POP: Mining POtential Performance of new fashion products via webly cross-modal query expansion

Christian Joppi*[1][0000−0003−4495−9515], Geri Skenderi*[2][0000−0001−9968−7727], and Marco Cristani[2,1][0000−0002−0523−6042]

[1] Humatics s.r.l
{name.surname}@sys-datgroup.com
[2] University of Verona
{name.surname}@univr.it

Abstract. We propose a data-centric pipeline able to generate exogenous observation data for the New Fashion Product Performance Forecasting (NFPPF) problem, i.e., predicting the performance of a brand-new clothing probe with no available past observations. Our pipeline manufactures the missing past starting from a single, available image of the clothing probe. It starts by expanding textual tags associated with the image, querying related fashionable or unfashionable images uploaded on the web at a specific time in the past. A binary classifier is robustly trained on these web images by confident learning, to learn what was fashionable in the past and how much the probe image conforms to this notion of fashionability. This compliance produces the POtential Performance (POP) time series, indicating how performing the probe could have been if it were available earlier. POP proves to be highly predictive for the probe's future performance, ameliorating the sales forecasts of all state-of-the-art models on the recent VISUELLE fast-fashion dataset. We also show that POP reflects the ground-truth popularity of new styles (ensembles of clothing items) on the Fashion Forward benchmark, demonstrating that our webly-learned signal is a truthful expression of popularity, accessible by everyone and generalizable to any time of analysis. Forecasting code, data and the POP time series are available at: https://github.com/HumaticsLAB/POP-Mining-POtential-Performance

Keywords: Computer Vision for Fashion, Data-centric Artificial Intelligence, Time Series Forecasting

1 Introduction

Forecasting the performance of a new clothing item is a crucial challenge for fashion companies [7,11]. A good forecast in terms of predicted sales or product popularity can greatly help optimize the supply chain [30] and minimize losses on

* indicates equal contribution.

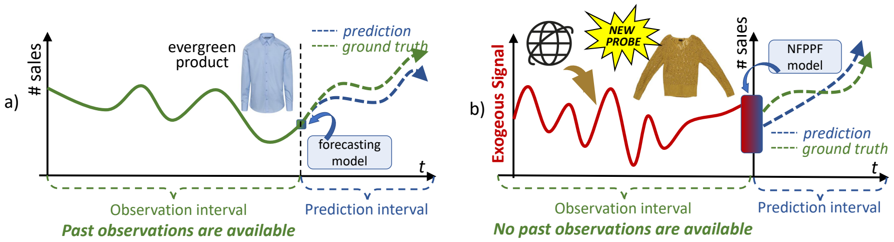

Fig. 1: a) A standard forecasting setup, where an evergreen item has past observations to exploit, *e.g.,* # sales; b) New Fashion Product Performance Forecasting (NFPPF) problem, where no past observations are available and exogenous data must be considered. Here we propose POP, the POtential Performance series, which is webly learned. The signals in b) appear on the same scale purely for visualization purposes, in reality this might not be the case.

multiple levels. Unfortunately, standard forecasting approaches require observations from the past to provide a forecast for the same product in the future [14,1] and this information is typically available for evergreen products only (Fig. 1a). In other cases, judgemental forecasts [14] from fashion professionals[3] are the only ones that can help. Starting from photos or realistic renderings, which we call *probe* images, they perform comparisons with trends as they surface and then infer the probe's success [30]. In this paper, we try to model this line of reasoning and create a data-centric [25] pipeline that is able to extract a highly predictive signal from the web, dubbed "POtential Performance" (POP), which can be fed into any NFPPF forecasting model as an additional variable and lead to more accurate forecasts. (see Fig. 1b).

Our cross-modal, query expansion based pipeline is sketched in Fig. 2. The input is a single probe image of the product to be analyzed, or a photorealistic rendering[4]. The pipeline first extracts textual tags from the probe automatically or by directly considering the associated technical sheet. The tag set is expanded with *positive* and *negative* tags that are used to perform a *time-dependent* query online, i.e., collecting images of "fashionable" and "unfashionable" items related to the tags, which have been uploaded during some specified K_{past} intervals in the past. These images are used to *confidently learn* [27] a binary classifier that captures what is fashionable VS unfashionable in that interval. This learning procedure prunes noisy images from both the positive and negative classes, resulting in a robust model. Subsequently, pruned positive images are projected into an embedding space by the learned model and compared with the (also projected) initial probe image, providing the K_{past}-long POP signal. The POP

[3] A commercial example is Trendstop https://www.trendstop.com/ and its "Trend Platform Membership" service

[4] Several such tools are available, for instance https://www.tg3ds.com/3d-fashion-design-tools.

signal indicates how popular the probe could have been over time if it were available earlier in the past.

Our approach should be cast in the field of *data-centric artificial intelligence* (DCAI) [25], since it automates the creation of high quality training data that can be used to improve any forecasting model which accommodates multivariate time series forecasting. POP has been tested on diverse state-of-the-art NFPPF algorithms that predict sales curves of new products on the recent VISUELLE fast-fashion dataset [34], providing superior performance when compared to other types of training signals. It has also been customized to deal with fashion styles (*i.e.,* ensembles of clothing items) on the Fashion Forward benchmark [1]. Fashion Forward (FF) calculates a popularity time series for an automatically extracted style based on the dataset properties and then applies standard forecasting algorithms. We substitute their popularity series with POP, reaching similar predictions despite relying only on an exogenous input. Surprisingly, on the Dresses partition of FF, we reach the absolute best, suggesting that POP can foresee the success of a potentially new fashion style. Summarizing, the contributions of this work are threefold:

1. The first data-centric strategy tailored to forecasting, used to create an exogenous observation signal which improves forecasts of the performance (number of sold items, popularity) of brand new clothing items with non-existent pasts.
2. A webly-learned method to freely collect information about fashion trends without relying on private or costly repositories.
3. Best overall results on all the tested NFPPF tasks.

The rest of the paper is organized as follows: related literature is analyzed in Sec. 2; the proposed approach is detailed in Sec. 3; experiments are reported in Sec. 4, and finally; concluding remarks are drawn in Sec. 5.

2 Related literature

NFPPF problem. The NFPPF problem has been deeply investigated in the fields of quantitative fashion design [29,3,16], marketing and social sciences [32,12], but is relatively new in the computer vision community. In both [9,33], the main idea is that new products will sell comparably to similar, older products; this similarity is exploited in [33] via textual tags only, while in [9] an autoregressive RNN model takes past sales, textual product attributes, and the product image as input, to forecast the item sales. The work in [34] focuses on the additional direction of checking the past to look for predictive exogenous signals. In particular, the authors exploit Google Trends, querying textual attributes related to the probe and embed the resulting trend into a Transformer-based [37] architecture, which considers images, text and other metadata. The authors also rendered accessible the first publicly available dataset for NFPPF, VISUELLE. In our paper we follow the idea of looking back to web data, but use images as the main representation of online fashionability, obtaining a richer exogenous

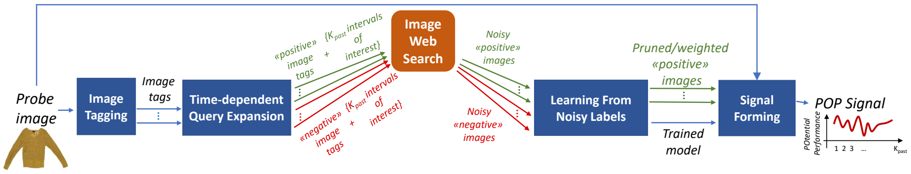
Probe image
Image Tagging
Image tags
Time-dependent Query Expansion
«positive» image tags + {K_past intervals of interest}
«negative» image tags + {K_past intervals of interest}
Image Web Search
Noisy «positive» images
Noisy «negative» images
Learning From Noisy Labels
Pruned/weighted «positive» images
Trained model
Signal Forming
POP Signal
POtential Performance
1 2 3 ... K_past

Fig. 2: Schematic pipeline of our approach; we start with a probe image and obtain the POtential Performance (POP) signal at the end. Along this pipeline, we sequentially process information in different modalities, thereby creating a *cross-modal signal*.

signal. Predicting the success of new fashion styles has never been taken into account, with past works [1,20,22] focusing on the standard forecasting setup.

Data-centric AI. Data-Centric AI [25] (DCAI) shifts the attention from the models to the data used to train and evaluate them. It is a topic whose importance is constantly growing in many AI communities [2,28,26][5], with important effects on CV & ML. In general, DCAI investigates methodologies for accelerating open-source dataset creation from lower-quality resources. Consequently, it is tightly coupled with learning on noisy data, which aims at producing consistent and low noise data samples, or removing labeling noise and inconsistencies from existing data [27,35,38]. Our methodology is data-centric, since it automates the creation of training data from a large amount of web resources, while removing labeling noise. Notably, it represents a novelty in the DCAI panorama, since it creates *temporally-dependent* training data, i.e., time series, as it is required by NFPPF and in general by forecasting tasks.

3 Methodology

The goal of our approach is to produce an exogenous variable that can aid a forecasting model in predicting the future performance of a product (sales, popularity). The input to our approach is the probe image $\mathbf{z}^{(t)}$, where $\mathbf{z}$ represents the new clothing item and t the *observation time*, which is the date from when we begin to look into the past. The output is the POP signal $S_{\mathbf{z}}^{(t)} = s_{\mathbf{z}}^{(t-K_{past})}, \dots, s_{\mathbf{z}}^{(t-k)}, \dots, s_{\mathbf{z}}^{(t-1)}$, defined for K_{past} time steps preceding t, where $k = 1, \dots, K_{past}$ and $s_{\mathbf{z}}^{(t-k)} \in \mathbb{R}$. In this paper, we describe the observation times in terms of weeks and set $K_{past} = 52$. This translates to looking one year prior to the observation time t, as typically done in fashion market analysis [36]. The next sections will sequentially detail the general pipeline of our approach, depicted in Fig. 2.

3.1 Image Tagging

The first operation is the extraction of textual tags $\{a_{\mathbf{z}}^{(j)}\}_{j=1,\dots,J}$ associated to $\mathbf{z}$. These tags should represent the clothing item with sufficient generality,

[5] https://datacentricai.org/.

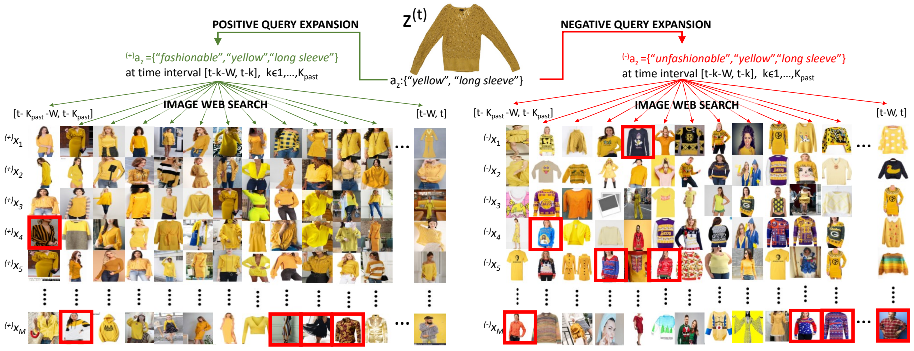
z(t)
POSITIVE QUERY EXPANSION
(+)a_z ={"fashionable","yellow","long sleeve"}
at time interval [t-k-W, t-k], kϵ1,...,K_past
a_z:{"yellow", "long sleeve"}
NEGATIVE QUERY EXPANSION
(-)a_z ={"unfashionable","yellow","long sleeve"}
at time interval [t-k-W, t-k], kϵ1,...,K_past
IMAGE WEB SEARCH
[t- K_past -W, t- K_past]
[t-W, t]
(+)x_1
(+)x_2
(+)x_3
(+)x_4
(+)x_5
(+)x_M
IMAGE WEB SEARCH
[t- K_past -W, t- K_past]
[t-W, t]
(-)x_1
(-)x_2
(-)x_3
(-)x_4
(-)x_5
(-)x_M

Fig. 3: Pipeline insights on *Time-dependent Query Expansion* (Sec. 3.2), *Image Web Search* (Sec. 3.3) and *Learning From Noisy Labels* (Sec. 3.4) steps. This figure reports a real world excerpt of the download and processing of N=2600 images ($N = 2(M \times K_{past})$, $M = 25$, K_{past}=52).

capturing at least categorical information (*e.g* "long sleeve") and a dominant color (*e.g* "yellow"). Empirically, we found these tags to work well while being easily obtainable. Category and color can be automatically extracted with high accuracy [19] or are usually contained in the technical data sheet accompanying the product, which is what we exploit in this work, as shown in Sec. 4.

3.2 Time-dependent Query Expansion

The second operation (detailed in Fig. 3 on a real example) performs two different textual query expansions, generating *positive expansions*, $\{a_{\mathbf{z}}^{(j)}\}_{j=1,\ldots,J} \cup J^{(+)}$ where the additional $J^{(+)}$ tags indicate attractive clothing items, and conversely for *negative expansions*. In this paper, we found the tags $J^{(+)}$ = "fashionable" and $J^{(-)}$ = "unfashionable" to be the most effective for positive and negative expansions, respectively. Alternatives as "best seller" and "unattractive" were considered, returning similar results.

Each expansion, either positive or negative, is associated to a particular $k = 1, ..., K_{past}$ for the time interval $[t-k-W, t-k]$, where W is a temporal window we wish to consider for the image search, also expressed in weeks. In our experiments we set $W = 4$, which translates to having a sliding window of size 4 and stride of 1 over the temporal axis. This allows the pool of downloaded images to disclose what are newly indexed items in relation to previous time steps, developing a temporal locality in the data pool. The precise value of W was chosen after an empirical evaluation over the range $1, ..., 12$.

3.3 Image Web Search

A given expanded textual query along with a time interval is fed into a web API request to gather M representative fashionable and unfashionable images ${}^{(+)}\{\mathbf{x}_i\}_{i=1,\ldots,M}^{(t-k)}$; ${}^{(-)}\{\mathbf{x}_i\}_{i=1,\ldots,M}^{(t-k)}$ that have been uploaded in the interval

$[t-k-W, t-k]$, for $k = 1, \ldots, K_{past}$. In particular, we adopt Google Image search, selecting the first $M = 25$ images returned, assuming the ordering of Google Images perfectly mirrors a genuine image relevance [17]. After the image web search phase, $M \times K_{past}$ fashionable and unfashionable images are collected respectively (as shown in Fig. 3). These images are then used to train a binary classifier θ, aimed at distinguishing fashionable from unfashionable images. Webly learning and supervision based on Google Images has been considered before in computer vision, especially for image classification and object detection [10,6,18]. POP goes one step further, merging visual and textual search while adding a time-dependent query expansion to create more discriminative image sets. Nevertheless, the labels assigned to the images from the query expansions might be noisy, therefore we apply a confident learning method.

3.4 Learning From Noisy Labels

In the following, we adapt the confident learning (CL) methodology specifically for our binary problem. For a broader overview, readers may refer to [27]. Let $\mathbf{X} = \{\mathbf{x}_i, \tilde{y}_i\}_{1 \ldots N}$ be our set of $N = 2(M \times K_{past})$ images with associated observed noisy binary labels $\tilde{y}_i \in \{$“fashionable”, “unfashionable”$\}$. CL assumes that a true, latent label $y_i^* \in \{$“fashionable”, “unfashionable”$\}$ exists for every sample. CL requires two inputs: 1) the out-of-sample $N \times 2$ matrix $\hat{\mathbf{P}}$ of predicted probabilities where $\hat{\mathbf{P}}_{i,h} = \hat{p}(\tilde{y}_i = h; \mathbf{x}_i, \theta)$ with θ a generic (binary) classifier initially trained on $\mathbf{X}$; 2) the set of noisy labels $\{\tilde{y}_i\}$. Subsequently, a robust 2×2 confusion matrix, called the *confident joint* matrix $\mathbf{C}_{\tilde{y},y^*}$, is computed[6]:

$$\begin{aligned} &\mathbf{C}_{\tilde{y},y^*}(h,l) = |\hat{\mathbf{X}}_{\tilde{y}=h,y^*=l}|, \text{with} \\ &\hat{\mathbf{X}}_{\tilde{y}=h,y^*=l} = \left\{ \mathbf{x} \in \mathbf{X}_{\tilde{y}=h} : \hat{p}(\tilde{y}=l; \mathbf{x}, \theta) \geq t_l \right\} \end{aligned} \tag{1}$$

where t_l is a threshold that represents the expected self confidence value for each class:

$$t_l = \frac{1}{|\mathbf{X}_{\tilde{y}=l}|} \sum_{x \in \mathbf{X}_{\tilde{y}=l}} \hat{p}(\tilde{y}=l; x, \theta) \tag{2}$$

In practice, $\mathbf{C}_{\tilde{y},y^*}$ counts only those elements which have been confidently classified in a particular class, where the term “confident” means with a probability that is higher than the average probability of an element belonging to that class. In simpler words, if samples labeled as belonging to class h tend to have higher probabilities because the model is over-confident about class h, then t_h will be proportionally larger. It also worth noting that Eq. 1 corresponds to a simplified version of the general building procedure of the confident joint matrix $\mathbf{C}_{\tilde{y},y^*}$ of [27], which nonetheless in our case is perfectly acceptable since we deal with binary classification and no *label collision* may happen, *i.e.,* the fact that a noisy label can correspond to a more than a single alternative class.

[6] We drop the index i for clarity.

On this robust confusion matrix, we estimate label errors from the off diagonal elements of $\mathbf{C}_{\tilde{y},y^*}(h,l)$. Wrongly labeled images are therefore pruned (indicated by the red boxes in Fig. 3), obtaining the cleaned fashionable and unfashionable images ${}^{(+)}\{x'_i\}_{i=1,\ldots,M'^{(t-k)}}^{(t-k)}; {}^{(-)}\{x'_i\}_{i=1,\ldots,M''^{(t-k)}}^{(t-k)}$, where $M'^{(t-k)}$ and $M''^{(t-k)}$ indicate that we can have a different number of positive and negative images, respectively, related to each $t-k$ time step, due to the noisy sample elimination. The classifier is retrained on the cleaned data, obtaining a robust trained model θ'. This procedure is data-centric and model agnostic; the specific θ used in this work is described in Sec. 4.

3.5 Signal Forming

The POP signal $S_{\mathbf{z}}^{(t)} = s_{\mathbf{z}}^{(t-K_{past})}, \ldots, s_{\mathbf{z}}^{(t-k)}, \ldots, s_{\mathbf{z}}^{(t-1)}$, is computed by considering the cleaned fashionable images ${}^{(+)}\{\mathbf{x}'_i\}_{i=1,\ldots,M'^{(t-k)}}^{(t-k)}$, the robust model θ', and the image $\mathbf{z}$, as follows:

$$s_{\mathbf{z}}^{(t-k)} = \frac{1}{M'^{(t-k)}} \sum_{i=1}^{M'^{(t-k)}} \frac{\langle \theta'\left({}^{(+)}\mathbf{x}'^{(t-k)}_i\right) \cdot \theta'\left(\mathbf{z}\right)\rangle}{\parallel \theta'\left({}^{(+)}\mathbf{x}'^{(t-k)}_i\right) \parallel \parallel \theta'\left(\mathbf{z}\right) \parallel} \tag{3}$$

where $\theta'\left(\mathbf{z}\right)$ indicates the extracted features of $\mathbf{z}$ from θ', and $\langle\cdot\rangle$ indicates the dot product. In other words, the signal value $s_{\mathbf{z}}^{(t-k)}$ is the average cosine similarity between the embedding of the probe image $\mathbf{z}$ and each fashionable image $\mathbf{x}'_{i^{(t-k)}}$ from the $M'^{(t-k)}$ downloaded images. An assessment of alternative signal forming options is shown in the supplementary material.

4 Experiments

In line with the general requirements of DCAI [25], we show how our automatically manufactured time series helps a forecasting model ψ achieve better results on a given task γ. The main idea behind our approach is that by knowing POP, the forecasting model can gain a context on the past which otherwise would be missing and therefore improve. To demonstrate this, we perform extensive evaluation on two tasks (and different forecasting models): *new fashion product sales curve prediction* [33,9,34], and *style popularity forecasting* [1]. We show ablative studies on the first task and an impressive outcome on the second.

The binary classifier θ for learning on noisy data (Sec. 3.4) is based on a ResNet50 [13], pre-trained on ImageNet [8], with two additional fully connected layers. During the confident learning procedure, we fine-tune its last convolutional block and fully connected layers for 50 epochs with a batch size of 64, using CE loss, following a 5-fold cross validation protocol. AdamW [21] is used as optimizer, with a learning rate of $1e-4$. The forecasting neural network models are all trained for 200 epochs with a batch size of 128 and L2 loss, using the AdaFactor [31] optimizer. The experiments are performed on two NVIDIA 3090 RTX GPUs.

4.1 Task 1: New Fashion Product Sales Curve Prediction

The output of a sales curve forecasting model for a probe clothing item $\mathbf{z}$ is a time series $O_{\mathbf{z}}^{(st)} = o_{\mathbf{z}}^{(st+1)}, \dots, o_{\mathbf{z}}^{(st+k)}, \dots, o_{\mathbf{z}}^{(st+K_{fut})}$ that indicates how many pieces of z will be sold starting at a particular time step st (typically the start of the season), for the next K_{fut} time steps.

We run our first set of experiments on the VISUELLE dataset [34]. For each available product, multi-modal information is provided: i) images, ii) text tags, iii) Google Trends, iv) sales curves. The evaluation protocol follows that of VISUELLE, simulating how a fast-fashion company deals with new products on two particular moments: the *first order setup* and the *release setup*. The former takes place when the company decides which products and how many pieces to order by looking at probe images. The latter is right before the season, and is useful to obtain an accurate forecast in order to plan stock replenishment. These two setups use 28 and 52 week long exogenous signals (originally Google Trends [34]), respectively.

Note that for the sake of fairness, we do not alter the training setup or models from [34], keeping the cardinality and the type of the training data fixed and *substituting only the Google Trends with our POP signal*. All the models are trained considering the 12-week long sales signals, whilst the evaluation is done on a 6-week horizon. This is shown to give the best predictions while simulating politics of real fashion companies [34].

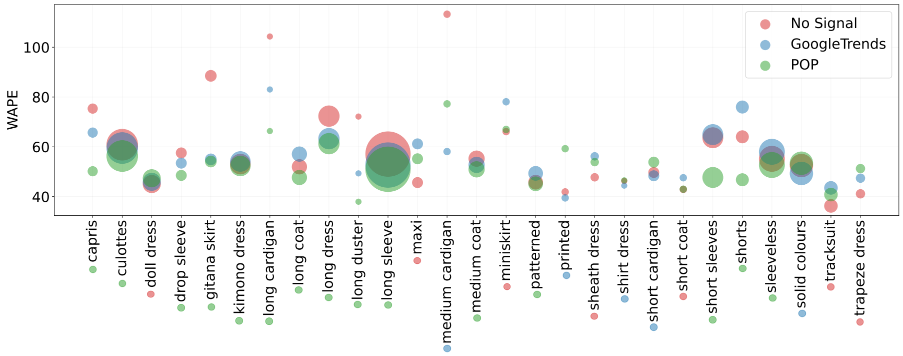

Fig. 4: Forecasting WAPE results per clothing category; the larger the blob, the higher the # of items in that category; the color below each category name indicates the type of training setup which gives the best WAPE.

We consider 5 algorithms (from oldest to newest): *Gradient Boosting* for forecasting [15], Concat Multi-Modal RNN [9] (*Concat MM RNN* in the tables), Residual Multi-Modal RNN [9] (*Residual MM RNN*), Cross-Attention RNN [9] (*X-Attention RNN*) and GTM Transformer [34] (*GTM Transf.*) We consider the Weighted Absolute Percentage Error (WAPE) as primary evaluation metric and

Table 1: Results on VISUELLE with the *first order setup*; “W” stands for WAPE, “M” for MAE. Lower is better for all metrics.

First Order Setup (K_{best} = 28 weeks)															
Exogenous Signal	***Gradient Boosting*** **[15] 2020**			***Concat MM RNN*** **[9] 2020**			***Residual MM RNN*** **[9] 2020**			***X-Attention RNN*** **[9] 2020**			***GTM Transformer*** **[34] 2021**		
	W	M	ERP	W	M	ERP	W	M	ERP	W	M	ERP	W	M	ERP
No Signal	64.10	35.02	0.43	63.31	34.41	0.42	64.26	34.92	0.44	59.49	32.33	0.38	56.62	30.93	0.37
Google Trends	64.29	35.12	0.43	64.11	34.84	0.43	68.11	37.02	0.47	58.70	31.90	0.38	56.83	31.05	0.35
POP Signal	**63.75**	**34.83**	**0.42**	**58.09**	**31.73**	**0.39**	**58.88**	**32.16**	**0.39**	**57.78**	**31.56**	**0.38**	**53.41**	**29.18**	**0.32**

Table 2: Results on VISUELLE with the *release setup*; “W” stands for WAPE, “M” for MAE. Lower is better for all metrics.

Release Setup (K_{best} = 52 weeks)															
Exogenous Signal	***Gradient Boosting*** **[15] 2020**			***Concat MM RNN*** **[9] 2020**			***Residual MM RNN*** **[9] 2020**			***X-Attention RNN*** **[9] 2020**			***GTM Transformer*** **[34] 2021**		
	W	M	ERP	W	M	ERP	W	M	ERP	W	M	ERP	W	M	ERP
No Signal	64.10	35.02	0.43	63.31	34.41	0.42	64.26	34.92	0.44	59.49	32.33	0.38	56.62	30.93	0.37
Google Trends	63.52	34.70	0.42	65.87	35.80	0.44	68.46	37.21	0.48	59.02	32.08	0.38	55.24	30.18	0.33
POP Signal	**63.38**	**34.62**	**0.42**	**57.43**	**31.37**	**0.36**	**58.38**	**31.89**	**0.39**	**57.36**	**31.33**	**0.36**	**52.39**	**28.62**	**0.29**

also compute the Mean Absolute Error (MAE) to demonstrate the error on an absolute scale [34]:

Note that the WAPE is not bounded by 100. Finally, we measure the similarity of the slope of the predicted curve with the ground truth using the *Edit distance with Real Penalty* (ERP) [5]. This metric counts the number of edit operations (insert, delete, replace) that are necessary to transform one series into the other. Because we are dealing with continuous values, a threshold ϵ=0.03 is used to decide if values are considered different and have to be edited.

The results are shown in Table 1 for the *first order setup* and in Table 2 for the *release setup*. As reference, we also report results *without* any exogenous series, to show the net value of these indicators. For all the algorithms and both setups, adding POP to the model boosts the performances over all the metrics, reaching the absolute best when coupled with GTM Transformer. On average, in the *first order setup*, we improve the WAPE by 3.42% over the Google Trends and by 3.21% over not using any exogenous signals. In the *release setup* we improve by 2.85% over the Google Trends and by 4.23% over not using exogenous signals. These results demonstrate how our data-centric approach can provide optimal forecasts by creating a highly-predictive signal of past popularity that is image-based, unlike Google Trends. The forecasts are performed on 497 products over different stores, meaning that these improvements can provide a large impact on the supply chain operations.

In Fig. 4 we show the WAPE *per clothing category*. We mostly perform better than the other training alternatives, yet some particular categories display limitations of our approach. These limitations arise due to the fact that the Image Tagging phase is assumed as flawless, since we rely on the technical sheet accompanying the probe image to extract the tags. The results per category (Fig. 4) display how possibly mislabeled categories, or categories labeled in a general manner ("solid colours","doll dress") may lead to misleading web images. As visible in Fig. 5, the related images from the web, both fashionable and not, are completely useless, since the tag of the category itself is misleading. In such cases, a robust automated category extraction could potentially lead to better results.

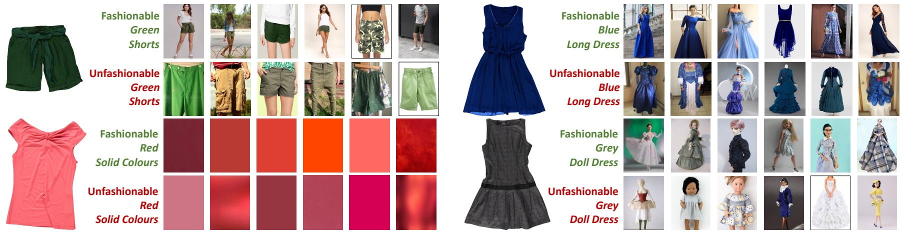

Fig. 5: Examples of VISUELLE items (seasons SS17 and SS18 on the left, SS19 and AI19 on the right, respectively) and the correspondent fashionable/unfashionable images from the web. Some web images can be misleading, due to the questionable category names of the VISUELLE dataset ("solid colours", "doll dress").

Ablation studies. In the following, we focus on alternative versions of our proposed pipeline, ablating the specific modules illustrated in Fig. 2. Table 3 contains all the results.

Time dependent query expansion

- *No expansion:* Images are queried with the original tags collected in the Image Tagging phase, without generating positive or negative expansions. This is equivalent to querying only with "color + category". The learning step is impacted directly, since no positive or negative classes are available for learning, therefore we use our backbone model to extract image features. For each image $\mathbf{z}^{(\mathbf{t})}$, the web images $\{\mathbf{x}_i\}_{i=1,\dots,M}^{(t-k)}$ that have been uploaded in the interval $[t-k-W, t-k]$, for $k = 1, \dots, K_{past}$ are collected. The signal forming Eq. 3 changes accordingly, using all the M downloaded images;
- *Misaligned past:* We modify the query expansions by looking one year earlier than the "correct" past. Given the observation time t of the probe $\mathbf{z}^{(t)}$, instead of looking backwards from $t-1$ weeks to $t - K_{past}$, we go from $t - 1 - K_{past}$ to $t - 2 \cdot K_{past}$.

With respect to all the alternative versions in this study, the *No expansion* ablation gives the worst result. POP provides an improvement of 0.73% and 1.06% WAPE for the *first order setup* and *release setup*, respectively. The *Misaligned*

Table 3: Alternative versions of our pipeline (Fig. 2) on both the *release* and *first order* setups; "W" stands for WAPE, "M" for MAE. Lower is better for all metrics.

Time Dependent Query Expansion				
	Release Setup		***First Order Setup***	
Strategy	**W**	**M**	**W**	**M**
No Expansion	53.12	29.02	54.47	29.77
Misaligned past	53.02	28.96	53.63	29.30
Learning With Noisy Labels				
	Release Setup		***First Order Setup***	
Strategy	**W**	**M**	**W**	**M**
No Learning	53.03	28.97	53.83	29.41
No Robust Learning	52.81	28.85	53.59	29.28
Symmetric Cross Entropy [38]	52.63	28.75	53.58	29.27
SELFIE [35]	52.56	28.71	53.51	29.23
POP	**52.39**	**28.62**	**53.41**	**29.18**

past yields slightly better results, but still performs worse than POP by 0.63% and 0.22% WAPE for the *first order setup* and *release setup*, respectively. This confirms that fashion has an evolution that changes year after year that we have to take into account.

Learning from noisy data

- *No learning*: A predefined image classification network is used to compute the distance among embeddings of the probe image with the positive, downloaded images. This is equivalent to ablating the "Learning from Noisy Data" phase of Fig. 2. It will highlight the importance of dealing with distances among embeddings which are specifically learned against distances coming from a general purpose network. We utilise the backbone of our binary classifier, specified in the introduction of Sec. 4;
- *No robust learning*: All of the downloaded positive and negative images are used to learn our binary classifier without pruning noisy data by confident learning;
- *Symmetric cross entropy* [38]: SCE is a robust classification loss; it adds to the standard cross entropy loss a *reverse cross entropy* term which assumes the predicted labels as ground truth, and the original labels as possibly faulty. In practice, it penalizes noisy labels, without removing any associated training data;
- *SELFIE* [35]: the key idea is to correct the label of noisy *refurnishable* samples with high precision, with the help of clean data which is defined as those samples within a mini-batch creating a small loss. Repeated training runs (dubbed "restarts") allow to use more training data, *i.e.,* noisy samples which have been corrected in their labels. In particular, we use 3 restarts, after which 1.1% of both fashionable and unfashionable items have been removed from the training data.

The results in Table 3 show slightly different performances, promoting the general idea of learning from webly data. *No learning* gives the worse performance, indicating that a fine tuning on the web data is beneficial (53.03 and 53.83 WAPE); when learning is done on the web data, there is some increase (52.81 and 53.59 WAPE); when learning is robust to label noise, with SCE, performances are better (52.63 and 53.58 WAPE); removing some outliers with SELFIE gives a further help (52.56 and 53.51 WAPE). Confident learning remains the best solution, with 52.39 and 53.41 WAPE, while removing 0.8% and 1.1% of fashionable and unfashionable items respectively, from the 14,500,200 images mined using our cross-modal pipeline.

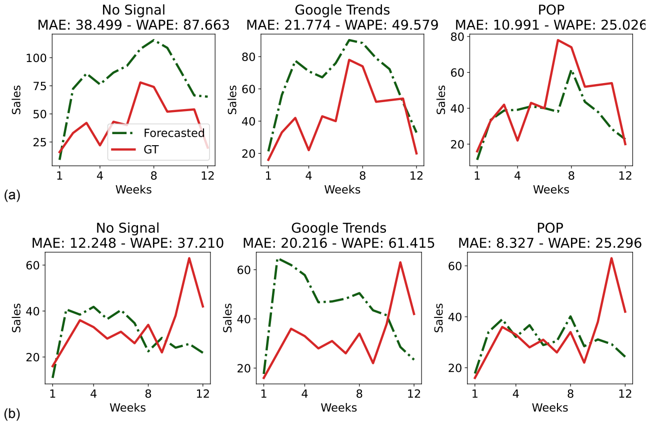

Fig. 6: Qualitative results for the sales forecast of two different products on VISUELLE, considering all 12 time-steps. In all cases, using POP provides better forecasts. In the bottom row (bottom right plot), we show a forecasting failure case, where the product is discounted in its final week of sales.

4.2 Task 2: Popularity Prediction Of Fashion Styles

The style popularity prediction task [1] is different from product sales forecasting in that it considers a popularity signal y based on multiple clothing items. In the literature [1,23], style is defined as a latent property of a set of clothing images that share some common visual features. Concretely, in Fashion Forward (FF) [1], Non-negative Matrix factorization is applied to extract K styles from the attribute extraction features [19] of all the product images. Formally, let $\mathbf{A} \in \mathbb{R}^{M \times N}$ indicate the confidence that each of the M visual attributes is contained in each of the N images. $\mathbf{A}$ can be factorized into two matrices with non-negative entries:

$$\mathbf{A} \approx \mathbf{W}\mathbf{H}, \mathbf{W} \in \mathbb{R}^{M \times K} \quad \text{and} \quad \mathbf{H} \in \mathbb{R}^{K \times N} \tag{4}$$

where $\mathbf{W}$ represents the confidence that each attribute is part of a style and $\mathbf{H}$ represents the confidence that each style is associated to an image. The popularity signal y for a style k is built by considering the interactions in the Amazon Reviews dataset [24] of all the items $\{\mathbf{z}\} \in A$ at time t, weighted by their style membership $H(k, z)$. For a detailed explanation, we refer to [1].

Table 4: Average results over all Fashion Forward [1] dataset partitions and specific results for the Dresses partition, where POP outperforms even the original GT style popularity time series (*Oracle*).

Global Average

Signals	**Mean**		**Last**		**Drift**		**AR**		**ARIMA**		**SES**	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Oracle	*0.136*	*0.170*	*0.093*	*0.114*	*0.174*	*0.222*	*0.271*	*0.403*	*0.136*	*0.167*	*0.094*	*0.116*
GoogleTrends	0.846	1.000	0.846	1.000	0.846	1.000	0.846	1.000	0.846	1.000	0.846	1.000
POP	0.152	0.192	0.116	0.144	0.182	0.229	0.281	0.418	0.235	0.293	0.125	0.156

Dresses

Signals	**Mean**		**Last**		**Drift**		**AR**		**ARIMA**		**SES**	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
Oracle	*0.155*	*0.197*	*0.130*	*0.158*	*0.203*	*0.263*	*0.307*	*0.409*	*0.173*	*0.209*	*0.129*	*0.157*
GoogleTrends	0.849	1.000	0.849	1.000	0.849	1.000	0.849	1.000	0.849	1.000	0.849	1.000
POP	**0.119**	**0.157**	**0.108**	**0.127**	**0.173**	**0.216**	**0.229**	**0.334**	**0.162**	**0.193**	**0.109**	**0.130**

To extend this problem to a NFPPF setup, we have to imagine we are evaluating the performance of a brand new style that does not have a past. The purpose of POP becomes replacing the original style popularity series. This means that POP has to be modified to deal with a style and not with a single clothing item, where two challenges are presented: 1) To verify how similar POP is to the ground truth popularity signal and; 2) To check if POP is highly predictive of the future popularity.

To deal with the first challenge, we consider for each style k the 2 textual attributes [19] w_1, w_2 (extracted from $\mathbf{W}$) with the highest confidence scores and use them for the time dependent query expansion. FF provides the only dataset for style forecasting where both images and product metadata are available. The task is to predict a popularity score on a yearly basis. The data ranges from $[2008 - 2013]$, but since Google Images returns little to no images for queries before 2010, we use the range $[2010 - 2013]$ in our experiments. We set $K_{past} = 208$, meaning that we investigate 4 years back. In this way we can create a weekly series for each year and use the average as the value representing the popularity for that year. As probe image to create our POP signal, we consider the top 10 images $\{\mathbf{z}\}$ that represent a style (based on their membership weight $H(k, z)$). Each image will lead to one POP signal, which we average together to obtain the *POP style signal*. This process is repeated for all the dataset partitions presented in FF. To deal with the second challenge, we adopt the best performing statistical forecasting techniques from Fashion Forward and feed them the style POP signal described above. For more details on the forecasting techniques, we refer the reader to [14,4]:

1. **Naive methods.** These methods infer by utilizing general information from the training data. *Mean* forecasts the future as the mean of past observations, while *Last* as the last observed value. *Drift* is the same as *Last*, but the forecasts change over time based on the global trend of the series;
2. **Auto Regressive and Moving Average methods.** These methods forecast using a linear combination of *some* past observations in a regression

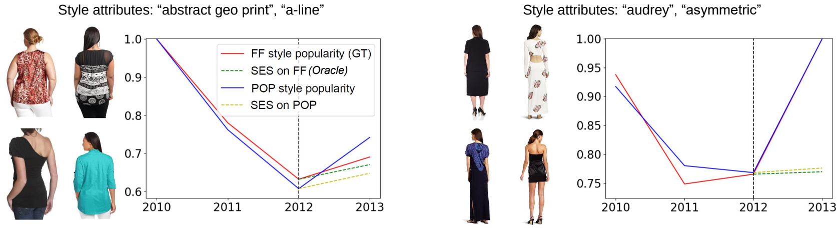

Fig. 7: Qualitative results on the forecasting of two different styles from FF, represented by their respective ”style-defining” images and top (automatically extracted) attributes. In both cases, POP and the ground-truth (GT) style popularity from Fashion Forward are substantially similar. The plot on the right shows a forecasting failure case, which holds for both POP and the GT).

framework. The most famous and representative method of this class is the *AutoRegressive Integrated Moving Average*(ARIMA) model.

3. **Simple Exponential Smoothing.** Stands for simple exponential smoothing, is a weighted average of previous observations where the weights decrease exponentially as we go further in the past.

Following the protocol of [1], all models are trained on all but the last timestep, which is used for testing. We utilise the mean absolute percentage error (MAPE) and the mean absolute error (MAE) to evaluate the forecasting accuracy on the last timestep of the signal stemming from FF. To provide an additional comparison, we show additional results using Google Trends as the substitute popularity time series [34]. Note that to obtain fair and comparable results, all the signals are rescaled in the range [0,1] using min-max normalization. The results are shown in Table 4, where *Oracle* refers to the original ground-truth style popularity series given as input to the forecasting models.

POP proves to be a natural substitute to the GT style popularity time series and it allows for optimal forecasts, providing better results than the GT signal itself for the *Dresses* partition. On the other hand, Google Trends are not able to convey such similarities, partially because searching only for the popularity of textual tags might not provide a series that is as predictive as ours.

5 Conclusion

Metaphorically, our approach performs a kind of “time travel”: it sends a fashion probe image in the past, before its launch in the market. It then models the popularity from that past point forward by relying on highly ranked web images, queried by using general textual tags related to the probe. The probe similarity with the past is then shown to be a good exogenous indicator for future performance. This pipeline provides a new, effective and data-centric scheme for NFPPF problems.

Acknowledgements. This work was supported by the Italian MIUR through the project ”Dipartimenti di Eccellenza 2018-2022”.

References

1. Al-Halah, Z., Stiefelhagen, R., Grauman, K.: Fashion forward: Forecasting visual style in fashion. In: ICCV (2017)
2. Anik, A.I., Bunt, A.: Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021)
3. Arvan, M., Fahimnia, B., Reisi, M., Siemsen, E.: Integrating human judgement into quantitative forecasting methods: A review. Omega **86** (2019)
4. Box, G., Jenkins, G., Reinsel, G., Ljung, G.: Time Series Analysis: Forecasting and Control. John Wiley & Sons (2015)
5. Chen, L., Ng, R.: On the marriage of lp-norms and edit distance. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (2004)
6. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1431–1439. IEEE Computer Society, Los Alamitos, CA, USA (dec 2015). https://doi.org/10.1109/ICCV.2015.168, https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.168
7. Cheng, W.H., Song, S., Chen, C.Y., Hidayati, S.C., Liu, J.: Fashion meets computer vision: A survey. ACM Computing Surveys (CSUR) **54**(4) (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009). https://doi.org/10.1109/CVPR.2009.5206848
9. Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S.S.K., Dwivedi, S., Raykar, V.: Attention based Multi-Modal New Product Sales Time-series Forecasting. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, Virtual Event CA USA (Aug 2020). https://doi.org/10.1145/3394486.3403362, https://dl.acm.org/doi/10.1145/3394486.3403362
10. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. vol. 2, pp. 1816–1823 Vol. 2 (2005). https://doi.org/10.1109/ICCV.2005.142
11. Fildes, R., Ma, S., Kolassa, S.: Retail forecasting: Research and practice. International Journal of Forecasting (2019)
12. Garcia, C.C.: Fashion forecasting: an overview from material culture to industry. Journal of Fashion Marketing and Management: An International Journal (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
14. Hyndman, R., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, Australia, 2nd edn. (2018)
15. Ilic, I., Görgülü, B., Cevik, M., Baydoğan, M.G.: Explainable boosted linear regression for time series forecasting. Pattern Recognition (2021)
16. Jeon, Y., Jin, S., Kim, B., Han, K.: Fashionq: An interactive tool for analyzing fashion style trend with quantitative criteria. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (2020)
17. Jing, Y., Baluja, S.: Visualrank: Applying pagerank to large-scale image search. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**, 1877–1890 (2008)

18. Li, J., Song, Y., Zhu, J., Cheng, L., Su, Y., Ye, L., Yuan, P., Han, S.: Learning from large-scale noisy web data with ubiquitous reweighting for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(5), 1808–1814 (2021). https://doi.org/10.1109/TPAMI.2019.2961910
19. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
20. Lo, L., Liu, C., Lin, R., Wu, B., Shuai, H., Cheng, W.: Dressing for Attention: Outfit Based Fashion Popularity Prediction. In: 2019 IEEE International Conference on Image Processing (ICIP) (Sep 2019). https://doi.org/10.1109/ICIP.2019.8803461, iSSN: 2381-8549
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
22. Ma, Y., Ding, Y., Yang, X., Liao, L., Wong, W.K., Chua, T.S.: Knowledge Enhanced Neural Fashion Trend Forecasting. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. ACM, Dublin Ireland (Jun 2020). https://doi.org/10.1145/3372278.3390677, https://dl.acm.org/doi/10.1145/3372278.3390677
23. Ma, Y., Ding, Y., Yang, X., Liao, L., Wong, W.K., Chua, T.S.: Knowledge enhanced neural fashion trend forecasting. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. ICMR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3372278.3390677, https://doi.org/10.1145/3372278.3390677
24. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 43–52. SIGIR '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2766462.2767755, https://doi.org/10.1145/2766462.2767755
25. Motamedi, M., Sakharnykh, N., Kaldewey, T.: A data-centric approach for training deep neural networks with less data. arXiv preprint arXiv:2110.03613 (2021)
26. Ng, A.: A chat with andrew on mlops: From model-centric to data-centric ai. https://www.youtube.com/watch?v=06-AZXmwHjo (May 2021)
27. Northcutt, C., Jiang, L., Chuang, I.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70** (2021)
28. Northcutt, C.G., ChipBrain, M., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. stat **1050** (2021)
29. Ren, S., Chan, H.L., Ram, P.: A comparative study on fashion demand forecasting models with multiple sources of uncertainty. Annals of Operations Research **257**(1) (2017)
30. Ren, S., Chan, H.L., Siqin, T.: Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. Annals of Operations Research **291**(1) (2020)
31. Shazeer, N., Stern, M.: Adafactor: Adaptive learning rates with sublinear memory cost. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/shazeer18a.html
32. Silva, E.S., Hassani, H., Madsen, D.Ø., Gee, L.: Googling fashion: forecasting fashion consumer behaviour using google trends. Social Sciences **8**(4) (2019)

33. Singh, P.K., Gupta, Y., Jha, N., Rajan, A.: Fashion Retail: Forecasting Demand for New Items. In: 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Jun 2019), http://arxiv.org/abs/1907.01960
34. Skenderi, G., Joppi, C., Denitto, M., Cristani, M.: Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends. arXiv preprint arXiv:2109.09824 (2021)
35. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning. PMLR (2019)
36. Sorger, R., Udale, J.: The fundamentals of fashion design. Bloomsbury Publishing (2017)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
38. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)