Research Statement: Scalable and Reliable Machine Learning with Language-guided Representation Learning

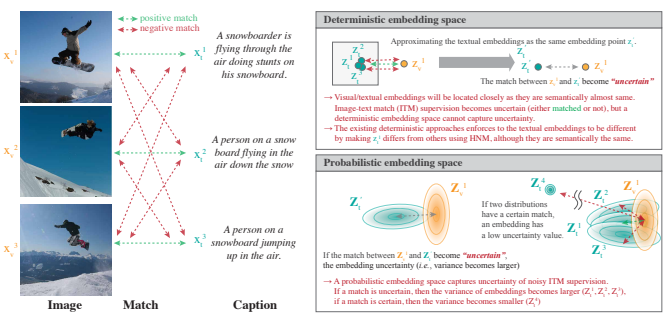
Sanghyuk Chun (sanghyuk.chun@gmail.com https://sanghyukchun.github.io/home)

Ensuring the real-world applicability of machine learning (ML) models poses a primary challenge, namely, the ability to generalize effectively to unseen scenarios encountered beyond the training phase. There are three prominent scenarios frequently encountered in practical applications: (1) when input data significantly differs from the training data; (2) when the model faces the target behavior beyond the scope of training targets, such as unexplored labels; and (3) when the application needs human opinions or subjective value judgments. Addressing these scenarios demands more than massive large-scale datasets; it needs the inclusion of human knowledge that extends beyond web-crawled content. However, how can we effectively integrate large-scale training and human knowledge guidance? To answer the question, my research aims to expand the knowledge of large-scale ML models by more controllability and interpretability, enabling human intervention to guide model behavior even after the training phase. In this research statement, I will explain my three research themes to achieve this goal: (1) Language-combined representation learning, (2) Reliable machine learning, and (3) Optimization techniques for large-scale ML. The first two themes aim to expand machine knowledge in terms of interpretability, controllability, and generalizability, while the last theme is towards practical machine learning algorithms at scale.

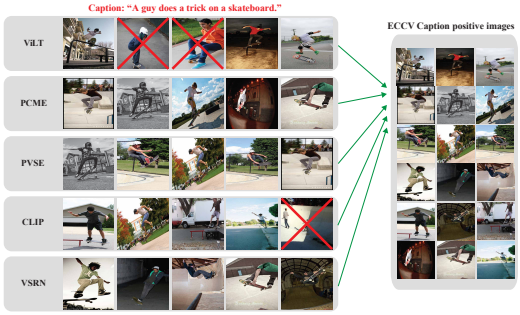
Language-combined Representation Learning

Language serves as the most natural method for encoding human knowledge. If our model can comprehend human language alongside the target modality, we can understand the model better by interventing the space with human language. How can we make a model comprehend human language alongside the target modality? One possible direction for learning language-combined representations is to encode the inputs to the shared embedding space. Despite the recent success of the joint embedding space approach (*e.g.*, CLIP), my recent works have shown that we cannot truly solve the problem by conventional deterministic approaches. Another possible line of research is to leverage recent strong generative models, such as diffusion models or large language models (LLMs). I believe that bridging the gap between generative models and representation learning remains an under-explored frontier with huge potential. In the remaining section, I will delve into the details of these two approaches.

Tackling multiplicity and false negatives of image-text matching tasks. As language descriptions are the product of conscious choices of the key relevant concepts to report from input data, language-combined representation learning methods often suffer from *the multiplicity* (or many-to-many problem) between modalities. In image-text matching (ITM) tasks, this problem is even more serious because there are abundant false negatives (FNs) in the dataset, where we treat the "aligned" image-text pair as the only positive. Assume that we have a "good" image (or text) encoder that understands the semantics of inputs. Then, we can assume that this encoder will map semantically similar images (or captions) to very close locations in the embedding space. In other words, we can approximate these semantically similar image (or caption) embeddings as one unified embedding. Unfortunately, as our dataset has a lot of FNs, the approximated unified image embedding and caption embedding will have a "uncertain" match, *i.e.*, the annotation of the match will be either positive or negative (See Figure 1a). My recent works address this problem by understanding and addressing the multiplicity problem by probabilistic representation learning, *e.g.*, PCME++ [1] and PCME [2]. In this paradigm, an input is mapped to a probabilistic distribution rather than a deterministic vector. This approach enhances the interpretability of datasets and user controllability (*e.g.*, understanding datasets by input uncertainty or uncertainty-based zero-shot prompt tuning). However, not only learning paradigm is required to tackle this problem; we also need a correct and reliable ITM benchmark. For example, in the ECCV Caption paper [3], I built an image-text matching benchmark that fixes numerous FNs in the COCO Caption dataset. In this work, I revealed that the COCO Caption evaluation set has $\times 3.6$ positive captions and $\times 8.5$ positive images compared to the original annotations. Another example is the RoCOCO benchmark [4], a robustness benchmark for the COCO image-text matching task; we showed that current VL models often retrieve captions with a different meaning (*e.g.*, changing "man" to "woman" or "umbrella" to "gun") as the best matching caption to the given query image. The overview of my representative works in addressing many-to-many correspondences caused by abundant FNs is illustrated in Figure 1.



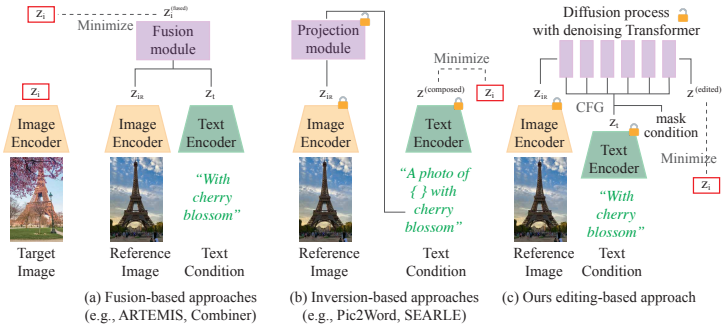
(a) False negatives (FNs) are the source of ambiguity; A probabilistic embedding approach for tackling the problem [1, 2].



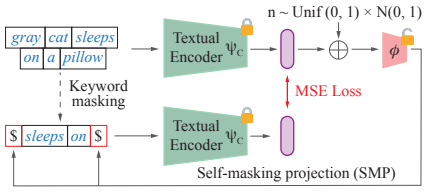
(b) Fixing ITM benchmarks from abundant FNs by machine-in-the-loop [3].

Figure 1: **Two approaches for tackling multiplicity and FNs in image-text matching (ITM) tasks.** (a) Probabilistic embeddings for solving the multiplicity by FNs – PCME (Chun *et al.*, 2021 [2]), PCME++ (Chun, 2023 [1]). (b) Fixed MS-COCO Caption evaluation annotations – ECCV Caption (Chun *et al.*, 2022 [3]).

Leveraging the knowledge of the powerful pre-trained models. The existing language-combined representation learning approaches focus on learning a specialized model with selectively collected text-aligned training datasets. However, it limits the generalizability of the models, *i.e.*, we cannot apply the method to the other types of datasets rather than the training dataset. One of my recent research interests is to address this problem by leveraging knowledge of strong pre-trained models, such as generative models (*e.g.*, diffusion models and large language models (LLMs) trained on extensive billion-scale web-curated datasets) or multi-modal joint embedding models (*e.g.*, CLIP and CLAP). To illustrate this approach, consider the task of composed image retrieval (CIR), which relies on triplets of an image query, a text query, and a target image. Here, obtaining such triplets can be prohibitively expensive and sometimes unfeasible. Therefore, existing CIR methods are trained only on small-scale triplet CIR datasets and struggle to adapt to in-the-wild retrieval scenarios. Moreover, these existing methods are not flexible; they cannot handle versatile conditions beyond a limited textual one (See Figure 2a). In our CompoDiff paper [5], we address these challenges by employing two key ideas: (1) a massive synthetic dataset with a fine-tuned LLM and StableDiffusion models and (2) a versatile latent diffusion model that takes various conditions (*e.g.*, negative text, masked condition, or combinations of them) by classifier-free guidance. Another interesting idea is to train methods only with language inputs by relying on powerful pre-trained VL models, such as CLIP. In LinCIR paper [6], we train a projection module from the CLIP textual latent embedding space to the token embedding space only using texts. This surprisingly simple method shows both efficiency and effectiveness on ZS-CIR tasks (See Figure 2b). I think there are a number of interesting directions for leveraging the power of recent powerful pre-trained models. These may include data augmentation with high fidelity and controllability of the recent generative models, the adaptation of a visual module to LLMs, or fine-tuning LLMs with non-language inputs.



(a) Leveraging powerful diffusion models for CIR tasks [5].



(b) Overview of simple language-only training for ZS-CIR [6].

Figure 2: **Enabling versatile and generalizable CIR by utilizing powerful pre-trained models.** (a) In the CompoDiff paper (Gu and Chun *et al.* [5]), we address the drawbacks of the existing methods originated by the cost of triplet collection and the inherent inflexibility by (1) synthesizing a massive dataset and (2) highly controllable diffusion models with various conditions. (b) In the LinCIR paper (Gu and Chun *et al.* [6]), we train a CIR model only using text inputs, resulting in a high scalability and effectiveness.

Machine Learning Reliability

Existing machine learning models cannot understand the problem itself [7]. This causes many realistic problems, such as discrimination by machines and poor generalizability to unseen (or minor) corruptions, environments, or groups. Current state-of-the-art machines only do "predict", rather than "logical thinking based on logical reasoning". As models prefer to learn by shortcuts [8, 7], just training models as usual will lead to biased models. I am interested in investigating these phenomena with various tools.

If it is difficult to make machines understand the problem itself, what can we do? Our model should not learn undesirable shortcut features [9, 10], or should be robust to unseen corruptions [11, 12, 13, 14] or significant distribution shifts [15, 16]. Also, we need to make a machine not discriminative to certain demographic groups [17, 18]. We expect a model to say "I don't know" when they get unexpected inputs [2, 1]. At least, we expect a model can explain why it makes such decisions [19, 20, 21, 22], how different model design choices will change model decisions [23, 24] and how it can be fixed (*e.g.*, more data collection? more annotations? or filtering?). My research focuses on expanding machine knowledge from "just prediction" to "logical reasoning". In recent years, I have concentrated on tackling various generalization downstream tasks, such as de-biasing [9, 10, 8, 7], domain generalization [15, 16], algorithmic fairness [17, 18], and adversarial robustness [23]. I think most of these tasks can be explained by Figure 3; we expect a model to focus on "signal S", but there is a spurious correlated feature "bias B". Different downstream tasks target different training dataset scenarios with different assumptions of the test signal-bias correlation. As these tasks aim to tackle situations when the training information and the evaluation information are not the same (*e.g.*, different test dataset distributions), I have strong research principles for these topics: (1) the methods should be theoretically sound [9, 17, 18, 15, 16], and (2) the evaluation benchmark should be designed in a correct way [3, 12, 21, 22].

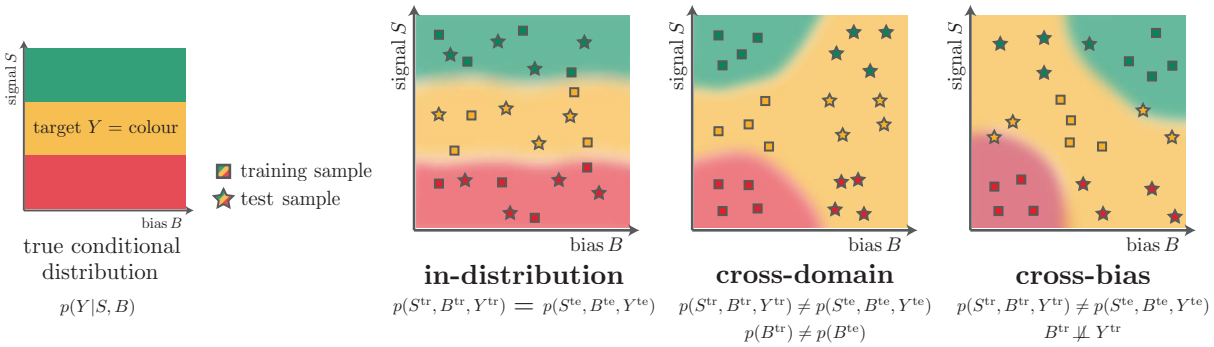


Figure 3: **Two representative learning scenarios for distribution shifts.** Cross-domain and cross-bias scenarios correspond to domain generalization and de-biasing tasks, respectively. Roughly speaking, algorithmic fairness can be viewed as a more targeted version of "cross-bias". The figure is from the ReBias paper [9].

As of now, I am more interested in solving the machine learning reliability or generalizability problem by language-combined representation learning. For example, a vision-language (VL) model trained on large-scale datasets can be easily generalized to various corruptions (*e.g.*, noise, blur), different domains (*e.g.*, cartoon, sketch), or less commonly happened situations (*e.g.*, car in the sky, elephant with zebra skin). Of course, a web-curated dataset can be biased to various types of discriminations (*e.g.*, web images with English captions might be biased to countries using English). However, in general, I think this approach can address many problems that we have targeted before. Similarly, we can expect to make a model more "explainable" by using language guiding, *e.g.*, by letting a model give a rationale for the decision in a human language. There have been a few primitive studies in this area, and I think this direction is worth exploring. The existing VL models cannot directly tackle the uncertainty issue (*e.g.*, estimating the correct confidence of the model prediction), but my previous attempts show that migrating the concept of the uncertainty estimation and VL training is somewhat possible in a small-scale dataset with neglectable sacrifices [2, 1]. Another interesting idea could be directly estimating its confidence as a language output (*e.g.*, 75% confident), which is similar to the XAI with human language. Algorithmic fairness is a possible limitation of the current VL training because it is (usually believed to be) originated from the training dataset. I think that we should keep working on algorithmic fairness for this approach. Overall, I think that language-combined representation learning, my first research theme, can be a potential candidate to target the machine learning reliability problem; while we should keep our eyes on algorithmic fairness to overcome the inherent discrimination of web-curated datasets.

Optimization Techniques for Large-scale ML

Last but not least, I have actively worked on developing general optimization techniques for large-scale ML models. Although we have a number of great practices for large-scale optimization, I believe we still need to develop more optimization techniques to solve our problems. My research emphasizes two key objectives: empirical impact and theoretical soundness. For the empirical impact, I aim to develop easy-to-use techniques that seamlessly function as plug-and-play solutions. I am also eager to develop theoretical sound methods, not based on heuristics and heavy parameter tuning. For example, AdamP [25] can be applied to any method as a plug-and-play solution with minimal changes to the source code. At the same time, we showed that AdamP solves the theoretical problem of the drastic effective learning rate decay problem of scale-invariant parameters. All of my image-text matching cross-modal retrieval works [2, 1] are based on AdamP because their ℓ_2 normalization layers make scale-invariant parameters, resulting in suboptimal results with the existing optimizers. Similarly, SWAD [15] is another plug-and-play solution for domain generalization (DG) algorithms and now the rule-of-thumb to achieve state-of-the-art performance on the benchmarks. In the paper, we have shown the theoretical relationship between flatness and domain generalization and proposed a practical modified version of SWA, a flatness-aware optimizer, for domain generalization tasks. Now, SWAD is essential for achieving a better DG performance, making researchers focus more on algorithms than the optimizer selection.

My research interests include wide areas, such as data augmentation (usually based on mixed sample data augmentation, such as CutMix [11]) [11, 26, 10], storage-efficient learning [27, 13, 28], optimizer [25, 15], network architecture [14, 29, 30], refined labels [13], and theoretical understanding for optimization techniques [26]. I aim to develop a globally applicable algorithm regardless of the dataset domain or architecture. For example, CutMix [11] is now one of the famous data augmentation methods widely used for applications in the vision or audio domain. In the AdamP paper [25], we showed that AdamP could improve the model performances in 13 different benchmarks ranging from vision tasks like classification, retrieval, and detection to language modeling and audio classification.

I also have worked on domain-specific optimization techniques by using assumptions on the structure of data, *e.g.*, the compositionality of Korean/Chinese letters [31, 32, 33, 34, 35], low- and high- frequency information for better audio understanding [36] and style transfer [37], or harmonic information for multi-source audio understanding [38, 39]. Although these approaches are not globally applicable as my main focus, I recognize that if we can properly model the human inductive bias to a model, we can drastically improve the trial-and-error of the model where we will need a tremendously large number of data points with data-driven methods; it is especially beneficial when the training data collection is expensive (*e.g.*, font generation [31, 32, 33, 34, 35]) or legally vague (*e.g.*, music modeling [19, 20, 38, 39]).

Similar to my perspective on machine learning reliability (the second theme), my current main interest lies in language-combined representation learning (the first theme). As one of the key factors of language-combined representation learning is large-scale optimization, I think that developing a theoretically sound and empirically easy-to-use optimization technique will be helpful for my main research theme.

References

[1] **Sanghyuk Chun**. Improved probabilistic image-text representations. *arXiv preprint arXiv:2305.18171*, 2023.

[2] **Sanghyuk Chun**, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021.

[3] **Sanghyuk Chun**, Wonjae Kim, Song Park, Minsuk Chang Chang, and Seong Joon Oh. ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for MS-COCO. In *ECCV*, 2022.

[4] Seulki Park, Daeho Um, Hajung Yoon, **Sanghyuk Chun**, Sangdoo Yun, and Jin Young Choi. RoCOCO: Robust benchmark MS-COCO to stress-test robustness of image-text matching models. *arXiv preprint arXiv:2304.10727*, 2023.

[5] Geonmo Gu, **Sanghyuk Chun**, HeeJae Jun, Yoohoon Kang, Wonjae Kim, and Sangdoo Yun. CompoDiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

[6] Geonmo Gu, **Sanghyuk Chun**, Yoohoon Kang, Wonjae Kim, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. *arXiv preprint arXiv:2312.01998*, 2023.

[7] **Sanghyuk Chun**, Kyungwoo Song, and Yonghan Jung. FAccT 2022 translation/dialogue tutorial: "shortcut learning in machine learning: Challenges, analysis, solutions", 2022.

[8] Luca Scimeca, Seong Joon Oh, **Sanghyuk Chun**, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. In *ICLR*, 2022.

[9] Hyojin Bahng, **Sanghyuk Chun**, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020.

[10] **Sanghyuk Chun** and Song Park. Styleaugment: Learning texture de-biased representations by style augmentation without pre-defined textures. *arXiv preprint arXiv:2108.10549*, 2021.

[11] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, **Sanghyuk Chun**, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[12] **Sanghyuk Chun**, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. In *ICML Workshop on Uncertainty and Robustness in Deep Learning.*, 2019.

[13] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and **Sanghyuk Chun**. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *CVPR*, 2021.

[14] Byeongho Heo, Sangdoo Yun, Dongyoon Han, **Sanghyuk Chun**, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *ICCV*, 2021.

[15] Junbum Cha, **Sanghyuk Chun**, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *NeurIPS*, 2021.

[16] Junbum Cha, Kyungjae Lee, Sungrae Park, and **Sanghyuk Chun**. Domain generalization by mutual-information regularization with pre-trained models. In *ECCV*, 2022.

[17] Sangwon Jung, **Sanghyuk Chun**, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *CVPR*, 2022.

[18] Sangwon Jung, Taeeon Park, **Sanghyuk Chun**, and Taesup Moon. Re-weighting based group fairness regularization via classwise robust optimization. In *ICLR*, 2023.

[19] Minz Won, **Sanghyuk Chun**, and Xavier Serra. Toward interpretable music tagging with self-attention. *arXiv preprint arXiv:1906.04972*, 2019.

[20] Minz Won, **Sanghyuk Chun**, and Xavier Serra. Visualizing and understanding self-attention based music tagging. In *ICML Workshop on Machine Learning for Music Discovery.*, 2019.

[21] Junsuk Choe, Seong Joon Oh, Seungho Lee, **Sanghyuk Chun**, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020.

[22] Junsuk Choe, Seong Joon Oh, **Sanghyuk Chun**, Seungho Lee, Zeynep Akata, and Hyunjung Shim. Evaluation for weakly supervised object localization: Protocol, metrics, and datasets. *arXiv preprint arXiv:2007.04178*, 2020.

[23] Jisung Hwang, Younghoon Kim, **Sanghyuk Chun**, Jaejun Yoo, Ji-Hoon Kim, and Dongyoon Han. Where to be adversarial perturbations added? investigating and manipulating pixel robustness using input gradients. *ICLR Workshop on Debugging Machine Learning Models*, 2019.

[24] Jaehui Hwang, Dongyoon Han, Byeongho Heo, Song Park, **Sanghyuk Chun**, and Jong-Seok Lee. Similarity of neural architectures based on input gradient transferability. *arXiv preprint arXiv:2210.11407*, 2022.

[25] Byeongho Heo, **Sanghyuk Chun**, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights. In *ICLR*, 2021.

[26] Chanwoo Park, Sangdoo Yun, and **Sanghyuk Chun**. A unified analysis of mixed sample data augmentation: A loss function perspective. In *NeurIPS*, 2022.

[27] Song Park, **Sanghyuk Chun**, Byeongho Heo, Wonjae Kim, and Sangdoo Yun. SeiT: Storage-efficient vision training with tokens using 1pixel storage. In *ICCV*, 2023.

[28] Saehyung Lee, **Sanghyuk Chun**, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *ICML*, 2022.

[29] Hwanjun Song, Deqing Sun, **Sanghyuk Chun**, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and Ming-Hsuan Yang. ViDT: An efficient and effective fully transformer-based object detector. In *ICLR*, 2022.

[30] Hwanjun Song, Deqing Sun, **Sanghyuk Chun**, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and Ming-Hsuan Yang. An extendable, efficient and effective transformer-based object detector. *arXiv preprint arXiv:2204.07962*, 2022.

[31] Junbum Cha, **Sanghyuk Chun**, Gayoung Lee, Bado Lee Lee, Seonghyeon Kim, and Hwalsuk Lee. Few-shot compositional font generation with dual memory. In *ECCV*, 2020.

[32] Junbum Cha, **Sanghyuk Chun**, Gayoung Lee, Bado Lee Lee, Seonghyeon Kim, and Hwalsuk Lee. Toward high-quality few-shot font generation with dual memory. *CVPR Workshop on AI for Content Creation*, 2020.

[33] **Sanghyuk Chun**, Song Park, Junbum Cha, Bado Lee, and Hyunjung Shim. Few-shot font generation with localized style representations and factorization. In *AAAI*, 2021.

[34] Song Park, **Sanghyuk Chun**, Junbum Cha, Bado Lee, and Hyunjung Shim. Few-shot font generation with weakly supervised localized representations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 1–17, aug 5555.

[35] Song Park, **Sanghyuk Chun**, Junbum Cha, Bado Lee, and Hyunjung Shim. Multiple heads are better than one: Few-shot font generation with multiple localized experts. In *ICCV*, 2021.

[36] Jang-Hyun Kim, Jaejun Yoo, **Sanghyuk Chun**, Adrian Kim, and Jung-Woo Ha. Multi-domain processing via hybrid denoising networks for speech enhancement. *arXiv preprint arXiv:1812.08914*, 2018.

[37] **Sanghyuk Chun**, Jaejun Yoo, Youngjung Uh, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019.

[38] Minz Won, **Sanghyuk Chun**, Oriol Nieto, and Xavier Serra. Automatic music tagging with harmonic cnn. In *Late Breaking Demo in the ISMIR*, 2019.

[39] Minz Won, **Sanghyuk Chun**, Oriol Nieto, and Xavier Serra. Data-driven harmonic filters for audio representation learning. In *ICASSP*, 2020.