_{enc} follows the first half of the CycleGAN generator [18]. We also extract features from 5 different scales to calculate the patch-based structure discriminator loss: the input RGB pixels, the first and second downsampling convolution features, and the first and fifth residual block features. We sample 256 random locations for each layer's features and apply a 2-layer MLP to obtain 256-dimension features as the final output for computing the multi-scale patch-wise contrastive loss.



(a) Distribution of video duration (b) Distribution of video categories

Fig. 1: Statistical analysis of the *Into the Wild* Dataset.

Fig. 2: A screenshot of AMT for rating the audio-visual correspondence.

***Into the Wild* dataset** We divide all of the videos into 3-seconds video clips, then uniformly sample 8 frames from each video clip to save as images, yielding a total of 454560 images and 56820 audios. We then randomly sample 20% audios as the test set.

The *Greatest Hits* dataset We first identify the videos by the type of object being hit on, and then only the outdoor videos are used for training: dirt, grass, gravel, leaf, and water, resulting in a total of 32172 images and 8043 audios. We then select 15% audios at random as the test set.

A.3 Evaluation Details

Audio-visual Correspondence (AVC) A two-stream network is utilized to compute AVC [1], with one stream extracting audio feature and the other extracting visual feature. Specifically, we apply OpenL3 [4] to obtain these features, and then compute the average cosine similarity for each image-audio pair. To be more explicit, we employ an "env" content type pre-trained model with 512-dimensional linear spectrogram representation.

Fréchet Inception Distance (FID) FID [6] is calculated by scaling the images to 299-by-299 using the PyTorch framework's bi-linear sampling, and then take

Table 1: Quantitative comparison for different pre-training methods on the *Into the Wild* dataset.

Pre-training Method	**Objective Evaluation**		
	AVC (↑)	FID (↓)	CLIP (↑)
Ours (from scratch)	0.820	34.139	0.238
+ SeLaVi [2]	0.822	32.882	0.242
+ Wav2CLIP [16]	**0.831**	**30.334**	**0.246**

the activation of the last average pooling layer of a pre-trained Inception V3 [15]. We adopt Clean-FID [12] to circumvent the issue that FID computation requires complicated and error-prone steps, such as the resizing functions in different libraries often produce inaccurate results.

Contrastive Language-Image Pre-Training (CLIP) [14] is computed by performing contrastive pre-training on a variety of image-text pairs. It's widely known for zero-shot prediction, but we use it as a feature extractor to compute the cosine similarity between images and labels in order to assess conversion quality. To calculate it, we leverage an off-the-shelf "ViT-B/32" CLIP model [14].

Amazon Mechanical Turk (AMT) In addition to the objective evaluations mentioned above, we employ AMT to study the relationship between audio and visual from a subjective standpoint, *i.e.*, human perspective. A screenshot of the demo page is shown in Figure 2. The MTurker is required to rank such correlations based on audios and images generated by our method and the baseline methods, with the best earning 4 points and the worst earning 1 point. Thus, the scores range from 1 to 4. Notably, twenty Mturkers were asked to rank a total of 1000 random samples from the test set in our case. The final scores are reported on average.

A.4 Additional Results

Additional qualitative comparisons Additional qualitative comparisons on our method to the baselines and ablations are shown in Figure 3. It turns out that our model produces better or competitive results, exhibiting its versatility compared to label-based baselines.

Additional generalization results Additional qualitative results of the generalization experiment are shown in Figure 4. These are accomplished by using images from the Places dataset [17] and the audios from the VGG-Sound dataset [3]. Our model is able to generate plausible images that match the content of the out-of-distribution audio.

Additional pre-training comparisons We also use Wav2CLIP [16], an audio representation learning method derived on CLIP [14], to fine-tune ADIS. To transfer knowledge, it employs a frozen image model to bridge the gap between a sophisticated language model and a scratch audio model. Wav2CLIP could be a better pre-training method for ADIS than SeLaVi [2] since it is implicitly exposed to numerous well-annotated image-text pairs. Table 1 shows the quantitative comparison results. It appears that Wav2CLIP surpasses both training from scratch and SeLaVi pre-training methods with respect to the AVC, FID, and CLIP metrics, indicating that it has a stronger representation ability than the others.

Fig. 3: Randomly selected qualitative results of our model, baselines and ablations. This is an extension of Figure 5 in the main paper.

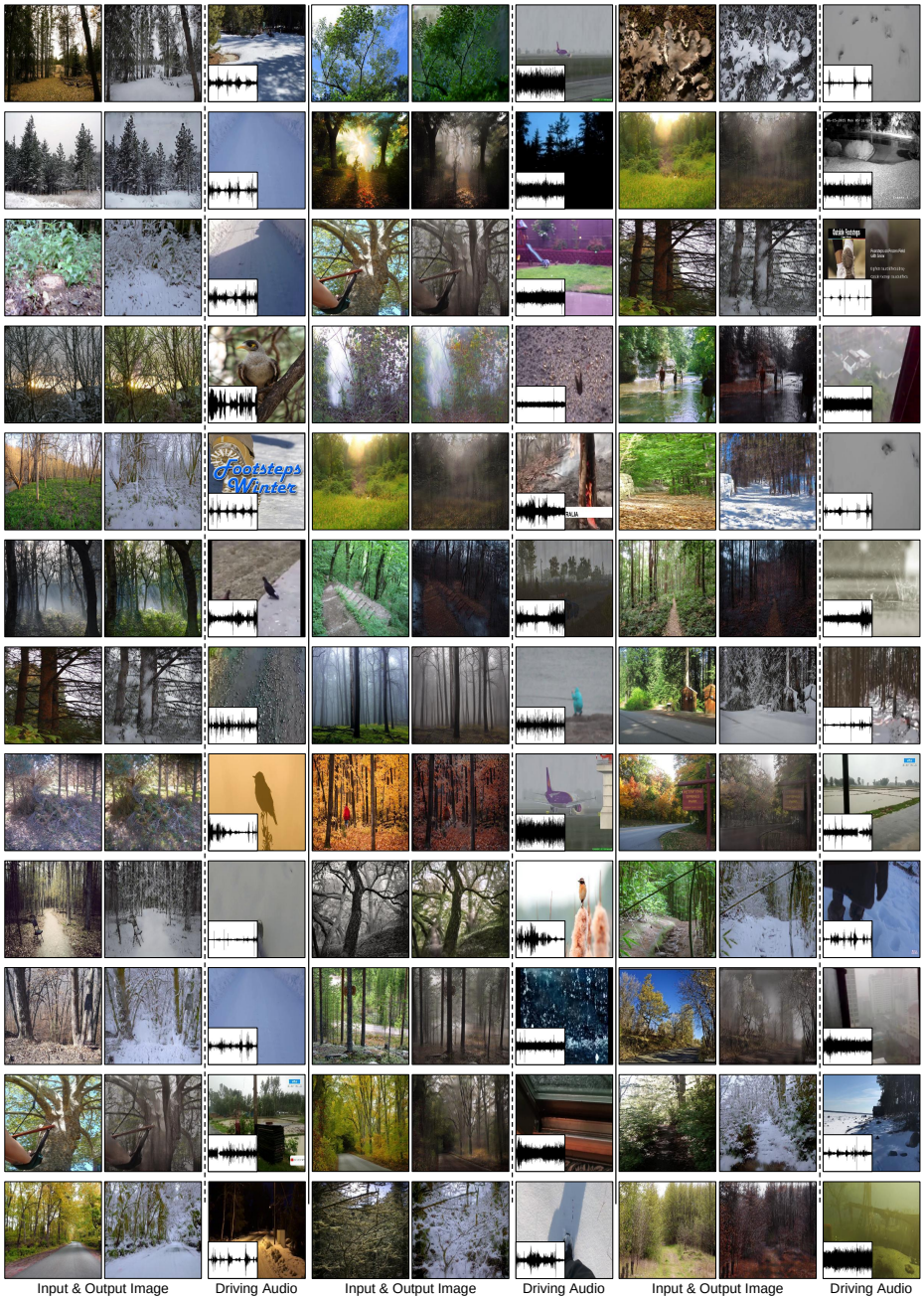


Fig. 4: Randomly selected qualitative results of generalization experiment. This is an extension of Figure 8 in the main paper.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017) 2
2. Asano, Y.M., Patrick, M., Rupprecht, C., Vedaldi, A.: Labelling unlabelled videos from scratch with multi-modal self-supervision. In: Advances in Neural Information Processing Systems (2020) 3, 4
3. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 3
4. Cramer, J., Wu, H.H., Salamon, J., Bello, J.P.: Look, listen, and learn more: Design choices for deep audio embeddings. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3852–3856. IEEE (2019) 2
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (2017) 2
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 1
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016) 1
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations (2015) 1
10. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017) 1
11. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision. pp. 319–345 (2020) 1
12. Parmar, G., Zhang, R., Zhu, J.Y.: On buggy resizing libraries and surprising subtleties in fid calculation. arXiv preprint arXiv:2104.11222 (2021) 3
13. Plakal, M., Ellis, D.: YAMNet. Jan 2020 [Online], available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet 1
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021) 3, 4
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 3
16. Wu, H.H., Seetharaman, P., Kumar, K., Bello, J.P.: Wav2clip: Learning robust audio representations from clip. arXiv preprint arXiv:2110.11499 (2021) 3, 4
17. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 3

18. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2223–2232 (2017) 1