CoordGAN: Self-Supervised Dense Correspondences Emerge from GANs

Jiteng Mu[1*], Shalini De Mello[2], Zhiding Yu[2], Nuno Vasconcelos[1],
Xiaolong Wang[1], Jan Kautz[2], Sifei Liu[2]
[1]UC San Diego, [2]Nvidia

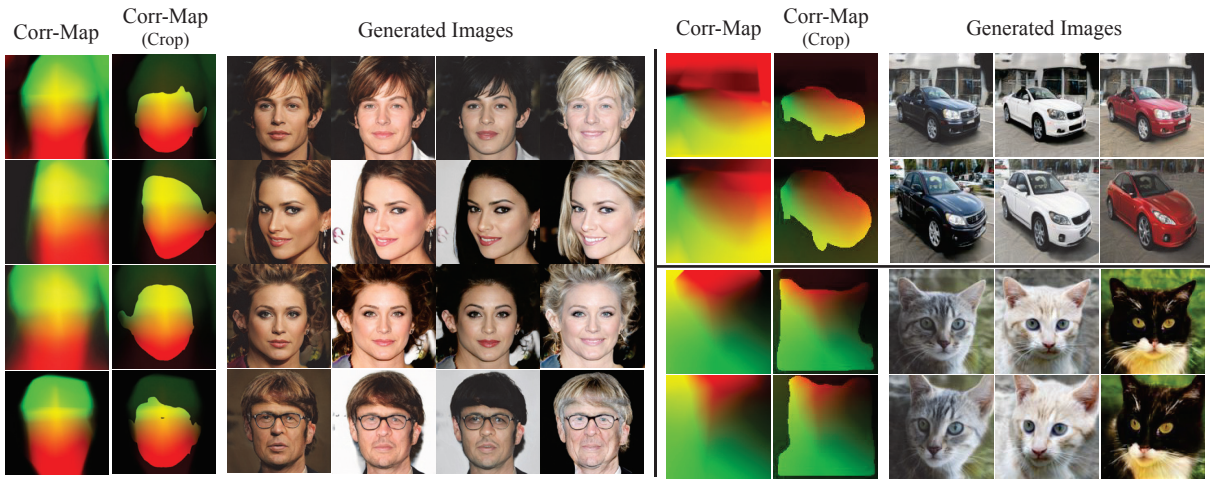

Figure 1. Images synthesized by the proposed CoordGAN for various object categories (left: faces; top-right: cars; bottom-right: cats) : each row displays images with the same structure but different textures; in each column, structure varies while keeping texture fixed. The correspondence maps (Corr-Map) controlling the structure of the synthesized images are shown in the first column of each row. For better visualization, we use off-the-shelf segmentation models to highlight the foreground areas of all the correspondence maps, as shown with Corr-Map (Crop).

Abstract

Recent advances show that Generative Adversarial Networks (GANs) can synthesize images with smooth variations along semantically meaningful latent directions, such as pose, expression, layout, etc. While this indicates that GANs implicitly learn pixel-level correspondences across images, few studies explored how to extract them explicitly. In this work, we introduce Coordinate GAN (CoordGAN), a structure-texture disentangled GAN that learns a dense correspondence map for each generated image. We represent the correspondence maps of different images as warped coordinate frames transformed from a canonical coordinate frame, i.e., the correspondence map, which describes the structure (e.g., the shape of a face), is controlled via a transformation. Hence, finding correspondences boils down to locating the same coordinate in different correspondence maps. In CoordGAN, we sample a transformation to represent the structure of a synthesized instance, while an independent texture branch is responsible for rendering appearance details orthogonal to the structure. Our approach can also extract dense correspondence maps for real images by adding an encoder on top of the generator. We quantitatively demonstrate the quality of the learned dense correspondences through segmentation mask transfer on multiple datasets. We also show that the proposed generator achieves better structure and texture disentanglement compared to existing approaches. Project page: https://jitengmu.github.io/CoordGAN/

1. Introduction

Generative Adversarial Networks (GANs) have achieved great success in synthesizing high-quality images [3, 20–22, 37], and many recent studies show that they also learn a rich set of interpretable directions in the latent space [40, 41]. Moving latent codes along a semantically meaningful direction (e.g., pose) generates instances with smoothly varying appearance (e.g., continually changing viewpoints), implying that GANs also implicitly learn which pixels or regions

[*]Work done while an intern at Nvidia.

are in correspondence with each other, from different synthesized instances.

On the other hand, dense correspondence is established between local semantically-similar regions, but with varying appearance (e.g., patches of two different eyes). Learning dense correspondence across images of one category remains challenging because labeling large-scale, pixel-level annotations is extremely laborious. While most existing works rely on supervised [7,11,17,39], or unsupervised [47] image classification networks, few have investigated how to learn dense correspondence from GANs.

In this work, we explore learning dense correspondence from GANs. Specifically, we aim to learn an explicit correspondence map, i.e., a pixel-level semantic label map. Since correspondence represents structure (e.g., shapes of facial components) and is independent of texture (e.g., global appearance like skin tone and texture), this task is highly relevant to disentanglement of structure and texture in GANs [1, 28, 33, 41, 45, 50]. Studies show that disentanglement of semantic attributes can be achieved by carefully searching for latent directions learned by GANs [12,41,50], but all attributes being factorized have to be identified by humans. Some recent advances [1, 28] demonstrate effective structure-texture disentanglement by improving the noise code input to GANs [1], or by applying spatial attention in the intermediate layers [28]. However, they either produce a relatively low resolution (e.g., 4×4) structure map [1], or do not produce it explicitly [28].

Our key idea is to introduce a novel coordinate space, from which pixel-level correspondence can be explicitly obtained for all the synthesised images of a category. Inspired by UV maps of 3D meshes [19, 27, 31], where shapes of one category are represented as deformations of one canonical template, in this work, we represent the dense correspondence map of a generated image as a warped coordinate frame transformed from a canonical 2D coordinate map. This enables the representation of a unique structure as a transformation between the warped and the canonical frames. We design a Coordinate GAN (CoordGAN) with structure and texture controlled via two independently sampled noise vectors. While the texture branch controls the global appearance via Adaptive Instance Normalization (AdaIN) [21], in the structure branch, we learn an MLP as the aforementioned transformation. This maps a sampled noise vector to a warped coordinate frame, which is further modulated in the generator to control the structure of the synthesized image in a hierarchical manner.

We adopt several objectives during training to ensure that the network learns accurate dense correspondence, i.e., (1) a texture swapping constraint to ensure the same structure for images with the same structure code but different texture codes; (2) a texture swapping constraint to ensure similar texture for images with the same texture code, but different structure codes. We also introduce a warping loss to further regularize the correspondence maps. In addition, we show that CoordGAN can be flexibly equipped with an encoder that produces dense correspondence maps for real images. We summarize our contributions as follows:

- We introduce a novel coordinate space from which dense correspondence across images of one category can be explicitly extracted. A warping function is introduced to learn this coordinate space.
- We propose CoordGAN, a disentangled GAN that generates dense correspondence maps and high-quality images, via a set of effective objectives.
- CoordGAN can be flexibly equipped with an encoder to produce the correspondence maps for real images. In other words, we also introduce a network (i.e., the encoder) that learns explicit structure representation.
- Experiments show that CoordGAN generates accurate dense correspondence maps and high-quality structure/texture editable images, for various categories.

2. Related Work

Disentangled GANs. Recent studies [12, 41, 50] show that rich semantically meaningful directions (e.g., pose, color, lighting, etc.) automatically emerge in GANs. To factorize these meaningful latent directions, a line of disentangled GANs [4, 5, 33, 34, 42] are proposed to synthesize images via multiple latent factors, where each factor controls a certain attribute, e.g., object shape or texture. Unlike [34, 42, 45] where human annotations (e.g., bounding boxes, surface normals, etc) are required, most related to ours are self-supervised disentanglement approaches [1, 28, 33]. Among them, Alharbi et al. [1] show that injecting hierarchical noise in the first layer of GANs leads to fine-grained spatial content disentanglement. Kwon et al. [28] further inject noise into multiple layers with diagonal spatial attention modules. However, the learned content code only captures coarse structure such as viewpoints, i.e., keeping the same content code and only modifying the texture code would change the subject's shape. In contrast, our method models finer structure that allows for generating images of the same identity with various textures.

Style Transfer. Style transfer [6, 10, 18, 35, 43] synthesizes a novel image by combining the content of one image with the texture of another one. Most related to ours is to swap texture between semantically-related regions of two images. E.g., Park et al. [35] learns a disentangled auto-encoder such that texture of corresponding regions can be swapped. In contrast, our work studies disentanglement of unconditional GANs and extracts dense correspondence between images explicitly.

Dense Correspondence. Identifying dense correspondence has been a challenging problem due to large shape and appearance variances. Most existing approaches are

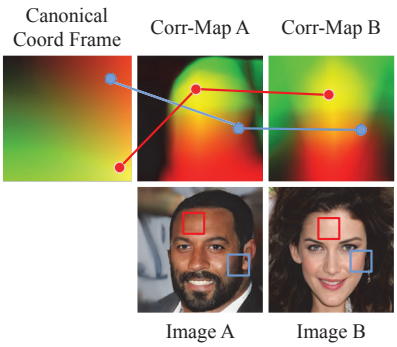

Figure 2. Correspondence in coordinate space. The correspondence maps (Corr-Map) establish dense correspondence between all synthesized images and the canonical coordinate frame.

based on discriminative networks, i.e., either supervised image classification [7, 11, 17, 25, 32, 39], or unsupervised image-level contrastive learning [47, 48]. Our work differs in that we investigate how to extract dense correspondence from GANs. Recently, several works [49, 52] show that semantics can be extracted from GANs via a linear classifier in a few-shot setting. However, these methods still require manual annotations for training the classifier. Inspired by these works, we move one step further to extract dense correspondence without using any annotated labels.

Concurrent Work. Peebles et al. [36] achieves visual alignment through equipping a pre-trained StyleGAN2 [23] with additional Spatial Transformer Network (STN) [16]. However, dense correspondence is only identified for part of the object. Differently, through disentanglement of structure and texture, the proposed CoordGAN automatically generates correspondence maps of full images and neither pretrained StyleGAN nor additional STN is required.

3. Dense Correspondence from CoordGAN

We design a structure-texture disentangled GAN such that dense correspondence can be extracted explicitly from the structural component, where the key component is to tie image structure to a coordinate space that is shared by all images. Specifically, the structure of each generated image is represented as a warped coordinate frame, transformed from a shared canonical 2D coordinate frame. This reduces finding correspondence between image pixels to locating the coordinates in corresponding warped coordinate frames, which are transformed from the same coordinates in the canonical frame. We call our model Coordinate GAN (CoordGAN).

Coordinate Map Representation. We define C as a 2D coordinate map of width W^c and height H^c. When $C(i,j) = (i,j)$, this denotes the *canonical coordinate map* (see Figure 2). Pixel locations and coordinates are normalized to the range $[-1,1]$. For example, $C(1,1) = (1,1)$ indicates the bottom right pixel of the coordinate map is of coordinate $(1,1)$. It is then possible to define a *warping function* $\mathcal{W} : (C, w) \rightarrow C^w$, parameterized by a code $w \in \mathbb{R}^N$, that maps C into a warped coordinate map C^w. Since the code w relates the pixel coordinates of the image to the canonical coordinate map, it can be seen as the representation of image structure. In particular, $C^w(i,j) = (k,l)$ implies that the pixel i,j of the image is in correspondence with the canonical coordinate k,l. Given the two images with codes w_1 and w_2, it is also possible to establish correspondence between them by seeking pixels of similar coordinates. Given pixel (i,j) of the image associated with coordinate $C^{w_1}(i,j)$, the corresponding pixel in another image of coordinate map C^{w_2} is,

$$T_{1,2}(i,j) = \arg\min_{p,q} ||C^{w_1}(i,j) - C^{w_2}(p,q)||^2, \quad (1)$$

where $T_{1,2}$ defines the forward transformation from warped coordinates C^{w_1} to C^{w_2}. In this way, a generative model for images that includes a warping function automatically establishes dense correspondence between all synthesized images, as shown in Figure 2. This can be useful for transferring properties between the images, such as semantic labels, landmark locations, image pixels, etc.

3.1. Overview

An overview of the proposed CoordGAN is presented in Figure 3. The CoordGAN is a generative model based on the structural coordinate map representation. The inputs to our model include two latent code vectors with dimension N: a structure code $z_s \in \mathcal{R}^N$ for modeling layouts and object structure, and a texture code $z_t \in \mathcal{R}^N$ for modeling texture, lighting, etc. The CoordGAN generator $G(z_s, z_t; \theta_{_G})$ is a mapping from these codes to the image space, with parameters $\theta_{_G}$. This is implemented by a combination of structure and texture mappings. A structure mapping network $w_s = \mathcal{S}(z_s; \theta_{\mathcal{S}})$ of parameters $\theta_{\mathcal{S}}$ maps the structure noise variable z_s into a structure code w_s, which is then used by a warping function $\mathcal{W}(C, w_s)$ to produce a warped coordinate map C^{w_s} for the image. A texture mapping network $w_t = \mathcal{T}(z_t, \theta_{\mathcal{T}})$ of parameters $\theta_{\mathcal{T}}$ maps the texture noise variable z_t into a texture code w_s. The modulated generator then produces an image with the mapping $\mathcal{A}$ parameterized by $\theta_{\mathcal{A}}$,

$$G(z_s, z_t; \theta_{_G}) = \mathcal{A}(C^{w_s}, w_t; \theta_{\mathcal{A}}), \quad (2)$$

where $\theta_{_G}$ includes $\theta_{\mathcal{S}}, \theta_{\mathcal{T}}$, and $\theta_{\mathcal{A}}$. The details of the various modules are discussed in the following sections.

3.2. Coordinate Warping Network

One major component in CoordGAN is the warping function. We propose a Coordinate Warping Network, which learns a transformation between the canonical and a warped coordinate frame, conditioned on a latent structure code w_s. While there exist several differentiable transformation functions, such as Thin Plate Splines (TPS), Spatial Transformation Network (STN) [16], and affinity ma-

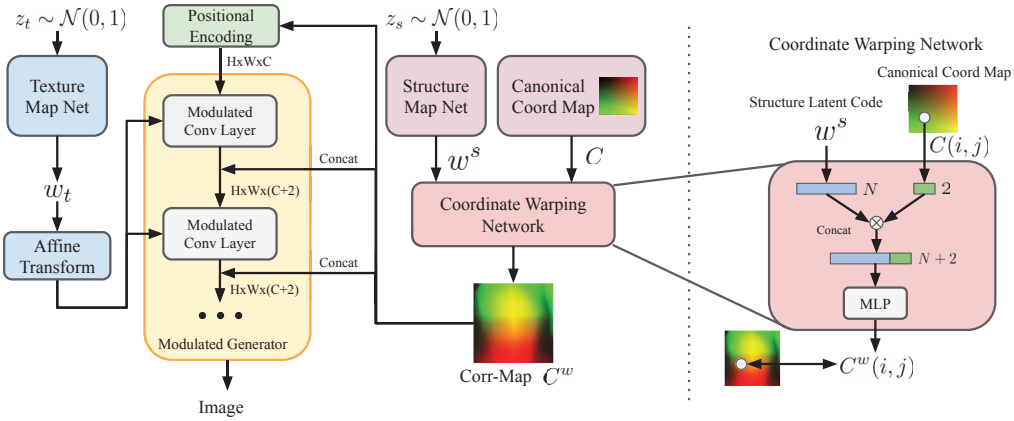

Figure 3. Overview of CoordGAN. CoordGAN mainly consists of a texture mapping network, a structure mapping network, a coordinate warping network, and a modulated generator. The coordinate warping network (on the right) takes the structure latent code and a canonical coordinate map and outputs a correspondence map, which is then fed into multiple layers of the modulated generator to synthesize images.

trix [30, 44], in CoordGAN this transformation is implemented with a MLP as

$$C^{w_s}(i,j) = \mathcal{W}(C(i,j), w_s) = \mathcal{P}([C(i,j), w_s], \theta_{\mathcal{P}}) \quad \forall i,j \quad (3)$$

where $\mathcal{P}$ is a three layer MLP of parameters $\theta_{\mathcal{P}}$ and $[C(i,j), w_s] \in \mathcal{R}^{N+2}$ is the concatenation of coordinate i,j from the canonical frame with the structure latent code w_s. In the supplementary materials, we show that the MLP is a learnable, conditional geometric transformation between the canonical coordinate frame and a warped coordinate frame.

The advantages of learning the transformation via a MLP are two folds. First, since an MLP is a continues function containing only linear projection layers and ReLUs, it preserves the order of the coordinates in the canonical coordinate frame, i.e., it ensures that the warping is diffeomorphic. Second, compared to TPS and STN, our design of $\mathcal{W}$ is generic and allows for more flexible deformation.

3.3. Incorporating Warping in CoordGAN

We introduce the rest of CoordGAN components. While our generator design is inspired by StyleGAN [21] (see Figure. 3), we discuss the major differences in the following.

Positional Encoding. Rather than inputting dense correspondence map directly to the generator, we map it via a positional encoding layer [2]. I.e., a Fourier embedding is obtained by the application of a 1×1 convolution followed by a sine function. The Fourier embedding serves as the first layer of the generator.

Mapping Networks $\mathcal{S}(\cdot)$ and $\mathcal{T}(\cdot)$. We use the same architecture as StyleGAN for the mapping network. Different from StyleGAN, we apply two independent mapping networks responsible for structure and texture, respectively.

Modulated Generator $\mathcal{A}(\cdot)$. We replace the learnable constant input of StyleGAN with the correspondence map. Since the latter has high resolution (i.e., 128×128), instead of gradually increasing spatial resolution, the spatial resolution is kept the same as the input Fourier embedding at all layers as shown in Figure 3. We inject the latent texture code w_t into different layers of the modulated generator, via weight modulation [22], to render appearance details at different levels. To balance the structure and texture inputs at an architectural level, the dense correspondence map is also concatenated with the features produced by multiple intermediate layers of the modulated generator. We found that, without this multi-layer modulation of dense correspondence map, the coordinate warping network can only learn coarse and inaccurate structure information (e.g., viewpoints of faces), as shown in Table 3.

3.4. Learning Objectives

To learn accurate correspondence maps and encourage the disentanglement of the latent space, such that z_s and z_t encode the image structure and texture separately, CoordGAN is trained with the following objectives.

Texture Swapping Constraint. To ensure the CoordGAN generates the same identity and image layout when the structure is fixed and only the texture code is modified, a texture swapping constraint is applied. Given a pair of synthesized images with a shared structure code z_s and different texture codes z_{t_1}, z_{t_2}, the texture swapping loss $\mathcal{L}_t$ is defined as the LPIPS [51] loss between the two synthesized images:

$$\mathcal{L}_t = \mathcal{L}_{LPIPS}(G(z_s, z_{t_1}; \theta_G), G(z_s, z_{t_2}; \theta_G)). \quad (4)$$

Structure Swapping Constraint. To encourage images that share the same texture code to have similarly looking textures, a structure swapping constraint is introduced. This consists of encouraging two images with the same texture code z_t but different structure codes z_{s_1} and z_{s_2} to have similar textures. Following [35], this is done with

a non-saturating GAN loss based on a patch discriminator D_{patch}:

$$\mathcal{L}_s = \mathbb{E}\Big[-\log\Big(D_{patch}\big(G(z_{s_1}, z_t; \theta_{_G}), G(z_{s_2}, z_t; \theta_{_G})\big)\Big)\Big]. \tag{5}$$

Warping Loss. A warping loss is defined to explicitly regularize the correspondence map. Given a pair of synthesized images $x_{_1} = G(z_{s_1}, z_{t_1}; \theta_{_G})$ and $x_{_2} = G(z_{s_2}, z_{t_2}; \theta_{_G})$, $x_{_1}$ is warped to the coordinate frame of $x_{_2}$ by transferring pixel colors according to Equation (1). In practice, similar to [30, 44, 46], we relax Equation (1) with affinity matrix to make the warping differentiable. This produces a warped image $x^w_{2,1}$. A warping loss based on the LPIPS loss [51],

$$\mathcal{L}_{warp} = \mathcal{L}_{LPIPS}(x^w_{2,1}, x_2), \tag{6}$$

is used to minimize the distance between $x^w_{2,1}$ and $x_{_2}$.

Chamfer Loss. Suppose a canonical coordinate map C is transformed to a warped coordinate map C^w, a Chamfer loss is implemented to avoid the collapse of the transformation,

$$\begin{aligned}\mathcal{L}_{cham} =& \frac{1}{|C|}\sum_{(i,j)\in C}\min_{(p,q)}||C(i,j) - C^w(p,q)||_2 \\ &+ \frac{1}{|C^w|}\sum_{(p,q)\in C^w}\min_{(i,j)}||C^w(p,q) - C(i,j)||_2.\end{aligned} \tag{7}$$

Overall Learning Objective. To generate realistic images, a standard GAN objective function $\mathcal{L}_{GAN}$ is applied to the synthesized images. Combining all the aforementioned loss objectives, the overall training objective is defined as

$$\begin{aligned}\mathcal{L}_G =& \lambda_t * \mathcal{L}_t + \lambda_s * \mathcal{L}_s + \lambda_{warp} * \mathcal{L}_{warp} \\ &+ \lambda_{cham} * \mathcal{L}_{cham} + \lambda_{GAN} * \mathcal{L}_{GAN},\end{aligned} \tag{8}$$

where $\lambda_t, \lambda_s, \lambda_{warp}, \lambda_{cham}, \lambda_{GAN}$ are coefficients used to balance the different losses.

3.5. Inverting CoordGAN via an Encoder

The CoordGAN can be equipped with an encoder to enable the extraction of dense correspondence from real images. Specifically, an encoder $E(\cdot; \theta_{_E})$ parameterized by $\theta_{_E}$ is introduced to map an image x to a pair of structure $w_{s,E}$ and texture $w_{t,E}$ latent codes. These latent codes are then input to the CoordGAN to synthesize a replica of the image. As observed in [38], embedding real images directly into $W+$ space rather than W space leads to better reconstruction. So for the texture branch, we design the encoder to output texture latent codes $w^+_{t,E}$ in $W+$ space as opposed to $w_{t,E}$ in W space. During training, we fix the generator while optimizing the encoder via latent consistency, reconstruction and texture swapping losses, which are described as follows.

Latent Consistency Loss. We introduce a latent consistency loss by feeding synthesized images back to the encoder and matching the distribution of encoder outputs to that originally produced by the mapping network. Suppose an image is synthesized with latent codes w_t, w_s, and correspondence map C^w. Inputting this image back into the encoder produces a pair of latent codes $w^+_{t,E}$ and $w_{s,E}$, and the correspondence map C^w_E. The latent consistency loss $\mathcal{L}_{con}$ is defined as

$$\mathcal{L}_{con} = \mathcal{L}_2(w_s, w_{s,E}) + \mathcal{L}_2(C^w, C^w_E), \tag{9}$$

where $\mathcal{L}_2(\cdot,\cdot)$ denotes the $\mathcal{L}_2$ loss.

Reconstruction Loss. This is a reconstruction loss for input real images, with L1 ($\mathcal{L}_1$) and LPIPS [51] ($\mathcal{L}_{LPIPS}$) components, defined as

$$\mathcal{L}_{rec} = \mathcal{L}_1(x, G(E(x))) + \mathcal{L}_{LPIPS}(x, G(E(x))), \tag{10}$$

Overall Learning Objective. The overall learning objective used for encoder training an encoder is

$$\mathcal{L}_E = \lambda_{con} * \mathcal{L}_{con} + \lambda_{rec} * \mathcal{L}_{rec} + \lambda_t * \mathcal{L}_t, \tag{11}$$

where $\lambda_{con}, \lambda_{rec}, \lambda_t$ are hyperparameters that balance the different losses.

We note that the encoder facilitates explicit structure representation learning for real images. It is significantly more efficient than optimization-based GAN-inversion methods, as no iterative inference is required.

4. Experiments

In this section, we show quantitative and qualitative results of models trained on the CelebAMask-HQ [29], Stanford Cars [26], and AFHQ-Cat [6] datasets. We train separate models on each dataset, using a resolution of 512×512 for the CelebAMask-HQ model and 128×128 for the other two. For CelebAMask-HQ, we first train CoordGAN with an output size of 128×128 and then append two upsampling layers to generate high-resolution images. Detailed network design and training hyper-parameters are described in the supplementary.

4.1. Evaluation on Dense Correspondence

We quantitatively demonstrate the quality of the extracted dense correspondence on the task of semantic label propagation. Given one reference image with semantic labels, its correspondence map is first inferred with the trained encoder. This establishes a mapping between the semantic labels and the correspondence map for that image. Another correspondence map is then inferred for a query image and the labels of the reference image are obtained with Equation (1). To align with the training stage, we relax Equation (1) with affinity matrix in practice.

Datasets and Metrics. We evaluate different methods on the CelebAMask-HQ [29] and DatasetGAN [52] datasets. We merge CelebAMask-HQ dataset labels and select 6 classes (eyes, nose, ear, mouth, face and eyebrow)

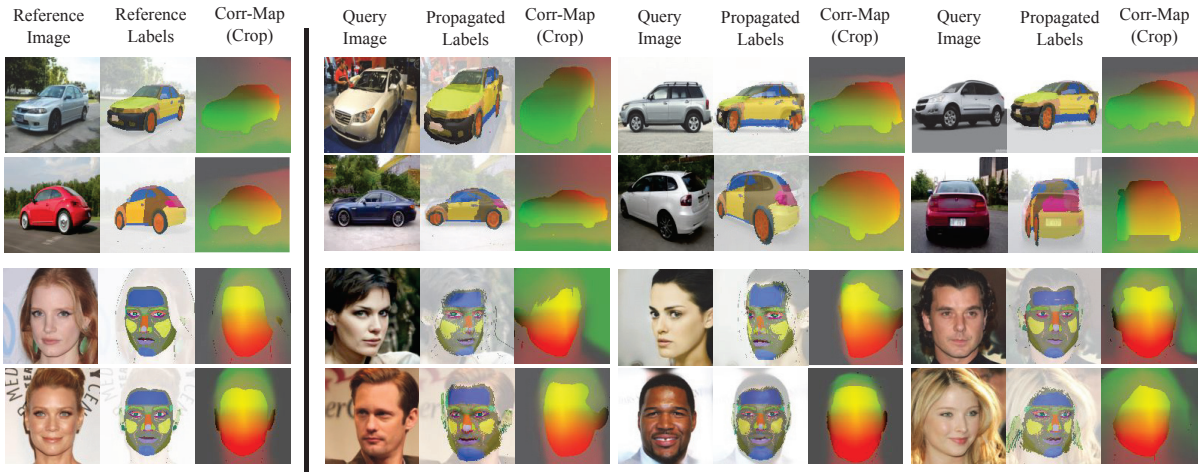

Figure 4. Qualitative results for semantic label propagation. In each row, given one reference image along with its semantic labels as shown on the left, the proposed approach predicts its correspondence map and propagates its segmentation mask to other query images on the right. For better visualization, we use the ground-truth masks to highlight the foreground areas of all the predicted correspondence maps, denoted with Corr-Map (Crop). Note that no ground-truth masks are used for actual label propagation.

for our evaluation. The DatasetGAN dataset contains detailed manually annotated labels for faces (34 classes) and cars (20 classes). For the DatasetGAN faces, we excluded neck and hair since they are not consistently visible for all images in the dataset. For all datasets, we randomly select 5 images as reference and another set as query images. Each reference image's semantic label is propagated to all query images and the mean intersection-over-union (IOU) with the ground-truth segmentation maps is computed for evaluation. We report the averaged score of these 5 runs.

Baselines. For all baseline models, we extract features from hidden layers and use nearest neighbor search to determine feature correspondences and propagate labels. We detail the features selected for label propagation below. We employ two sets of baselines. The first set comprises of transfer learning based methods with either supervised ImageNet pre-training, e.g., ResNet50 [14] or self-supervised contrastive learning based pre-training, e.g., MoCo [13] pre-trained on ImageNet [8] and VFS [48] pre-trained on Kinetics video dataset [24]. For all these methods, ResNet50 [14] is employed as the backbone and the pre-trained models are directly tested on our task without fine-tuning. We follow [47, 48] and use the Res-block 4 features for label propagation as it is shown that Res-block 4 gives the best pixel-level correspondences. Another set of baselines is based on auto-encoders, such as Swapping Auto-encoder [35] and Pix2Style2Pix [38]. Both methods are trained on the same datasets as ours. For Swapping Auto-encoder, the structure branch features are used for label propagation. For Pix2Style2Pix encoder, the Res-block 4 features are used for label propagation. All methods are evaluated with input image resolution of 128, except for Pix2Style2Pix where the input image size is set to 256 following the original paper.

	CelebA-HQ	DGAN-face	DGAN-car
Resnet50 [14]	39.48	11.05	11.07
Moco [13]	36.19	10.00	9.53
VFS [48]	38.10	8.55	6.88
Swap AE [35]	24.73	5.48	5.37
Pix2Style2Pix [38]	48.50	20.36	10.77
CoordGAN	**52.25**	**23.78**	**13.23**

Table 1. IOU comparison for label propagation. Our method shows the best semantic label propagation results among all baseline methods.

Quantitative Results. As reported in Table 1, the proposed CoordGAN outperforms all self-supervised baselines across all three datasets on the task of semantic segmentation label propagation. The most related approach is Pix2Style2Pix, which also learns an encoder for a pre-trained StyleGAN2 model. While Pix2Style2Pix encoder features contain both structure and texture information, CoordGAN correspondence maps, with only structure information, still achieve better label propagation performance. These results suggest that CoordGAN learns much accurate correspondence than the other methods.

Qualitative Results. We visualize both the coordinate maps and the propagated segmentation labels in Figure 4. On the left, several reference images from the DatasetGAN dataset are shown along with their semantic labels. On the right, we show the propagation results for different query test images. The predicted correspondence maps for both the reference and query images are color-coded and masked with the foreground ground-truth semantic labels for better visualization. Note that this is only for visualization, no ground-truth masks are used for the actual label propagation. Note that our method produces precise label propagation results for both frontal and profile query faces. For

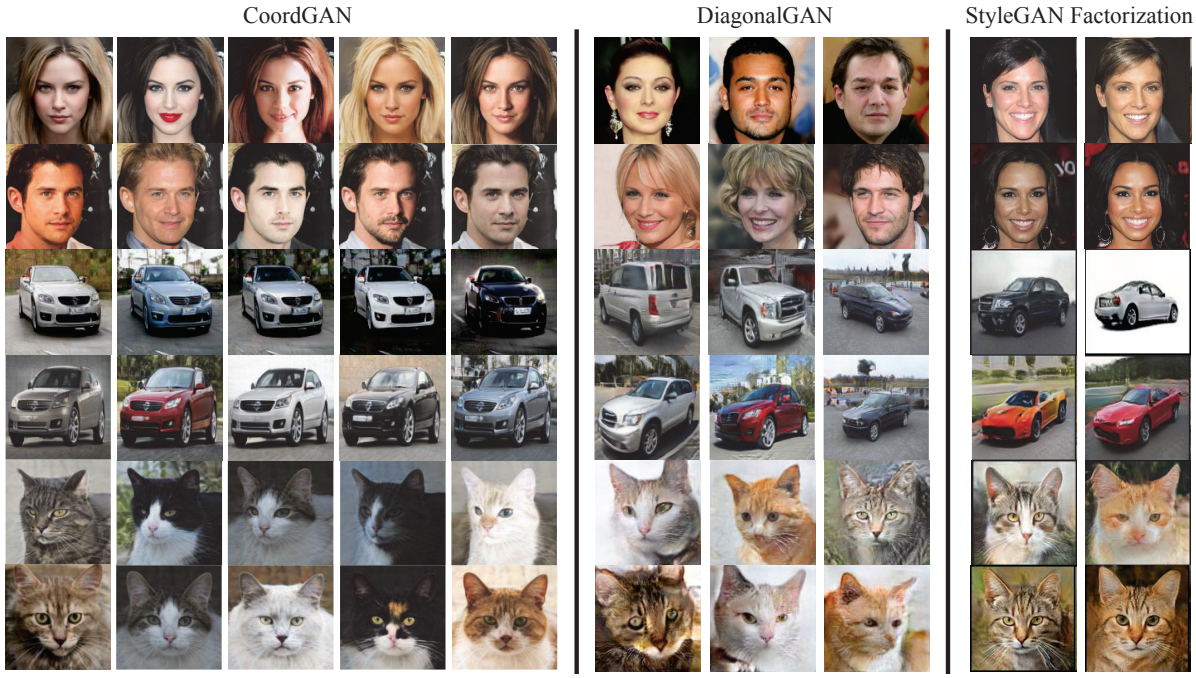

Figure 5. Qualitative comparison for texture swapping. From top to bottom: models trained on CelebAMask-HQ, Stanford Cars, and AFHQ-cat datasets. For CoordGAN and DiagonalGAN, images shown in each row are generated with the same structure code and diverse texture codes. For GAN Factorization, images in each row are generated with random perturbations along the identified eigen-vector directions. It is apparent that CoordGAN preserves structure better when only texture codes are modified.

	CelebA-HQ			Stanford Cars		AFHQ-cat	
	LPIPS ↓	Arcface ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓
StyleGAN2 [23]	-	-	8.21	-	16.20	-	21.02
DiagonalGAN [28]	0.58	0.79	11.16	0.61	18.09	0.55	17.63
CoordGAN	**0.22**	**0.38**	16.16	**0.21**	24.27	**0.27**	23.62

Table 2. Texture swapping comparison. The lowest LPIPS and Arcface feature distances of CoordGAN suggest better structure preservation when the texture code is varied.

cars, this is even more challenging, considering the large differences in viewpoints and scales. For example, in extreme cases where the reference car is viewed from the front and the query car from the back, no correspondence exist. Surprisingly, even in cases where the reference car is observed from the side and the query car from the rear, CoordGAN still matches the labels reasonably well. We conjecture this is because it learns a reasonable prior for the category, by observing many instances and densely associating them during training.

4.2. Identity-preserving Texture Swapping

We analyze disentanglement of structure and texture of CoordGAN by generating images with the same structure code but different texture codes (i.e., texture swapping) and evaluating the structural consistency of the outputs. We focus on the generator and do not use an encoder in these experiments.

Metrics. To quantitatively examine different methods, we use the ArcFace [9] face identity loss and the LPIPS [51] loss to evaluate disentanglement and structure preservation performance, and FID [15] score for measuring the perceptual image quality of the generated images. ArcFace computes a feature-level cosine similarity loss between two faces. It can be used to measure whether the face identity is preserved since the smaller the loss is, the more likely both images capture the same identity. LPIPS [51] measures whether two images have similar image layouts.

Baselines. CoordGAN is compared against two baselines: DiagonalGAN [28] and GAN Factorization [41]. DiagonalGAN achieves state-of-the-art performance for StyleGAN-based structure and texture disentanglement. Similar to CoordGAN, it uses separate structure and texture codes as inputs. To generate texture-swapped images, we sample a structure code and different texture codes, and then compute the structural similarity among the images synthesized using the aforementioned metrics. GAN Factorization exploits SVD to identify semantically meaningful latent directions across different GAN layers. The paper suggests that the final layers of the GAN are mainly responsible for controlling texture. Therefore, we generate texture-swapped images with GAN Factorization by adding perturbations along the computed eigen-vectors of the last two convolution layers of a pre-trained StyleGAN2.

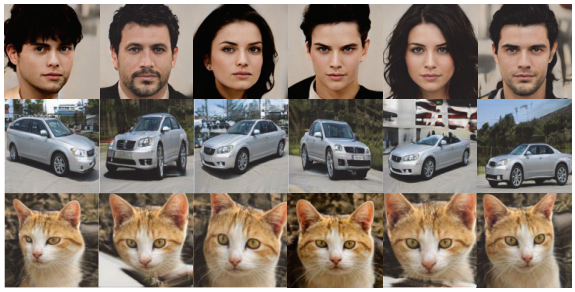

Figure 6. Qualitative results for structure swapping. Images shown in each row are generated with the same texture code and diverse structure codes.

Results. As shown in Table 2, CoordGAN outperforms the baselines by a significant margin for all disentanglement metrics (ArcFace and LPIPS) on all object categories. This suggests that it successfully preserves the fine-grained image structure independent of the input texture. Note that ArcFace is only available for human faces. The FID score is computed over 10,000 generated images for all methods, for reference. Note that, as discussed in [1, 28], a slight decrease in the FID score is observed due to the strong disentanglement constraints enforced.

In Figure 5, each row shows diverse texture-swapped images generated by fixing the structure code and varying the texture code. The DiagonalGAN changes the subject's identity completely. This becomes more clear when testing on cars, where the viewpoint is ambiguous and scale can vary. Results suggest that its disentangled content code only captures coarse structural information, such as the rough image layout and viewpoint. In contrast, CoordGAN successfully maintains both the coarse and fine-grained image structure and only varies appearances, on all datasets. For GAN factorization, while potentially possible to exhaustively search for the latent eigen-vectors that only modify image textures, it is not easy to finely control the appearance of the synthesized images.

4.3. Structure Swapping

To further demonstrate CoordGAN successfully disentangles structure and texture, in this section, we synthesize images of the same texture code and various structure codes (i.e., structure swapping). As show in Figure 6, from top to bottom, we show synthesized images of models trained separately on CelebAMask-HQ, Stanford Cars, and AFHQ-cat datasets. It is clear that images in each row show similar textures (e.g., hair/face colors for humans, grey cars, orange cats) with diverse structural variations (e.g., viewpoint, scale, shape, layout, etc). The again confirms that CoordGAN learns a disentangled representation where the structure code and the texture code capture different attributes of a synthesized image. More visualizations are included in the supplementary materials.

	Disentanglement		Correspondence	
	LPIPS↓	Arcface ↓	CelebA-HQ	DGAN-face
CoordGAN	0.10	0.32	52.25	23.78
w/o struc-mod	0.32	0.73	48.59	20.01

Table 3. Ablation on structure modulation. We show that incorporating the structure modulation is essential to a good disentanglement and correspondence performance (measured by IOU).

4.4. Ablation Studies

We ablate different architectures w.r.t the structure branch, i.e., feeding the correspondence map (1) only to the first layer of CoordGAN (w/o struc-mod), or (2) to modulate multiple layers, as discussed in Section 3.3. Both models are trained to synthesize images of resolution 128×128 on the CelebAMask-HQ dataset. Table 3 shows that the proposed structure modulation design is crucial to achieve a good disentanglement of structure and texture. This confirms that a non-trivial architecture design is needed to embed the structure information and highlights the importance of the proposed balanced structure and texture modulation. More studies on objectives are included in the supplementary materials.

5. Discussion

Conclusion. In this work, we show that it is possible to train GANs so that dense correspondence can automatically emerge. We propose a novel disentangled GAN model, CoordGAN, that produces dense correspondence maps represented by a novel coordinate space. This is complemented by an encoder for GAN inversion, which enables the generation of dense correspondence for real images. Experimental results show that CoordGAN generates accurate dense correspondence maps for a variety of categories. This opens up a new door for learning dense correspondences from generative models in an unsupervised manner. We qualitatively and quantitatively demonstrate that CoordGAN successfully disentangles the structure and texture on multiple benchmark datasets.

Limitations and Future Work. The current proposed model is restricted to learn correspondence within the same category, since it requires the coordinate maps transformed from the same canonical space. While we can potentially infer the 3D viewpoints from the coordinate map (as visualized in Figure 4), we have not explicitly modelled the 3D structure in our representation. A future extension of this work can be to learn a 3D UV coordinate map instead of a 3D map to represent the underlying structure.

Acknowledgements. This work was supported, in part, by grants from NSF IIS-1924937, NSF IIS-2041009, NSF CCF-2112665 (TILOS), and gifts from Qualcomm and Amazon.

References

[1] Yazeed Alharbi and Peter Wonka. Disentangled image generation through structured noise injection. In *CVPR*, pages 5133–5141, 2020. 2, 8

[2] Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In *CVPR*, pages 14278–14287, 2021. 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[4] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 2

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, pages 2172–2180, 2016. 2

[6] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 2, 5

[7] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *arXiv preprint arXiv:1606.03558*, 2016. 2, 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 7

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 2

[11] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *ICCV*, pages 1831–1840, 2017. 2, 3

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020. 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 7

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 28:2017–2025, 2015. 3

[17] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *ECCV*, pages 351–366, 2018. 2, 3

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 2

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, pages 371–386, 2018. 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1, 2, 4

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 1, 4

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116, 2020. 3, 7

[24] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 6

[25] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. FCSS: fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, pages 616–625, 2017. 3

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, Sydney, Australia, 2013. 5

[27] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *ICCV*, pages 2202–2211, 2019. 2

[28] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation. In *ICCV*, pages 13980–13989, 2021. 2, 7, 8

[29] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5

[30] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *arXiv preprint arXiv:1909.11895*, 2019. 4, 5

[31] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, pages 677–693. Springer, 2020. 2

[32] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, pages 4462–4471, 2020. 3

[33] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, pages 7587–7596, 2019. 2

[34] Utkarsh Ojha, Krishna Kumar Singh, and Yong Jae Lee. Generating furry cars: Disentangling object shape and appearance across multiple domains. In *ICLR*, 2021. 2

[35] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*, 2020. 2, 4, 6

[36] William S. Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A. Efros, and Eli Shechtman. Gan-supervised dense visual alignment. *CoRR*, abs/2112.05143, 2021. 3

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[38] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021. 5, 6

[39] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017. 2, 3

[40] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9243–9252, 2020. 1

[41] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, pages 1532–1540, 2021. 1, 2, 7

[42] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, pages 6490–6499, 2019. 2

[43] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016. 2

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 4, 5

[45] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 318–335, 2016. 2

[46] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pages 2566–2576, 2019. 5

[47] Taihong Xiao, Sifei Liu, Shalini De Mello, Zhiding Yu, Jan Kautz, and Ming-Hsuan Yang. Learning contrastive representation for semantic correspondence. *arXiv preprint arXiv:2109.10967*, 2021. 2, 3, 6

[48] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. *arXiv preprint arXiv:2103.17263*, 2021. 3, 6

[49] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *CVPR*, pages 9351–9360, 2021. 3

[50] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *IJCV*, pages 1451–1466, 2021. 2

[51] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 4, 5, 7

[52] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, pages 10145–10155, 2021. 3, 5