DAVID: Dual-Attentional Video Deblurring

Junru Wu[†], Xiang Yu[‡], Ding Liu[*], Manmohan Chandraker[§‡] and Zhangyang Wang[†]

[†]Texas A&M University
[‡] NEC Laboratories America
[§]University of California, San Diego
[*]ByteDance AI Lab

{sandboxmaster,atlaswang}@tamu.edu, {xiangyu,manu}@nec-labs.com, dingliu2@illinois.edu

| Input | DVD (noalign) | DAVID (Ours) | Groud Truth |

Figure 1: Visual comparison on Challenging DVD Dataset with DVD (noalign) [35].

Abstract

Blind video deblurring restores sharp frames from a blurry sequence without any prior. It is a challenging task because the blur due to camera shake, object movement and defocusing is heterogeneous in both temporal and spatial dimensions. Traditional methods train on datasets synthesized with a single level of blur, and thus do not generalize well across levels of blurriness. To address this challenge, we propose a dual attention mechanism to dynamically aggregate temporal cues for deblurring with an end-to-end trainable network structure. Specifically, an internal attention module adaptively selects the optimal temporal scales for restoring the sharp center frame. An external attention module adaptively aggregates and refines multiple sharp frame estimates, from several internal attention modules designed for different blur levels. To train and evaluate on more diverse blur severity levels, we propose a Challenging DVD dataset generated from the raw DVD video set by pooling frames with different temporal windows. Our framework achieves consistently better performance on this more challenging dataset while obtaining strongly competitive results on the original DVD benchmark. Extensive ablative studies and

qualitative visualizations further demonstrate the advantage of our method in handling real video blur.

1. Introduction

Mobile phones, high internet bandwidths and social media have led to a recent spurt in video acquisition and sharing. However, videos of dynamic scenes, or those captured through hand-held devices, often display spatially and temporally varying blur patterns. The source of blur can be camera shake, depth variation, object motions or a combination of them, which manifest in complex patterns such as jittering, jumping or ghosting artifacts. Restoration of spatial structure and image sharpness is an ill-posed problem, for which single-image deblurring methods have been proposed relying on statistical [8, 12] or learned priors [24, 33, 38, 48, 19]. But video deblurring remains relatively less-studied, since effectively and efficiently modeling the inherent temporal dynamics among consecutive video frames is challenging.

Recent works on video deblurring usually work on a fixed temporal scale. For example, a fixed number of blurry frames are stacked as inputs to the network in [35, 13]. However, the

spatio-temporal scale of blurs can vastly vary. Camera shake usually leads to short-term, spatially uniform and temporally uncorrelated blurs [44, 35], while object motion causes long-term, spatially localized and temporally smooth blurs [27]. Such heterogeneity in blur patterns is unlikely to be well represented by a single model in the above approaches, and calls for the selection or modulation of temporal scales.

This paper proposes a Dual-Attentional VIdeo Deblurring (**DAVID**) framework, aiming to simultaneously model temporal dynamics among consecutive frames and handle various levels of spatially heterogeneous blur in real videos. We start by constructing a compact backbone module based on a U-Net variant [32], to infer the sharp center frame from a number of consecutive blurry frames. Using this network module as a basic building block, we propose two levels of attention modules to hierarchically reason about blur compositions and levels in both temporal and spatial dimensions.

We determine the optimal scale of temporal dependency through an **internal attention** module that takes as input several consecutive blurry frames, while multiple backbone branches work on different temporal scales to generate different sharp frame estimates. The resultant sharp frame estimates are adaptively aggregated according to the attention map, which is inferred by another branch in this module. Further, an **external attention** module handles various levels of spatially heterogeneous blur by determining how to adaptively aggregate the outputs of multiple internal attention modules, each of which is designed for a specific blur level. The architecture is illustrated in Figure 2. To the best of our knowledge, this is the first work in video deblurring that exploits hierarchical attention.

Our proposed method achieves consistently better PSNR compared to others on public datasets such as DVD. Further, we create a new video deblurring dataset, called *Challenging Deep Video Deblurring* (Challenging DVD), to introduce stronger blur variations by synthesis. Compared to the original Deep Video Deblurring (DVD) dataset proposed in [35], we witness performance drops for state-of-the-art methods when applied to Challenging DVD. Extensive ablation studies and visual results on real blur videos further indicate that **DAVID** can adaptively deal with different levels of blurriness, leading to perceptible improvement in quality.

In summary, our main contributions are:

- A novel Dual-Attentional Video Deblurring (**DAVID**) framework, which jointly takes into account heterogeneous blur information in temporal and spatial dimensions for video deblurring, with an external attention and several internal attention modules.

- A new video deblurring dataset, Challenging Deep Video Deblurring (Challenging DVD), that introduces stronger blur composition and level variations.

- Extensive experiments on both Challenging and original DVD datasets, to demonstrate that DAVID achieves consistently better or competitive results, respectively.

2. Related Work

Image Deblurring: Image deblurring can be traced back to the traditional methods using different types of priors, such as total variation (TV) [12], sparsity [8], heavy-tailed gradient prior [34], and l_0-norm gradient prior [46]. Many early approaches [15, 17, 9, 10] proposed to estimate the blur kernel and then apply deconvolution, in which the estimated kernel quality significantly impacted the result. [28] presented a simple and effective dark channel prior. [30] showed that a simple low-rank model significantly reduces blur even without using any kernel information. Segmentation information [11] was also investigated as an accompanying cue for motion.

In recently years, CNN-based image deblurring methods have achieved success [45]. [36] proposed to directly estimate motion blur fields by CNN. [33] adopted a coarse-to-fine manner to stack multiple CNNs to analyze the blur formation. [47] proposed to estimated motion blur kernel using deep learning. [24] presented multi-scale loss function that mimics conventional coarse-to-fine approaches in training a multi-scale deblurring CNN. [18] introduced generative adversarial networks (GANs) to obtain sharp and realistic-looking images. [38] introduced a scale-recurrent network to exploit the multi-scale spatial information. Recent works also shows that restoration of blurry image can be used to facilitate high-level task such as object detection [50], image classification[41] and image segmentation [42].

Video Deblurring: In the video case, temporal variations critically determine the blur effect, and provide additional clues for deblurring. Early video deblurring methods [23, 5] compensated sharp details of the current frame by nearby frames, via patch matching, motion flow and frame alignment. They however failed easily when dealing with large movements. [7] applied optical flow to warp nearby frames and fused them in the Fourier domain, but suffered from unreliable flow estimation when occlusions or outliers are present. [29] proposed to simultaneously deblur stereo videos and estimate the scene flow, where the motion cues from the scene flow and blur information can complement each other.

In deep learning, video deblurring has so far received relatively limited attentions. [35] proposed DeBlurNet (DBN) on accumulating information across frames, where neighboring frames were stacked as inputs to predict the clean central frame. [13] introduced a spatio-temporal recurrent network that adaptively enforced temporal consistency between consecutive frames. Segmentation information has also been incorporated with video deblurring when available [31].

Attention Model: Attention in general serves as a learnable guidance, to re-allocate available processing resources towards the most informative input components. It has shown promise in language translation [40], object recognition [1],
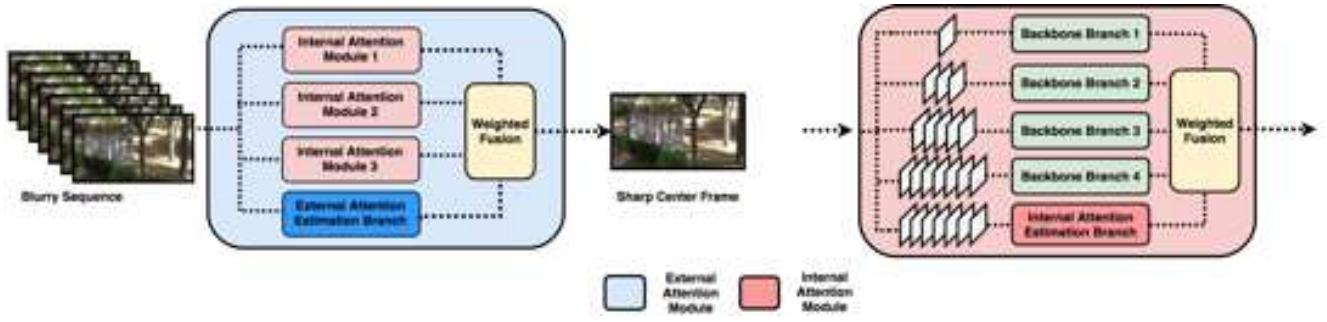
Figure 2: The overview of the proposed DAVID framework. Left: the structure of the external attention module. Right: the structure of the internal attention module. Each internal attention module is designed for a specific blur level. Each backbone branch in the same internal attention module works on a specific temporal scale.

image generation [51] and person re-id [4], meta-learning [2]. For video-based applications, [49] proposed a temporal attention network to aggregate video frames for face recognition. [20] developed attention-based temporal modulation for video super-resolution, where the predictions made within different temporal windows were adaptively fused with a learned pixel-level attention map. We also note that a few others [43, 16] discussed the concept of hierarchical attention, but with completely different contexts and motivations from ours.

3. The Proposed Model: DAVID

3.1. Network Design

The overall architecture of our DAVID model is illustrated in Figure 2. Taking a set of $2N-1$ consecutive blurry video frames, DAVID aims to output the center sharp frame, *i.e.*, the N-th frame. Each internal attention module determines the optimal temporal scale for removing blur, via adaptively aggregating multi-scale temporal information from several *backbone branches*. The external attention then estimates the spatially global blur level, and make soft assignments for several *internal attention modules*, each of which is dedicated for a specific blur level. Finally, their outputs are fused adaptively under the guidance of external attention, to generate the sharp frame.

Backbone Branch: Our backbone branch adopts an encoder-decoder U-Net, which is very popular for many image restoration tasks [21, 22, 37, 38], as detailed in Figure 3. The encoder consists of five blocks. Each block contains two consecutive convolutional layers, followed by a max-pooling layer, which downsamples the feature maps by half. The decoder is formulated in a mirroring way, with five blocks. Each one contains a bilinear upsampling layer followed by two convolutional layers. The feature maps after each block is upsampled by a factor of two and thus the final output achieves the same size as the input. Skip connections are adopted to aggregate the feature maps from the encoder to the decoder on each spatial scale, to better utilizes features from multiple scales.

As the building block of DAVID, we carefully fine-tune the backbone to obtain remarkable effectiveness: as shown in experiments later, a single backbone branch is already able to outperform [35] on the DVD dataset. Its performance will be further boosted as the temporal information is added.

Internal Attentional Module: As shown in Figure 4, an internal attention module $\mathcal{A}_n$ is designed to coordinate a number of backbone branches. Each backbone branch handles a different temporal scale: The D_i backbone branch takes $2i+1$ consecutive blurry input frames (centered at the current frame) to predict the center sharp estimate. As a result, the convolutional filters in the first layer are customized to have $(2i+1) \times r$ channels, where r denotes the channel number of each input frame.

Besides the backbone branches, we design an *internal attention estimation branch* to learn the selectivity on different temporal scales according to blur information, by predicting pixel-level aggregation weights for each branch's output. In practice, the internal attention estimation branch takes all $2N-1$ frames as inputs, and outputs the pixel-level weight maps on all N possible temporal scales. Considering the computation cost and efficiency, we adopt a shallower architecture of two downsampling blocks for the estimation branch. Eventually, each backbone branch's output D_i is pixel-wisely multiplied with its corresponding weight map from the internal attention estimation branch, and these products are summed up to form the final output of the internal attention module. This operation is expressed in Equation 1:

$$\hat{I}_n = \sum_{i=1}^{N} \mathcal{A}_{n,i} \otimes D_i \tag{1}$$

$\mathcal{A}_{n,i}$ denotes the internal attention map for backbone branch i. D_i is the output from backbone branch i. $\otimes$ represents pixel-wise multiplication. N is the total number of backbone branches and $\hat{I}_n$ is the recovered output from internal attention module.

External Attentional Module: In DAVID, each internal attention module focuses on one global background blur
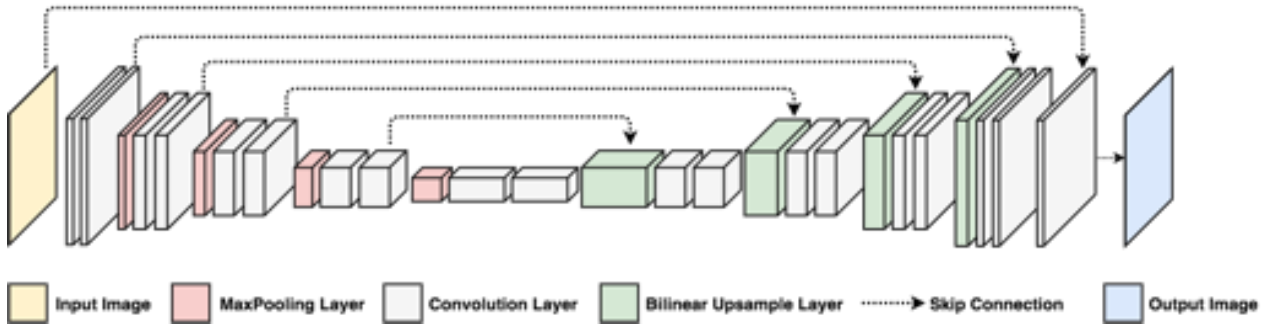
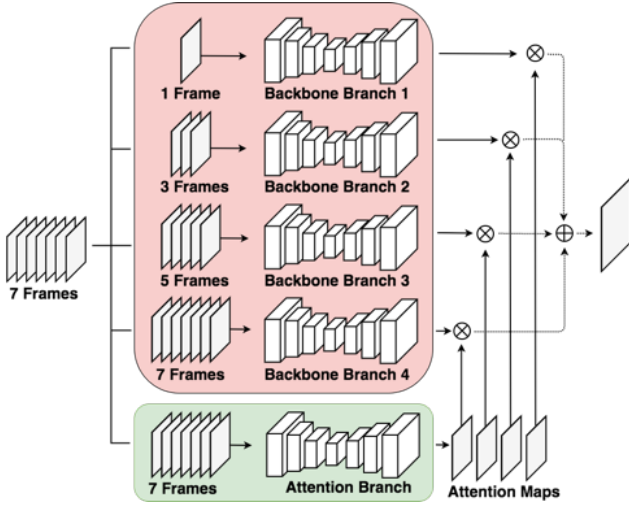Figure 3: A proposed backbone branch for DAVID.

Figure 4: An internal-attentional module for DAVID.

level. The assumption on the blur level further needs to be aggregated in a probabilistic way so that we can better model the spatially variant blur effect. Thus, an external attention mechanism $\mathcal{A}_e$ is proposed on top of the internal attention modules $\mathcal{A}_n$. Similar to the internal attention module, we design an *external attention branch* to predict a pixel-level weight map for each internal attention module, with all $2N-1$ frames as its input. Its implementation also refers to a backbone branch architecture with only two downsampling blocks. The final deblurring result by DAVID is the pixel-wise weighted summation result of all internal attention module outputs, as in Equation 2.

$$\hat{I}_e = \sum_{j=1}^{M} \mathcal{A}_{e,j} \otimes \hat{I}_{n,j} \qquad (2)$$

$\mathcal{A}_{e,j}$ is the external attention map for the j-th internal attention module output $\hat{I}_{n,j}$. M is the total number of internal modules and $\hat{I}_e$ the final output from the external module.

3.2. Challenging DVD Dataset

To train data-driven video deblurring models, a large number of pairs of blurry and sharp frames are needed. Although it is possible to use a beam-splitter and multiple cameras to build a special capturing system, this setup would be challenging to construct robustly, and can face many calibration issues. [35] proposed to collect real-world sharp videos at very high frame rate, and synthetically create blurred ones by averaging consecutive short-exposure images to approximate a longer exposure [39]. The authors collected 71 videos at 240 frames per seconds (fps), with 3-5s average length. In order to simulate realistic blurs at 30fps, they subsampled every eighth frame to create the 30fps ground truth sharp video, while averaging 7 neighboring frames (centered at the corresponding ground truth frame) to generate the blurred video at 30fps. The video deblurring model is then trained to recover the 30fps ground truths from the 30fps blurred versions. We refer to this resulting dataset as the Deep Video Deblurring (DVD) dataset, which has been the most popular benchmark for video deblurring algorithms.

To mimic more challenging real-world blurs, we create a new **Challenging DVD dataset**, aiming to cover a wider spectrum of blur variations. This is achieved by introducing more stochasticity during blurry video synthesis. Using the same high fps video of DVD, we randomly choose every 3, 7, 11 or 15 frames to average into the blurry frame,[1] instead of only averaging 7 frames as in the DVD dataset.[2] We dubbed them C-DVD-3, C-DVD-7, C-DVD-11 and C-DVD-15, respectively. The modification leads to the new Challenging DVD dataset that contains dynamic, and often more severe (when 11 or 15 frames averaged) blurs. Experiments manifest that almost all video deblurring methods witness performance drops when applied to Challenging DVD, compared to performance on DVD. The value of Challenging DVD dataset is justified later by our model's generalization evaluation on real-world blurry videos when trained on it.

4. Experiments

In this section, we firstly explain the detail implementation and the training protocols. Then, we compare DAVID

---

[1]We choose 3, 7, 11, 15 here, since we hope that they can represent several different blur levels from light to severe. Those choices are empirical, in no way unique, and may be further tuned for better practical performance.

[2][35] suggest to use optical flow to interpolate additional frames in order to smoothen the averaging. However, we observe that the classical optical flow often introduce visible "ghost" artifacts in the synthetic blurry frames. We thus use a neural network based frame interpolation model [26] to replace that, observing improved blur quality.

to several video-based state-of-the-arts on the Challenging DVD dataset, to highlight our method's effectiveness on different levels of blur. The performance of DAVID on the original DVD dataset is also reported for reference comparison. We then investigate the ablative functions for each of our proposed modules, *i.e.*, single back bones, internal attention modules. Visualization of our attention modules is provided for further insightful analysis. Finally, we display qualitative results of DAVID on real blurry videos and compare to the state-of-the-arts.

4.1. Multi-Phase Training for DAVID

To train our Dual-Attention Video Deblurring (**DAVID**) framework, good initializations are needed for both backbone branches and internal/external attention modules. We propose a multi-phase strategy to smoothly conduct training.

Phase 1: Pre-train each backbone branch: We first pre-train each backbone branch D_i. Specifically, the backbone branches D_1, D_2, D_3, D_4 are independently trained to take in 1, 3, 5, 7 frames, respectively.

Phase 2: Pre-train the internal attention module: After Phase 1, we assemble each pre-trained backbone branch into the internal attention module. The internal attention modules $\mathcal{A}_{n,1}$, $\mathcal{A}_{n,2}$, $\mathcal{A}_{n,3}$ are trained separately, with samples from C-DVD-3, C-DVD-7, and C-DVD-11, respectively, where each represents a different level of blur that is temporally pooled from 3, 7, 11 frames. To train each $\mathcal{A}_i$, we train the internal attention branch with a learning rate $1e^{-5}$ while fixing the weight of backbone branches. After 100 epochs, we unfreeze the backbone branches and fine-tune entire $A_{n,i}$ and D_i with a smaller learning rate $2e^{-6}$ until convergence. In this way, we enforce each $A_{n,i}$ to focus on a specific level of blur. The learning rate and hyper-parameters are grid searched based on a validation set.

Phase 3: Joint training with the external attention module: We load all the pre-trained internal attention modules from Phase 2 and fix the weights for each of them. We fine-tune the external attention branch with initial learning rate $1e^{-5}$. After 200 epochs, we unfreeze the attention branch in each of the internal attention module, jointly fine-tune the internal/external attention branches for 200 epoch with a learning rate $2e^{-6}$. Finally, we unfreeze all the backbone branches in each of the internal attention module and jointly fine-tune the entire DAVID model with a learning rate $2e^{-6}$ for another 200 epochs.

Implementation Details We implement our backbone branch with PyTorch. By default, our model has 3 internal attention modules, each consists of 4 single backbone branches, that take 1, 3, 5, 7 consecutive input frames, respectively. Please refer to the detail network structure design in the supplementary material. During training, we apply random cropping, random horizontal and vertical flipping as data augmentation. Batch size is set to 16 across all the train-ing phases. We adopt Adam [14] solver with L_2 loss. Initial learning rate is set 0.0001. The momentum and weight decay are fixed as 0.9 and 0, respectively.

4.2. Evaluation on the Challenging DVD Dataset

The original DVD dataset only considers a fixed blur level, *i.e.*, synthesized by averaging 7 frames. We aim to highlight the capability of deblurring models in dealing with different levels of blur. In Challenging DVD dataset, we purposely synthesize the blur of multiple levels by averaging the frames with different window size such as 3, 7, 11 or 15. We also form a 10-fold testing split that has no data overlap according to [35]. Our testing set consist of 297, 349, 739, 524 samples for C-DVD-3, C-DVD-7, C-DVD-11, C-DVD-15, respectively. We compare DAVID with two state-of-the-art video deblurring methods, WFA [6] and DVD[35]; both are also fine-tuned on the Challenging DVD dataset.

In Table 1, PSNR is averaged over all frames for each video. We clearly see that all the methods show performance degradation compared to performance on original DVD set (see Table 3), suggesting that our newly synthesized dataset is indeed more challenging. However, DAVID shows consistently superior performance as the dual attention module can dynamically gather the temporal information and thus handle different levels of blur.

4.3. Ablation Study on Challenging DVD

4.3.1 Effect on Different Model Components

In Table 1, under the same number of averaged frames, *i.e.*, On C-DVD-11, we consistently observe that single backbone models are always sub-optimal compared to the one with internal attention module, indicating the effectiveness to aggregate the temporal information across the consecutive video frames. Meanwhile, Table 1 shows that with the same setting, *i.e.* all single backbone setting or all with internal attention setting, the larger the window size of averaging, the better the performance is, which suggests that larger window size is better for information gathering to conduct deblur. Moreover, our Dual Attention model further outperforms any internal attention model. This is achieved by further adding an external attention module, which aims at gathering the spatial correlation across different internal attention models.

4.3.2 Effect on Stacked Frames

We observe in Table 1 that larger temporal window size is always beneficial than smaller window size. This motivates us to investigate how the temporal scale influences the overall performance. We fix a single backbone model, *i.e.*, single backbone trained on C-DVD-11, and prepare the testing data by stacking frames of 1, 3, 5, 7 and 9 respectively. Table 2 shows the ablative numbers with respect to the number of stacked frames. As the stacking number increases, we observe the performance increases. The larger window size

Methods	1	2	3	4	5	6	7	8	9	10	Average
WFA[6]	25.16	29.42	32.73	23.88	22.91	29.77	26.67	25.36	31.10	23.68	26.34
DVD (no align)[35]	27.37	32.00	34.93	25.88	25.02	31.68	28.87	27.47	33.23	25.79	28.52
Single (C-DVD-3)	26.91	32.77	37.44	24.85	24.93	31.43	28.51	26.76	32.88	24.61	28.52
Single (C-DVD-7)	29.69	35.19	39.11	27.13	26.68	34.10	31.05	29.06	36.60	27.65	31.03
Single (C-DVD-11)	29.72	35.63	38.99	28.15	26.90	34.28	31.19	29.33	36.55	28.44	31.31
Internal Att (C-DVD-3)	27.34	33.11	37.74	25.11	25.20	31.84	28.95	27.05	33.29	24.85	28.86
Internal Att (C-DVD-7)	30.05	35.57	39.25	27.45	26.95	34.41	31.37	29.27	36.85	27.93	31.32
Internal Att (C-DVD-11)	30.17	35.95	39.03	28.42	27.20	34.48	31.49	29.38	36.73	28.67	31.56
Dual Att (C-DVD-3-7-11)	**30.67**	**36.23**	**39.26**	**28.58**	**27.41**	**34.82**	**31.73**	**29.50**	**37.35**	**28.94**	**31.84**

Table 1: PSNR results on the full Challenging DVD testing set. Best results are shown in bold.
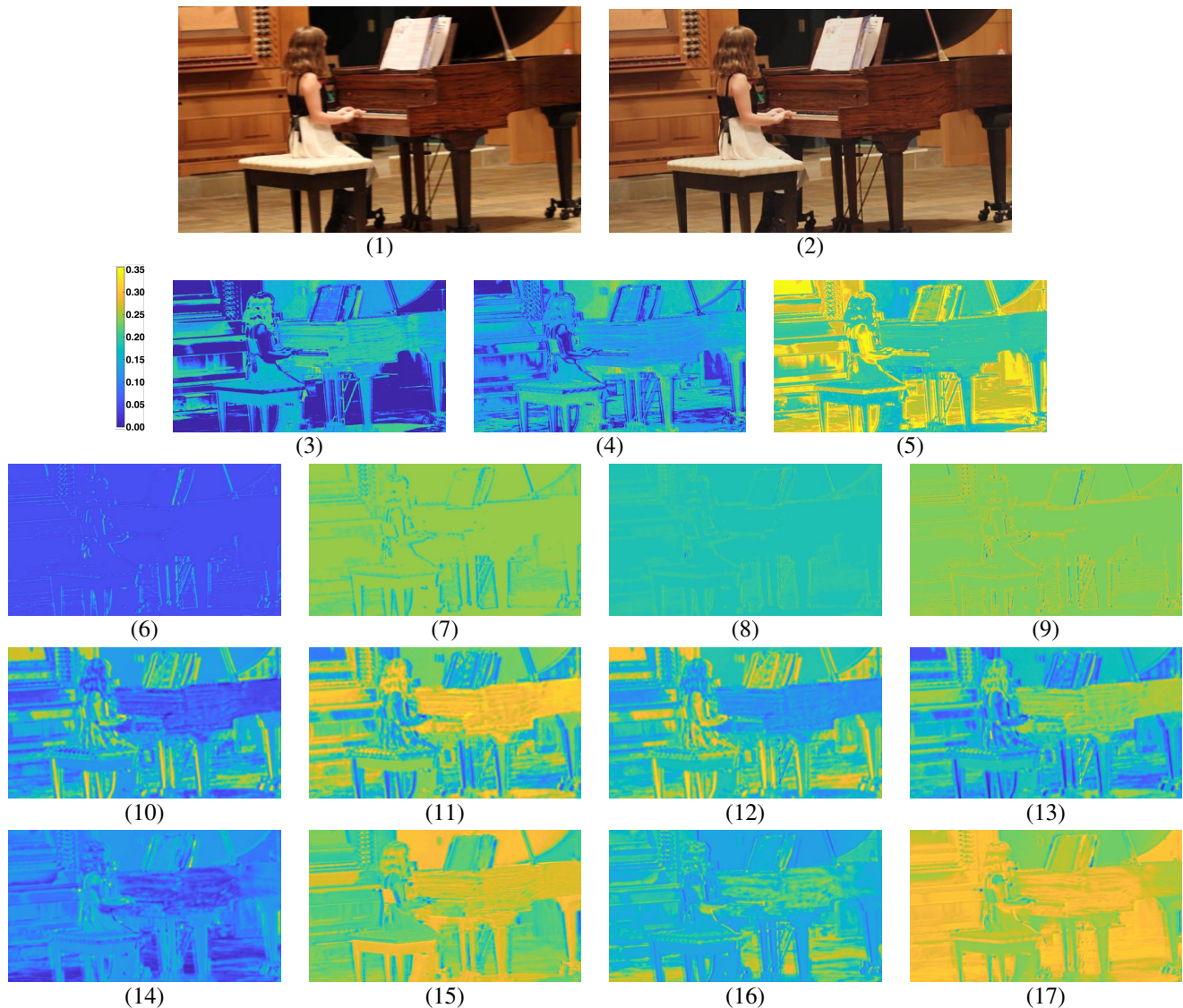

Figure 5: Visualization of external and internal attention maps on real-world blur dataset proposed in [35]. (1): blurry frame; (2): Deblurred frame; (3)-(5): attention maps of external attention branch 1-3; (6)-(9): attention maps of internal attention branch 1-3 in the first internal attention module; (10)-(13): attention maps of internal attention branch 1-3 in the second internal attention module; (14)-(17): attention maps of internal attention branch 1-3 in the third internal attention module.
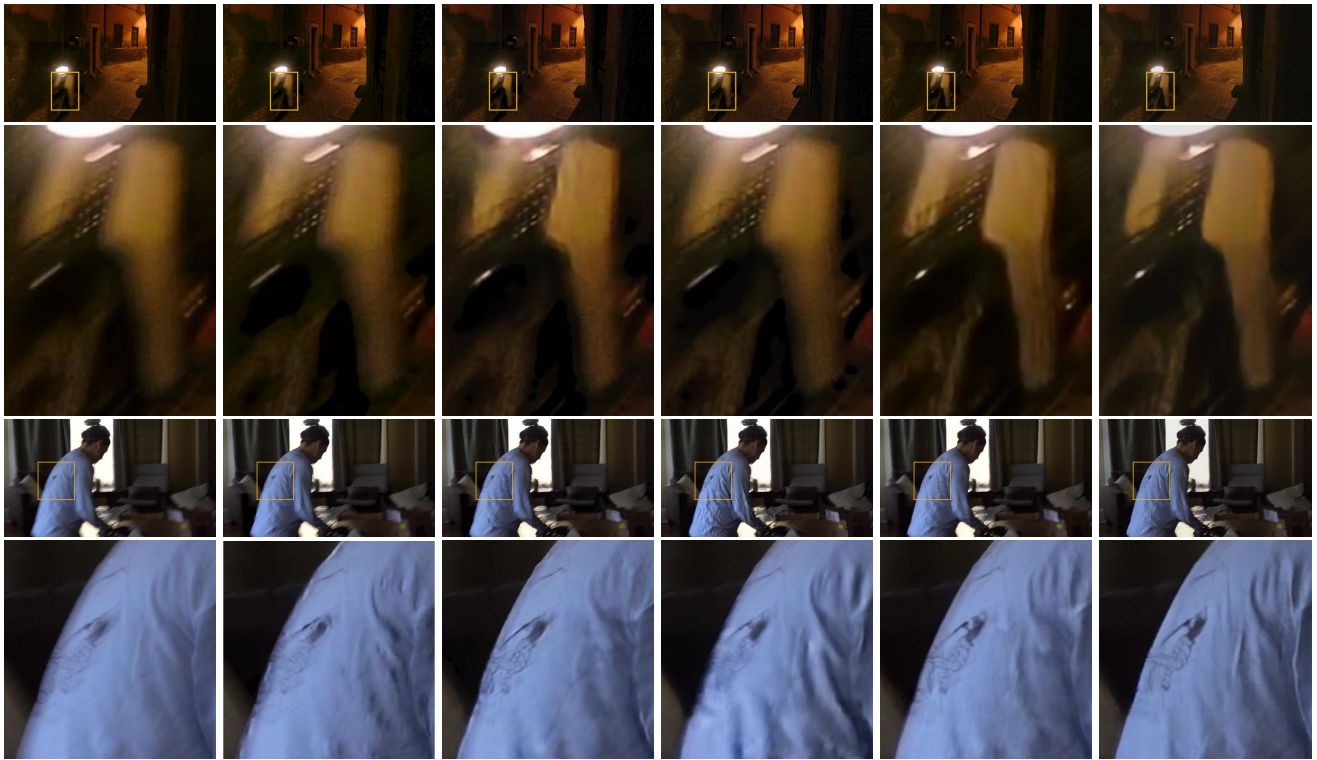
(a) Blurry image	(b) DVD (single)	(c) DVD (noalign)	(d) DVD (flow)	(e) DAVID (DVD)	(f) DAVID (C-DVD)

Figure 6: Visual result comparison on real blurry videos. The second and fourth rows depict enlarged local areas of the first and third rows, respectively, where blur artifacts could be most easily observed and compared around the boundary of the front fender (second row), and the texture on clothes (fourth row).

stacked frames	1	3	5	7	9
C-DVD-11	28.91	29.80	29.81	**29.84**	29.82

Table 2: Effect of multiple stacked frames as the input for the single backbone branch, trained on C-DVD-11 subset, tested on C-DVD-11 testing set.

indeed provide more temporal information for deblurring. However, when the number continues to increase, we see the performance is saturated. We observe the similar trend for other single backbone models trained on C-DVD-3 and C-DVD-7. It is because too large window size does not provide more information as two frames far apart might be irrelevant. Thus, we empirically find the optimal number of stacked frames (7 in our experimental setting) and apply it to conduct all our experiments.

4.4. Attention Map Visualization

To better understand the mechanism of our dual attention model, we visualize the attention maps given an input frame in Figure 5. The frame is from the real blurry video "piano" in the qualitative testing set provided by [35]. Inside the sample frame, there are multiple blur sources, including hand-held camera shake (which have caused global scene movement in video), and the local motion blurs caused by the player's head and hand moment. In this specific case, the camera movement is much stronger than object movement.

Figure 5 (3)-(5) show the external attention maps for Internal Att 3, Internal Att 7 and Internal Att 11 respectively. All three emphasize different detail structural information, which suggests that the external attention modules **do not degenerate** into selecting one out of the three. Notice that the third map shows the largest response magnitude, indicating that the external attention module favors more on the averaging 11 frames channel. We further acquire that the blur caused by global hand-held movement is closest to the blur by averaging 11 frames. Such selectivity also shows to be spatially variant, which aligns with our hypothesis. For example, the third channel (averaging 11 frames) seems to account for most global blurs (in the majority area of the non-moving background); meanwhile, the first channel (averaging 3 frames) captures more responses in the area of the player's head and hands, which are moving objects.

Each group of internal attention maps then tend to further decompose the blur information into finer scales that possess multiple-scale temporal correlations. (6)-(9) show mostly spatial low frequency response, which indicates that the backbone branches trained on averaging 3 frames input captures less motion. In contrast, (10)-(13) and (14)-(17) show response maps with plenty of the edge and region structural information, indicating that the backbone branches trained

Methods	1	2	3	4	5	6	7	8	9	10	Average
PSDEBLUR	24.42	28.77	25.15	27.77	22.02	25.74	26.11	19.71	26.48	24.62	25.08
WFA[6]	25.89	32.33	28.97	28.36	23.99	31.09	28.58	24.78	31.30	28.20	28.35
DVD (single)[35]	25.75	31.15	29.30	28.38	23.63	30.70	29.23	25.62	31.92	28.06	28.37
DVD (noalign)[35]	27.83	33.11	31.29	29.73	25.12	32.52	30.80	27.28	33.32	29.51	30.05
DVD (flow)[35]	28.31	33.14	30.92	29.99	25.58	32.39	30.56	27.15	32.95	29.53	30.05
IFI-RN[25]	-	-	-	-	-	-	-	-	-	-	30.80
Reblur2Deblur[3]	-	-	-	-	-	-	-	-	-	-	31.37
Ours (Single)	29.95	34.98	33.12	31.61	26.09	33.35	31.45	27.79	35.66	30.11	31.41
Ours (Internal Att)	30.28	35.20	33.25	31.81	26.37	33.47	31.61	28.00	35.75	30.37	31.65
Ours (External Att)	30.38	35.27	33.33	31.99	26.45	33.55	31.64	28.03	35.79	30.44	31.71
Ours (Dual Att)	**30.68**	**35.61**	**33.59**	**32.19**	**26.78**	**33.87**	**31.96**	**28.35**	**36.14**	**30.73**	**31.99**

Table 3: PSNR comparison on the Original DVD dataset. Best results are shown in bold.

on averaging 7 and 11 frames capture most motion and blur information for this specific video.

4.5. Evaluation on Original DVD Dataset

Original DVD dataset provides training data synthesized only by averaging 7 frames. It is not most suited for the temporal-spatially varying blur case that DAVID targets. However, as a general public benchmark, we report PSNR performance on DVD to provide a reference comparison. As shown in Table 3, our single backbone already outperforms the state-of-the-art Reblur2Deblur model [3]. It also surpasses [35] by 2dB, and more than 3.5dB over [6]. Further equipped with the proposed attention modules, our DAVID model yields stronger PNSR results with a 0.58dB gain over the single backbone. Notice that the gain is understandably smaller compared to the Challenging DVD case, since the blur here is not variant by averaging merely 7 frames.

4.6. Qualitative Results on Real Blurry Videos

Besides the evaluation on synthetic data, we visually illustrate the effectiveness of our DAVID model on the real videos. Figure 6 shows two examples from the qualitative testing set provided in [35]. We compare three versions of author-provided DVD models: single, no-align and flow; and two DAVID versions: one trained on the original DVD dataset, the other on the Challenging DVD (C-DVD) dataset.

As shown from the input zoom-in crop region, real videos have rich heterogeneous blurs at local regions. We observe that results from DVD (single) present significant ringing artifact, *i.e.* on the front tire cover of the motorcycle and on the shirt's edge. DVD (no-align) alleviates but still with certain artifacts alongside the shirt's edge and the front fender. DVD (flow) for the first example still leaves unattended blurs while for the second example produces extra fake wrinkle at the bottom of the shirt. In contrast, DAVID (DVD) and DAVID (C-DVD) shows far fewer artifacts while preserving the sharp structural information.

Moreover, when comparing DAVID (DVD) to DAVID (C-DVD), we still observe some slight artifact for the former, *i.e.*, blurry boundary region under the light of the first example. Therefore with the same proposed model, training on more complicated synthetic video blur can help real-world

generalization, which justifies the necessity of the proposed Challenging DVD dataset.

4.7. Discussion on Temporal dependency

We use sets of raw frames and average them to synthesize the blurry central frames, which has no overlap with each other. However, these sets still share temporal content dependency, for which we consider the consistency across a wider range of temporal scales, *e.g.* consistent motion of foreground objects, which is commonly exploited in video restoration literature [20]. Thus, we consider the temporal dependency as how many of the "sets" could be involved for restoring one central frame, rather than exposure period of a single set. Exploiting temporal dependency means adaptive and selective fusion of those different scales by internal attention. Certainly, central and neighboring frames are important. But wider temporal dependency is not negligible. In Table 2, when stacking more blurred frames (more "sets"), the performance improves up to 7 stacking frames. Our other empirical results also endorses the importance of exploiting longer temporal dependency.

5. Conclusions

This work proposed the DAVID framework for blind video deblurring. The internal attention model is trained to adaptively select the temporal scales, while the external attention model further spatially aggregates the output from each internal module at the pixel-level. Different from previous synthetic data focusing on only a single blur level, we propose a Challenging DVD dataset (C-DVD) which incorporates multiple levels of blur by pooling video frames with different window sizes. Experiments on both datasets demonstrate that DAVID achieves better PSNR performance compared to several state-of-the-arts. Qualitatively, we further validate the advantage of our method in recovering sharp appearance with fewer artifacts, on real blurry videos. Ablation study shows that the dual attentions mutually benefit with learned structurally meaningful attention maps. Our future work will develop the unsupervised adaption of DAVID to real-world blurry videos.

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 2

[2] Y. Cao, T. Chen, Z. Wang, and Y. Shen. Learning to optimize in swarms. In *Advances in Neural Information Processing Systems*, pages 15018–15028, 2019. 3

[3] H. Chen, J. Gu, O. Gallo, M.-Y. Liu, A. Veeraraghavan, and J. Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *ICCP*, 2018. 8

[4] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8351–8361, 2019. 3

[5] S. Cho, J. Wang, and S. Lee. Video deblurring for hand-held cameras using patch-based synthesis. *TOG*, 31(4):64, 2012. 2

[6] M. Delbracio and G. Sapiro. Burst deblurring: Removing camera shake through fourier burst accumulation. In *CVPR*, 2015. 5, 6, 8

[7] M. Delbracio and G. Sapiro. Hand-held video deblurring via efficient fourier aggregation. *arXiv preprint arXiv:1509.05251*, 2015. 2

[8] W. Freeman, F. Durand, Y. Weiss, and A. Levin. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009. 1, 2

[9] S. Harmeling, H. Michael, and B. Schölkopf. Space-variant single-image blind deconvolution for removing camera shake. In *NIPS*, 2010. 2

[10] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Scholkopf. Fast removal of non-uniform camera shake. In *ICCV*, 2011. 2

[11] T. Hyun Kim, B. Ahn, and K. Mu Lee. Dynamic scene deblurring. In *ICCV*, 2013. 2

[12] T. Hyun Kim and K. Mu Lee. Segmentation-free dynamic scene deblurring. In *CVPR*, 2014. 1, 2

[13] T. H. Kim, K. M. Lee, B. Schölkopf, and M. Hirsch. Online video deblurring via dynamic temporal blending network. In *ICCV*, 2017. 1, 2

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] R. Köhler, M. Hirsch, B. Mohler, B. Schölkopf, and S. Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *ECCV*, 2012. 2

[16] A. Kosiorek, A. Bewley, and I. Posner. Hierarchical attentive recurrent tracking. In *NIPS*, 2017. 3

[17] D. Krishnan, T. Tay, and R. Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, 2011. 2

[18] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *arXiv preprint arXiv:1711.07064*, 2017. 2

[19] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8878–8887, 2019. 1

[20] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, 2017. 3, 8

[21] D. Liu, B. Wen, X. Liu, Z. Wang, and T. S. Huang. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284*, 2017. 3

[22] X.-J. Mao, C. Shen, and Y.-B. Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016. 3

[23] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum. Full-frame video stabilization with motion inpainting. *TPAMI*, 28(7):1150–1163, 2006. 2

[24] S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 2

[25] S. Nah, S. Son, and K. M. Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8102–8111, 2019. 8

[26] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017. 4

[27] J. Pan, Z. Hu, Z. Su, H.-Y. Lee, and M.-H. Yang. Soft-segmentation guided object motion deblurring. In *CVPR*, 2016. 2

[28] J. Pan, D. Sun, H. Pfister, and M.-H. Yang. Blind image deblurring using dark channel prior. In *CVPR*, 2016. 2

[29] L. Pan, Y. Dai, M. Liu, and F. Porikli. Simultaneous stereo video deblurring and scene flow estimation. In *CVPR*, 2017. 2

[30] W. Ren, X. Cao, J. Pan, X. Guo, W. Zuo, and M.-H. Yang. Image deblurring via enhanced low-rank prior. *IEEE Transactions on Image Processing*, 25(7):3426–3437, 2016. 2

[31] W. Ren, J. Pan, X. Cao, and M.-H. Yang. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1077–1085, 2017. 2

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[33] C. Schuler, M. Hirsch, S. Harmeling, and B. Scholkopf. Learning to deblur. *TPAMI*, (1):1–1, 2016. 1, 2

[34] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. In *TOG*, volume 27, page 73. ACM, 2008. 2

[35] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[36] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, 2015. 2

[37] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 3

[38] X. Tao, H. Gao, Y. Wang, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. *arXiv preprint arXiv:1802.01770*, 2018. 1, 2, 3

[39] J. Telleen, A. Sullivan, J. Yee, O. Wang, P. Gunawardane, I. Collins, and J. Davis. Synthetic shutter speed imaging. In *Computer Graphics Forum*, 2007. 4

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[41] R. G. VidalMata, S. Banerjee, B. RichardWebster, M. Albright, P. Davalos, S. McCloskey, B. Miller, A. Tambo, S. Ghosh, S. Nagesh, et al. Bridging the gap between computational photography and visual recognition. *arXiv preprint arXiv:1901.09482*, 2019. 2

[42] S. Wang, B. Wen, J. Wu, D. Tao, and Z. Wang. Segmentation-aware image denoising without knowing true segmentation. *arXiv preprint arXiv:1905.08965*, 2019. 2

[43] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li. Hierarchical attention network for action recognition in videos. *arXiv preprint arXiv:1607.06416*, 2016. 3

[44] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. *IJCV*, 98(2):168–186, 2012. 2

[45] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *NIPS*, 2014. 2

[46] L. Xu, S. Zheng, and J. Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, 2013. 2

[47] X. Xu, J. Pan, Y.-J. Zhang, and M.-H. Yang. Motion blur kernel estimation via deep learning. *IEEE Transactions on Image Processing*, 27(1):194–205, 2018. 2

[48] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 251–260, 2017. 1

[49] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 3

[50] Y. Yuan, W. Yang, W. Ren, J. Liu, W. J. Scheirer, and Z. Wang. UG^{2+} track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments. *arXiv preprint arXiv:1904.04474*, 2019. 2

[51] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 3